Basic probability and statistics

Philip Schulz
Christian Schaffner

last modified: September 6, 2015

Contents

Chapter 1

Basic Probability And Combinatorics

Notational conventions

In this script we make use of certain notational conventions. We **bold-face** newly introduced technical terms on first mention. Those are the terms whose definitions you are expected to know by heart in this and following courses. *Italics* serve the purpose of highlighting passages in the script but also to discriminate linguistic examples from the rest of the text. Occasionally, we will point to online references outside of this script. The corresponding links are coloured in blue and you are encouraged to click them.

We denote sets with uppercase letters and overload notation by using $|\cdot|$ as both a function that yields the cardinality of a set and the length of a sequence. Besides using standard notation for set union and intersection we denote the complement of a set S with respect to another set X by $S \backslash X$.

1.1 Introduction

1.1.1 Why study probability theory?

The fact that you have picked up this script and started reading it demonstrates that you already have some interest in learning about probability theory. This probably means that you also have some conception of what probability theory is and what to do with it. Nevertheless, we will take the opportunity to quickly give you some additional motivations for studying probability theory.

This script is all about formalizing the notion of probability. In particular, we are interested in giving a formal interpretation to statements like "A is more probable than B". Let us take a simple example to demonstrate

why this is useful: Suppose it is Monday and you have a date scheduled for Friday. Obviously you want to impress your date. Unluckily, however, you have tendency to be broke come weekends. The decision you have to make now is whether to take your date to a fancy restaurant (the impressive but expensive option) or to just go for drinks (the cheaper option). On what basis can you make this decision? Well, you can ask yourself whether it is more likely that you are broke on Friday night or not. If you think that you being broke is more probable than you going for drinks, otherwise you opt for the fancy restaurant.

The above is an example where we have used the intuitive notion of probability to assist us in decision making. The first part, the computation of the probabilities of events (e.g. you being broke or not) is something that we are going to develop in some detail in this script. The second part, the development of a so-called *decision rule* (e.g. to plan for the circumstances that are most probable to occur in the future) is something that will be covered in later courses.

Here is a second example of what one can do with probability theory. Assume you want to invest in the stock market. You will be putting in some money now and then you want to cash in on your gains (or losses) in ten years time, say. Notice that this time around simply asking whether it is more probable that your stock has risen or fallen in price is not enough. Even if your stock is worth more in ten years than it was when you bought it, the absolute increase may be so miniscule that you could have found much better investment options that would have yielded more gains. Worse even, if your gain is a smaller percentage of your original capital than the overall inflation that occurred during the ten years of your investment, you will actually have incurred a loss in terms of pure market power! So instead of asking whether or not your stock will be worth more than what it was when you first bought it, you should rather ask how much of an absolute gain you can expect from your investment. This second application of probability theory, the computation of expectations over real values, is something we are going to cover in this script, as well.

Alright, we hope that this has gotten you excited for the rest of the script. Let's get going!

1.2 Sample spaces and events

The whole of probability theory is based on assigning probability values to elements of a **sample space**. The members of the sample space are referred to as **outcomes** or **samples**.

Definition 1.1 (Sample Space) *A sample space is any Borel set* Ω*. We denote the members of a sample space by* $\omega \in \Omega$*.*

Standard examples of sample spaces are the flipping of a coin and the rolling of a die. Formally, the sample space of a die roll is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The sample space of a coin toss would consist of heads and tails. However, it is often more convenient to represent outcomes numerically. In the context of this course, we will achieve this by imposing any total order on the sample space and then identifying the outcomes with the positions they occupy in the corresponding ordered list. In this spirit we let the sample space of a coin toss be $\Omega = \{1, 2\}$ where 1 represents heads and 2 represents tails, say (the other way around would be just as fine).

More generally, we denote a sample space with n members as $\Omega = \{1, \ldots, n\}$. A useful metaphor that we will often use is to think of generating an outcome from a sample space as a blind draw from an urn with n balls that are numbered and possibly coloured but otherwise indistinguishable. The rolling of a die, for example, corresponds to drawing a ball from an urn with balls numbered 1 to 6. A somewhat more involved example is that of writing an English sentence of six words, for example the sentence: *To be or not to be*. The process of writing this sentence can be conceptualized as drawing six balls from an urn that contains balls corresponding to words in the English language[1]. Note that this will be a rather large urn as the vocabulary of the English language has already exceeded 1 million words.

In our sample spaces as defined above, it is easy to distinguish individual outcomes. However, often times we do not care about the outcomes themselves but about properties that some of them share. In the die example we might be only interested in whether the outcome is even or odd. Transferring this scenario to the urn metaphor we would colour the balls with odd numbers green and the balls with even numbers red. Again, any other colours are just as fine. All that matters is that we can discriminate a member of $E = \{2, 4, 6\}$ from a member of $O = \{1, 3, 5\}$. We do *not* need to discriminate between the outcomes that are members of the same set! In this particular setting E and O are the **events** that we are interested in.

Definition 1.2 (Event) *An event A is any subset $A \subseteq \Omega$.*

Events are what usually interests us in probability theory. Just as with outcomes, we can also define the notion of an event space.

[1] This is obviously a very unrealistic conception of how English sentences are written as it totally ignores the fact that the words in a sentence are dependent on each other and have to be placed in a particular order.

