Basic Probability and Statistics

Philip Schulz
Christian Schaffner

last modified: September 21, 2015

Contents

Contributors

While we strive to continuously update this script and keep it on an acceptable level of grammaticality and mathematical correctness, it is unavoidable that some mistakes creep in. We are therefore utterly grateful to our contributors who have helped improving the script and would like to acknowledge their contributions here.

- Philip Michgelsen has corrected a mistake in the definition of event spaces in chapter 1.

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 1.4 *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 1.5 *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 1.6 (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 1.7 (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 1.8 *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 1.9 (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 1.10 (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 1.11 (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 1.12 *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 1.13 (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.

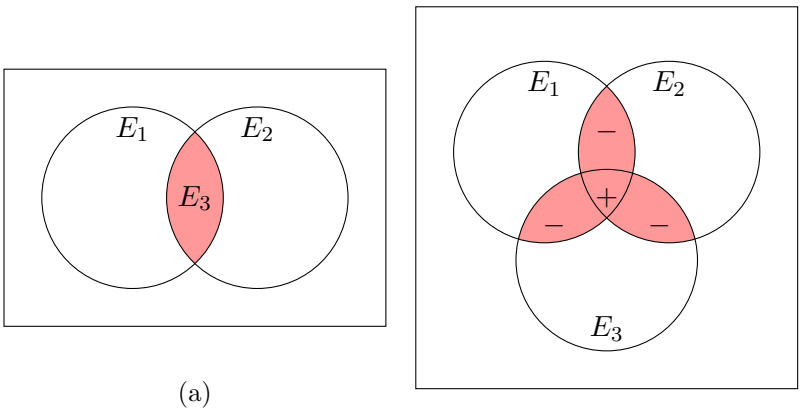
It is important to point out that we just chose the uniform probability measure as the one that seems "natural" for a die roll. However, nobody is forcing us to do so. In fact, Definition 2.3 allows us to impose arbitrary probability measures.

Exercise 2.4 *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 2.5 (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 2.6 (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 2.7 *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 2.8 *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 2.9 (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 2.9.

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 2.10 *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 2.11 (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 2.12 *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).

Chapter 3

Random Variables and Their Properties

3.1 What is a Random Variable?

Random variables are a tremendously useful concept in probability theory. In most situations of science and general life we are not really interested in the actual outcome of a random experiment. Instead, we care about particular properties of that outcome.

For starters, let us talk about the weather. Like most people living in the north of Europe, your choice of clothing probably depends strongly on the weather. Let us assume that you have a thick winter jacket that you put on when it is cold, a light soft-shell jacket that you wear when temperatures are mild and that you simply wear a T-shirt or a sweater when it is warm. Let us also assume that the forecasting site you consult every morning reports the temperatures up to one decimal. The question is: does it matter to you whether it is 5.4 or 7.3 degrees centigrade outside when you make your decision about what to wear? Most likely you are only interested in whether it is cold, mild or warm. This may obviously vary according to your own perception of warmth and cold. Here we will assume that it is cold whenever the temperature is below 10 degrees, mild between 10 and 20 degrees (inclusive) and warm whenever the temperature rises above 20 degrees. So what you are really interested in, at least as far as clothing is concerned, is whether the temperature t is $t < 10$ or $10 \leq t \leq 20$ or $t > 20$. What we are looking for is a way of transforming outcomes from the temperature scale into judgements of perceived temperature.

Another example is the lottery. Let us again consider the German lottery where you have to correctly predict 6 out of 49 numbers all of which are drawn *independently* and *uniformly at random* (you should be able to understand the first term by now; after reading this chapter, you will also understand the second). You get the main prize if you predict all these items

correctly. However, pay-outs start from three correctly predicted numbers and increase for each additional number that you got right. So what do you really care about when playing the lottery? Does it matter to you whether or not you correctly predicted the number 39? Probably not. All you care about is how many of the numbers you got right.

