Chapter 3

Language of Probabilities: Probability Models

Mathematics is a tool for reasoning. It is meant to help us see things clearly and refine our reasoning to the point that they are free from logical sloppiness. Probability theory is the mathematical language developed for reasoning about chance and randomness.

The basis of mathematical approach to science is the use of *mathematical models*. In order to reason about a natural phenomenon, we first build a mathematical model that captures the essence of that natural phenomenon. Once we have a mathematical model, we can use mathematical tools (such as algebra, calculus, combinatorics, etc.) to analyze the model and make predictions about it. Finally, we need to translate the result of our analysis and predictions back to see what they tell us about the original natural phenomenon.

In probability theory, the natural phenomena we wish to reason about are those which involve chance and randomness. Such phenomena can often be described as *random experiments*.

3.1 Examples

We start with a few simple examples of random experiments and their mathematical models. These first examples are so simple that writing down mathematical models for them may seem redundant. However, the examples are meant to help us familiarize ourselves with the language and the logic of using it. Once we become familiar with the language, we can apply it to more and more complex scenarios where its power becomes apparent.

Example 3.1.1 (Flipping a fair coin)**.** Consider the random experiment of flipping a fair coin. The experiment has two possible *outcomes*: either the coin comes up *heads* or it comes up *tails*. Let us represent these two outcomes with $\mathtt{H}$ (standing for heads) and $\mathtt{T}$ (standing for tails). The set

$$\Omega := \{\mathtt{H}, \mathtt{T}\}$$

is called the *sample space* for the experiment. Since the coin is fair, the two possible outcomes are equally likely. We express this by writing

$$\mathbb{P}(\mathtt{H}) = 1/2 \qquad \text{(read as: "The probability of } \mathtt{H} \text{ is } 1/2.\text{")}$$
$$\mathbb{P}(\mathtt{T}) = 1/2 \qquad \text{(read as: "The probability of } \mathtt{T} \text{ is } 1/2.\text{")}$$

The function $\mathbb{P}$ is the *measure of probabilities*: it tells us the probability of each possible outcome.

Ⓠ What exactly is the interpretation of $\mathbb{P}(\mathtt{H})$? How should we understand the statement "$\mathbb{P}(\mathtt{H}) = 1/2$"?

A good way to think about $\mathbb{P}(\mathtt{H})$ is as the "idealized frequency" of the occurrence of the outcome $\mathtt{H}$ when we repeat the experiment many many times. More specifically,

- If we repeat the experiment n times (where n is very large), then

$$\mathbb{P}(\mathtt{H}) \approx \frac{\langle \text{number of heads} \rangle}{n}\,.$$

Note that for any fixed n (say $n = 10000$), the frequency of heads need not be exactly $1/2$, but we expect it to be close to $1/2$. Furthermore, the larger we choose n, the closer we expect the frequency of heads to be to $1/2$. We

can imagine that, as the number of experiments goes to infinity, the frequency of heads converges to $1/2$. The probability $\mathbb{P}(\mathtt{H})$ refers to this limit value.

The sample space Ω and the measure of probabilities $\mathbb{P}$ together fully describe our mathematical model of the random experiment of flipping a fair coin. Alas, this random experiment is so simple we cannot do much with our model. ○

Example 3.1.2 (Flipping a biased coin)**.** Let us again consider the random experiment of flipping a coin, but this time suppose that the coin is *biased* (i.e., not fair), meaning that the outcomes are not equally likely. The possible outcomes are the same as before, so the sample space is again

$$\Omega := \{\mathtt{H}, \mathtt{T}\} .$$

What can we say about the measure of probabilities? The fact that the coin is biased means that the two possible outcomes `H` and `T` are not equally likely. Without further information about the amount of the bias, the best we can say is that

$$\mathbb{P}(\mathtt{H}) + \mathbb{P}(\mathtt{T}) = 1 .$$

As before, $\mathbb{P}(\mathtt{H})$ (the probability of `H`) and $\mathbb{P}(\mathtt{T})$ (the probability of `T`) are understood as "idealized frequencies" in many repeated experiments. The two probabilities add up to 1 because the frequencies of the two possible outcomes in repeated experiments add up to 1.

If we do not know the value of $\mathbb{P}(\mathtt{H})$, we can still treat it as a parameter. If we denote the value of $\mathbb{P}(\mathtt{H})$ by p, then $\mathbb{P}(\mathtt{T}) = 1 - p$. We now have a complete (parametric) model of the random experiment of flipping a coin. Note that based on the value of the parameter p, this model encompasses both the case in which the coin is biased and the case in which the coin is fair. Namely, $p = 1/2$ if the coin is fair and $p \neq 1/2$ if the coin is biased. If $p > 1/2$, the coin is biased towards showing `H` more often, and if $p < 1/2$, the coin is biased towards showing `T` more often. ○

Example 3.1.3 (Rolling a die)**.** After flipping a coin, the next simplest example of a random experiment is perhaps the experiment of rolling an ordinary 6-sided die. The experiment has 6 possible outcomes, depending on which side of the die is shown. The sample space (i.e., the set of all possible outcomes) is

$$\Omega := \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\} .$$

If the die is fair (i.e., not rigged), the possible outcomes are all equally likely, hence the appropriate measure of probabilities is

$$\mathbb{P}(\boxdot) = \mathbb{P}(\boxdot) = \cdots = \mathbb{P}(\boxdot) = 1/6 .$$

This completes the description of our model of a fair die.

Q What is the probability that the number shown on the die is even?

A1 $1/2$. This is rather obvious based on symmetry: the number shown on the die is either even (three outcomes) or odd (three outcomes), and there is no reason why one of the two possibilities should be more likely than the other.

While the above answer is valid, it does not really use our mathematical model. Here is how we can find the answer using the model:

A2 The *event* "the dies shows an even number" can be represented by a subset of the possible outcomes

$$A := \{\boxdot, \boxdot, \boxdot\} ,$$

which consists of those outcomes that realize the event (see Figure 3.1). The event happens if and only if

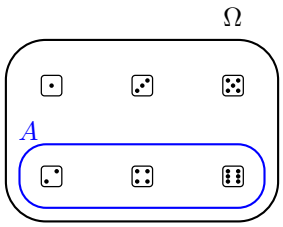


Figure 3.1: Venn diagram for the outcomes in Example 3.1.3

the outcome of the experiment falls within this set. The probability of this event is simply the sum of the probabilities of the individual outcomes within this event:

$$\mathbb{P}(A) = \mathbb{P}(\boxdot) + \mathbb{P}(\boxdot\!\boxdot) + \mathbb{P}(\boxdot\!\boxdot\!\boxdot) = 1/6 + 1/6 + 1/6 = 1/2 \ ,$$

which is consistent with the other answer. ○

Example 3.1.4 (Flipping a coin twice)**.** Consider the experiment of flipping a coin twice. The sample space consists of four outcomes

$$\Omega \coloneqq \{\mathtt{HH}, \mathtt{HT}, \mathtt{TH}, \mathtt{TT}\} \ ,$$

indicating the result of the first and the second flips. We do not assume the coin to be fair or unfair. Instead, we use a parameter p to denote the chance of getting a head in each single flip (as in Example 3.1.2).

Ⓠ What is the appropriate measure of probabilities in this example?

A

$$\mathbb{P}_p(\mathtt{HH}) = p^2 \ , \qquad \mathbb{P}_p(\mathtt{HT}) = p(1-p) \ ,$$
$$\mathbb{P}_p(\mathtt{TH}) = (1-p)p \ , \qquad \mathbb{P}_p(\mathtt{TT}) = (1-p)^2 \ .$$

Why? In order to justify $\mathbb{P}_p(\mathtt{HH}) = p^2$, imagine repeating this experiment many many times. In a fraction of about p of these repeated experiments, the first flip shows a head. Among those times in which the first flip shows a head, in a fraction of about p, the second flip also shows a head. Thus, in overall, in a fraction of about $p \times p$ of all the repeated experiments, both flips show heads. The other probability assignments can be argued similarly.

As concrete examples, if $p = 1/2$, we have

$$\mathbb{P}_{1/2}(\mathtt{HH}) = \mathbb{P}_{1/2}(\mathtt{HT}) = \mathbb{P}_{1/2}(\mathtt{TH}) = \mathbb{P}_{1/2}(\mathtt{TT}) = 1/4 \ ,$$

but if $p = 1/3$, we have

$$\mathbb{P}_{1/3}(\mathtt{HH}) = 1/9 \ , \qquad \mathbb{P}_{1/3}(\mathtt{HT}) = \mathbb{P}_{1/3}(\mathtt{TH}) = 2/9 \ , \qquad \mathbb{P}_{1/3}(\mathtt{TT}) = 4/9 \ .$$

This completes the description of the mathematical model. Note that we have used the subscript p in $\mathbb{P}_p$ to emphasize the dependence of the measure of probabilities on the parameter p.

Let us use the model to answer some questions.

Ⓠ What is the probability that in both flips, the coin shows the same side?

A The event of interest is $A \coloneqq \{\mathtt{HH}, \mathtt{TT}\}$, and its probability is

$$\mathbb{P}_p(A) = \mathbb{P}_p(\mathtt{HH}) + \mathbb{P}_p(\mathtt{TT}) = p^2 + (1-p)^2 \ .$$

Ⓠ What is the probability that in the first flip, the coin comes up heads?

A1 (Short answer) p. Remember that our original assumption was that each single time we flip the coin, the chance of it coming up heads is p. Whether we flip the coin a second time afterwards or not is irrelevant.

A2 (Answer using the model) The event of interest is $B \coloneqq \{\mathtt{HH}, \mathtt{HT}\}$, and its probability is

$$\mathbb{P}_p(A) = \mathbb{P}_p(\mathtt{HH}) + \mathbb{P}_p(\mathtt{HT}) = p^2 + p(1-p) = p(p+1-p) = p \ .$$

Fortunately, this is consistent with the previous short answer. Our model seems to be working well.

Now, consider the two events A and B mentioned above.

Ⓠ What is the probability that either A or B happens?

The event of interest in this case can be represented by $A \cup B$ (see Figure 3.2). This is the set of all outcomes which realize either A or B (or both).[1]

A1 Observe that $A \cup B = \{\mathtt{HH}, \mathtt{HT}, \mathtt{TT}\}$. Hence,

$$\mathbb{P}_p(A \cup B) = \mathbb{P}_p(\mathtt{HH}) + \mathbb{P}_p(\mathtt{HT}) + \mathbb{P}_p(\mathtt{TT}) = p^2 + p(1-p) + (1-p)^2 \ .$$

[1] In the language of mathematics, "or" is non-exclusive. So, when we say "A or B happens", we include the possibility that both A and B happen.

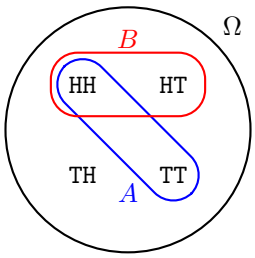


Figure 3.2: Venn diagram for the outcomes in Example 3.1.4

A2 Note that $A \cup B = \Omega \setminus \{\texttt{TH}\}$. Therefore,

$$\mathbb{P}_p(A \cup B) = 1 - \mathbb{P}_p(\texttt{TH}) = 1 - (1-p)p \ .$$

To justify the first equality, imagine we repeat the experiment n times, where n is very large. Let $N_n(A \cup B)$ denote the number of times in which the event $A \cup B$ occurs, and observe that $N_n(A \cup B) = n - N_n(\texttt{TH})$ (see Figure 3.2). Based on our interpretation of the probabilities as "idealized frequencies",

$$\mathbb{P}_p(A \cup B) \approx \frac{N_n(A \cup B)}{n} = \frac{n - N_n(\texttt{TH})}{n} = 1 - \frac{N_n(\texttt{TH})}{n} \approx 1 - \mathbb{P}_p(\texttt{TH}) \ .$$

In the limit of $n \to \infty$, the approximations become accurate and we get equality.