Definition 1.3 (Event space) *An event space associated with a sample space Ω is a set $\mathcal{A}$ such that*

1. *$\mathcal{A}$ is non-empty*
2. *If $A \in \mathcal{A}$ then $A \subseteq \Omega$*
3. *If $A \in \mathcal{A}$ then $\Omega \setminus A \in \mathcal{A}$*
4. *If $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$*

Notice that since $\emptyset \subseteq S$ for any set S we always have $\Omega \in \mathcal{A}$ by item 3.

Exercise 1.1 *You can also arrive at the conclusion that $\Omega \in \mathcal{A}$ always holds in a different (and arguably more cumbersome) way. How so?*

The fact that event spaces are closed under the set complement operation is very convenient. Say I organized a dinner party and invited 10 people. The day after you ask me if more than 8 people actually showed up. I just answer that I was very disappointed that my friends Mary and Paul did not come. Although I did not directly address your question you know that the answer is negative. After all, I informed you that the complement event of the event you asked about had occurred.

Exercise 1.2 *In the above party example, what is the sample space? What is the smallest possible event space that is necessary to model the situation just described?*

In general, we will not worry too much about constructing an event space every time we encounter a new problem. The **power set** of the sample space conveniently happens to fulfil all the requirements we have for event spaces, so we will just always use it. Thus, all we will ever need to worry about is the construction of sample spaces since we now know how to construct event spaces from them in a simple manner. In case you are a bit rusty, here is a reminder of what a power set is.

Definition 1.4 (Power Set) *The power set $\mathcal{P}(S)$ of any set S is defined as $\mathcal{P}(S) := \bigcup_{s \subseteq S} s$.*

In general, this leaves us with the pair $(\Omega, \mathcal{P}(\Omega))$. For outcomes in a sample space, let us stress again an important difference, namely that $\omega \in \Omega$ but $\{\omega\} \in \mathcal{A}$.

1.3 Some basic combinatorics

Combinatorics is the mathematics of counting. Counting is of course a very basic problem that may be solved by just looking at each element of a

set. However, this naïve procedure is often unreasonably time consuming. Moreover, it does not allow us to make general statements about sets of any size, i.e. sets of size n.

In order to assess the size of our sample spaces, we would like to make such general statements. The reason is that when we are dealing with probability we often start from **uniform probabilities** on the sample space where by uniform probability we simply mean the value $\frac{1}{|\Omega|}$. This is the probability we will assign to each and every $\omega \in \Omega$. We now say that all the elements in our sample space are equally probable. Note that at this point we are using probabilities solely for the purpose of motivating combinatorics which is kind of a hack because we haven't even told you yet what a probability is. However, we hope that you find the idea of uniform probabilities somewhat intuitive.

Let us start from scratch: What is the cardinality (size) of the sample space of a die roll? It is 6 because $|\{1, 2, 3, 4, 5, 6\}| = 6$. Now what if we roll two dice? The sample space for each individual die is already known. Let us call it Ω_1. The sample space for the rolling of two dice is then just the Cartesian product of two such sample spaces, i.e. $\Omega_2 = \Omega_1 \times \Omega_1 = \{(x, y) | x \in \Omega_1, y \in \Omega_1\}$. Since the cardinality of the Cartesian product of two sets S and S' is $|S| \times |S'|$ we conclude that $|\Omega_2| = |\Omega_1 \times \Omega_1| = |\Omega_1| \times |\Omega_1| = |\Omega_1|^2 = 36$.

Unsurprisingly, this method of performing a draw from the same sample space (urn) multiple times generalizes to any number of times $n > 2$. Nicely enough, it also generalizes to sets of different sizes (again by the Cartesian product argument from above). However, we have to impose one important restriction on the use of this technique: it may only be applied when the sample spaces are independent, i.e. when the outcome of one space does not affect the outcome of the other. Often times, we will simply assume that this is the case, though.

The technique of inferring the size of a complex sample space from the sizes of the sample spaces it is constructed from is known as the **basic principle of counting**.

Definition 1.5 (Basic principle of counting) *The basic principle of counting states that if two draws from sample spaces of size M and N respectively are performed independently of each other then the sample space composed from them has size $M \times N$.*

Exercise 1.3 *Let us assume that a football game is played for strictly 90 minutes. Both teams start with 11 players. A red card to a player results in that player being sent off the pitch. According to the rules of football, the game is stopped prematurely when either team has only 6 or fewer players remaining on the pitch. We are now interested in how many possible situations (we assume that situations occur in one-minute*

intervals) there are in which the game still progresses, one or more red cards have been issued and exactly four goals have been scored. Give the corresponding sample space and its size.

Note that up to now we have implicitly assumed that we would put every drawn ball back into the urn. This is also referred to as **sampling with replacement**. Let us now look at problems for **sampling without replacement**, i.e. problems where we are shrinking our sample space at each draw. One class of such problems is known as **permutation** problems.