Observe that one possible way of modelling the sample space for the lottery situation is the following:

$$\Omega = \{(b_1, \ldots, b_6, c_1, \ldots, c_6) \,|1 \leq b_1 < b_2 < \ldots < b_6 \leq 49; \\ c_1, \ldots, c_6 \in \{0, 1\}\} \tag{3.1}$$

This rather cumbersome expression states that we draw 6 balls from an urn of 49 balls without replacement (and then sort them). We also add indicators of whether we predicted each ball drawn correctly. Note that we did not model what exactly our prediction was, but only if we predicted a number correctly. So again the question arises of how we can transform outcomes from this sample space into the events that we actually care about (namely how many balls we predicted correctly)?

After this little motivating section, let us formally define what a random variable is.

Definition 3.1 (Real random variable) *A (real) random variable (RV) X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a function*

$$X : \Omega \to \mathbb{R}$$

where we require that for all $r \in \mathbb{R}$, the set $A = \{\omega | X(\omega) \leq r\}$ is in the event space $\mathcal{A}$.

Notice that the naming of the random variable is arbitrary but by convention we will often use the capital letters X, Y, Z. Also, as you can see, RVs are defined on an entire probability space. The reason is that we want to have a probability measure associated with our sample space. In fact, this probability measure is the reason why we can see random variables as taking on their values in a random way. Moreover, since we can only measure events from our event space, the condition on RVs ensures that we can in fact measure the probability of each real value which a RV may map an outcome to. In technical terms, we are thus making sure that the pre-image of the RV is *measurable*.

Let us define a random variable for the weather situation.

$$X(t) = \begin{cases} 1 & t < 10\,, \\ 2 & 10 \leq t \leq 20\,, \\ 3 & t > 2\,. \end{cases} \tag{3.2}$$

In principle we can easily determine the probability of each of the values of this RV. However, we did not bother to define a sample space and associated probability measure, mostly because we would have to restrict the possible temperatures to those that we think are reasonable. Nevertheless, this is something that you can do yourself.

Exercise 3.2 *Define an upper and lower bound for temperature values (make sure you can actually get all three perceived degrees of warmth). These bounds will lead you to a discrete sample space since we postulated that the website we consult only reports temperatures up to one decimal place after the comma. Then impose any probability measure on the associated event space. Finally, compute the probability of each of the values of the RV X from Equation (3.2).*

We can do the same for the lottery. Our random variable would then look as follows:

$$X(\omega) = \sum_{i=1}^{6} c_i \tag{3.3}$$

Whenever a RV assumes a particular value, we write $X = x$. Thus, if X in Equation (3.3) assumes the value 4, i.e. we guessed four balls correctly, we would write $X = 4$. And here comes the function that computes the probability that this event happens.

Definition 3.3 (Probability distribution) *The discrete probability distribution of a RV X is denoted by P_X and is defined as the function*

$$P_X(X = x) := \mathbb{P}(\{\omega \mid X(\omega) = x\})$$

In the literature, often one of the two mentions of X is dropped from the notation, resulting in $P(X = x)$ or $P_X(x)$.

What is $P(X = 4)$ for the lottery example? Recall that there are $a = \binom{49}{6}$ sequences of balls that we can draw (6 out of 49 balls). In our modelling of the scenario, we also have to take into account whether or not we correctly predicted each ball. Notice that the indicator c_i is a function of our predictions. If our prediction $v_1, \ldots, v_6$ are ordered in ascending fashion, then

$$c_i = \begin{cases} 1 & \text{if } v_i = b_i \\ 0 & \textit{otherwise} \end{cases}$$

The unfairness of the lottery comes from the fact that the player has to pick out of a large number of sequences only very few of which will have more than three numbers in common with the sequence that ends up being

drawn. Let us apply this intuition to our example. There are $\binom{6}{4}\binom{43}{2}$ ways in which your predictions can get exactly 4 numbers right. This is because there are $\binom{6}{4}$ ways to guess 4 out of 6 numbers correctly and $\binom{43}{2}$ ways to choose any two out of the 43 numbers that were not drawn. In total there are

$$\binom{6}{4}\binom{43}{2} = 13545 \tag{3.4}$$

ways to agree on four balls with some lottery draw. This might seem like a lot. However, if we divide this by the total number of possible lottery draws, we find that $P(X = 4) = 0.000969$ which is very little.

We have just seen examples where the RV takes on a specific value. We can of course also make statements like $X < x$ and $X \geq x$. This is a way of expressing that a RV assumes a value above or below a certain threshold. The way to get the probability of this event is through the cumulative distribution function.

