1 Fundamentals

Many natural situations involve some notion of randomness or unpredictability. You might think of rolling a dice, or flipping a coin, or checking the Google share price. In this first section we begin our study of probability theory, which aims to model such behaviour mathematically.

1.1 Sample spaces, outcomes and events

Every probabilistic model begins with an *experiment* which produces elements from a set Ω.

- The set Ω is called the *sample space*. It represents the possible results of the experiment.
- Each element $\omega \in \Omega$ is called an *outcome*. The experiment always produces exactly one outcome.
- Each subset of the sample space $A \subseteq \Omega$ is called an *event*. If the experiment produces an outcome $\omega \in A$ then we say that the event A *occurs*.

In probability, the first step is often to decide on a sample space to represent the process we are interested in. Let's see some examples of how to do this.

Example 1.1 (Coins)**.** Tossing a coin is a simple experiment. Here we can take the sample space $\Omega = \{H, T\}$, with the outcome H representing 'heads' and the outcome T representing 'tails'. In this case Ω is small enough to list all events: $\emptyset$, $\{H\}$, $\{T\}$, $\{H, T\}$.

A different experiment is to toss a coin three times. We can write each outcome as $(\omega_1, \omega_2, \omega_3)$, where $\omega_i \in \{H, T\}$ is the result of the i-th coin toss. This gives 8 possible outcomes:

$$(T,T,T) \quad (T,T,H) \quad (T,H,T) \quad (H,T,T) \quad (H,H,T) \quad (H,T,H) \quad (T,H,H) \quad (H,H,H).$$

Note that order matters here, e.g. $(T,T,H) \neq (H,T,T)$. The sample space Ω is then

$$\Omega = \big\{(\omega_1, \omega_2, \omega_3) : \omega_i \in \{H,T\} \text{ for all } i = 1,2,3\big\} = \big\{H,T\big\} \times \big\{H,T\big\} \times \big\{H,T\big\} = \big\{H,T\big\}^3.$$

Already there are too many events to list (there are $2^8 = 256$), but here are some examples:

- $A_1 = \big\{(T,T,T),(H,H,H)\big\}$ is the event that 'all three tosses land on the same side'.
- $A_2 = \big\{(T,T,H),(T,H,T),(H,T,T),(H,H,H)\big\}$ the event 'an odd number of heads appear'.
- $A_3 = \big\{(T,T,T),(H,H,T),(T,H,H)\big\}$ is an event as $A_3 \subseteq \Omega$, but I don't see a nice description.

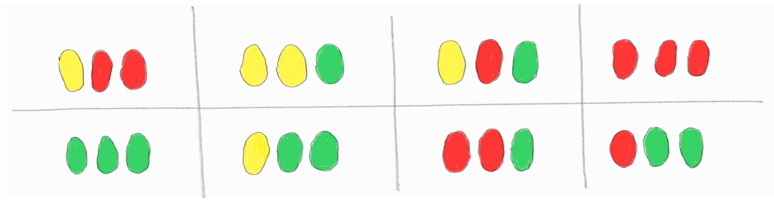
Example 1.2 (Cards)**.** Selecting a card from a standard deck is an experiment. There are 52 possible outcomes and the sample space

$$\Omega = \big\{A\clubsuit, K\clubsuit, \ldots, 2\clubsuit, A\diamondsuit, K\diamondsuit, \ldots, 2\diamondsuit, A\spadesuit, K\spadesuit, \ldots, 2\spadesuit, A\heartsuit, K\heartsuit, \ldots, 2\heartsuit\big\}.$$

There are a huge number of events in this case (about 4.5×10^{15}). Here are some natural ones:

- $A_1 = \{Q\clubsuit, Q\diamondsuit, Q\spadesuit, Q\heartsuit\}$ is the event that 'we select a queen'.
- $A_2 = \{J\diamondsuit, J\heartsuit, Q\diamondsuit, Q\heartsuit, K\diamondsuit, K\heartsuit\}$ is the event that 'we select a red picture card'.
- $A_1 \cup A_2 = \{J\diamondsuit, J\heartsuit, Q\diamondsuit, Q\heartsuit, K\diamondsuit, K\heartsuit, Q\clubsuit, Q\spadesuit\}$ is the event that 'we select a red picture card or a queen'.
- $A_2 \setminus A_1 = \{J\diamondsuit, J\heartsuit, K\diamondsuit, K\heartsuit\}$ for the event 'we select a red picture card which is not a queen'.

Example 1.3 (Pick 'n' Mix)**.** Without looking, we choose three sweets together from a bag containing yellow, green and red sweets. As the sweets are chosen together, we only distinguish the sweets by their colours (so here there is no order). Here are a couple of outcomes:



Again we can represent each outcome as a tuple $(\omega_1, \omega_2, \omega_3)$, where ω_1 is the number of yellow sweets in the sample, and ω_2 (ω_3, respectively) is the number of red (green, respectively) sweets in the sample.

We can set the sample space

$$\Omega = \{(\omega_1, \omega_2, \omega_3) \in \mathbb{Z}^3 : \omega_1 + \omega_2 + \omega_3 = 3, \omega_i \geq 0\}.$$

One could consider the event:

- $A_1 = \{(0,3,0), (0,2,1), (0,1,2), (0,0,3)\}$ that there are 'no yellow sweets in the sample', or
- $A_2 = \{(1,1,1)\}$ that 'sweets of all colours appear in the sample'.