Definition 1.6 (Permutation) *A permutation on a set S is a bijection $\sigma : S \to S : s \mapsto \sigma(s)$.*

Often times people also use the word permutation to refer to the image of a set under a permutation. What we need permutations for in practice is the reordering of ordered sets (which we will call lists). For example the permutations of the list $L = (1, 2, 3)$ are:

- $\sigma_1 = \{1 \mapsto 1, 2 \mapsto 2, 3 \mapsto 3\}$ $\sigma_1(L) = (1, 2, 3)$
- $\sigma_2 = \{1 \mapsto 1, 2 \mapsto 3, 3 \mapsto 2\}$ $\sigma_1(L) = (1, 3, 2)$
- $\sigma_3 = \{1 \mapsto 2, 2 \mapsto 1, 3 \mapsto 3\}$ $\sigma_1(L) = (2, 1, 3)$
- $\sigma_4 = \{1 \mapsto 2, 2 \mapsto 3, 3 \mapsto 1\}$ $\sigma_1(L) = (2, 3, 1)$
- $\sigma_5 = \{1 \mapsto 3, 2 \mapsto 1, 3 \mapsto 2\}$ $\sigma_1(L) = (3, 1, 2)$
- $\sigma_6 = \{1 \mapsto 3, 2 \mapsto 2, 3 \mapsto 1\}$ $\sigma_1(L) = (3, 2, 1)$

The way to think about a permutation as a draw from an urn is to look at each of the positions in the list in turn and insert an element from S. Since a permutation is a bijection, we can only use each $s \in S$ exactly once. This is precisely what it means to sample without replacement. Once a ball is drawn, it is removed from the urn. Let us make this effect concrete in the above example. For position one we have three elements to choose from. Hence we are dealing with a sample space of size 3. Position two still leaves us 2 choices, giving us a sample space of size 2. Finally, the element in the last position is totally determined as we are dealing with a sample space of size 1.

Applying the basic principle of counting we now know that there are $3 \times 2 \times 1$ permutations of the list $(1, 2, 3)$. Incidentally, this proves our above example to be correct. More generally, if we have to reorder a list with n distinct elements (or draw without replacement from an urn with n numbered balls), there are $n \times (n-1) \times \ldots \times 2 \times 1$ permutations. Since this is pretty painful to write down we introduce a more succinct notation, provided by the **factorial** function.

Definition 1.7 (Factorial) *The factorial $n!$ of a non-negative natural number $n \in \mathbb{N}$ is defined recursively as*

- $0! = 1$
- $k! = k \times (k-1)!$ *for* $0 < k \leq n$

From the above discussion we can now conclude that the number of permutations on a set or list of size n is $n!$.

We can also define the notion of a k-permutation on a set S of size n such that $k < n$. This means we are still drawing without replacement but we do not fully empty the urn. The reasoning for how many of those k-permutations there are remains exactly the same. There are $n \times (n-1) \times (n-k+2) \times (n-k+1)$ such permutations (make sure you understand why!). In order to ease notation we can again sneak in the factorial through multiplying this number with 1 in disguise. Concretely, we write

$$\begin{aligned} & n \times (n-1) \times \ldots \times (n-k+2) \times (n-k+1) \times 1 \\ =& n \times (n-1) \times \ldots \times (n-k+2) \times (n-k+1) \times \frac{(n-k)!}{(n-k)!} \\ =& \frac{n!}{(n-k)!} \end{aligned}$$

for the number of k-permutations on a set of size n.

We will not see k-permutations all that often in this script but they constitute a helpful stepping stone to another concept that will be of crucial importance. Let us draw k balls from an urn with n balls where $k \leq n$ and disregard the order in which we draw them. A classical example of such a setting would be the lottery where you are only interested in the balls drawn but not in the order in which they were drawn. We already know that for a set of k balls there are $\frac{n!}{(n-k)!}$ orders in which we can draw them, as this is a k-permutation on our urn. Now, though, we need to get rid off the different orderings. This is to say that we want to count each set of k balls that we can draw only once and not once per permutation of it. Luckily, we know how many permutations of a set of size k there are, namely $k!$. Thus we divide out this number of permutations, yielding $\frac{n!}{(n-k)! \times k!}$ as the number of possible ways to draw k *different* balls from an urn with n balls. At this point we should take a break and pat our own backs. After all, we have just derived one of the most important combinatorial formulas, which is known as the **binomial coefficient**.

Definition 1.8 (Binomial co-efficient) *The binomial coefficient $\binom{n}{k}$ is defined as*

$$\binom{n}{k} := \frac{n!}{(n-k)! \times k!}$$

for $0 < n, 0 \leq k \leq n$. It counts the number of ways to sample k distinct elements from a set with a total of n elements without regard to the order in which they are drawn. For this reason, it is pronounced "n choose k".

Exercise 1.4 *In the German lottery you have to bet on a set of 6 numbered balls to be drawn out of a total of 49 balls. Assuming that each ball is equally likely to be drawn, what is the chance of an individual bet to win the jackpot? The Dutch lottery is slightly more involved. They also draw an additional coloured ball from 6 coloured balls. In order to win the jackpot you need to have the number-colour combination right. What is your chance here?*

The binomial coefficient will become crucially important later on. A common application, that you will see in this and other courses is counting the number of bit strings with certain properties. A bit is a variable that can take on values in $\{0, 1\}$. By the basic principle of counting there are 2^n bit strings of length n. How many bit strings of length 5 are there that contain exactly 3 ones? Well, there are $2^5 = 32$ bit strings of that length in total and $\binom{5}{3} = 10$ of them contain exactly three ones. Unsurprisingly, this is the same number of 5-bit strings with exactly 2 zeros. The moral lesson here is that $\binom{n}{k} = \binom{n}{n-k}$ as can be easily seen from the definition. Some other trivia about the binomial coefficient are that $\binom{n}{0} = \binom{n}{n} = 1$. Again, this follows directly from the definition. Somewhat trickier is the fact that $\binom{n}{1} = \binom{n}{n-1} = n$. Can you derive this?