A3 A third way to compute the probability of $A \cup B$ is to write

$$\mathbb{P}_p(A \cup B) = \mathbb{P}_p(A) + \mathbb{P}_p(B) - \mathbb{P}_p(A \cap B) = \cancel{p^2} + (1-p)^2 + p - \cancel{p^2} = p + (1-p)^2 \ .$$

The first equality is the probabilistic form of the *inclusion-exclusion principle*. In order to justify it, again imagine that we repeat the experiment n times (n very large). Let $N_n(A \cup B)$ denote the number of times in which the event $A \cup B$ occurs. By the (standard) inclusion-exclusion principle, we have

$$N_n(A \cup B) = N_n(A) + N_n(B) - N_n(A \cap B) \ .$$

Hence,

$$\begin{aligned} \mathbb{P}_p(A \cup B) \approx \frac{N_n(A \cup B)}{n} &= \frac{N_n(A) + N_n(B) - N_n(A \cap B)}{n} \\ &= \frac{N_n(A)}{n} + \frac{N_n(B)}{n} - \frac{N_n(A \cap B)}{n} \approx \mathbb{P}_p(A) + \mathbb{P}_p(B) - \mathbb{P}_p(A \cap B) \ . \end{aligned}$$

○

Exercise. To answer the last question in the last example, we used three different reasonings and obtained three apparently different values for the probability $\mathbb{P}_p(A \cup B)$. Verify that the three answers are indeed the same.

The next example is somewhat more interesting. It contains a question whose answer cannot be easily guessed without the help of the mathematical model.

Example 3.1.5 (Flipping until a head comes up)**.** Consider the same coin we used in the previous example. As before, let p denote the chance of getting a head in one flip of the coin. We perform the following experiment: we repeat flipping the coin until the coin comes up heads for the first time.

This experiment has infinitely many possible outcomes, which can be represented by H, TH, TTH, and so on. The sample space is

$$\Omega := \{\texttt{H}, \texttt{TH}, \texttt{TTH}, \texttt{TTTH}, \ldots\} \ .$$

Q What is the appropriate measure of probabilities for these outcomes?

A

$$\begin{aligned} \mathbb{P}(\texttt{H}) &= p \\ \mathbb{P}(\texttt{TH}) &= (1-p)p \\ \mathbb{P}(\texttt{TTH}) &= (1-p)^2 p \\ \vdots \qquad &\qquad \vdots \\ \mathbb{P}(\texttt{T}^n\texttt{H}) &= (1-p)^n p \qquad \text{(for every } n \geq 0\text{)} \end{aligned}$$

These can be justified as in Example 3.1.4.

We now have a complete model of the random experiment: the sample space Ω and the measure of probabilities $\mathbb{P}$.

Q What is the probability that we get an even number of T's before the first H?

A1 The event of interest is

$$E \coloneqq \{\mathtt{H}, \mathtt{T}^2\mathtt{H}, \mathtt{T}^4\mathtt{H}, \ldots\} .$$

Thus,

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(\mathtt{H}) + \mathbb{P}(\mathtt{T}^2\mathtt{H}) + \mathbb{P}(\mathtt{T}^4\mathtt{H}) + \ldots \\ &= p + (1-p)^2 + (1-p)^4 p + \ldots .\end{aligned}$$

This is a geometric series[2] with starting term p and common ratio $(1-p)^2$. It converges to $p/[1-(1-p)^2]$. Hence,

$$\mathbb{P}(E) = \frac{p}{1-(1-p)^2} .$$

Note that in case the coin is fair (i.e., $p = 1/2$), the above answer gives $\mathbb{P}(E) = 2/3$. Is it surprising that the answer in this case is not $1/2$? The Venn diagram in Figure 3.3 may help you understand this. Another

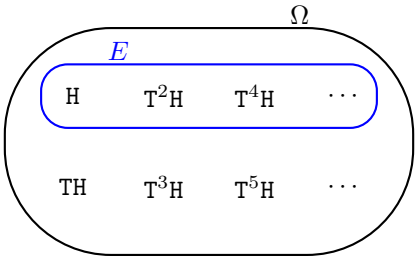


Figure 3.3: Venn diagram for the outcomes in Example 3.1.5

explanation is given by the following alternative answer.

A2 Consider the opposite event that "we get an odd number of T's before the first H." This event is described by the complement of E, that is, the set

$$E^{\mathsf{c}} = \Omega \setminus E = \{\mathtt{TH}, \mathtt{T}^3\mathtt{H}, \mathtt{T}^5\mathtt{H}, \ldots\} .$$

Observe that in order for E^{c} to happen, the first flip must show a T, and after the first flip, there must be an even number of T's before the first H. We can symbolically express this observation by writing $E^{\mathsf{c}} = \mathtt{T}E$. The chance of having a tail in the first flip is $1-p$, and the chance of realizing the event E starting from the second flip is $\mathbb{P}(E)$. Since the result of the first flip does not in any way affect what happens after the first flip, we obtain

$$\mathbb{P}(E^{\mathsf{c}}) = (1-p)\,\mathbb{P}(E) .$$

Now, remember that the probabilities of all the outcomes must add up to 1, hence

$$\mathbb{P}(E^{\mathsf{c}}) + \mathbb{P}(E) = 1 .$$

Solving the above two equations for $\mathbb{P}(E)$ and $\mathbb{P}(E^{\mathsf{c}})$, we obtain

$$\mathbb{P}(E) = \frac{1}{2-p} .$$

This solution will become clearer once we formulate the concept of the independence of events later in this chapter. ○

Exercise. Verify that the two answers in the last example are the same.

[2] A quick review of geometric series comes in Interlude 3.A.

Interlude 3.A (Review of geometric series)**.** A *geometric series* is the sum of an infinite number of terms in which the ratio between every two consecutive terms is the same. For instance, in the series

$$S = 3 + \frac{3}{2} + \frac{3}{4} + \frac{3}{8} + \cdots ,$$

each term is half the previous term. The following trick can be used to find the value the sum. Observe that

$$S = 3 + \frac{1}{2}\Big(\underbrace{3 + \frac{3}{2} + \frac{3}{8} + \cdots}_{S}\Big) .$$

Hence, $S = 3 + (1/2)S$. Solving this equation, we obtain $S = 6$.

More generally, a geometric series has the form

$$S = a + ar + ar^2 + ar^3 + \cdots = \sum_{k=0}^{\infty} ar^k , \qquad (\approx\approx)$$

where a is the starting term and r is the common ratio of the consecutive terms. Note that a geometric series does not always converge, for instance when $a = 1$ and $r = -2$, the geometric series

$$1 - 2 + 4 - 8 + 16 + \cdots$$

diverges. The series $(\approx\approx)$ converges if and only if $-1 < r < 1$.[3] When it converges, the value of the series $(\approx\approx)$ can be found as in the example above. ◇

3.2 Probability model of a random experiment

Having seen a few examples of probability models for random experiments, let us go through the general concepts of such models in a more systematic fashion.

Terminology.

- The set of all possible outcomes in a random experiment is called the sample space and is often denoted by the Greek letter Ω.
- An event regarding the experiment is described by a subset of the sample space.
- Two events are mutually exclusive (i.e., they cannot occur simultaneously) if and only if they are disjoint as sets (i.e., do not share elements).[4]
- A measure of probabilities, often denoted by $\mathbb{P}$, is a function that assigns numbers to events.

Interpretation. The probability of an event $A \subseteq \Omega$ is understood as the "idealized frequency" of the occurrence of A if we repeat the experiment many many times.[5] More specifically, suppose that we repeat the experiment n times, and denote by $N_n(A)$ the number of times in which A occurs. Then,

$$\mathbb{P}(A) \approx \frac{N_n(A)}{n}$$

when n is large, and the approximation becomes more and more accurate as $n \to \infty$.

Axioms for consistency. In order to have a meaningful and self-consistent model, the measure of probabilities must satisfy a number of axioms:

I. For each event $A \subseteq \Omega$, $0 \leq \mathbb{P}(A) \leq 1$.

II. $\mathbb{P}(\Omega) = 1$.

III. $\mathbb{P}(\varnothing) = 0$.

IV. (additivity) If A and B are *disjoint* events (i.e., $A \cap B = \varnothing$), then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

[3] The rigorous proof of this can be found in most calculus books.

[4] We will use the two terms *mutually exclusive* and *disjoint* interchangeably.

[5] Later on you may notice that, in some scenarios, this interpretation does not make much sense. Still, the "idealized frequency" interpretation is a good general idea to think about probabilities.

More generally,

V. (countable additivity) If $A_1, A_2, \ldots$ is a finite or infinite sequence of *pairwise disjoint* events, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots$$

Q What is the interpretation of each of the above axioms?

A The first three are easy to interpret. To understand the *additivity* axiom, imagine repeating the experiment n times, where n is very large. If A and B are disjoint, they cannot occur simultaneously, hence $N_n(A \cup B) = N_n(A) + N_n(B)$. Therefore,

$$\mathbb{P}(A \cup B) \approx \frac{N_n(A \cup B)}{n} = \frac{N_n(A) + N_n(B)}{n} = \frac{N_n(A)}{n} + \frac{N_n(B)}{n} \approx \mathbb{P}(A) + \mathbb{P}(B) \,,$$

where the approximations become accurate at the limit of $n \to \infty$. The *countable additivity* axiom is a natural extension of the *additivity* axiom.

Usual steps in reasoning about probabilities. In order to mathematically reason about a random experiment, we often take the following steps:

(1) Identify the sample space.

(2) Assign probabilities to those events for which the probabilities are self-evident.

(2.5) Verify the consistency of the probabilities.

(3) Extract information from the model by mathematical reasoning.

(4) Interpret the extracted information.

The first two steps concern building a mathematical model for the experiment, step (3) involves solving purely mathematical problems, and step (4) is about translating back the solutions of the mathematical problems to see what they tell us about the original random experiment. Step (2.5) is handled in more advanced courses in probability theory or measure theory. Recognizing the distinction between these steps often helps us see things more clearly.

The following is an example in which assigning self-evident probabilities and verifying the consistency of the model is far from trivial.

Example 3.2.1 (Drawing a number from an interval at random)**.** Consider the experiment of picking a number from the interval $[0, 1]$ completely at random. For instance, the experiment could involve spinning a wheel (similar to that in a game of roulette) as in Figure 3.4. The picked number would then be the angle between the reference mark and the arrow divided by 2π.

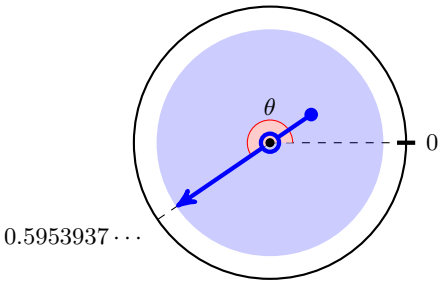


Figure 3.4: A spinning wheel device for picking a number between 0 and 1 at random. The device consists of a wheel (blue circle) which can rotate freely relative to a fixed frame (black circle). The wheel has an arrow drawn on it, which can be positioned at any angle θ relative to a reference mark on the frame. In order to pick a random number, we spin the wheel. The picked number is $\theta/(2\pi)$ once the wheel stops.

The sample space is the set of all real numbers between 0 and 1, that is,

$$\Omega := [0, 1] \,.$$

As in Example 3.1.5, this experiment has infinitely many possible outcomes. However, unlike that example, here the sample space is uncountable.[6]

[6]That is, the set is so large that its elements cannot be enumerated in a sequence. If you have not seen this before, you can check out Cantor's diagonal argument.

Ⓠ What is the appropriate measure of probabilities?

Ⓐ Identifying the measure of probabilities for this experiment is more complicated than in the previous examples, because there are way too many events. However, the probabilities of some events are self-evident. Consider the event that the picked number falls within an interval $[a, b]$, where $0 \le a \le b \le 1$. We expect that, if we repeat the experiment many many times, then the proportion of times in which the picked number falls within the interval $[a, b]$ will be roughly proportional to the length of the interval. Therefore, the probability of the event $[a, b]$ must be

$$\mathbb{P}([a,b]) = \frac{\langle \text{length of } [a,b] \rangle}{\langle \text{length of } [0,1] \rangle} = \frac{b-a}{1-0} = b - a \ .$$

In more advanced courses in probability theory or measure theory, it will be shown that these probabilities (for all the closed sub-intervals of $[0, 1]$) are consistent with one another, and that the probability of virtually any other event can be deduced from the probabilities of the closed sub-intervals of $[0, 1]$.