Example 1.4 (Darts)**.** Consider the experiment of throwing a dart at a dartboard. The typical radius ≈ 9 inches, so we could represent the centre of the board with the origin $(0,0)$ and represent outcomes as points $(x,y) \in \mathbb{R}^2$ with $\sqrt{x^2+y^2} \leq 9$. The sample space is then

$$\Omega = \{(x,y) \in \mathbb{R}^2 : \sqrt{x^2+y^2} \leq 9\}.$$

Regions of the dartboard correspond to events here (e.g. '17' or 'triple 10'). For example, $A_1 = \{(x,y) \in \Omega : \sqrt{x^2+y^2} \leq 1/4\}$ would represent the event that 'the dart hits the bullseye'.

Remark 1.5. Here are some takeaways from these examples:

- Set operations are often useful in building new events from old ones, e.g. $A \cup B$, $(A \cap B^c) \setminus C, \ldots$. See Example 1.2 above.
- Experiments can sometimes be broken down into several smaller experiments, and Cartesian product notation can then be useful in describing the sample spaces, e.g. see Example 1.1.

Please see 'Background' if you would like a refresher on set operations.

1.2 Probability distributions

The sample space Ω forms half of a probabilistic model and the other half is given by a *probability distribution*. This is a function which assigns a number $\mathbb{P}(A) \in [0,1]$ to each event $A \subseteq \Omega$. Intuitively $\mathbb{P}(A)$ can be thought of as a measurement of 'how likely' the event A is to occur; probability 0 represents 'extremely unlikely' and probability 1 represents 'extremely likely'.

Definition 1.6 (Probability distribution)**.** A *probability distribution* or *probability measure* on a sample space Ω is a function $\mathbb{P}$ which assigns a number $\mathbb{P}(A)$ to each event $A \subseteq \Omega$, so that:

(i) $\mathbb{P}(A) \in [0,1]$ for every event $A \subseteq \Omega$;

(ii) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$;

(iii) (a) for every sequence of pairwise disjoint events $A_1, A_2, \ldots, A_k \subseteq \Omega$ we have

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

(b) for every sequence of pairwise disjoint events $A_1, A_2, \ldots \subseteq \Omega$ we have

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

Remark 1.7. Properties (i)–(iii) provide the 'ground rules' of probability [1] and we will see that many intuitive facts follow from them (see Proposition 1.19 below). Let us discuss them individually.

- The condition $\mathbb{P}(A) \in [0,1]$ for all $A \subseteq \Omega$ is a convention. It is also important in practice; if your calculate a probability of -0.4 or 2 then something has gone wrong. :(
- Recall that an experiment always results in exactly one outcome $\omega \in \Omega$. For any outcome ω we have $\omega \notin \emptyset$ and $\omega \in \Omega$, and so property (ii) agrees with the intuition that probability 0 represents an unlikely event, while probability 1 represents an extremely likely event.
- $\mathbb{P}(A)$ is sometimes thought of as a 'mass' associated to A. Condition (iii)(a) says that if we start pairwise disjoint events $A_1, A_2, \ldots, A_k$ then the mass associated with $\bigcup_{i=1}^{k} A_i$ equals the the sum of the individual masses; the same reasoning applies to (iii)(b) [2].

Before looking at examples of probability distributions, we first state a lemma which gives a simpler description of probability distributions in many cases.

We say a sample space Ω is *discrete* if Ω is either finite or countable [3]. Below we will sometimes need to sum real numbers over the elements of a discrete set Ω, i.e. $\sum_{\omega \in \Omega} a_\omega$ for some real numbers a_ω. Please see the 'Background' file if this notation is unfamiliar.

[1] Conditions (i)–(iii) are (essentially) *Kolmogorov's axioms*, named after the Russian mathematician who stated them.

[2] Property (iii)(a) can in fact be deduced from the other properties. It's a nice exercise!

[3] Please see 'Background' if you would like a little more information on countability. The main infinite discrete set we will need is $\Omega = \mathbb{N} = \{1, 2, 3, \ldots\}$.

Lemma 1.8. Suppose that Ω is a discrete sample space.

(a) If $\mathbb{P}$ is a probability distribution on Ω then

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}), \qquad \text{for all events } A \subseteq \Omega.$$

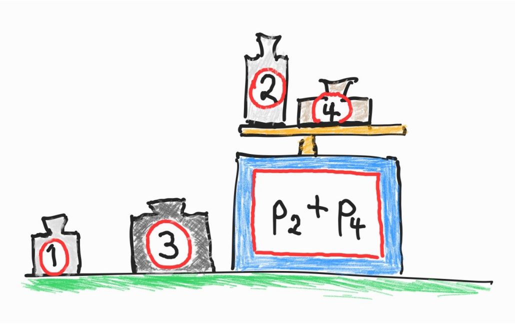
(b) Let $p_\omega \geq 0$ for all $\omega \in \Omega$ with $\sum_{\omega \in \Omega} p_\omega = 1$. Then the function $\mathbb{P}$ defined by

$$\mathbb{P}(A) := \sum_{\omega \in A} p_\omega, \qquad \text{for all events } A \subseteq \Omega,$$

is a probability distribution on Ω. (In particular, $\mathbb{P}(\{\omega\}) = p_\omega$ for all $\omega \in \Omega$.)

Remark 1.9.

- Given $\omega \in \Omega$ we often write $\mathbb{P}(\omega)$ for $\mathbb{P}(\{\omega\})$. The values $\{\mathbb{P}(\omega) : \omega \in \Omega\}$ are called the *weights* of the distribution.
- Lemma 1.8 (a) shows that probability distributions on discrete sample spaces are very simple – to find the probability of an event A, just sum up the weights of the outcomes in A.
- Lemma 1.8 (b) gives an easy way to build probability distributions on discrete sample spaces Ω.