We can straightforwardly generalize the idea of the binomial coefficient to choosing more than just one set of objects. This means that instead of just looking at red versus non-red balls, say, we now distinguish between all the colours in our urn. For our strings this means that we move away from bit strings to strings with large alphabets, e.g. strings written in the English alphabet (which has 26 letters). Let's say we have r red, b blue, g green and y yellow balls in our urn such that $n = r + b + g + y$ is the total number of balls in the urn. How many different colour sequences can we draw? Well, we first arrange the r red balls in r out of n positions. This can be done in $\binom{n}{r}$ ways. We then place the b blue balls in $\binom{n-r}{b}$ ways. Next, we place the g green balls in $\binom{n-r-b}{g}$ ways. Finally, we place the remaining yellow balls deterministically in the remaining positions since $\binom{n-r-b-g}{y} = \binom{y}{y} = 1$. We compute the total number of arrangements as

$$\binom{n}{r}\binom{n-r}{b}\binom{n-r-b}{g}\binom{n-r-b-g}{y} = \tag{1.1}$$

$$\frac{n!}{r! \times (n-r)!} \times \frac{(n-r)!}{b! \times (n-r-b)!} \times \frac{(n-r-b)!}{g! \times (n-r-b-g)!} \times 1 = \tag{1.2}$$

$$\frac{n!}{r!b!g!y!} \tag{1.3}$$

Observe that the last equality follows because many of the factorials cancel and because we know that $n - r - b - g = y$. We have now worked with only four colours, but the general case follows directly by induction on the number of colours (with the binomial coefficient as base case). Thus, we can define the **multinomial coefficient**.

Definition 1.9 (Multinomial co-efficient) *The multinomial coefficient for choosing k sets of objects with size m_k from a total of $0 < n = \sum_{i=1}^{k} m_i$ objects is*

$$\frac{n!}{\prod_{i=1}^{k} m_i!}$$

Further material

For a slow and thorough introduction to combinatorics, see Faticoni (2013): Combinatorics. At the ILLC, there is a biannual course on combinatorics, taught by Ronald de Wolf. Online, Princeton also offers a course on combinatorics.

Chapter 2

Axiomatic Probability Theory

2.1 Axioms of Probability

In the previous chapter, we have introduced sample spaces and event spaces. We would like to be able to express that certain events are more (or less) likely than others. Therefore, we are going to measure the probability of events in a mathematically precise sense.

Definition 2.1 (Finite Measure) *A* finite measure *is a function* $\mu : \mathcal{S} \to \mathbb{R} : S \mapsto \mu(S)$ *that maps elements from a countable set of sets* $\mathcal{S}$ *(formally a* σ*-algebra) to real numbers. Such a measure has the following properties:*

1. $\mu(S) \in \mathbb{R}$ *for* $S \in \mathcal{S}$,
2. $\mu\left(\bigcup_{i=1}^{\infty} S_i\right) = \sum_{i=1}^{\infty} \mu\left(S_i\right)$ *for disjoint sets* $S_1, S_2, \ldots$.

Notice that we are restricting ourselves to finite measures here, i.e. the value of the measure can never be infinite. This restriction makes sense as probabilities are finite as well. Property 2 is known as *countable additivity*.

Let $S = \bigcup_{i=1}^{n} S_i$ for some positive natural number n and disjoint S_i and $S_j = \emptyset$ for $j > n$. By countable additivity, we then get

$$\mu(S) = \mu(\bigcup_{i=1}^{\infty} S_i) = \mu\left(\bigcup_{i=1}^{n} S_i \cup \bigcup_{j=n+1}^{\infty} \emptyset\right) = \sum_{i=1}^{n} \mu(S_i) + \sum_{j=n+1}^{\infty} \mu(\emptyset) \tag{2.1}$$

Since the S_i are disjoint, we must have $\mu(S) = \sum_{i=1}^{n} \mu(S_i)$ and it follows that $\mu(\emptyset) = 0$. We conclude that the empty set has measure 0 for all

measures. Furthermore, we also see from the above derivation that countable additivity implies finite additivity, i.e. $\mu(S) = \sum_{i=1}^{n} \mu(S_i)$ for finite positive n (again, this only holds if the S_i are disjoint).

Examples of measures are not hard to find. In fact, we have already seen a measure, namely the function $|\cdot|$ that counts the elements of a set (check yourself that it really is a measure). Another measure is the Dirac-measure that is related to the characteristic function of a set. While the characteristic function tells you whether any object belongs to a given set, the Dirac-measure tells you whether any set contains a given object. Let us call the object in question a. Then its Dirac measure $\delta_a(S) = 1$ iff $a \in S$ and 0 otherwise (check yourself that the Dirac-measure indeed is a measure).