This is the end of the modeling stage. Let us see how the probabilities of some simple events can be deduced from the probabilities of intervals.

Ⓠ What is the probability that the picked number is is either $\ge 2/3$ or $\le 1/3$?

Ⓐ The event of interest is is $[0, 1/3] \cup [2/3, 1]$. Since the events $[0, 1/3]$ and $[2/3, 1]$ are disjoint (i.e., mutually exclusive), we have

$$\begin{aligned} \mathbb{P}\left(\text{the picked number is either } \ge 2/3 \text{ or } \le 1/3\right) &= \mathbb{P}\left([0,1/3] \cup [2/3,1]\right) \\ &= \mathbb{P}([0,1/3]) + \mathbb{P}([2/3,1]) = 1/3 + 1/3 = 2/3 \ . \end{aligned}$$

Ⓠ What is the probability that the picked number is $1/2$?

Ⓐ The event of interest is $[1/2, 1/2]$, hence

$$\mathbb{P}\left(\text{the picked number is } 1/2\right) = \mathbb{P}([1/2,1/2]) = 1/2 - 1/2 = 0 \ .$$

Ⓠ What is the probability that the picked number falls in the *open* interval (a, b)? (Assume: $0 \le a \le b \le 1$.)

Ⓐ We have $[a, b] = (a, b) \cup [a, a] \cup [b, b]$, hence

$$\mathbb{P}\left((a,b)\right) = \mathbb{P}([a,b]) - \mathbb{P}([a,a]) - \mathbb{P}([b,b]) = \mathbb{P}([a,b]) - 0 - 0 = b - a \ .$$

Note that, in the modeling stage, we could have declared the probability of the open intervals (a, b) as self-evident, too. However, that would have been redundant because, as the above computation shows, the probability of open intervals can be deduced from the probability of closed intervals.

Here is a somewhat more challenging question:

Ⓠ What is the probability that the picked number is rational?

Ⓐ

$$\mathbb{P}\left(\text{the picked number is rational}\right) = \mathbb{P}(\mathbb{Q} \cap [0,1]) = 0 \ .$$

To see why, recall that the set of rational numbers is countable. Let $q_1, q_2, q_3, \ldots$ be an enumeration of all the rational numbers in $[0, 1]$. As we saw above, for each k,

$$\mathbb{P}\left(\text{the picked number is } q_k\right) = \mathbb{P}([q_k, q_k]) = 0 \ .$$

By the countable additivity axiom,

$$\mathbb{P}(\mathbb{Q} \cap [0,1]) = \mathbb{P}(\{q_1, q_2, \ldots\}) = \mathbb{P}(q_1) + \mathbb{P}(q_2) + \cdots = 0 \ .$$

○

Some basic facts. The following facts about probability models have rather intuitive interpretations. Nevertheless, they can be deduced directly from the axioms of consistency via logical reasoning alone.

- (complement) For every event $A \subseteq \Omega$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (see Figure 3.5).

  Interpretation:
  - A^c is the event that A does not happens.

- (inclusion-exclusion principle) For every two events $A, B \subseteq \Omega$ (not necessarily disjoint),

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

  (see Figure 3.7a).

  Interpretation:
  - $A \cup B$ is the event that either A or B happens.
  - $A \cap B$ is the event that both A and B happen.

  Similarly, for every three events $A, B, C \subseteq \Omega$,

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

  (see Figure 3.7b).

  This can be extended to any finite number of events.

- If $A, B \subseteq \Omega$ are two events such that $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$ (see Figure 3.6).

  Interpretation:
  - $A \subseteq B$ means that whenever A happens, B also happens.

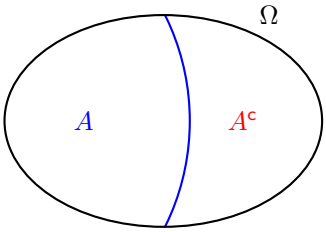


Figure 3.5: An event and its complement. Event A^c happens if and only if A does not happen.

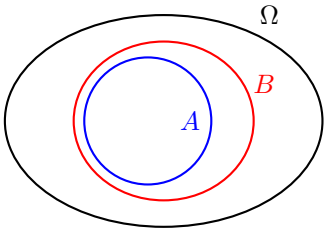
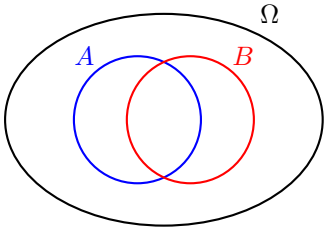
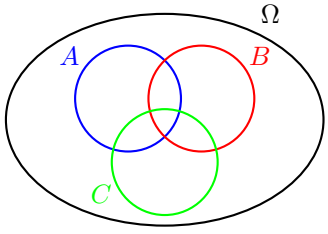


Figure 3.6: Two events. Whenever A happens, B also happens.



(a)



(b)

Figure 3.7: Two Venn diagrams

3.3 Conditional probabilities and independence

Suppose we learn some partial information about the outcome of a random experiment. How can we incorporate that information in our measure of probabilities? We now discuss how this is done in the language of probability models.

	female	male
CS	10	8
other	7	8

Table 3.1: Distribution of the students in a class (Example 3.3.1)

Example 3.3.1 (Students in a class)**.** Suppose there are 33 students in a class, out of which 17 are female and 16 male. Among the female students, 10 are computer science majors and 7 have other majors, whereas among the male students, 8 are computer science majors and 8 have other majors. This information is summarized in Table 3.1.

During office hours, the instructor hears a knock on the door of his/her office and knows that it must be a student from this class.

Ⓠ What is the chance that the student is a computer science major?

A The sample space, Ω, is the set of all 33 students in the class. Since there is no reason to believe otherwise, let us assume that all the students are equally likely to show up during office hours. So the measure of probabilities, $\mathbb{P}$, assign probability $1/|\Omega| = 1/33$ to each student. The event of interest, E, is the set of all 18 computer science majors in the class. Thus,

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{18}{33} .$$

Suppose now that the instructor opens the door and sees that the student is male.

Ⓠ After learning this additional information, what is the chance that the student is a computer science major?

A The information reduces the set of possibilities to the set of all 16 male students, which we call B. The revised probability is

$$\mathbb{P}(E \mid B) = \frac{|E \cap B|}{|B|} = \frac{8}{16} = 1/2 .$$

The revised probability $\mathbb{P}(E \mid B)$ is referred to as the *conditional probability of E given B*. ○

Example 3.3.2 (Flipping a coin twice)**.** Let us get back to the random experiment of flipping a coin twice. We do not assume the coin to be fair or unfair, so we use a parameter p to denote the chance of getting a head in one flip of the coin. The sample space for the experiment is

$$\Omega := \{\texttt{HH}, \texttt{HT}, \texttt{TH}, \texttt{TT}\} ,$$

and the measure of probabilities is given by

$$\mathbb{P}(\texttt{HH}) = p^2 , \qquad \mathbb{P}(\texttt{HT}) = p(1-p) ,$$
$$\mathbb{P}(\texttt{TH}) = (1-p)p , \qquad \mathbb{P}(\texttt{TT}) = (1-p)^2 .$$

Suppose someone performs this experiment in secret, and without revealing the exact outcome of the experiment, tells us that the coin has shown the same side in both flips.

Ⓠ Knowing this partial information about the outcome, what is the probability that the coin has come up heads in the first flip?

A The event of interest is described by the set

$$E := \{\texttt{HH}, \texttt{HT}\} \qquad \text{(the first flip shows a head)}$$

while the partial information says that the event

$$C := \{\texttt{HH}, \texttt{TT}\} \qquad \text{(both flips show the same side)}$$

has happened. How should we revise the probability of E once we learn that C has happened?

Our guide is again thinking of probabilities as "idealized frequencies". The revised probability of E knowing C should approximately be the frequency of the occurrence of E among those times in which C has happened, if we repeat the experiment many many times. More specifically, imagine repeating the experiment n times, where n is very large. Here is an example of a sequence of outcomes for these repeated experiments:

$$\begin{matrix} \overset{E}{\overset{C}{\text{H}}\ \text{H}} & \text{T}\ \text{H} & \text{T}\ \text{H} & \overset{E}{\overset{C}{\text{H}}\ \text{H}} & \overset{C}{\text{T}}\ \text{T} & \text{H}\ \text{T} & \text{H}\ \text{T} & \overset{C}{\text{T}}\ \text{T} & \overset{E}{\overset{C}{\text{H}}\ \text{H}} & \text{H}\ \text{T} & \cdots \end{matrix}$$

As before, let $N_n(C)$ denote the number of times in which C has occurred. Then, the revised probability is

$$\mathbb{P}(E \mid C) \approx \frac{N_n(E \cap C)}{N_n(C)} = \frac{N_n(E \cap C)/n}{N_n(C)/n} \approx \frac{\mathbb{P}(E \cap C)}{\mathbb{P}(C)} \,.$$

In the limit $n \to \infty$, the approximations become accurate. In the current example,

$$\mathbb{P}(E \mid C) = \frac{\mathbb{P}(E \cap C)}{\mathbb{P}(C)} = \frac{p^2}{p^2 + (1-p)^2} \,.$$

This is what we call the *conditional probability of E given C*. ○

Terminology. In general, if $E, C \subseteq \Omega$ are two events in a probability model, then the *conditional probability of E given C* is the quantity

$$\mathbb{P}(E \mid C) \coloneqq \frac{\mathbb{P}(E \cap C)}{\mathbb{P}(C)} \,.$$

This is how we should revise the probability of E once we learn that C has happened. Note that the conditional probability $\mathbb{P}(E \mid C)$ makes sense *only* if $\mathbb{P}(C) > 0$.

Some basic facts about conditional probabilities. Conditional probabilities satisfy properties analogous to unconditional probabilities. The following facts all have intuitive interpretations, and can be logically deduced from the definition of conditional probabilities and the basic axioms and facts about the measure of probabilities. Let $C \subseteq \Omega$ be an event with $\mathbb{P}(C) > 0$.

- For each event $A \subseteq \Omega$, $0 \leq \mathbb{P}(A \mid C) \leq 1$.
- $\mathbb{P}(\Omega \mid C) = 1$. (In fact, $\mathbb{P}(C \mid C) = 1$.)
- $\mathbb{P}(\varnothing \mid C) = 0$. (In fact, $\mathbb{P}(A \mid C) = 0$ whenever $A \cap C = \varnothing$.)
- (additivity) If A and B are *disjoint* events, then $\mathbb{P}\left(A \cup B \,\middle|\, C\right) = \mathbb{P}(A \mid C) + \mathbb{P}(B \mid C)$.
- (countable additivity) If $A_1, A_2, \ldots$ is a finite or infinite sequence of *pairwise disjoint* events, then
$$\mathbb{P}\left(A_1 \cup A_2 \cup \cdots \,\middle|\, C\right) = \mathbb{P}(A_1 \mid C) + \mathbb{P}(A_2 \mid C) + \cdots$$
- (complement) For every event $A \subseteq \Omega$, $\mathbb{P}(A^{\mathsf{c}} \mid C) = 1 - \mathbb{P}(A \mid C)$.
- (inclusion-exclusion principle) For every two events $A, B \subseteq \Omega$ (not necessarily disjoint),
$$\mathbb{P}(A \cup B \mid C) = \mathbb{P}(A \mid C) + \mathbb{P}(B \mid C) - \mathbb{P}(A \cap B \mid C) \,.$$
Similar identities hold for more than two events.
- If $A, B \subseteq \Omega$ are two events such that $A \subseteq B$, then $\mathbb{P}(A \mid C) \leq \mathbb{P}(B \mid C)$.

Let us now consider a variant of Example 3.3.1.

Example 3.3.3 (Students in a class; variant of Example 3.3.1)**.** Consider another class with 21 students, 9 of which are female and 12 male. As before, some students are computer science majors and some have other majors. The number of students in each category is given in Table 3.2.