A sketch proof is included at end of this section in an Appendix, but you might like to try to prove this for yourself first. In the meantime... on to some examples!

1.2.1 Uniform distributions

Definition 1.10 (Uniform distribution)**.** Let Ω be a finite set. The *uniform distribution* on Ω is the function $\mathbb{P}$ which for all $A \subseteq \Omega$ is given by [4]

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

Remark 1.11.

- The uniform distribution is the most familiar distribution, and it is the natural probability distribution to use whenever each outcome in a finite sample space is *equally likely to occur.*
- From Definition 1.10 we see that calculating $\mathbb{P}(A)$ for the uniform distribution reduces to a problem of *counting*; we need to find $|A|$ and $|\Omega|$.

[4] As usual, here $|A|$ denotes the number of elements in the set A.

- We should check *it is* a probability distribution. As Ω is discrete (finite), we can use Lemma 1.8 (b). For each $\omega \in \Omega$ let $p_\omega := 1/|\Omega| \geq 0$, noting that $\sum_{\omega\in\Omega} p_\omega = 1$. The distribution given by these weights is $\mathbb{P}(A) = \sum_{\omega\in A} p_\omega = |A|/|\Omega|$ for each $A \subseteq \Omega$, i.e. the uniform distribution.

Example 1.12 (Cards)**.** Consider the card example from Example 1.2 above. If the deck of cards is well shuffled, each of the possible 52 outcomes from Ω should be equally likely to occur and so taking the uniform distribution $\mathbb{P}$ on Ω is the natural choice. For

- $A_1 = \big\{Q\clubsuit, Q\diamondsuit, Q\spadesuit, Q\heartsuit\big\}$ this gives $\mathbb{P}(A_1) = 4/52 = 1/13$; and
- $A_2 = \big\{J\diamondsuit, J\heartsuit, Q\diamondsuit, Q\heartsuit, K\diamondsuit, K\heartsuit\big\}$ this gives $\mathbb{P}(A_2) = 6/52 = 3/26$.

Example 1.13 (Dice)**.** We roll two (distinguishable) fair dice ⚂ ⚅. Here we can take the sample space $\Omega = \{1, \dots, 6\}^2$. If both dice are fair, the corresponding distribution is the uniform distribution on Ω, and so $\mathbb{P}(\omega) = 1/|\Omega| = 1/36$ for each outcome $\omega \in \Omega$. Some examples:

- A_1, the event that 'the sum of the two dice is 11 or more'. Here $A_1 = \{(5,6),(6,5),(6,6)\}$ and $\mathbb{P}(A_1) = 3/36 = 1/12$.
- A_2, the event that 'the dice differ by ≥ 4'. Here $A_1 = \{(1,6),(1,5),(2,6),(5,1),(6,2),(6,1)\}$ and so $\mathbb{P}(A_1) = 6/36 = 1/6$.

1.2.2 Non-uniform discrete probability distributions

While uniform distributions are very natural, several important probability distributions are not of this form. We will meet many of these in later sections, but here are a few examples for the moment.

Example 1.14 (Coins)**.** Consider the sample space $\Omega = \{H, T\}$ from Example 1.1. Set $p_T = p \in [0,1]$ and $p_H = 1 - p$, so that $\sum_{\omega\in\Omega} p_\omega = p_H + p_T = 1$. Lemma 1.8 (b) then gives a probability distribution $\mathbb{P}$ on Ω with $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{T\}) = p$, $\mathbb{P}(\{H\}) = 1 - p$ and $\mathbb{P}(\Omega) = p + (1-p) = 1$. If $p \neq 1/2$ then this distribution is not uniform.

Another example is given by the three coin experiment from Example 1.1, where $\Omega = \big\{(\omega_1, \omega_2, \omega_3) : \omega_i \in \{H, T\} \text{ for all } i = 1, 2, 3\big\}$. Here to each $\omega = (\omega_1, \omega_2, \omega_3) \in \Omega$ we assign weight

$$p_\omega := p^t(1-p)^{3-t}, \quad \text{where } t = |\{i \in \{1,2,3\} : \omega_i = T\}|.$$

Clearly $p_\omega \geq 0$ for all $\omega \in \Omega$ and you should check that $\sum_{\omega\in\Omega} p_\omega = 1$. By Lemma 1.8 (b) this gives a probability distribution given by $\mathbb{P}(A) = \sum_{\omega\in A} p_\omega$ for $A \subseteq \Omega$. Considering the events in Example 1.1,

- $A_1 = \big\{(T,T,T),(H,H,H)\big\}$ we have $\mathbb{P}(A_1) = p_{(T,T,T)} + p_{(H,H,H)} = p^3 + (1-p)^3$.
- $A_2 = \big\{(T,T,H),(T,H,T),(H,T,T),(H,H,H)\big\}$ we have $\mathbb{P}(A_2) = 3p^2(1-p) + (1-p)^3$.

Example 1.15 (Letters)**.** A bag contains 10 tiles with letters which spell the word 'statistics'. We select a tile from the bag at random and look at the resulting letter. A sample space for this experiment is $\Omega = \{a, c, i, t, s\}$. As each tile is equally likely to occur, the uniform distribution *on the tiles* is the correct choice. This gives $\mathbb{P}(a) = \mathbb{P}(c) = 1/10$, $\mathbb{P}(i) = 2/10$ and $\mathbb{P}(s) = \mathbb{P}(t) = 3/10$, a non-uniform distribution on Ω.