Apart from these examples, there is one measure, however, that is going to be the star of the rest of this script, namely the **probability measure**.

Definition 2.2 (Probability measure) *A probability measure $\mathbb{P} : \mathcal{A} \to \mathbb{R}, A \mapsto \mathbb{P}(A)$ on an event space $\mathcal{A}$ associated with a sample space Ω has the following properties:*

1. $\mathbb{P}(A) \geq 0$ *for all* $A \in \mathcal{A}$,
2. $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *for disjoint events* $A_1, A_2, \ldots$,
3. $\mathbb{P}(\Omega) = 1$.

Notice that we only added Property 3 to the general definition of a measure. Hence, a **probability** (the value that the probability measure assigns to an event) will always lie in the real interval $[0, 1]$. The above three axioms for a probability measure are often referred to as *axioms of probability* or *Kolmogorov axioms* after their inventor Andrey Kolmogorov.

We have already discussed uniform probabilities in the previous chapter. We can now formally explain what we meant by that. The uniform probability measure $\mathbb{P}$ has the property that $\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}$ for all $\omega \in \Omega$. At this point, the distinction between sample and event spaces becomes important. We cannot measure the elements of a sample space, only the elements of an event space! Recall our convention that we will always assume that $\mathcal{A} = \mathcal{P}(\Omega)$ which obviously contains a singleton for each element in Ω. Using this assumption, the uniform probability measure is indeed well-defined. Whenever we talk about *uniform probability*, we either mean the uniform probability measure or, more often, the real value $\frac{1}{|\Omega|}$ to which this measure uniformly evaluates.

In order to create a tight relationship between a sample space, an event space and a probability measure, we introduce the concept of a **probability space**. Probability spaces are also known as **(probabilistic) experiments**.

Definition 2.3 (Probability space) *A probability space is a triple $(\Omega, \mathcal{A}, \mathbb{P})$, consisting of a sample space Ω, an event space $\mathcal{A}$ and a probability measure $\mathbb{P}$.*

If we roll a die, for example, we have the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and, by convention, the event space $\mathcal{A} = \mathcal{P}(\Omega)$. If we add the uniform probability measure, we have constructed a *probabilistic experiment*. We can use it to answer a couple of questions. For example, we might wonder about the probability of obtaining an even number. By Property 2 of our definition, this probability is given by

$$\mathbb{P}(\{2, 4, 6)\}) = \mathbb{P}(\{2\} \cup \{4\} \cup \{6\}) \tag{2.2}$$

$$= \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \tag{2.3}$$

Notice that this calculation is rather cumbersome. After all, we might just have evaluated $\mathbb{P}(\{2, 4, 6\})$ directly. This is because by convention we have $\mathcal{A} = \mathcal{P}(\Omega)$ which certainly contains $\{2, 4, 6\}$. Since the probability measure is defined on $\mathcal{A}$, it must map $\{2, 4, 6\}$ to some real number. However, the above calculation points to an interesting fact. In order to fully specify a probability measure, is suffices to specify the measure on the singleton sets of the event space. By countable additivity, this assignment already specifies the measure on the entire event space, as we can construct any event as a countable union of singletons.

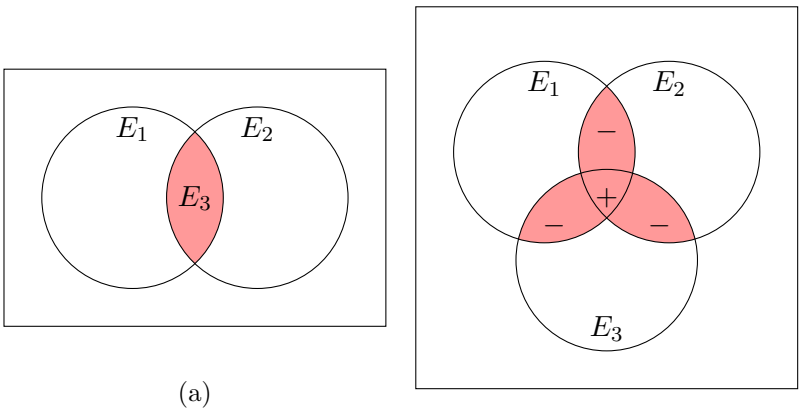
It is important to point out that we just chose the uniform probability measure as the one that seems "natural" for a die roll. However, nobody is forcing us to do so. In fact, Definition 2.3 allows us to impose arbitrary probability measures.

Exercise 2.1 *Let us consider a rigged die. Take $(\Omega, \mathcal{A}, \mathbb{P})$ with Ω and $\mathcal{A} = \mathcal{P}(\Omega)$ as in the uniform die-roll example before, but use the probability measure specified by*
$\mathbb{P} = \{(\{1\}, 0), (\{2\}, \frac{1}{12}), (\{3\}, \frac{1}{6}), (\{4\}, \frac{1}{6}), (\{5\}, \frac{1}{3}), (\{6\}, \frac{1}{4})\}$.