Let us repeat the exercise we did in Example 3.3.1 for this new class.

	female	male
CS	3	4
other	6	8

Table 3.2: Distribution of the students in a class (Example 3.3.3)

Q What is the chance that the a random student from this class is a computer science major?

A As before, the model can be described with

- (sample space) Ω: set of all 21 students,
- (measure of probabilities) $\mathbb{P}(\omega) = 1/|\Omega| = 1/21$ for each outcome $\omega \in \Omega$ (i.e., all outcomes are equally likely).

We are interested in the probability of the event

- (event of interest) E: set of all 7 CS majors.

We have

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{7}{21} = 1/3 \,.$$

Suppose now that we have a partial information about the outcome.

Q What is the chance that the a random student from this class is a computer science major, if we know that the student is male?

A The partial information tells us that the event

- (condition) B: set of all 12 male students

has happened. We therefore want to know the conditional probability of E given B:

$$\mathbb{P}(E \mid B) = \frac{|E \cap B|}{|B|} = \frac{4}{12} = 1/3 \,.$$

Note that the two probabilities with or without the partial information turned out to be the same. This is not surprising: the ratio of CS to other is the same among both male and female students. In other words, whether a randomly chosen student from this class is a CS major or not is *statistically independent* of whether he/she is male or female. ○

Example 3.3.4 (Flipping a coin twice)**.** Let us get back to the random experiment of flipping a coin with parameter p twice. The sample space for the experiment is

$$\Omega := \{\texttt{HH}, \texttt{HT}, \texttt{TH}, \texttt{TT}\} \,,$$

and the measure of probabilities is given by

$$\mathbb{P}(\texttt{HH}) = p^2 \,, \qquad \mathbb{P}(\texttt{HT}) = p(1-p) \,,$$
$$\mathbb{P}(\texttt{TH}) = (1-p)p \,, \qquad \mathbb{P}(\texttt{TT}) = (1-p)^2 \,.$$

Consider the following two events:

- $A := \{\texttt{HH}, \texttt{HT}\}$ (in the 1st flip, the coin comes up heads),
- $B := \{\texttt{HH}, \texttt{TH}\}$ (in the 2nd flip, the coin comes up heads).

Intuitively, we know that the two events A and B are independent of each other: the result of the first flip does not in any way affect the result of the second flip. In the language of our model, this independence is reflected in the identity $\mathbb{P}(B \mid A) = \mathbb{P}(B)$, which can be directly verified:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{p^2}{p} = p = \mathbb{P}(B) \ .$$

Suggestion. Make sure you follow the 2nd and the 4th equalities. What are the values of $\mathbb{P}(A \cap B)$, $\mathbb{P}(A)$ and $\mathbb{P}(B)$? ○

In general, the idea that two events $A, B \subseteq \Omega$ in a probability model are independent can be conveniently expressed by the identity

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \ . \tag{$\perp_1$}$$

Although this identity appears asymmetric, the condition of independence is symmetric. Namely, since $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\,\mathbb{P}(B)$, the condition ($\perp_1$) is (essentially) equivalent to the condition

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \ . \tag{$\perp_2$}$$

The latter condition is however superior: in addition to treating A and B symmetrically, condition ($\perp_2$) makes sense even if $\mathbb{P}(B) = 0$. Note that when $\mathbb{P}(B) = 0$, the conditional probability $\mathbb{P}(A \mid B)$ is meaningless. For this reason, we adopt ($\perp_2$) as the definition of statistical independence.

Terminology. Two events $A, B \subseteq \Omega$ in a probability model are said to be (statistically) *independent* if

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

This means that whether A has happened or not does not provide us with any information about the occurrence of B, and vice versa.

Some basic facts about independence. The following facts all have intuitive interpretations, and can be logically deduced from the definition of conditional probabilities and the basic axioms and facts about the measure of probabilities. Let $A, B \subseteq \Omega$ be two events.

- If $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ then A and B are independent. The two conditions are equivalent if $\mathbb{P}(B) > 0$.
- If A and B are independent, then so are any of the pairs (A^{c}, B), (A, B^{c}) and $(A^{\mathsf{c}}, B^{\mathsf{c}})$. (In words, saying that A and B are independent is the same thing as saying that A^{c} and B are independent, and so on.)

Let us conclude this section with a more interesting example regarding conditional probabilities.

Example 3.3.5 (Flipping a coin 3 times)**.** Consider the experiment of flipping a coin 3 times in a row. As usual, we let p be the bias parameter of the coin, indicating the chance that, in one flip, the coin comes up heads. The sample space for this experiment is

$$\Omega \coloneqq \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{HTT}, \mathtt{THH}, \mathtt{THT}, \mathtt{TTH}, \mathtt{TTT}\}$$

which can more concisely be written as[7]

$$\Omega \coloneqq \{\mathtt{H}, \mathtt{T}\} \times \{\mathtt{H}, \mathtt{T}\} \times \{\mathtt{H}, \mathtt{T}\} \ .$$

The reasonable measure of probabilities for this experiment is given by

$$\mathbb{P}(\mathtt{HHH}) = p^3, \quad \mathbb{P}(\mathtt{HHT}) = p^2(1-p), \quad \cdots, \quad \mathbb{P}(\mathtt{TTT}) = (1-p)^3 \ .$$

Suppose someone performs this experiment in secret, and tells us that, in these three flips, the coin has come up tails twice and heads once, but does not tell us in which order.

Ⓠ Conditioned on this partial information, what is the probability that, in the first flip, the coin has come up heads?

[7] Here $\times$ denotes the *Cartesian product* of sets. Recall that the Cartesian product of two sets A and B, denoted by $A \times B$, is the set of all ordered pairs (a, b), where $a \in A$ and $b \in B$.

A The event of interest ("the 1st flip is a head") is

$$E \coloneqq \{\texttt{HHH}, \texttt{HHT}, \texttt{HTH}, \texttt{HTT}\} = \{\texttt{H}\} \times \{\texttt{H}, \texttt{T}\} \times \{\texttt{H}, \texttt{T}\} \ ,$$

while the condition ("two tails and one head") can be described by the event

$$C \coloneqq \{\texttt{HTT}, \texttt{THT}, \texttt{TTH}\} \ .$$

We have

$$\mathbb{P}(E \mid C) = \frac{\mathbb{P}(E \cap C)}{\mathbb{P}(C)} = \frac{p(1-p)^2}{\underbrace{p(1-p)^2 + (1-p)p(1-p) + (1-p)^2 p}_{3p(1-p)^2}} = 1/3 \ .$$

At first, it might come as a surprise that the answer does not depend on the parameter p. However, upon a second inspection, this makes sense: the condition C is realized by three possible outcomes, all of which are equally likely. Therefore, it is natural that, with the knowledge that C has happened, the chance of each of the three possible outcomes is $1/3$. ○

3.4 Counting without counting

In calculating probabilities, we often need to count things. For instance, we may need to know the total number of possible outcomes in an experiment, or the total number of outcomes realizing a certain event. If the number of objects is relatively small, we can count them literally one by one, although being systematic would help us avoid mistakes, and paying attention to symmetries could help us do the counting in smarter ways. However, in many scenarios which we encounter, the number of things we want to count is so large that counting them one by one is not feasible. To handle such cases, mathematicians have further developed the idea of being systematic and using symmetries into a range of techniques that allow us do the counting even in cases where the numbers are enormous.[8] In this section, we review some of such techniques that are pertinent to the kind of problems in probability theory we encounter in this course.

3.4.1 Review examples

Example 3.4.1 (Clothing combinations[9])**.** Suppose someone has 3 pairs of pants, 5 shirts and 7 pairs of socks.

Q What is the total number of ways in which he/she can dress?

A $3 \times 5 \times 7 = 105$. ○

In general:

Fact. If we have two collections of objects A (with m elements) and B (with n elements), then the number of pairs (a, b) where a is an object from A and b is an object from B is $m \times n$.[10]

Example 3.4.2 (Freshmen and mentors)**.** There are 10 freshmen students just entering the university. The university wants to assign 10 volunteered seniors as mentors for these freshmen, in such a way that each freshman has got a mentor and each senior is a mentor of only one freshman.

Q In how many ways can this assignment be done?

A Imagine assigning mentors to the freshmen one after another. There are 10 options for the first freshman. Once a mentor is assigned to the first freshman, there remains 9 options for the second freshman. Once a mentor is assigned to the second freshman, there remains 8 options for the third freshman, and so on. Thus, in total, there are $10 \times 9 \times 8 \times \cdots \times 1 = 10!$ ways we can assign mentors to the freshman. ○

In general:

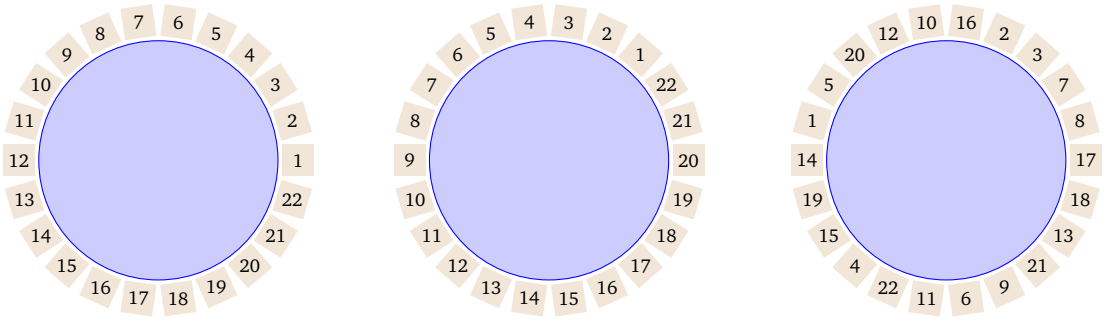
[8] There is an entire branch of mathematics known as *combinatorics* (more specifically, *combinatorial enumeration*) dedicated to counting.
[9] This example is taken from the book *Elementary Probability for Applications* by Rick Durrett.
[10] In concise mathematical notation: $|A \times B| = |A| \times |B|$.

Fact. The number of ways one can match n distinguishable objects of category 1 with n distinguishable objects of category 2 in a one-to-one fashion is $n!$.

Example 3.4.3 (Sitting around a circular table)**.** In a meeting of the Arab League, the representatives of the 22 member states sit around a circular table. Suppose we are tasked to assign the 22 seats around the table to these 22 representatives.

Q In how many possible ways can we assign the seats if we only care about the relative position of the representatives (in particular, who sits next to whom) and not the actual seats they take?

See Figure 3.8 for a clarification of the question.



(a) (b) (c)

Figure 3.8: Three possible assignments of seats to the 22 representatives (see Example 3.4.3). The representatives are numbered from 1 to 22 in a fixed way (say, Algeria is 1, Bahrain is 2, and so on). The two assignments (a) and (b) are equivalent (because the relative positions in both are the same) and should be counted as one. Assignment (c) is however different from the other two.

A1 There are 22! ways to assign exact seats to the 22 representatives. However, many of these assignments are equivalent. For instance, if we ask all representatives to shift one seat to the right, we get an equivalent assignment. In fact, each of these 22! assignments is equivalent to precisely 22 other assignments. Therefore, the total number of non-equivalent assignments is $22!/22 = 21!$.

A2 Pick one of the representatives arbitrarily, say the representative for Lebanon. Note that the exact position of the Lebanese representative is irrelevant, so we can ask him/her to sit in a seat of his/her choosing. Once the seat of the Lebanese representative is fixed, there remains 21 seats which need to be assigned to the remaining 21 representatives. This can be done in $21!$ different ways. ○

3.4.2 Drawing balls from a jar

Suppose we have a jar containing n distinguishable balls (see Figure 3.9, left). We would like to draw a sample of k balls from this jar. In how many ways can we do this? The description of the problem is rather ambiguous:

- Do we draw the balls with or without replacement? In other words, each time we draw a ball, do we put it back before drawing the next ball, or do we keep the drawn balls outside the jar?
- Does the order matter? Do we take note of which balls were chosen before which, or only which balls were chosen in overall?