Example 1.16. Suppose that you have an experiment with sample space $\Omega = \mathbb{N}$ and that $\mathbb{P}$ is a probability distribution on Ω, with $\mathbb{P}(n) = 2 \cdot \mathbb{P}(n + 1)$ for all $n \in \mathbb{N}$. Determine the weights of $\mathbb{P}$ and hence find the probability of the event A that 'the outcome of the experiment is at least 3'.

Our first step is to find the weights of the distribution, i.e. find $\mathbb{P}(n)$ for all $n \in \Omega$. As the sample space $\Omega = \mathbb{N}$ is discrete, we can apply Lemma 1.8 to find

$$\mathbb{P}(A) = \sum_{n \in A} \mathbb{P}(n),$$

for all $A \subseteq \Omega$. In particular, as $\mathbb{P}(\Omega) = \mathbb{P}(\mathbb{N}) = 1$ by Definition 1.6, we obtain $\sum_{n\in\mathbb{N}} \mathbb{P}(n) = 1$.

Now we use the condition given above to note that, for $n \geq 2$ we have

$$\mathbb{P}(n) = 2^{-1} \cdot \mathbb{P}(n - 1) = \ldots = 2^{-(n-1)} \cdot \mathbb{P}(1). \tag{1}$$

Combined, this gives

$$1 = \sum_{n\in\mathbb{N}} \mathbb{P}(n) = \sum_{n\in\mathbb{N}} 2^{-(n-1)} \cdot \mathbb{P}(1) = 2 \cdot \mathbb{P}(1)$$

where the final equality used the formula for the sum of geometric series [5]. This gives $\mathbb{P}(1) = 1/2$ and so, by (1), the weights are given by $\mathbb{P}(n) = 2^{-n}$ for all $n \in \mathbb{N}$.

Lastly, to calculate $\mathbb{P}(A)$ we could find the sum

$$\mathbb{P}(A) = \sum_{n\in\{3,4,..\}} 2^{-n} = 2^{-2} = 0.25,$$

where the second equality is again given by summing a geometric series [6].

1.2.3 A non-discrete probability distribution

The following gives an indication of some of the differences between probability distributions on discrete and non-discrete spaces. We will see more examples later in Section 4.

Example 1.17 (Darts)**.** We throw a dart at a dartboard as in Example 1.4. Recall that the sample space $\Omega = \left\{(x, y) \in \mathbb{R}^2 : \sqrt{x^2 + y^2} \leq 9\right\}$ and events $A \subseteq \Omega$ are regions of the dartboard.

As an entirely useless player, my darts are equally like to go anywhere on the board. Here it is natural to relate the probability of an event A to the *area* of the board corresponding to the region A, i.e.

[5] This fundamental formula states that $\sum_{n=0}^{\infty} x^n = (1 - x)^{-1}$ for $|x| < 1$, and was proven in 1RAC.

[6] In fact, in this case it is easier to note that $A^c = \{1, 2\}$ and so $\mathbb{P}(A^c) = \sum_{n\in\{1,2\}} 2^{-n} = 0.75$. By Proposition 1.19(i) below it follows that $\mathbb{P}(A) = 0.25$.

$\mathbb{P}(A) = c{\cdot}\mu(A)$ for some constant $c > 0$, where $\mu(A)$ is the area of the region A [7]. As $c{\cdot}\mu(\Omega) = \mathbb{P}(\Omega) = 1$ by Definition 1.6(ii), we should take $c = 1/\mu(\Omega) = 1/(\pi 9^2)$. The probability of the event A_1 'bullseye' from Example 1.4 is then $\mathbb{P}(A_1) = (\pi 0.25^2)/(\pi 9^2) = 1/1296$.

Remark 1.18.

- Note $\mathbb{P}(\omega) = 0$ for every outcome $\omega \in \Omega$ in Example 1.17. This is very different from the discrete case – we can't calculate the probability of events by just 'summing' up the outcomes as in Lemma 1.8, since $\mathbb{P}(\Omega) = 1$ but $\mathbb{P}(\omega) = 0$ for all $\omega \in \Omega$.
- Integration ($\int$) will take the place of summation ($\sum$) when studying distributions on non-discrete spaces. We will wait until the second half of the course to study such distributions.
- Although $\mathbb{P}(\omega) = 0$ for all $\omega \in \Omega$ in Example 1.17, the experiment always produces *some* outcome. Probability 0 events should therefore be viewed as 'extremely unlikely' rather than 'impossible'.

1.3 Tools to calculate probabilities

We have seen above (Remark 1.5) that some events are built using set operations on other events, e.g. $A \cup B$, $B^c \setminus (A \cap C)$, ... Given this, it is natural to ask how to calculate probabilities for events formed by such operations. The following proposition proves a number of bounds of this kind.

The proof below is entirely based on the properties from Definition 1.6 – you're encouraged to try to prove (i)–(v) for yourself first using this definition.

Proposition 1.19. Let $\mathbb{P}$ be a probability distribution on sample space Ω. Then the following hold:

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for all $A \subseteq \Omega$.

(ii) $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for all $A, B \subseteq \Omega$.

(iii) $\mathbb{P}(A) \leq \mathbb{P}(B)$ for all $A, B \subseteq \Omega$ with $A \subseteq B$.

(iv) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ for all $A, B \subseteq \Omega$.