1. *Verify that $\mathbb{P}$ is indeed a probability measure.*
2. *Compute the probability of obtaining a number strictly smaller than 5 in this experiment.*

2.2 Probability of Arbitrary Unions of Events

We have seen how to compute probabilities of events if they can be formed as unions of *disjoint* events. The natural question to ask is what to do if we want to compute the probability of the *union of non-disjoint events*. In order to reason about this problem, we first take a step back and think



(a)

(b)

Figure 2.1: 2.1a: Two overlapping events E_1 and E_2. Their intersection (the coloured region) gets counted twice if we add up their probabilities.
2.1b: Venn diagram with 3 events. First we deduct $E_1 \cap E_2, E_1 \cap E_3, E_2 \cap E_3$ in order to prevent double counting and then we add in $E_1 \cap E_2 \cap E_3$. Deductions and additions are indicated by pluses and minuses.

about the outcomes of our probability space. We know that each event with non-zero probability contains at least one outcome (since $\mathbb{P}(\emptyset) = 0$, we can safely ignore the empty event). Let us assume that we take the union of events E_1 and E_2 with $E_1 \cap E_2 = E_3 \neq \emptyset$. This means that the outcomes in E_3 are contained in both E_1 and E_2. This situation is illustrated in Figure 2.1a. If we were to simply add up the probabilities of E_1 and E_2, we would effectively count the contribution of the outcomes in E_3 twice. We would hence get an overestimate of the actual value of $\mathbb{P}(E_1 \cup E_2)$. In order to avoid this we will need to subtract the probability of E_3 one time. This leads us to the following formulation:

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2) \tag{2.4}$$

Notice that this is fully general in that it is true even if E_1 and E_2 were disjoint. In that case, their intersection would be empty. We can generalize this principle to the (countable) union of an arbitrary number of events. This will give us a principled way of calculating the probability of any union of events. This calculation technique is know as the **Inclusion-Exclusion principle**.

Theorem 2.1 (Inclusion-Exclusion principle) *The probability of any (countable) union of events $E_1, \ldots, E_n$ can be computed as*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} (-1)^{i+1} \left(\sum_{j_1 < \ldots < j_i} \mathbb{P}\left(E_{j_1} \cap \ldots \cap E_{j_i}\right) \right) \tag{2.5}$$

We are going to proceed with a combinatorial proof of the Inclusion-Exclusion principle. It is very elegant but invokes the binomial theorem. For completeness sake we will prove the binomial theorem at the end of this chapter. For now, just trust us that it exists and is correct.

Proof We are going to focus on a particular outcome ω that is contained in m events which we call without loss of generality $E_1, \ldots, E_m$ for some $m < n$. Notice that we can safely neglect all events which do not contain ω, since ω is not going to contribute to their probability.

For all the E_i, $1 \leq i \leq m$ in which ω is contained, it is certainly true that ω is also contained in their intersections. The Inclusion-Exclusion-principle adds up or subtracts the probabilities of intersections of a given size. Notice that any intersection of more than m events will not contain ω as we intersect with at least one event that does not contain ω. Thus, we only need to consider intersections of our m ω-containing sets.

When $i = 1$ the intersection is trivial, as it just consists of one event. How many ways are there to pick one out of m events? The answer is $\binom{m}{1}$. This is the number of times that ω contributes to the overall probability. At this point we have an overestimate of that probability (compare this to Figure 2.1a). Next we subtract the probabilities of the mutual intersections ($i = 2$). By the same reasoning as before, the contribution of ω is deducted $\binom{m}{2}$ times which gives us an underestimate since $\binom{m}{1} \leq \binom{m}{2}$ for $m \geq 3$. Since we are adding and subtracting in alternation, we will now keep flip-flopping between under- and over estimates. After considering all intersections of up to m sets, we should get the correct result, however.

What we want to prove is that the right-hand side of (2.5) counts ω's contribution to the overall probability exactly once (because this is what happens on the left hand-side of (2.5)). That is, we have to prove that

$$1 = \sum_{i=1}^{m} (-1)^{i-1} \binom{m}{i} \tag{2.6}$$

We are right on our way towards exploiting the binomial theorem. Let us first state it.

$$(p+q)^m = \sum_{i=0}^{m} \binom{m}{i} p^i q^{n-i} \tag{2.7}$$

Setting $p = (-1)$ and $q = 1$, and multiplying both sides with (-1), we obtain

$$-(-1+1)^m = -\sum_{i=0}^{n} \binom{m}{i} (-1)^i$$

which can be rewritten as

$$0 = -1 + \sum_{i=1}^{n} \binom{m}{i} (-1)^{i+1}, \tag{2.8}$$

because $\binom{m}{0} = 1$. Equation (2.8) implies (2.6) which we needed to prove. □

At this point we have done our fair share of math and found out how to calculate the probability of a union of events. We should ask ourselves what the probability of a union of events even tells us. Observe that an event occurs whenever we draw an outcome from our sample space that is contained in that event. By taking the union of events $E_1, \ldots, E_n$ we form a new event E that (possibly) contains more outcomes than each of the original events. Thus, the probability of the E will be higher than (or the same as) the probability of each of $E_1, \ldots, E_n$. What we are measuring then, is the probability that *any* of the events $E_1, \ldots, E_n$ occur. Crucially, we do not care anymore which one of them occurs.

What we are missing is a way to express the probability that a given number of events occur *together*. This concept is so important that we have a dedicated name for it, that of **joint probability**.