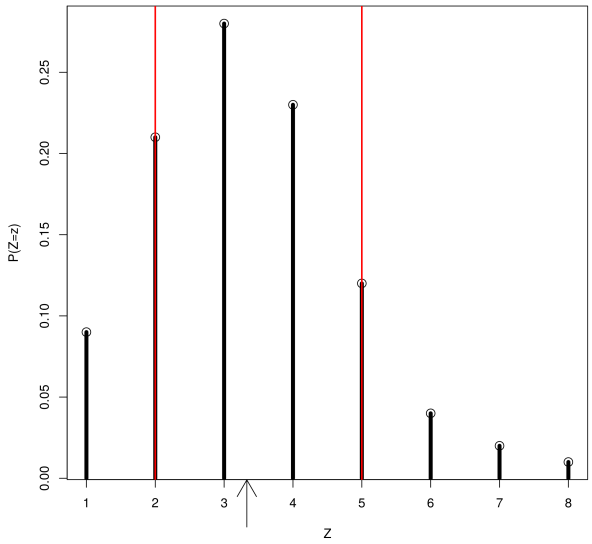


Figure 3.1: Probability distribution of a random variable Z with 8 possible values, given by $P(Z = 1) = 0.09, P(Z = 2) = 0.21, P(Z = 3) = 0.28, P(Z = 4) = 0.23, P(Z = 5) = 0.12, P(Z = 6) = 0.04, P(Z = 7) = 0.02$, and $P(Z = 8) = 0.01$. The arrow indicates the expectation. This is a visualisation of the spike that we can plug underneath the centre of mass.

Definition 3.4 (Cumulative distribution function) *The **cumulative distribution function (cdf)** of a random variable X is given by*

$$F_X(a) := P(X \leq a) = \sum_{x \leq a} P(X = x)$$

The cumulative distribution function is helpful in two ways. First of all, it allows to get the probability that X takes on a value in a certain range. Second, it has applications in sampling algorithms, some of which we might see during the programming part of the course. A plot of the cdf of Z is given in Figure 3.2. We can recognize that it is the cdf of a discrete distribution because it is discontinuous. If the underlying probability distribution was continuous, so would be its cdf.

To better understand how we can compute the probability that $a \leq X \leq b$ for some real values a, b let us take a look at Figure 3.1. There we see a plot of the probability distribution of a random variable Z. The red lines indicate the interval that we are interested in. As a warm-up exercise, let us first compute $P(Z \leq 2)$ and $P(Z > 5)$. By looking at the plot we easily see which areas we need to consider, namely the bars to the left of 2 and to the right of 5. The height of the bars corresponds to their probability.

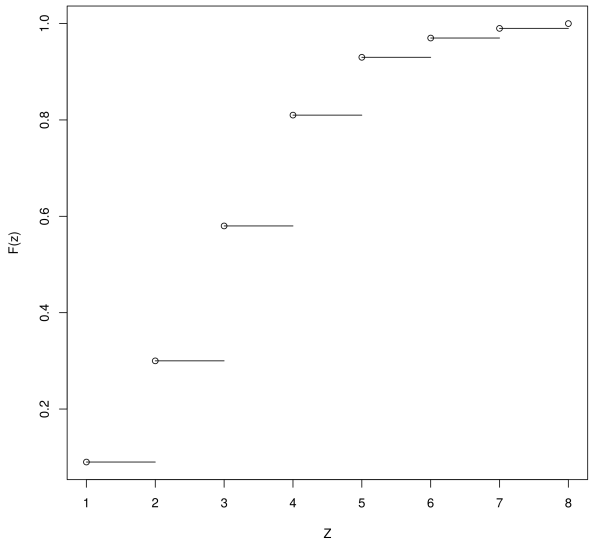


Figure 3.2: Cumulative probability distribution of Z. Observe that it stays constant in values that Z cannot take on. This cdf being discontinuous is an indicator that the underlying probability distribution is discrete.