(v) $\mathbb{P}\left(\bigcup_{i=1}^k A_i\right) \leq \sum_{i=1}^k \mathbb{P}(A_i)$ for all $A_1, \ldots, A_k \subseteq \Omega$. [8]

Proof. To prove (i) note that the events A and A^c are always pairwise disjoint. By Definition 1.6 (iii)(a) with $k = 2$ this gives

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1.$$

The second equality holds as $\Omega = A \cup A^c$ and the last equality holds as $\mathbb{P}(\Omega) = 1$ by Definition 1.6 (ii). Rearranging proves (i), as requried.

[7] We do not prove that this gives a probability distribution here, as there are some subtleties.

[8] The infinite version of this property is also true.

To prove (ii) note that for any $A, B \subseteq \Omega$ the events $A \cap B$ and $B \setminus A$ are pairwise disjoint. By Definition 1.6 (iii)(a) it follows that

$$\mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) = \mathbb{P}\big((A \cap B) \cup (B \setminus A)\big) = \mathbb{P}(B),$$

where the final equality uses $(A \cap B) \cup (B \setminus A) = B$. Rearranging gives (ii).

To see (iii) note that if $A \subseteq B$ then $A \cap B = A$. By (ii) we obtain

$$\mathbb{P}(B) - \mathbb{P}(A) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}\left(B \setminus A\right) \geq 0,$$

since $\mathbb{P}\left(B \setminus A\right) \geq 0$ by Definition 1.6(i). It follows that $\mathbb{P}(B) \geq \mathbb{P}(A)$.

To see (iv) note that A and $B \setminus A$ are pairwise disjoint and $A \cup (B \setminus A) = A \cup B$. By Definition 1.6 (iii)(a) with $k = 2$ we find

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

Also by (ii) we have $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$. Combining both equalities gives (iv).

For (v) the case $k = 2$ follows from (iv) since

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2),$$

where the inequality holds as $\mathbb{P}(A_1 \cap A_2) \geq 0$ by Definition 1.6 (i). The general case can be proven by induction on k, using that $\bigcup_{i=1}^{k} A_i = \left(\bigcup_{i=1}^{k-1} A_i\right) \cup A_k$. □

Example 1.20. Let $A, B \subseteq \Omega$ be events satisfying $\mathbb{P}\left(A\right) = 0.2, \mathbb{P}\left(B\right) = 0.7$ and $\mathbb{P}\left(A \cup B\right) = 0.8$. Proposition 1.19 then allows us to compute many other probabilities. For example:

$$\begin{aligned} \mathbb{P}\left(A^c\right) &= 1 - \mathbb{P}\left(A\right) = 0.8, \\ \mathbb{P}\left(A \cap B\right) &= \mathbb{P}\left(A\right) + \mathbb{P}\left(B\right) - \mathbb{P}\left(A \cup B\right) = 0.1, \\ \mathbb{P}\left(A \setminus B\right) &= \mathbb{P}\left(A\right) - \mathbb{P}\left(A \cap B\right) = 0.1. \end{aligned}$$

Example 1.21. We are interested in the weather tomorrow. Consider the sample space

$$\Omega = \{\text{hot, cold}\} \times \{\text{dry, rainy}\}.$$

Let $C \subseteq \Omega$ be the event that there is cold weather, i.e. $C = \{(\text{cold}, \text{dry}), (\text{cold}, \text{rainy})\}$. Let the events H for hot weather, R for rainy weather and D for dry conditions be similarly defined. Suppose that $\mathbb{P}\left(C\right) = 0.6$, $\mathbb{P}\left(\{(\text{hot}, \text{rainy})\}\right) = 0.3$ and $\mathbb{P}\left(\{(\text{cold}, \text{rainy})\}\right) = 0.5$. What can we say about the remaining probabilities?

- $\mathbb{P}\left(R\right) = \mathbb{P}\left(\{(\text{cold}, \text{rainy}), (\text{hot}, \text{rainy})\}\right) = \mathbb{P}\left(\{(\text{cold}, \text{rainy})\}\right) + \mathbb{P}\left(\{(\text{hot}, \text{rainy})\}\right) = 0.8$.
- By Proposition 1.19(i) $\mathbb{P}\left(H\right) = 1 - \mathbb{P}\left(C\right) = 0.4$ and $\mathbb{P}\left(D\right) = 1 - \mathbb{P}\left(R\right) = 0.2$.
- By Proposition 1.19(ii), $\mathbb{P}\left(\{(\text{hot}, \text{dry})\}\right) = \mathbb{P}\left(H \setminus \{(\text{hot}, \text{rainy})\}\right) = 0.4 - 0.3 = 0.1$, and $\mathbb{P}\left(\{(\text{cold}, \text{dry})\}\right) = \mathbb{P}\left(C \setminus \{(\text{cold}, \text{rainy})\}\right) = 0.6 - 0.5 = 0.1$.

As this determines $\mathbb{P}(\omega)$ for all $\omega \in \Omega$, we know the distribution by Lemma 1.8(a).