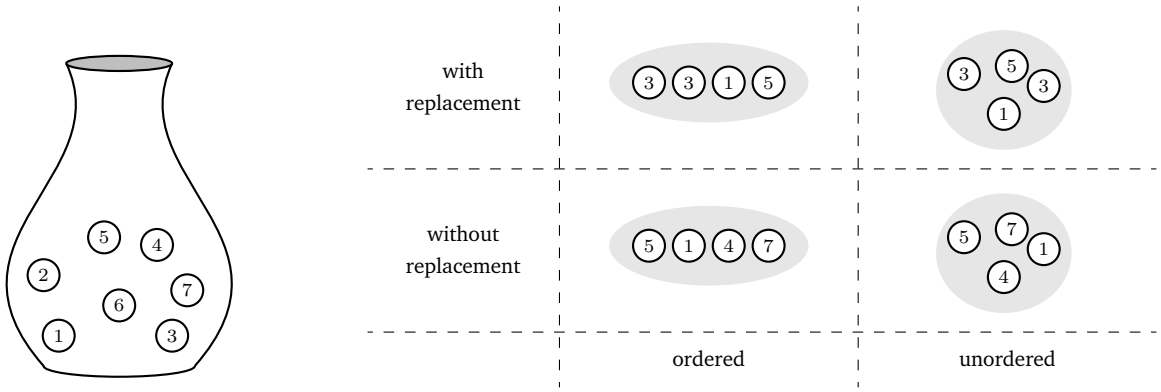
Depending on the answers to these two questions, we have four different possibilities for what we mean by "drawing a sample of k balls from a jar with n distinguishable balls". Figure 3.9 contains examples of these four types of sample in the case $n = 7$ and $k = 4$.

Various scenarios in probability and statistics can be likened to the act of drawing a samples from a jar containing balls.

Example 3.4.4 (Student committee)**.** A student club with 100 members needs an executive committee with 4 members: a chair, a vice-chair, a secretary and a treasurer. The act of choosing the committee members can be likened to the act of drawing a sample of 4 balls from a jar with 100 balls.

Q Which of the 4 types of sample do we have here?

A The sample is *without replacement*, because one student cannot be chosen for two different committee roles. The sample is also *ordered*, because the committee members all have distinct roles. ○



A jar with 7 distinguishable balls Four different types of samples

Figure 3.9: Drawing a sample of size 4 from a jar with 7 balls

In this subsection, we are concerned with counting the number of possible ways we can draw a sample of size k from a jar with n distinguishable balls, in each of these four cases. Let us consider the four cases one by one. Use Figure 3.9 as a guiding example.

Q What is the number of ways we can draw an ordered sample of size k with replacement from a jar with n distinguishable balls?

A n^k. Imagine drawing the k balls one after another, replacing each ball back in the jar after it is drawn. There are n choices for the first ball, n choices for the second ball and so on. In overall, there are $\underbrace{n \times n \times \cdots \times n}_{k \text{ times}} = n^k$ possibilities.

Q What is the number of ways we can draw an ordered sample of size k without replacement from a jar with n distinguishable balls?

A $\dfrac{n!}{(n-k)!}$. Again, imagine drawing the k balls one after another, but this time keeping the drawn balls outside the jar. There are n choices for the first ball, $n-1$ choices for the second ball (because one ball is already out), $n-2$ choices for the third ball, and so on. Hence, there are $n \times (n-1) \times (n-2) \times \cdots \times (n-k+1) = n!/(n-k)!$ possibilities in total.

Q What is the number of ways we can draw an unordered sample of size k without replacement from a jar with n distinguishable balls?

A $\binom{n}{k}$. Since we are considering samples without replacement, the drawn balls have to be distinct. Since we do not care about the order, we can draw the k balls at the same time. Thus, the sample is simply a subset of the balls in the jar which has k elements. How many k-element subsets does an n-element set have? That is precisely what $\binom{n}{k}$ stands for.

The remaining case is the trickiest.

Q What is the number of ways we can draw an unordered sample of size k with replacement from a jar with n distinguishable balls?

A Consider one such sample. Since the sample is drawn with replacement, the same ball can be drawn more than once. Since we do not care about the order, in order to specify the sample, we only need to tell how many times each of the balls was chosen. For instance, in the sample in Figure 3.9 (top-right), ball number 1 is chosen once, ball number 3 is chosen twice and ball number 5 is chosen once. The other balls are each chosen zero times. Let us represent this with the following string of characters $|$ and $\star$:

$$\star||\star\star||\star||$$

Q Can you guess how this string corresponds to the above sample?

A The four $\star$'s represent the drawn balls. The six $|$'s distinguish between the balls in the following manner:

- The star before the first $|$ represents the one time in which ball number 1 was drawn.

- There is no star between the first and the second $|$'s indicating that ball number 2 was not drawn.
- The two stars between the second and the third $|$'s represent the two times in which ball number 2 was drawn.
- ...
- There is no star after the sixth $|$ indicating that ball number 7 was not drawn.

In the same fashion, any unordered sample of k balls with replacement from a jar with n distinguishable balls can be represented using a string consisting of k copies of $\star$ and $n-1$ copies of $|$:

- The number of stars before the first $|$ indicates the number of times ball number 1 was drawn.
- The number of stars between the first and the second $|$ indicates the number of times ball number 2 was drawn.
- ...
- The number of stars after the last $|$ indicates the number of times ball number n was drawn.

This representation is faithful: each sample is represented by one and only one such string, and each such string corresponds to one and only one sample. Therefore, the number of samples of size k from a jar with n distinguishable balls is the same as the number of strings consisting of k copies of character $\star$ and $n-1$ copies of character $|$.

Q How many such strings are there?

A $\binom{n-1+k}{k}$. Each such string consists of $n-1+k$ characters, k of which are $\star$'s and the remaining are $|$'s. In order to identify one such string, it is enough to identify the positions of the k $\star$'s. There are $\binom{n-1+k}{k}$ choices for the position of the $\star$'s. Alternatively, we could identify each string with the position of its $n-1$ copies of $|$. There are $\binom{n-1+k}{n-1}$ choices for the position of $|$'s. However, note that $\binom{n-1+k}{k} = \binom{n-1+k}{n-1}$.

Table 3.3 summarizes the answers to the above counting problems.

	ordered	unordered
with replacement	n^k	$\binom{n-1+k}{k}$
without replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Table 3.3: Number of ways to draw a sample of k balls from a ball with n distinguishable balls

3.4.3 Allocating tokens to boxes

Suppose we have k tokens (or candies) which we would like to place into n distinguishable boxes. In how many ways can we do this? As in the previous subsection, the description is incomplete:

- Are the tokens distinguishable or all of the same type?
- Are we allowed to put more than one token in one box? In other words, is the allocation with or without exclusion? ("With exclusion" means each box can have at most one token; "without exclusion" means no restriction on the number of tokens per box.)

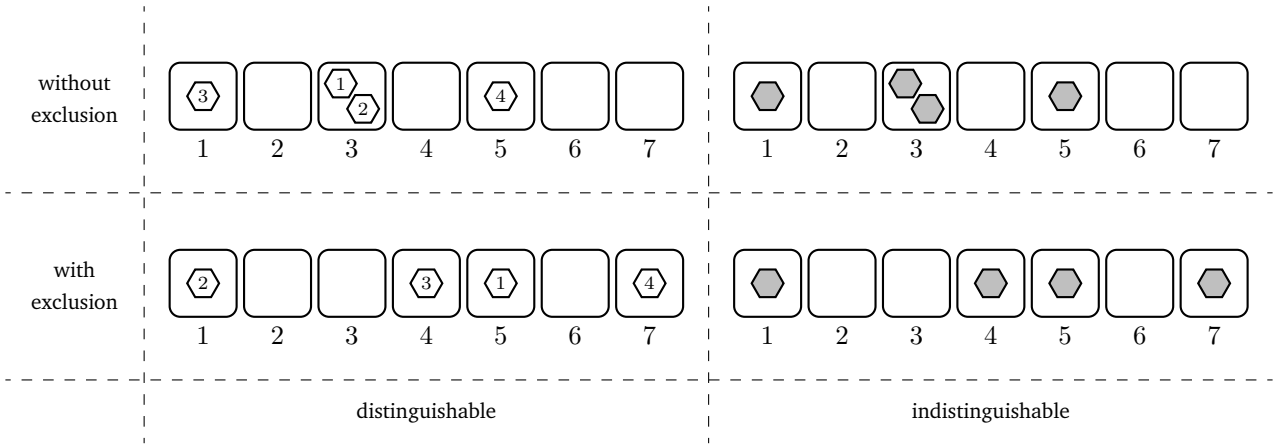
Depending on the answers to these two questions, we have four different variants of the problem. Figure 3.10 contains examples of these four types of allocations in the case $n = 7$ and $k = 4$.

There is a perfect analogy between the act of allocating k tokens to n distinguishable boxes and the act of drawing a sample of n balls from a jar with n distinguishable balls.

Q Can you see this analogy? (*Hint:* Compare Figures 3.9 and 3.10.)

A The boxes correspond to the balls in the jar. The tokens correspond to the balls drawn from the jar.

With similar reasonings as in the case of drawing a sample of balls from a jar (or simply using the above analogy), we can count the number of ways once can allocate k balls into n distinguishable boxes in each of the four different scenarios. Table 3.4 summarizes the answers to these counting problems.



Four different types of allocation

Figure 3.10: Allocating 4 tokens to 7 distinguishable boxes.

	distinguishable	indistinguishable
without exclusion	n^k	$\binom{n-1+k}{k}$
with exclusion	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Table 3.4: Number of ways to allocate k tokens to n distinguishable boxes

3.4.4 Using counting in probability theory

Let us now see some examples in which, in order to find the probability of an event, we need to count things.

Example 3.4.5 (Flipping a coin n times)**.** Consider the experiment of flipping a coin n times. As before, we let p be the bias parameter of the coin, indicating the chance that, in one flip, the coin comes up heads.

Ⓠ What is the probability that we get exactly k heads and $n-k$ tails?

🄰 It might be easier to start with a concrete example, say $n=5$ and $k=2$. In this case, the sample space is

$$\Omega := \{\texttt{HHHHH}, \texttt{HHHHT}, \ldots, \texttt{TTTTT}\}$$

which can be more concisely written as[11]

$$\Omega := \{\texttt{H},\texttt{T}\}^5 .$$

The measure of probabilities is given by

$$\mathbb{P}(\texttt{HHHHH}) = p^5 , \quad \mathbb{P}(\texttt{HHHHT}) = p^4(1-p) , \quad \ldots , \quad \mathbb{P}(\texttt{TTTTT}) = (1-p)^5 .$$

Let E denote the event that exactly 2 out of 5 flips show heads and 3 show tails. This event is realized precisely when one of the following 10 outcomes occurs:

```
HHTTT  THHTT  TTHHT  TTTHH
HTHTT  THTHT  TTHTH
HTTHT  THTTH
HTTTH
```

Observe that the probability of each of these 10 outcomes is $p^2(1-p)^3$. Thus,

$$\mathbb{P}(E) = \underbrace{p^2(1-p)^3 + p^2(1-p)^3 + \cdots + p^2(1-p)^3}_{10 \text{ times}} = 10p^2(1-p)^3 .$$

[11] The notation $\{\texttt{H},\texttt{T}\}^5$ refers to the Cartesian product of $\{\texttt{H},\texttt{T}\}$ five times with itself, that is, $\{\texttt{H},\texttt{T}\} \times \{\texttt{H},\texttt{T}\} \times \{\texttt{H},\texttt{T}\} \times \{\texttt{H},\texttt{T}\} \times \{\texttt{H},\texttt{T}\}$.

Let us now consider the general scenario (n and k fixed but arbitrary). The sample space is

$$\Omega := \{\mathtt{H}, \mathtt{T}\}^n \ .$$

The measure of probabilities can be described as follows. For any outcome $\omega \in \Omega$,

$$\mathbb{P}(\omega) := p^{\#\mathtt{H}(\omega)}(1-p)^{\#\mathtt{T}(\omega)} \ ,$$

where $\#\mathtt{H}(\omega)$ denotes the number of H's, and $\#\mathtt{T}(\omega)$ denotes the number of T's shown in ω.