We go on to compute the probability that Z is smaller or equal to 2 and that Z is bigger or equal to 5:

$$P(Z \leq 2) = P(Z = 1) + P(Z = 2) = 0.09 + 0.21 = 0.3 \tag{3.5}$$

$$\begin{aligned} P(Z > 5) &= P(Z = 6) + P(Z = 7) + P(Z = 8) \\ &= 0.04 + 0.02 + 0.01 = 0.07 \end{aligned} \tag{3.6}$$

It is straightforward to compute $P(2 < Z \leq 5)$. If we again look at Figure 3.1, we see that this is the area delimited by the red lines. Thus, $P(Z \leq 5)$ will be an overestimate of this probability. The amount by which it is an overestimate is exactly $P(Z \leq 2)$. Thus we get $P(2 < Z \leq 5) = P(Z \leq 5) - P(Z \leq 2)$.

Exercise 3.5 *We claim that you can compute $P(2 < Z \leq 5)$ based on $P(Z \leq 2)$ and $P(Z > 5)$ from above. What result do you get?*

In the case of discrete random variables, i.e. the ones that we are dealing with here, there is another function that is of great interest, namely the probability mass function.

Definition 3.6 (Probability mass function) *The **probability mass function (pmf)** of a random variable X is given by*

$$p(x) := P(X = x)$$

The pmf is notationally more convenient than the probability distribution but also less general. For example, the pmf does not allow us to express that X falls in a range of values. In order to avoid confusion, we will mostly use X's probability distribution here. Be aware, however, that most papers that you are going to read in the future will use the pmf instead. Regardless of whether we are using the probability distribution or the pmf, we will call the values in $\mathbb{R}$ to which they assign positive probability, their support.

Definition 3.7 *The **support** of a random variable X is given by*

$$\text{supp}(X) := \{x \in \mathbb{R} \mid P(X = x) > 0\}$$

3.2 Expectation and Variance

There are two major properties of probability distributions that are used to describe them. One is the center of mass, the point at which the probability mass of the distribution is split into half. If you think of the support of P as a plank and of the probabilities as weights, then the center of mass is the point at which you could put an infinitely thin spike into the plank from underneath such that the plank maintains perfect balance. In Figure 3.1 this so-called expectation is indicated by an arrow that can be interpreted as visualising the spike that we can plug underneath the probability plank.

Definition 3.8 (Expectation) *The **expectation** of a random variable X with respect to the distribution P is defined as*

$$\mathbb{E}[X] := \sum_{x \in \text{supp}(X)} xP(X = x)\,.$$

Exercise 3.9 *Compute the expectation of the random variable Z from the previous section.*

It is important to point out that the expectation of a random variable is a real number which does not need to in the support of X. This is for example the case in Figure 3.1 where the expectation is fractional although all support values are integers.

The expectation comes with an interesting property that is called the **linearity of expectation**. Basically, whenever we multiply X with a constant a, the expectation will also be scaled by that constant. Moreover,

when we add a constant b to X the expectation will also increase/decrease by b (depending on whether b is positive or negative). Let us prove that!

$$\mathbb{E}[aX+b] = \sum_{x \in \text{supp}(X)} (ax+b)P(X=x) \tag{3.7}$$

$$= a \sum_{x \in \text{supp}(X)} xP(X=x) \; + \; b \sum_{x \in \text{supp}(X)} P(X=x) \tag{3.8}$$

$$= a\mathbb{E}(X) + b \tag{3.9}$$

The equality in (3.9) follows from the fact that the sum of the probabilities of the support of a distribution is 1. The linearity of expectation is extremely useful. Basically, if you know the expectation of a RV you also know the expectation of its multiples. Furthermore, by adding a constant, we basically shift each element in the support. However, this shifting does not do anything surprising as the centre of mass shifts by the same amount.

We can actually take the expectation of any quantity, but it only has an effect on quantities that can be cast as functions of X. The expectation of a quantity that is constant with respect to X will just be that quantity itself. More formally:

$$\mathbb{E}[b] = \sum_{x \in \text{supp}(X)} bP(X=x) = b \tag{3.10}$$

A class of functions of X that people are often interested in are the moments. Moments are the powers of X, so X itself is the first moment, X^2 is the second moment and so on. We will not further discuss moments here, but it is useful to at least know what they are in case you find the expression in a paper.