Definition 2.4 (Joint probability) *The joint probability of a (countable) set of events $\{E_1, \ldots, E_n\}$ is defined as*

$$\mathbb{P}(E_1 \cap \ldots \cap E_n)$$

Sometimes one also finds the alternative notation

$$\mathbb{P}(E_1, \ldots, E_n)$$

Wow, that was simple! We don not event need to prove another rule for calculating the joint probability. After all, we already know how to take the intersection of sets. Annoyingly, one problem remains: our definition of event spaces does not guarantee that they contain the intersections of their members. Or does it? Well, let us see whether we can "paraphrase" what an intersection is.

$$E_1 \cap E_2 = \Omega \backslash ((\Omega \backslash E_1) \cup (\Omega \backslash E_2)) \tag{2.9}$$

All the operations on the right hand side are defined for events spaces. We have thus solved our problem since we have shown that we can indeed

do intersection in event spaces. To convince yourself that this is correct, you may want to consult Figure 2.1a. Alternatively, you may also just realise that this is an instance of DeMorgan's laws which you should know from set theory. Notice that we do not claim that this is the only valid "paraphrase". Feel free to find others, if you like!

2.3 Probability of Complements of Events

At this point we are capable to do most probabilistic computations that we will encounter in this course. From here on, it is all about making our lives easier. For example, how would you solve the following problem.

Exercise 2.2 *You are observing a panel of 200 light bulbs and you know that at least one of them will light up once you press a button. What is the probability that any except the 87th bulb will light up? Note: this is a conceptual exercise. For the very keen ones, you can obtain the probability for each bulb to be turned on by typing the following into the Python interpreter:*

```
import numpy

probabilities = numpy.random.rand(1,200)
print probabilities/probabilities.sum()
```

The point of the above exercise is that it will be awfully cumbersome to compute the probability of the union of the singletons E_i where $1 \leq i \leq 200$ and $i \neq 87$. On the other hand we can easily look up $\mathbb{P}(E_{87})$. The question is whether we can exploit this simpler calculation to help us answer the original question. Here we will again make use of the properties of event spaces. For any event E in our event space we also have $\Omega \backslash E$ in the same space. Furthermore, E and $\Omega \backslash E$ are disjoint which by our probability axioms means that we can simply add up their probabilities if we want to calculate the probability of their union. But what's the union of E and $\Omega \backslash E$? It's exactly Ω. From axiom 3 we know that $\mathbb{P}(\Omega) = 1$. By simple algebraic manipulations we find that

$$\mathbb{P}(\Omega \backslash E) = 1 - \mathbb{P}(E) \tag{2.10}$$

Thus if we want to find the probability that any but the $87th$ bulb will light up, we simply compute the probability that the $87th$ bulb will light up will light up and subtract that from 1. This is a rather general strategy to simplify calculations whenever the probability of an event is hard to compute. Maybe the probability of the complement of that event will be easier to compute.

Exercise 2.3 *Show that in general*

$$\mathbb{P}(E_1 \backslash (E_1 \cap E_2)) = \mathbb{P}(E_1) - \mathbb{P}(E_1 \cap E_2)$$

2.4 Conditional Probability and Independence

After we have seen how to measure the probability of events, we are going to introduce another tremendously important concept, that of **conditional probability** measures.

Definition 2.5 (Conditional probability measure) *The probability of an event E_i conditioned on another event E_j with $\mathbb{P}(E_j) > 0$ is defined as*

$$\mathbb{P}(E_i|E_j) := \frac{\mathbb{P}(E_i \cap E_j)}{\mathbb{P}(E_j)}$$

Before we get into the math of conditional probabilities, let us try to understand the meaning of this concept. When we are computing the conditional probability of an event E_i, we re-scale with the probability of the conditioning event E_j. If $E_j \neq \Omega$, $\mathbb{P}(E_j)$ might be smaller than 1. Thus, this rescaling assumes *that E_j has already occurred*. In other words, we are excluding all outcomes that are not in E_j from further consideration (even though they may be in E_i). The interpretation of conditional probabilities is that they are the probabilities of events assuming that another event has already occurred.

Another interpretation is that when working with a conditional probability measure, we are in fact working in a new probability space, where $\Omega_{new} = E_2$, i.e. our new sample space is the conditioning event. Notice that this also means that our probability measure will change and become the measure from Definition 2.5.

Here comes the cool part: although we have introduced a new concept, all the properties of probability measures that we know by now will seamlessly carry over to conditional probabilities, if we can prove that the conditional probability measure is a probability measure according to our axioms.

Exercise 2.4 *Use the axioms from Definition 2.2 to prove that $\mathbb{P}(\cdot|E_j)$ is a probability measure.*

We will make use of conditional probabilities quite a lot in this course. We will later see a way in which they help us to decompose joint probability distributions. For now, we are going to focus on the fact that they are also related to the idea of independence of events.

Definition 2.6 (Independence) *Two events E_1, E_2 are said to be independent if*

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) \times \mathbb{P}(E_2)$$

Independence of two events is denoted as $E_1 \perp E_2$.

This definition relates to conditional probabilities in the following way: assume that $E_1 \perp E_2$. Then we get

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} = \frac{\mathbb{P}(E_1) \times \mathbb{P}(E_2)}{\mathbb{P}(E_2)} = \mathbb{P}(E_1)\,. \tag{2.11}$$

Hence, independence of two events $E_1 \perp E_2$ is equivalent with $\mathbb{P}(E_1|E_2) = \mathbb{P}(E_1)$.