Let E denote the event that exactly k out of n flips show heads and $n-k$ show tails. Note that the probability of each individual outcome that realizes E is $p^k(1-p)^{n-k}$. Therefore,

$$\mathbb{P}(E) = \underbrace{p^k(1-p)^{n-k} + p^k(1-p)^{n-k} + \cdots + p^k(1-p)^{n-k}}_{\text{number of elements in } E} = |E| \cdot p^k(1-p) \ .$$

Thus in order to find the probability of E, we need to count the number of individual outcomes in E.

Q What is the number of outcomes in E?

A $\binom{n}{k}$. Each outcome in E can be represented by a sequence consisting of k H's and $n-k$ T's. Each such sequence is uniquely identified by the positions of the k H's in the sequence. The number of ways we can choose k positions out of n positions is $\binom{n}{k}$.

We conclude that

$$\mathbb{P}(E) = \binom{n}{k} p^k(1-p)^k \ .$$

○

Exercise. In the above example, the number of heads can be any integer between 0 and n. These possibilities are mutually exclusive and together exhaust the entire sample space. Therefore, by the additivity of the measure of probabilities, their probabilities must add up to 1. Based on the above computation, we find that

$$\sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} = 1 \ .$$

Verify the latter identity directly based on the properties of the binomial coefficients. (*Hint:* Recall the Binomial Theorem.)

Example 3.4.6 (blue balls, red balls; with replacement)**.** Suppose that we have a jar with N balls, K of which are blue, and the remaining $N-K$ are red (see Figure 3.11 for the case $N=7$ and $K=4$). At random, we draw an ordered sample of n balls with replacement from the jar.

Q What is the probability of getting k blue balls and $n-k$ red balls?[12]

A1 Let us start by building a model for this experiment. Since we only care about the color of the balls, we can represent each possible outcome by the sequence of colors we see. For instance, if $n=5$, one possible outcome would be

$$\mathtt{BRBBR} \ ,$$

indicating that the first drawn ball is blue, the second is red, the third and the fourth are blue, and the fifth is red. The sample space can thus be expressed as[13]

$$\Omega := \{\mathtt{B}, \mathtt{R}\}^n \ .$$

Observe that each time we draw a ball, the chance of getting a blue ball is K/N and the chance of getting a red ball is $(N-K)/N$. Furthermore, since we replace each ball before drawing the next balls, the color of the ball in each draw does not in any way affect the color of the following balls. Thus, for each outcome $\omega \in \Omega$, the measure of probability should be

$$\mathbb{P}(\omega) := \left(\frac{K}{N}\right)^{\#\mathtt{B}(\omega)} \left(\frac{N-K}{N}\right)^{\#\mathtt{R}(\omega)} \ ,$$

[12] Let us emphasize that, in general, mathematical notation is *case-sensitive*: N and n do not refer to the same thing. In fact, mathematical notation is also *font-sensitive*: P and $\mathbb{P}$ could refer to two different things.

[13] As before, $\{\mathtt{B}, \mathtt{R}\}^n$ denote the set of all strings of length n with characters B and R.

where $\#\mathtt{B}(\omega)$ denotes the number of blue balls drawn, and $\#\mathtt{R}(\omega)$ denotes the number of red balls drawn in ω. This completes the description of the model.

Let E denote the event that we draw exactly k blue balls and $n-k$ red balls. Note that, according to our model, each individual outcome in E has probability $(K/N)^k((N-K)/N)^{n-k}$. Hence,

$$\mathbb{P}(E) = |E| \cdot \left(\frac{K}{N}\right)^k \left(\frac{N-K}{N}\right)^{n-k}.$$

It remains to count the number of elements in E.

Q What is the number of outcomes in E?

A $\binom{n}{k}$. Each outcome in E is represented by a sequence consisting of k B's and $n-k$ R's. Each such sequence is uniquely identified by the positions of the k B's in the sequence. The number of ways we can choose k positions out of n positions is $\binom{n}{k}$.

We conclude that

$$\mathbb{P}(E) = \binom{n}{k}\left(\frac{K}{N}\right)^k \left(\frac{N-K}{N}\right)^{n-k}.$$

A2 Observe that there is an analogy between this example and the example of flipping a coin n times in a row (Example 3.4.5). Drawing a ball from the jar can be likened to flipping a coin. Drawing a blue ball corresponds to getting a head, and drawing a red ball corresponds to getting a tail. In order for the analogy to be perfect, the parameter of the coin has to be $p := K/N$. Now, the event of drawing k blue balls and $n-k$ red balls corresponds to the event of getting k heads and $n-k$ tails. Thus, based on Example 3.4.5, we immediately see that

$$\mathbb{P}(E) = \binom{n}{k}p^k(1-p)^{n-k} = \binom{n}{k}\left(\frac{K}{N}\right)^k \left(\frac{N-K}{N}\right)^{n-k}.$$

○

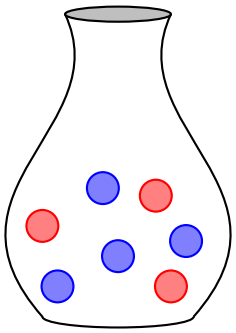


Figure 3.11: A jar with 4 blue balls and 3 red balls

Example 3.4.7 (blue balls, red balls; without replacement)**.** This is a variant of the previous example. Again, we have a jar with N balls, K of which are blue, and the remaining $N-K$ are red (Figure 3.11). This time, we draw a random unordered sample of n balls without replacement from the jar.

Q What is the probability of getting k blue balls and $n-k$ red balls? (Assume that $k \le K$ and $n-k \le N-K$ to make sure there are enough balls of each color.)

A The model can be described as follows

- (sample space) Ω: set of all n-element subsets of the N balls in the jar,
- (measure of probabilities) $\mathbb{P}(\omega) = 1/|\Omega|$ for each outcome $\omega \in \Omega$ (i.e., all outcomes are equally likely).

We are interested in the probability of the event

- (event of interest) E: set of all n-element subsets of the N balls in the jar containing k blue balls and $n-k$ red balls.

Since all outcomes are equally likely, we have

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

It remains to count the number of outcomes in Ω and in E.

Ⓠ What is the total number of outcomes in Ω?

A $\binom{N}{n}$. That is precisely the number of n-element subset of an N-element set.

Ⓠ What is the number of outcomes in E?

A $\binom{K}{k}\binom{N-K}{n-k}$. Each outcome in E contains k blue balls and $n-k$ red balls. There are $\binom{K}{k}$ ways to choose k blue balls from among the K blue balls in the jar. Similarly, there are $\binom{N-K}{n-k}$ ways to choose $n-k$ red balls from among the $N-K$ blue balls in the jar. Thus, the total number of possibilities is $\binom{K}{k}\binom{N-K}{n-k}$.

We conclude that

$$\mathbb{P}(E) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} .$$

○

Example 3.4.8 (Flipping a coin until k heads)**.** Let us again perform an experiment with a coin. As usual, the bias parameter of the coin is denoted by p. We repeat flipping the coin until we get k heads. For instance, if $k = 3$, a few possible outcomes of the experiment would be

```
H T T H H
H H H
T T T H T H T H
```

In this case, each possible outcome can be described by a string of characters `H` and `T` containing exactly 3 `H`'s and with the requirement that the last character is a `H`.

Ⓠ What is the probability that the k'th head comes up at the n'th flip?

For instance, if $k = 3$ and $n = 7$, the event we are interested in has happened if the outcome is

```
T T H T H T H
```

but has not happened if the outcome is

```
H T T H H .
```

A Let us start by describing a model for the experiment.

- (sample space) Ω: the set of all sequences of `H`'s and `T`'s that end with an `H`, and in which the number of `H`'s is k.
- (measure of probabilities) $\mathbb{P}$: For each possible outcome $\omega \in \Omega$,

$$\mathbb{P}(\omega) \coloneqq p^{\#\mathtt{H}(\omega)}(1-p)^{\#\mathtt{T}(\omega)} ,$$

  where $\#\mathtt{H}(\omega)$ denotes the number of `H`'s, and $\#\mathtt{T}(\omega)$ denotes the number of `H`'s in ω.

We are interested in the probability of the event

- (event of interest) E: set of all sequences in Ω which have length n.

Note that each individual outcome in E has probability $p^k(1-p)^{n-k}$ (k heads and $n-k$ tails). Therefore,

$$\mathbb{P}(E) = |E| \cdot p^k(1-p)^{n-k} .$$

Thus, it remains to count the number of outcomes in E.

Q What is the number of outcomes in E?

A $\binom{n-1}{k-1}$. Observe that an outcome is in E, if and only if

- The n'th flip shows a H,
- Among the first $n-1$ flips, there are $k-1$ H's.

The number of ways we can choose $k-1$ positions out of $n-1$ positions is $\binom{n-1}{k-1}$.

We conclude that

$$\mathbb{P}(E) = \binom{n-1}{k-1} p^k (1-p)^{n-k} .$$

○

Curiosity Exercise. In the above example, the k'th head will come at one of the times $n = k$, or $n = k+1$, or $n = k+2$, or .... These possibilities are mutually exclusive and together exhaust the entire sample space. Therefore, by the countable additivity of the measure of probabilities, their probabilities must add up to 1. Based on the above computation, we find that

$$\sum_{n=k}^{\infty} \binom{n-1}{k-1} p^k (1-p)^{n-k} = 1 .$$

Can you verify the latter identity directly based on the properties of the binomial coefficients? (*Hint:* Check out the Negative Binomial Theorem.)

3.5 More on conditional probabilities

3.5.1 Chain rule and principle of total probability

Recall that, the conditional probability of an event A given another event B is

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

Rearranging this identity, we get

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

In words: the probability that both A and B happen is the same as the probability that B happens times the probability that A happens given that B has happened. Recalling the interpretation of probabilities as "idealized frequencies" in repeated experiments, the latter identity has a simple interpretation.

Q What is the interpretation of the above identity in terms of "idealized frequencies"?

A Suppose we repeat the experiment n times, where n is very large. As usual, let $N_n(A \cap B)$ denote the number of times in which $A \cap B$ happens. Then, the above identity is simply the idealized version of the following

$$\mathbb{P}(A \cap B) \approx \frac{N_n(A \cap B)}{n} = \frac{N_n(B)}{n} \cdot \frac{N_n(A \cap B)}{N_n(B)} \approx \mathbb{P}(B)\, \mathbb{P}(A \mid B)$$

in the limit $n \to \infty$. In words, in many many repeated experiments, "the fraction of times in which both A and B happen" is the product of "the fraction of times in which B happens" and "the fraction of times in which A happens among those times in which B happens."

The identity ($\circledcirc$) is known as the *chain rule* of conditional probabilities. It can be extended to more than two events. For instance, in the case of three events A, B, C (see Figure 3.12), we have

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(C)\, \mathbb{P}(B \mid C)\, \mathbb{P}\left(A \,\middle|\, (B \cap C)\right) . \qquad (\circledcirc\circledcirc)$$

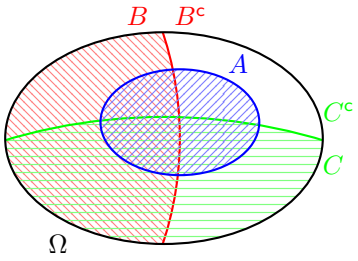


Figure 3.12: Venn diagram for three events A, B and C

Chain rule of conditional probabilities. If $A_1, A_2, \ldots, A_n$ are events in Ω, then

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}\left(A_3 \,\middle|\, (A_1 \cap A_2)\right) \cdots \mathbb{P}\left(A_n \,\middle|\, \bigcap_{k=1}^{n-1} A_k\right). \qquad (\text{oooooo})$$

Let us see an example of using the chain rule in action.

Example 3.5.1 (Boxes and balls)**.** Suppose we have two boxes containing blue and red balls (Figure 3.13).

- In box #1, there are k_1 blue balls and ℓ_1 red balls,
- In box #2, there are k_2 blue balls and ℓ_2 red balls.