After we have discussed the centre of mass, it is also interesting to look at how far the outcomes are spread around the average. Assume two random variables X and Y with

- $P(X=-100) = 0.5$; $P(X=100) = 0.5$
- $P(Y=-1000) = 0.5$; $P(Y=1000) = 0.5$

The spread of Y is obviously going to be greater, although both have the same expectation. The quantity usually used to assess the spread of a RV is the variance.

Definition 3.10 (Variance) *The variance of a RV X is given by*

$$var(X) := \mathbb{E}[\,(X - \mathbb{E}[X])^2\,].$$

The expression for $var(X)$ is rather scary-looking but let us try to make sense of it. The inner part is the difference between each value of the random variable and the expectation. Thus the outer expectation is just a weighted sum of differences between values in the support and the expectation. However, since values of X can be smaller or greater than the expectation, this sum will be 0. Therefore we square it, leaving us with only positive values. In summary, the variance is the expectation of the squared differences between the expectation and the values in the support of X.

Computing the variance as an expectation can be cumbersome. We show how it can be done more efficiently.

$$var(X) = \mathbb{E}[\,(X - \mathbb{E}[X])^2\,] \tag{3.11}$$

$$= \sum_{x \in \text{supp}(X)} P(X = x)(x - \mathbb{E}[X])^2 \tag{3.12}$$

$$= \sum_{x \in \text{supp}(X)} P(X = x)(x^2 - 2x\mathbb{E}[X] + \mathbb{E}[X]^2) \tag{3.13}$$

$$= \sum_{x \in \text{supp}(X)} P(X = x)x^2 - \sum_{x \in \text{supp}(X)} P(X = x)2x\mathbb{E}[X] + \mathbb{E}[X]^2 \tag{3.14}$$

$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \tag{3.15}$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2\,. \tag{3.16}$$

For most intents and purposes we will use (3.16) to compute the variance.

Exercise 3.11 *Compute the variance of the RV Z from the previous section.*

Although variance is not linear, we can still try to figure out what happens when we multiply a RV by a constant or add a constant to it.

$$var(aX + b) = \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] \tag{3.17}$$

$$= \mathbb{E}[(aX + b - a\mathbb{E}[X] - b)^2] \tag{3.18}$$

$$= \mathbb{E}[a^2(X - \mathbb{E}[X])^2] \tag{3.19}$$

$$= a^2 var(X) \tag{3.20}$$

Exercise 3.12 *Give an alternative proof of $var(aX + b) = a^2 var(X)$ by using $var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.*

What we see that the variance is not affected by adding a constant to the RV. This makes sense, as constant additions just shift the expectation. However, the relation of each individual value to the expectation is not affected by that operation. On the other hand, when we multiply each value of the RV by a constant, the variance is scaled by the square of that

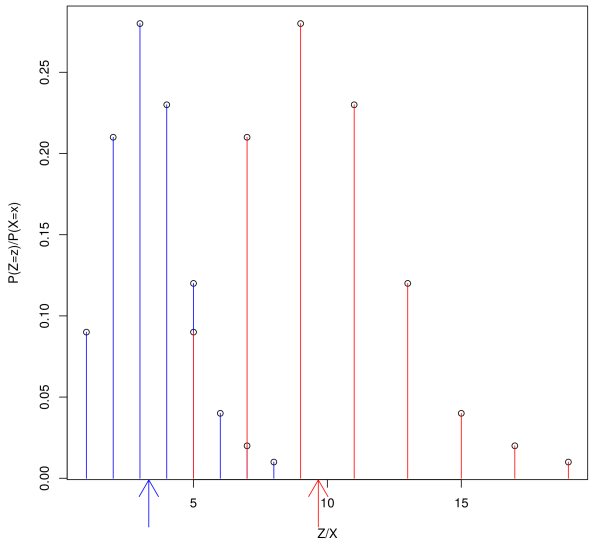


Figure 3.3: Plot of RVs Z and $X = 2Z + 3$. The arrows indicate the expectation of each distribution.

constant. Again, this is not surprising. Remember that we used squaring in our definition of the variance to turn all distances from the expectation into positive numbers. If we multiply an RV by a constant a, its spread will increase in both directions by a factor of a. Since we again want only positive differences, we better square that increased spread. This intuitively justifies the above equalities.