Inclusion-exclusion principle. We are often given events $A_1, \ldots, A_k \subseteq \Omega$ and would like to find $\mathbb{P}(\bigcup_{i=1}^k A_i)$. If $k = 2$ then Proposition 1.19(iv) shows us how to do this:

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

This is in fact a special case of the *inclusion-exclusion principle.* Before stating the general formula, let us first consider the case of three events A_1, A_2, A_3. We have

$$\begin{aligned}\mathbb{P}\left(A_1 \cup A_2 \cup A_3\right) = \mathbb{P}\left(A_1\right) + \mathbb{P}\left(A_2\right) + \mathbb{P}\left(A_3\right) - \mathbb{P}\left(A_1 \cap A_2\right) - \mathbb{P}\left(A_1 \cap A_3\right) - \mathbb{P}\left(A_2 \cap A_3\right) \\ + \mathbb{P}\left(A_1 \cap A_2 \cap A_3\right).\end{aligned}$$

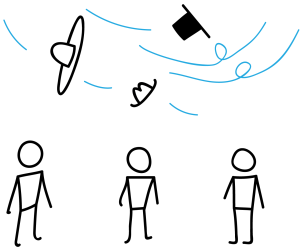
This follows from repeated application of Proposition 1.19(iv):

$$\begin{aligned}\mathbb{P}\left(\left(A_1 \cup A_2\right) \cup A_3\right) &= \mathbb{P}\left(A_1 \cup A_2\right) + \mathbb{P}\left(A_3\right) - \mathbb{P}\left(\left(A_1 \cup A_2\right) \cap A_3\right) \\ &= \mathbb{P}\left(A_1\right) + \mathbb{P}\left(A_2\right) - \mathbb{P}\left(A_1 \cap A_2\right) + \mathbb{P}\left(A_3\right) - \mathbb{P}\left(\left(A_1 \cap A_3\right) \cup \left(A_2 \cap A_3\right)\right) \\ &= \mathbb{P}\left(A_1\right) + \mathbb{P}\left(A_2\right) + \mathbb{P}\left(A_3\right) - \mathbb{P}\left(A_1 \cap A_2\right) - \mathbb{P}\left(A_1 \cap A_3\right) - \mathbb{P}\left(A_2 \cap A_3\right) \\ &\quad + \mathbb{P}\left(A_1 \cap A_2 \cap A_3\right).\end{aligned}$$

More generally, using induction on n one can prove that for events $A_1, A_2, \ldots, A_n \subseteq \Omega$ we have

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{\emptyset \neq J \subseteq \{1,\ldots,n\}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{j \in J} A_j\right).$$

A beautiful application of this identity is to the following problem:



At a graduation n people throw their hats in the air. The wind scrambles the hats, and each person receives a random hat back. What is the probability that nobody receives their own hat?

The sample space Ω for the problem is the set of possible assignments of the hats. As each person receives a random hat back, the uniform distribution on Ω is again the correct choice here. Letting p_n denote the required probability, inclusion-exclusion can be used to prove

$$p_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots + (-1)^n \cdot \frac{1}{n!}.$$

In particular as $n \to \infty$, we obtain $p_n \to e^{-1}$, as $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$ for all $x \in \mathbb{R}$.

For a proof using inclusion-exclusion consult, e.g. Example 1.27 in *Introduction to Probability* by Anderson, Seppäläinen, Valkó.

1.4 Counting and approximating

Counting often plays a significant role in calculating probabilities (see Remark 1.11). In this subsection we consider some useful counting principles and some estimation techniques.

1.4.1 Combinatorics

A number of different types of counting problems frequently crop up in practice. They are best illustrated by examples.

I) How many different combinations are there on a 4-digit padlock with each digit in $\{0, 1, \ldots, 9\}$?

II) How many ways can a rugby coach choose 15 players from a pool of 23 if she assigns a specific field position to each player?

III) How many ways are there to choose 4 books out of 20 when going on holiday?

IV) How many ways can we choose 4 scoops of ice-cream among the flavours chocolate, strawberry, vanilla, banana and hazelnut?

Abstracting, these four questions ask for the number of different ways that $k \geq 1$ elements can be chosen from a set of $n \geq 1$ elements where one distinguishes (a) whether the sample is taken with or without repetition, and (b) whether the elements in the sample are ordered or not ordered.

To answer these questions we recall the following notions:

- Given $n \in \mathbb{N}$ the nth *factorial* is defined as $n! := n \cdot (n-1) \cdots 2 \cdot 1$. We also define $0! = 1$.
- For $n = 0, 1, \ldots$ and $0 \leq m \leq n$, the quantities $\binom{n}{m} := \frac{n!}{m!(n-m)!}$ are called *binomial coefficients*. We also set $\binom{n}{m} = 0$ for $m < 0$ or $m > n$.

Given $n \geq k \geq 0$ it is easy to check

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

This identity is often represented using Pascal's triangle [9]. We also recall the binomial theorem [10]:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \quad n \geq 0,\ x, y \in \mathbb{R}.$$

[9] See e.g. Pascal's triangle on Wikipedia.

[10] This appears in 1RAC and 1AC, where a proof by induction is given.

Now suppose that we choose k elements out of n.

(i) If we choose with repetition then there are n choices for each element. Tracking order this gives n^k choices.

(ii) Choosing without repetition, there are n choices for the first element, $n-1$ choices for the second, and so on. Tracking order this gives $n(n-1)\cdots(n-k+1) = n!/(n-k)!$ distinct possibilities.

(iii) Choosing without repetition and without order, each selection in (ii) gives exactly $k!$ unordered choices. Thus there are $n!/k!(n-k)! = \binom{n}{k}$ possibilities here.

(iv) The fourth case is left to the combinatorics module.

	With repetition	Without repetition
Ordered	n^k	$\frac{n!}{(n-k)!}$
Unordered	1AC	$\binom{n}{k}$

Example 1.22. Let us come back to the examples I) - III).

I) This asks for the number of ordered samples of size $k = 4$ chosen with repetition from a set of size $n = 10$. The answer is $10^4 = 1000$.