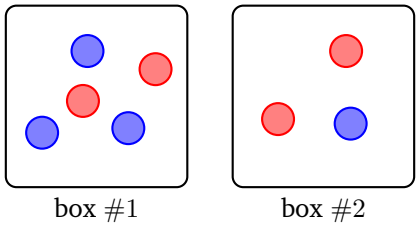


Figure 3.13: Two boxes with blue and red balls in them (see Example 3.5.1)

We perform the following two-stage random experiment:

I. We first choose one of the boxes at random.

II. We then draw a ball from the chosen box at random.

Q What is the chance that the picked ball is blue?

Before answering the question, let us emphasize that, depending on the values of k_1, ℓ_1, k_2, ℓ_2, the balls may *not* be equally likely to be drawn. For instance, if box #1 has only one blue ball, and box #2 has one blue ball and one red ball (i.e., $k_1 = 1$, $\ell_1 = 0$, $k_1 = 1$ and $\ell_1 = 1$), then it is clear that the chance that the blue ball from box #1 is picked is $1/2$, whereas the chance that the blue ball from box #2 is picked is $1/4$.

A An intuitive way to find the probability is to consider the tree of possibilities as in Figure 3.14. At the beginning, each of the two boxes has $1/2$ probability to be chosen. If box #1 is chosen, then the chance of drawing a blue ball is $\frac{k_1}{k_1+\ell_1}$. If box #2 is chosen, then the chance of drawing a blue ball is $\frac{k_2}{k_2+\ell_2}$. Therefore,

$$\mathbb{P}\left(\text{a blue ball is drawn}\right) = \frac{1}{2} \cdot \frac{k_1}{k_1+\ell_1} + \frac{1}{2} \cdot \frac{k_2}{k_2+\ell_2}.$$

Let us take a moment to contemplate on what the above computation amounts to in the language of probability models. This will reveal a general trick that can then be used in other scenarios.

As the sample space, we can choose[14]

$$\Omega := \left\{\mathtt{1B}, \mathtt{1R}, \mathtt{2B}, \mathtt{2R}\right\}.$$

The event that the ball is blue is $E := \{\mathtt{1B}, \mathtt{2B}\}$. In order to find the probability of E, we divide the possibilities based on whether box #1 is chosen or box #2. To be specific, let $C := \{\mathtt{1B}, \mathtt{1R}\}$ be the event that box #1 is chosen, and note that C^{c} is simply the event that box #2 is chosen. Observe that

$$E = (E \cap C) \cup (E \cap C^{\mathsf{c}})$$

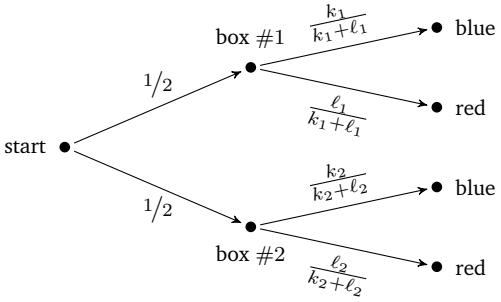


Figure 3.14: The tree of possibilities for the experiment in Example 3.5.1

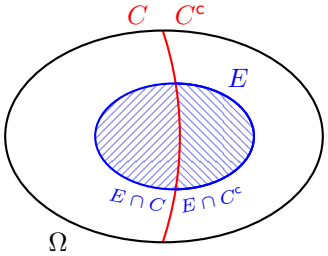


Figure 3.15: An event E can be written as $E = (E \cap C) \cup (E \cap C^{\mathsf{c}})$.

(see Figure 3.15). Since $E \cap C$ and $E \cap C^{\mathsf{c}}$ are disjoint, their probabilities add up to the probability of E, that is,

$$\mathbb{P}(E) = \mathbb{P}(E \cap C) + \mathbb{P}(E \cap C^{\mathsf{c}}) \, .$$

Now, using the chain rule, we can write

$$\mathbb{P}(E \cap C) = \mathbb{P}(C)\,\mathbb{P}(E \mid C) \qquad \text{and} \qquad \mathbb{P}(E \cap C^{\mathsf{c}}) = \mathbb{P}(C^{\mathsf{c}})\,\mathbb{P}(E \mid C^{\mathsf{c}}) \, .$$

Note that, based on the description of the experiment,[15]

$$\begin{aligned}
\mathbb{P}(C) &\coloneqq \big\langle \text{chance that box \#1 is chosen} \big\rangle = 1/2 \, , \\
\mathbb{P}(C^{\mathsf{c}}) &\coloneqq \big\langle \text{chance that box \#2 is chosen} \big\rangle = 1/2 \, , \\
\mathbb{P}(E \mid C) &\coloneqq \big\langle \text{chance of drawing a blue ball given that box \#1 is chosen} \big\rangle = \frac{k_1}{k_1 + \ell_1} \, , \\
\mathbb{P}(E \mid C^{\mathsf{c}}) &\coloneqq \big\langle \text{chance of drawing a blue ball given that box \#2 is chosen} \big\rangle = \frac{k_2}{k_2 + \ell_2} \, .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}(E) &= \mathbb{P}(E \cap C) + \mathbb{P}(E \cap C^{\mathsf{c}}) \\
&= \mathbb{P}(C)\,\mathbb{P}(E \mid C) + \mathbb{P}(C^{\mathsf{c}})\,\mathbb{P}(E \mid C^{\mathsf{c}}) \\
&= \frac{1}{2} \cdot \frac{k_1}{k_1 + \ell_1} + \frac{1}{2} \cdot \frac{k_2}{k_2 + \ell_2} \, ,
\end{aligned}$$

which is the same as what we obtained earlier. ○

The idea of breaking down the possibilities based on whether an event C has happened or not can be generalized as follows.

Principle of total probability. Let A be an event, and suppose that $F_1, F_2, F_3, \ldots$ is a finite or countably infinite collection of events that partition the sample space (see Figure 3.16).[16] Then,

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}(A \cap F_1) + \mathbb{P}(A \cap F_2) + \mathbb{P}(A \cap F_3) + \cdots \\
&= \mathbb{P}(F_1)\,\mathbb{P}(A \mid F_1) + \mathbb{P}(F_2)\,\mathbb{P}(A \mid F_2) + \mathbb{P}(F_3)\,\mathbb{P}(A \mid F_3) + \cdots \, .
\end{aligned}$$

This identity can sometimes help us calculate $\mathbb{P}(A)$.

[14] Alternatively, we could choose a more refined sample space by keeping track of which blue or red ball from the chosen box is drawn.

[15] These four equations implicitly describe the measure of probabilities for our model. They should be understood as part of the modeling rather than mathematical reasoning.

[16] Mathematically, saying that $F_1, F_2, F_3, \ldots$ partition Ω means that $F_1, F_2, F_3, \ldots$ are disjoint and their union is Ω. The interpretation of this is that the events $F_1, F_2, F_3, \ldots$ are mutually exclusive (i.e., no two of them can occur simultaneously) and together exhaust all the possibilities in Ω.

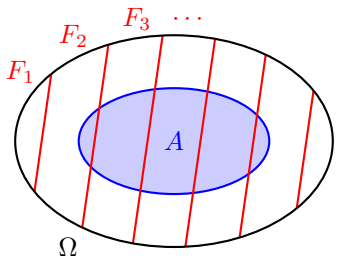


Figure 3.16: The events $F_1, F_2, F_3, \ldots$ form a (finite or countably infinite) partition of the sample space.

Example 3.5.2 (Flipping a coin until two heads)**.** We conduct the following experiment using a coin with bias parameter p: we repeat flipping the coin until we get 2 heads.

Ⓠ What is the probability that the second head comes up right after the first head?

Suggestion. Before trying to find the probability systematically, can you make a guess about what the answer should be?

A The idea is to break down the set of possibilities based on the time in which the first head comes up (see Figure 3.17). Namely, let Ω denote the sample space. Consider the following events:

- $E \coloneqq \langle$the 2nd head comes up right after the 1st head$\rangle$,
- $C_1 \coloneqq \langle$the 1st head come up in the 1st flip$\rangle$,
- $C_2 \coloneqq \langle$the 1st head come up in the 2nd flip$\rangle$,
- $C_3 \coloneqq \langle$the 1st head come up in the 3rd flip$\rangle$,
- $\ldots$
- $C_n \coloneqq \langle$the 1st head come up in the n'th flip$\rangle$,
- $\ldots$

We want to find $\mathbb{P}(E)$. Note that $C_1, C_2, C_3, \ldots$ are mutually exclusive and together exhaust all the possibilities in Ω. In other words, they partition Ω. Therefore,

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(E \cap C_1) + \mathbb{P}(E \cap C_2) + \mathbb{P}(E \cap C_3) + \cdots \\ &= \mathbb{P}(C_1)\,\mathbb{P}(E \mid C_1) + \mathbb{P}(C_2)\,\mathbb{P}(E \mid C_2) + \mathbb{P}(C_3)\,\mathbb{P}(E \mid C_3) + \cdots\ .\end{aligned}$$

Ⓠ What is the value of $\mathbb{P}(E \mid C_n)$ for each $n = 1, 2, 3, \ldots$?

A Note that $\mathbb{P}(E \mid C_n)$ is simply the chance of getting a head at the $(n+1)$st flip given that the first head has come up at the n'th flip. Since the results of the first n flips in no way affect the result of the $(n+1)$st flip, we have $\mathbb{P}(E \mid C_n) = p$. This is the case for each $n = 1, 2, 3, \ldots$.[17]

Therefore,

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(C_1)p + \mathbb{P}(C_2)p + \mathbb{P}(C_3)p + \cdots \\ &= \big[\underbrace{\mathbb{P}(C_1) + \mathbb{P}(C_2) + \mathbb{P}(C_3) + \cdots}_{1}\big]p = p\ .\end{aligned}$$

Note that, although we could find the value of $\mathbb{P}(C_n)$ for each $n = 1, 2, \ldots$, we did not need to do so. That $\mathbb{P}(C_1) + \mathbb{P}(C_2) + \mathbb{P}(C_3) + \cdots = 1$ follows from the fact that $C_1, C_2, C_3, \ldots$ partition Ω, and hence their probabilities must add up to 1.

Is the above answer consistent with your initial guess? ○

[17] Note again that this is a mere translation of the description of the model into the language of the model. It should be understood as part of the modeling rather than mathematical reasoning.

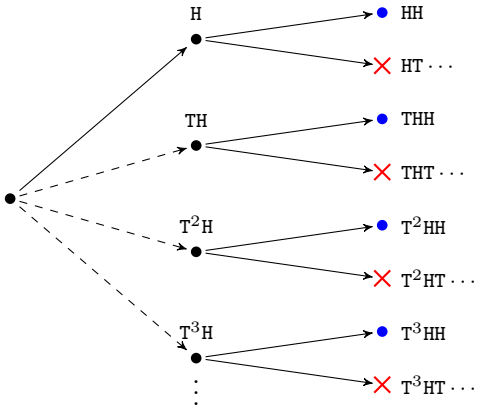


Figure 3.17: The tree of possibilities for the experiment in Example 3.5.2. The blue nodes indicate the possibilities in which E happens. The red crosses indicate the branches of possibilities in which E does not happen.

3.5.2 Bayes' rule

Example 3.5.3 (Mammograms[18])**.** Approximately 1% of women aged 40–50 have breast cancer. Physicians use mammogram tests to detect breast cancer. The test is however not perfect.

- For a woman with breast cancer, the chance of getting a positive test is 92%, while there is an 8% chance of a negative test (i.e., a false negative).
- For a woman without breast cancer, the chance of getting a negative test is 90%, while there is a 10% chance of a positive test (i.e., a false positive).

This information is summarized in Table 3.5.

	positive test	negative test	
with cancer	92%	8%	chance of false negative
without cancer	10%	90%	chance of false positive

Table 3.5: Chances of positive and negative tests for women with or without breast cancer (Example 3.5.3)

Suppose that a woman aged 40–50 has received a positive result on her mammogram test.

Ⓠ What is the chance that she has breast cancer?

A Consider the following two events regarding a random woman aged 40–50 who undergoes the mammogram test:

- $A \coloneqq \langle$the test is positive$\rangle$,
- $B \coloneqq \langle$the woman has breast cancer$\rangle$.

We are looking for $\mathbb{P}(B \mid A)$.

Ⓠ What is the available information in terms of A and B?