Exercise 2.5 *Prove that $E_1 \perp E_2$ is also equivalent with $\mathbb{P}(E_2|E_1) = \mathbb{P}(E_2)$.*

Independence will prove to be a useful concept in later chapters. More precisely, we will often just *assume* that two events (or random variables – see the next chapter) are independent. Although such an independence assumption might not always hold in practice, it will allow us to formulate much simpler probabilistic models.

2.5 A Remark on the Interpretation of Probabilities*

This concludes our introduction of axiomatic probability theory. We know that a probability is a real number in $[0, 1]$. For all that we are going to do in this course (and in most follow-up courses) this is fully sufficient. However, some of you may wonder what a "natural" interpretation of probabilities would be. There are two dominating views on that. One postulates that if we were to take A LOT (read: almost infinitely many) samples from a sample space, the probability of an event is its frequency amongst these samples divided by the total number of samples taken. For those of you who know limits, this principle can be formalized as $\mathbb{P}(E) = \lim_{n\to\infty} \frac{\#E}{n}$. This view is known as the *frequentist view*.

The second view postulates that probabilities are an expression for degrees of belief. Basically, if you assign $\mathbb{P}(E)$ to an event E, then $\mathbb{P}(E)$ is the strength of your personal belief that E will occur. This latter view is known as the *Bayesian view*.

Which conception of probability you choose is a philosophical matter and does not really impact the math. That is why we will not care about this issue in this course. However, it is useful to at least be aware of these two views (if only to appear knowledgeable in a conversation you may have with your philosopher friends).

2.6 The Binomial Theorem

The binomial theorem from Equation 2.7 is actually not that hard to prove. We will do so by induction. As a base case we choose $m = 0$. Then the equality is easy to see.

$$(p+q)^0 = 1 = \binom{0}{0} p^0 q^0 \tag{2.12}$$

Next, we assume that the theorem holds for $m = n$. What we want to show is that it also holds for $m = n + 1$. We achieve this by algebraic manipulation.

$$(p+q)^{n+1} = (p+q)^n \times (p+q) \tag{2.13}$$

$$= (p+q)^n p + (p+q)^n q \tag{2.14}$$

$$= p\sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} + q\sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \tag{2.15}$$

$$= \sum_{i=0}^{n} \binom{n}{i} p^{i+1} q^{n-i} + \sum_{i=0}^{n} \binom{n}{i} p^i q^{n+1-i} \tag{2.16}$$

$$= \sum_{j=1}^{n+1} \binom{n}{j-1} p^j q^{n+1-j} + \sum_{i=0}^{n} \binom{n}{i} p^i q^{n+1-i} \tag{2.17}$$

$$= \binom{n}{n} p^{n+1} q^{(n+1)-(n+1)} + \sum_{k=1}^{n} \binom{n}{k-1} p^k q^{n+1-k} \tag{2.18}$$

$$+ \binom{n}{0} p^0 q^{n+1} + \sum_{k=1}^{n} \binom{n}{k} p^k q^{n+1-k}$$

$$
\begin{aligned}
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\binom{n}{k} + \binom{n}{k-1} \right) p^i q^{n+1-k} && (2.19)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\frac{n!}{k!(n-k)!} + \frac{n!}{(k-1)!(n-k+1)!} \right) p^i q^{n+1-k} && (2.20)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\frac{n!(n+1-k)}{k!(n+1-k)!} + \frac{n!k}{k!(n-k+1)!} \right) p^k q^{n+1-k} && (2.21)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \left(\frac{n!(n+1)}{k!(n+1-k)!} \right) p^i q^{n+1-k} && (2.22)\\
&= q^{n+1} + p^{n+1} + \sum_{k=1}^{n} \binom{n+1}{k} p^k q^{n+1-k} && (2.23)\\
&= \sum_{i=0}^{n+1} \binom{n}{k} p^k q^{n-k} && (2.24)
\end{aligned}
$$

Let us clarify some parts of the proof. We use the induction hypothesis to expand the terms in Line 2.15. In Line 2.17, we switch the variable i in the first summand to $j = i + 1$. The reason why we do this is because we want to achieve congruence with the exponents of the second summand. In the following line we uniformly name the variables k. Since k has to run over a common range, we chop off the ends of both sums that stick out. In the first sum of line 2.17 that is the summand that corresponds to $j = n+1$ and in the second sum it is the summand that corresponds to $i = 0$. We pull out both of them in line 2.18 and then collapse the sums in line 2.19. The following lines are basically just an exercise in manipulation fractions. The jump from the second-to-last to the last line is allowed because

$$q^{n+1} = \binom{n+1}{0} p^0 q^{n+1-0}$$

and

$$p^{n+1} = \binom{n+1}{n+1} p^{n+1} q^{(n+1)-(n+1)}$$

which are exactly the quantities that we need to add to make our sum reach from 0 to $n + 1$. This completes the proof.

Further Reading

A very quick and dirty introduction to measure theory is provided by Maya Gupta and can be found here. If you are looking for something more extensive that also motivates event spaces and the like you may want to take a look at this script by Ross Leadbatter and Stamatis Cambanis (which has also been published as a book).