II) This asks for the number of ordered samples of size $k = 15$ chosen without repetition from a set of size $n = 23$. The answer is $23!/8! \approx 6.4 \cdot 10^{17}$.

III) This asks for the number of unordered samples of size $k = 4$ chosen without repetition from a set of size $n = 20$. The answer is $\binom{20}{4} = 4845$.

Example 1.23 (Children toys)**.** We give 15 different toys to five children, where each child receives three toys. In how many ways can this be done?

There are $\binom{15}{3}$ choices of toys for the first child (unordered sample, without repetition). Then, there remain $\binom{12}{3}$ choices for the second child (same argument). Continuing in this way gives the following for the total number of choices:

$$\binom{15}{3} \cdot \binom{12}{3} \cdot \binom{9}{3} \cdot \binom{6}{3} \cdot \binom{3}{3} = \frac{15!}{(3!)^5} = 168168000.$$

Let's see a famous example to see how this crops up in probability.

Example 1.24 (Birthday paradox)**.** Many people have heard of the fact that in a group of 23 people, the probability that at least two share their birthdays is roughly 1/2. Let's see why.

Assume k people have been chosen at random. We set $\Omega = \{1, \dots, 365\}^k$ where $\omega = (\omega_1, \dots, \omega_k) \in \Omega$ corresponds to the outcome that, for $i = 1, \dots, k$ the i-th person's birthday falls on day ω_i [11]. All

[11] Apologies to the leap year babies; for simplicity we assume there are 365 days in the year.

outcomes in Ω seem equally likely, so we take $\mathbb{P}$ to be the uniform distribution on Ω [12]. The event A that all k birthdays are *distinct* is then

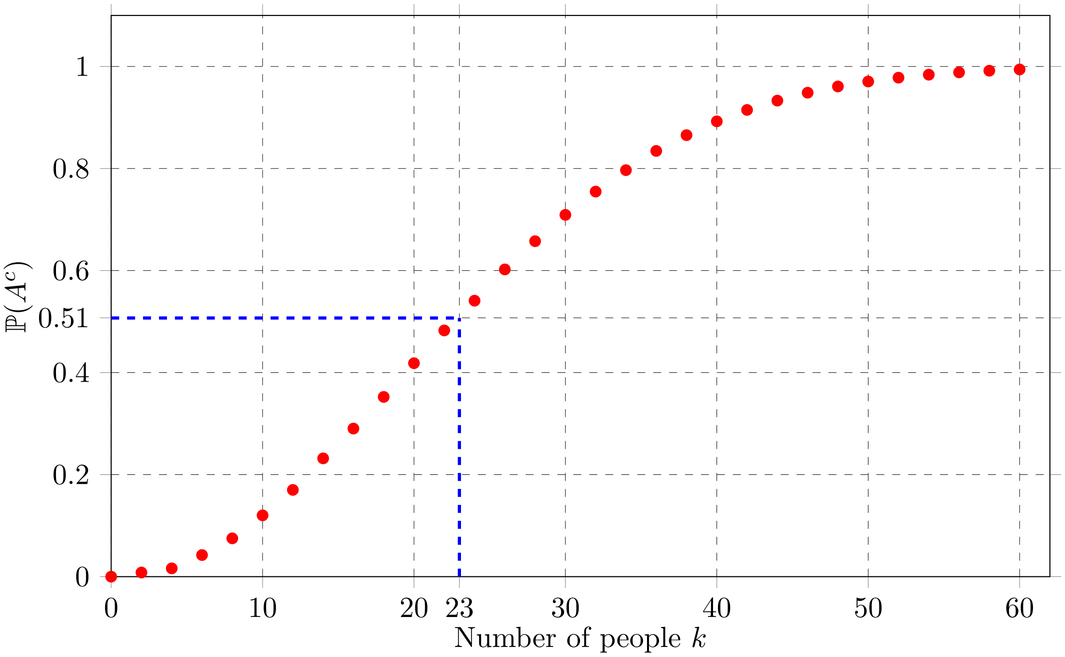
$$A = \big\{(\omega_1, \ldots, \omega_k) \in \Omega : \omega_i \neq \omega_j \text{ for all } i \neq j\big\}.$$

The event that some birthday is shared among the k people is then just A^c.

As $\mathbb{P}$ is uniform, $\mathbb{P}(A) = |A|/|\Omega|$. To finish note that $|\Omega|$ is the number of possible choices of k elements from 365 with order and with repetition, while $|A|$ is number of possible choices of k elements from 365 with order and without repetition. This gives

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{\frac{365!}{(365-k)!}}{365^k} = \frac{365 \cdot 364 \cdots (365-k+1)}{365^k} = \prod_{i=0}^{k-1}\left(1 - \frac{i}{365}\right). \tag{2}$$

Taking $k = 23$ we then find that $\mathbb{P}(A) = 0.492\ldots$, giving $\mathbb{P}(A^c) = 0.507\ldots$ by Proposition 1.19(i).



1.4.2 Product rule

Selecting an ordered k elements with repetition from a set X can be viewed as selecting an element from X^k. Above we argued that if $|X| = n$ then $|X^k| = n^k$. This is a special case of the *product rule*: if $X_1, X_2, \ldots, X_k$ are finite sets, then $|X_1 \times \cdots \times X_k| = |X_1| \cdots |X_k| = \prod_{i=1}^k |X_i|$.

[12] Although it seems natural to assume a uniform distribution over the 365 days of the year, this is not completely accurate – births tend to peak in Summer between July and October.