A

$$\mathbb{P}(B) = 0.01\,, \qquad \mathbb{P}(A \mid B) = 0.92\,, \qquad \mathbb{P}(A^{\mathsf{c}} \mid B) = 0.08\,,$$
$$\mathbb{P}(A \mid B^{\mathsf{c}}) = 0.10\,, \qquad \mathbb{P}(A^{\mathsf{c}} \mid B^{\mathsf{c}}) = 0.90\,.$$

[18] This example is taken from the book *Elementary Probability for Applications* by Rick Durrett.

The conditional probability $\mathbb{P}(B \mid A)$ can be found as follows:

$$\begin{aligned}
\mathbb{P}(B \mid A) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} && \text{(definition)} \\
&= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\mathsf{c}})} && \text{(principle of total probability)} \\
&= \frac{\mathbb{P}(B)\,\mathbb{P}(A \mid B)}{\mathbb{P}(B)\,\mathbb{P}(A \mid B) + \mathbb{P}(B^{\mathsf{c}})\,\mathbb{P}(A \mid B^{\mathsf{c}})} && \text{(chain rule)} \\
&= \frac{0.01 \times 0.92}{0.01 \times 0.92 + 0.99 \times 0.1} \\
&= \frac{92}{1082} \approx 8.50\% \,.
\end{aligned}$$

It might come as a surprise that the above chance is so small (less than 10%). The intuitive reason behind this is that there are many more healthy women with positive tests than there are women with cancer and positive tests. In particular, on average

- only 100 out of 10000 women have breast cancer, of which 92 receive positive test results, whereas
- 9900 out of 10000 women do not have breast cancer, of which still 990 receive positive test results.

Thus, among women who receive positive test results, only a fraction of $92/(990+92) \approx 8.50\%$ actually do have cancer. ○

The trick we used in the above example to calculate $\mathbb{P}(B \mid A)$ knowing the conditional probabilities of the opposite type is known as *Bayes' rule*.[19]

Bayes' rule. For every two events E and F,

$$\mathbb{P}(F \mid E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(E)} = \frac{\mathbb{P}(F)\,\mathbb{P}(E \mid F)}{\mathbb{P}(E)} \,.$$

This is useful trick for finding $\mathbb{P}(F \mid E)$ if what is available to us is the opposite conditional probabilities $\mathbb{P}(E \mid F)$ and $\mathbb{P}(E \mid F^{\mathsf{c}})$. The denominator $\mathbb{P}(E)$ can potentially be found using the principle of total probability, by writing it as $\mathbb{P}(F)\,\mathbb{P}(E \mid F) + \mathbb{P}(F^{\mathsf{c}})\,\mathbb{P}(E \mid F^{\mathsf{c}})$.

Example 3.5.4 (Three factories[20])**.** The computer chips used by an engineering company are made by three factories: 10% by factory $\#1$, 30% by factory $\#2$, and 60% by factory $\#3$. The chips are occasionally defective: 8% of the chips made by factory $\#1$, 1% of those made by factory $\#2$, and 2% of those made by factory $\#3$ are defective. This information is summarized in Table 3.6.

	share of chips produced	fraction of defective chips
factory $\#1$	10%	8%
factory $\#2$	30%	1%
factory $\#3$	60%	2%

Table 3.6: Computer chips used by an engineering company (Example 3.5.4)

Suppose that an engineer from the company encounters a defective chip.

Ⓠ What is the chance that the chip is from factory $\#1$?

Given the higher rate of defects among the chips made by factory $\#1$, it is tempting to jump into conclusion and blame factory $\#1$ for the observed defect. However, the other two factories provide a bigger portion of the chips used by the company, so we need to be careful.

[19] Named after statistician Thomas Bayes (1701–1761).
[20] This example is taken from the book *Elementary Probability for Applications* by Rick Durrett.

A Consider the following events regarding a random chip from the company:

- $D \coloneqq \langle$the chip is defective$\rangle$,
- $F_1 \coloneqq \langle$the chip is made in factory $\#1\rangle$,
- $F_2 \coloneqq \langle$the chip is made in factory $\#2\rangle$,
- $F_3 \coloneqq \langle$the chip is made in factory $\#3\rangle$.

We are looking for $\mathbb{P}(F_1 \mid D)$, while the information summarized in Table 3.6 contains the values of $\mathbb{P}(F_1)$, $\mathbb{P}(F_2)$, $\mathbb{P}(F_3)$, $\mathbb{P}(D \mid F_1)$, $\mathbb{P}(D \mid F_2)$ and $\mathbb{P}(D \mid F_3)$.

Note that

$$\mathbb{P}(F_1 \mid D) = \frac{\mathbb{P}(F_1 \cap D)}{\mathbb{P}(D)} = \frac{\mathbb{P}(F_1 \cap D)}{\mathbb{P}(F_1 \cap D) + \mathbb{P}(F_2 \cap D) + \mathbb{P}(F_2 \cap D)} \,.$$

The three terms involved can be calculated using the chain rule as follows:

$$\begin{aligned}
\mathbb{P}(F_1 \cap D) &= \mathbb{P}(F_1)\,\mathbb{P}(D \mid F_1) = 0.1 \times 0.08 = 0.008 \,, \\
\mathbb{P}(F_2 \cap D) &= \mathbb{P}(F_2)\,\mathbb{P}(D \mid F_2) = 0.3 \times 0.01 = 0.003 \,, \\
\mathbb{P}(F_3 \cap D) &= \mathbb{P}(F_3)\,\mathbb{P}(D \mid F_3) = 0.6 \times 0.02 = 0.012 \,.
\end{aligned}$$

Therefore,

$$\mathbb{P}(F_1 \mid D) = \frac{8}{8+3+12} = \frac{8}{23} \approx 34.8\% \,.$$

Similarly,

$$\begin{aligned}
\mathbb{P}(F_2 \mid D) &= \frac{3}{8+3+12} = \frac{3}{23} \approx 13.0\% \,, \\
\mathbb{P}(F_3 \mid D) &= \frac{12}{8+3+12} = \frac{12}{23} \approx 52.2\% \,.
\end{aligned}$$

○

3.6 More on independence

3.6.1 Independence of more than two events

Recall that saying that two events A and B are (statistically) independent is conveniently captured by the identity

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \,. \tag{$\perp : 2$}$$

When $0 < \mathbb{P}(A) < 1$, the independence of A and B can be equivalently expressed by the identity

$$\mathbb{P}(B \mid A) = \mathbb{P}(B \mid A^{\mathsf{c}}) \,. \tag{$\perp' : 2$}$$

Exercise. Assuming $0 < \mathbb{P}(A) < 1$, verify that the two conditions $(\perp : 2)$ and $(\perp' : 2)$ are equivalent.

In this section, we want to extend the concept of independence to more than two events.

Q What do we mean when we say that three events A, B, C are independent?

Example 3.6.1 (Flipping a coin three times)**.** Consider the experiment of flipping a coin three times in a row. Intuitively, we know that the results of the three flips are independent of one another.

Q How can we express this in the language of probabilities?

A Consider the three events

- $A \coloneqq \langle$the 1st flip shows a head$\rangle$,
- $B \coloneqq \langle$the 2nd flip shows a head$\rangle$,
- $C \coloneqq \langle$the 3rd flip shows a head$\rangle$.

What we mean by the independence of these three events is the following:

(i) The result of the 2nd flip is not affected by whether A has happened or not, that is,

$$\mathbb{P}(B \mid A) = \mathbb{P}(B \mid A^{\mathsf{c}}) \, .$$

(ii) The result of the 3rd flip is no affected by whether either of A and B has happened or not, that is,

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

$\bigcirc$

The two conditions in the above example can be taken as the definition of independence of three events. The following equivalent definition has the advantage that it makes sense even when some of the events involved have probability 0 (in which case, conditional probabilities are meaningless).

Terminology. Three events $A, B, C \subseteq \Omega$ in a probability model are said to be (statistically) *independent* if the following $2^3 = 8$ identities hold:

$$\boxed{\begin{array}{c} \mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) \, , \\ \mathbb{P}(A^{\mathsf{c}} \cap B \cap C) = \mathbb{P}(A^{\mathsf{c}})\,\mathbb{P}(B)\,\mathbb{P}(C) \, , \\ \mathbb{P}(A \cap B^{\mathsf{c}} \cap C) = \mathbb{P}(A)\,\mathbb{P}(B^{\mathsf{c}})\,\mathbb{P}(C) \, , \\ \mathbb{P}(A^{\mathsf{c}} \cap B^{\mathsf{c}} \cap C) = \mathbb{P}(A^{\mathsf{c}})\,\mathbb{P}(B^{\mathsf{c}})\,\mathbb{P}(C) \, , \\ \mathbb{P}(A \cap B \cap C^{\mathsf{c}}) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C^{\mathsf{c}}) \, , \\ \mathbb{P}(A^{\mathsf{c}} \cap B \cap C^{\mathsf{c}}) = \mathbb{P}(A^{\mathsf{c}})\,\mathbb{P}(B)\,\mathbb{P}(C^{\mathsf{c}}) \, , \\ \mathbb{P}(A \cap B^{\mathsf{c}} \cap C^{\mathsf{c}}) = \mathbb{P}(A)\,\mathbb{P}(B^{\mathsf{c}})\,\mathbb{P}(C^{\mathsf{c}}) \, , \\ \mathbb{P}(A^{\mathsf{c}} \cap B^{\mathsf{c}} \cap C^{\mathsf{c}}) = \mathbb{P}(A^{\mathsf{c}})\,\mathbb{P}(B^{\mathsf{c}})\,\mathbb{P}(C^{\mathsf{c}}) \, . \end{array}} \qquad (\perp : 3)$$

This is equivalent to the condition that the following $1 + 3 = 4$ identities hold:

$$\boxed{\begin{array}{c} \mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C) \, , \\ \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \, , \\ \mathbb{P}(A \cap C) = \mathbb{P}(A)\,\mathbb{P}(C) \, , \\ \mathbb{P}(B \cap C) = \mathbb{P}(B)\,\mathbb{P}(C) \, . \end{array}} \qquad (\perp' : 3)$$

Exercise. Can you verify that the two definitions $(\perp : 3)$ and $(\perp' : 3)$ are equivalent? Under what conditions, are these two definitions equivalent to the one given in Example 3.6.1?

The independence of more than three events can be formulated in an analogous fashion.

3.6.2 Two counter-examples

The following example shows that in order for three events to be independent, it is not enough for them to be pairwise independent.

Example 3.6.2 (Pairwise independent but not independent)**.** Consider the following three events concerning the birthdays of three friends Bassam, Leyla and Omar:

- $A \coloneqq \langle$Bassam and Leyla have the same birthday$\rangle$,
- $B \coloneqq \langle$Leyla and Omar have the same birthday$\rangle$,
- $C \coloneqq \langle$Omar and Bassam have same birthday$\rangle$.

Assuming that the birthdays of the three friends are random and independent of one another, the three events A, B, C are *pairwise* independent (why?). However, the three events are clearly not independent, because $C \subseteq A \cap B$. This illustrates that the first equality in $(\perp' : 3)$ is essential for the independence of A, B and C. $\bigcirc$

The next example shows that the identity $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$ alone does not capture the independence of A, B, C.

Example 3.6.3 (One identity not sufficient)**.** An jar contains 8 balls numbered $1, 2, \ldots, 8$. The first 4 balls are blue, and the rest are red. We pick a ball from the jar at random. Consider the three events

- $A \coloneqq B \coloneqq \{1, 2, 3, 4\} = \langle$the ball is blue$\rangle$,

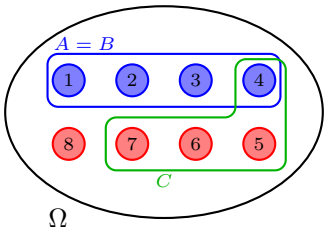


Figure 3.18: The sample space of Example 3.6.3.

- $C \coloneqq \{4, 5, 6, 7\}$

(see Figure 3.18). Clearly, the three events are not independent. Nevertheless, observe that

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{8} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(B)\ .$$

This shows that pairwise independence (i.e., the second, third and fourth equalities in $(\perp' : 3)$) is essential for the independence of A, B and C. ○