1.4.3 A little on approximations

The factorial function $n! = n(n-1)\cdots 1$ tends to appear quite a lot in many areas of mathematics, and particularly in probability (e.g. see 1.4.1 Combinatorics above). It is sometimes helpful to have an approximation for this behaviour. Given two sequences (a_n) and (b_n) we write $a_n \sim b_n$ if a_n/b_n tends to 1 as $n \to \infty$.

Theorem 1.25 (Stirling's approximation)**.** *We have*

$$n! \sim \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n. \tag{3}$$

In other words, as n gets larger, the expression on the right in (3) becomes an excellent approximation for $n!$. Stirling's approximation is amazingly accurate, even for quite small values of n. For example, even when $n = 10$ the two terms in (3) differ by less than 1%.

The following inequality is another tool, which can be helpful in upper bounding products [13]:

$$\text{we have } 1 + x \leq e^x \text{ for all } x \in \mathbb{R}. \tag{4}$$

To see why, let's return to the birthday problem above. In (2) we obtained the expression

$$\mathbb{P}(A) = \prod_{i=0}^{k-1}\Big(1 - \frac{i}{365}\Big).$$

By (4), taking $x = -i/365$, we have $1 - i/365 \leq e^{-i/365}$. For $i \leq k \leq 365$, both terms here are positive, which gives

$$\mathbb{P}(A) = \prod_{i=0}^{k-1}\Big(1 - \frac{i}{365}\Big) \leq \prod_{i=0}^{k-1} e^{-i/365} = e^{-\sum_{i=0}^{k-1} i/365} = e^{-\frac{1}{365}\binom{k}{2}},$$

using the famous identity that $\sum_{i=0}^{\ell} i = \binom{\ell+1}{2}$ for all $\ell \in \mathbb{N}$. Putting in $k = 23$ this gives $\mathbb{P}(A) \leq e^{-253/365} = 0.499...$. The actual value of $\mathbb{P}(A) = 0.492...$ so we're pretty close to the truth! [14]

Most important takeaways from this section.

You should:

- be able to construct sample spaces and describe events in simple scenarios involving dice, coins and urns.
- know the definition and be able to use the properties of probability distributions.
- feel at ease with the uniform distribution.
- know the inclusion-exclusion formula for three events.
- be familiar with factorials and binomial coefficients and their applications to counting.

[13] If you would like to prove this, show that the function $g(x) = e^x - x - 1$ satisfies (i) $g(0) = 0$, (ii) g is decreasing on $(-\infty, 0]$, and (iii) g is increasing on $[0, \infty)$.

[14] The inequality (4) is more accurate for x close to 0, as in this example. In the other direction, it also holds that $e^{x-x^2} \leq 1 + x$ but for $x \in [-1/2, 0]$, and you could use this to find a *lower bound* for $\mathbb{P}(A)$ above. :)

Appendix: Proof of Lemma 1.8

Proof of Lemma 1.8 (non-examinable). We prove this when Ω is finite, so that $\Omega = \{\omega_1, \ldots, \omega_m\}$ for some $m \in \mathbb{N}$. The proof for the countable case is very similar.

To prove (a) first suppose that $\mathbb{P}$ is a probability distribution as in Definition 1.6. Given an event $A \subseteq \Omega$ with $|A| = \ell \leq m$, we can write $A = \{\omega_{i_1}, \ldots, \omega_{i_\ell}\}$. It follows that

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{j=1}^{\ell} \{\omega_{i_j}\}\right) = \sum_{j=1}^{\ell} \mathbb{P}\left(\{\omega_{i_j}\}\right) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}),$$

where the second equality follows from Definition 1.6 (iii)(a) and the final equality is just a relabelling. This completes the proof of (a).

To prove (b), suppose that we are given $p_\omega \geq 0$ for all $\omega \in \Omega$ with $\sum_{\omega \in \Omega} p_\omega = 1$ and set $\mathbb{P}(A) := \sum_{\omega \in A} p_\omega$ for all $A \subseteq \Omega$. Then $\mathbb{P}(\emptyset) = \sum_{\omega \in \emptyset} p_\omega = 0$ and $\mathbb{P}(\Omega) = \sum_{\omega \in \Omega} p_\omega = 1$ so Definition 1.6 (ii) holds. Also, given any event $A \subseteq \Omega$ we have

$$0 \leq \mathbb{P}(A) = \sum_{\omega \in A} p_\omega \leq \sum_{\omega \in \Omega} p_\omega = 1,$$

using that $p_\omega \geq 0$ for the first and second inequalities, the definition of $\mathbb{P}(A)$ for the first equality and that $\sum_{\omega \in \Omega} p_\omega = 1$ for the second. Thus Definition 1.6 (i) holds. Lastly, if $A_1, \ldots, A_k \subseteq \Omega$ are pairwise disjoint sets then

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{\omega \in \cup A_i} p_\omega = \sum_{i=1}^{k} \Big(\sum_{\omega \in A_i} p_\omega \Big) = \sum_{i=1}^{k} \mathbb{P}(A_i).$$

This proves Definition 1.6(iii)(a). To see (iii)(b) note, since Ω is finite, if $A_1, A_2, \ldots$ are pairwise disjoint then there is $k \in \mathbb{N}$ such that $A_i = \emptyset$ for all $i > k$. But then $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{k} A_i$ and $\mathbb{P}(A_i) = 0$ for all $i > k$ and so (iii)(b) follows from (iii)(a). □