To gain a visual understanding of what happens when a RV gets scaled and a constant gets added to it, take a look at Figure 3.3. The random variable Z is the same as in Figure 3.1 and its probability distribution is plotted in blue. Figure 3.3 also shows the probability distribution of another RV $X = 2Z + 3$ in red. As predicted by our calculations above, the expectation (indicated by the red arrow) shifts by two times its original value plus three. Furthermore, note that the addition of three does not have any influence on the variance of X. The relative distance between the elements in the support of X is the same as the relative distance between the elements in the support of Z. However, the distance between the smallest and largest value in the support has doubled, hence increasing the variance by a factor of $2^2 = 4$.

3.3 Joint and conditional distributions

Up to now, we have only looked at cases that could be treated with one random variable. Most interesting problems involve several RVs, however. We introduce the concept of jointly distributed random variables. What this means is that there is a distribution over tuples of values, each from a different RV.

Definition 3.13 (Joint probability distribution) *Random variables $X_1, \ldots, X_n$ (**abbreviated as** X_1^n) are said to be **jointly distributed** with probability distribution $P_{X_1^n}$ if*

$$P(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(\{\omega \mid X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n\}).$$

From the above definition, we see that the random variables need to share an underlying sample space. This also gives us more insight into the real power of RVs: in applications, we usually do not care at all about the underlying sample space but only about quantities captured by the RVs. Therefore, we define a probability distribution over the random variables of interest. Since by definition the distribution has to fulfil conditions that allow us to interpret the distribution as a probability measure on a probability space, we can use the distribution without worrying about the underlying sample space.

Joint distributions are of particular interest because they make it possible to recover the probability distribution of each individual RV as well as of smaller joint distributions. The process by which this recovery can be accomplished is known as **marginalization**. Say we are given a joint distribution P_{XY}. How can we determine the probability that $X = x$? Our sample space is carved up by the values that X and Y can take on. We are interested in the probability of the subset $E = \{\omega | X(\omega) = x\}$. Assuming that Y can take on n values $y_1, y_2, \ldots, y_n$, let us define $F_i = \{\omega | Y(\omega) = y_i\}$ for $1 \leq i \leq n$. Some (possibly all) of the F_i will overlap with E. Thus we can partition E into $E \cap F_1, \ldots, E \cap F_n$. By countable additivity, we simply need to sum up the probabilities of these intersections. Observe that those probabilities correspond exactly to $P(X = x, Y = y_1), \ldots, P(X = x, Y = y_1)$. Hence, we get to the following equality:

$$P(X = x) = \sum_{i=1}^{n} P(X = x, Y = y_i) \tag{3.21}$$

If we set $X = (Z_1, \ldots, Z_m)$, we can recover $P_{Z_1^m}$ from $P_{Z_1^m Y}$ in the same way. Likewise, if we set $Y = (Z_1, \ldots, Z_m)$ we get the generalization of (3.21). We will assume that each Z_j can assume n_j values.

$X\backslash Y$	1	2	3
1	0.15	0.08	0.07
2	0.01	0.2	0.03
3	0.28	0.17	0.09

Table 3.1: Joint probability table for X and Y.

$$P(X = x) = \sum_{j_1=1}^{n_1} \dots \sum_{j_m=1}^{n_m} P(X = x, Z_1 = z_{j_1}, \dots, Z_m = z_{j_m}) \tag{3.22}$$

For exercise purposes, it often helps to draw a joint probability table and do marginalization from there.

Exercise 3.14 *Find the marginal distributions P_X and P_Y in table 3.1.*

We have already seen how the joint probability of random variables links back the the probability space that underlies them. This link makes it easy to import certain concepts that we have gotten to know in previous chapters. In particular, we can define **conditional probability distributions**.

Definition 3.15 (Conditional probability distribution) *Let X, Y be random variables with joint distribution P_{XY}. Let y be such that $P(Y = y) > 0$. The probability of $X = x$ conditioned on $Y = y$ is given by*

$$P(X = x|Y = y) := \frac{P(X = x, Y = y)}{P(Y = y)}.$$

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

Exercise 3.16 *Compute the conditional probability distributions $P_{X|Y=2}$ and $P_{Y|X=1}$ from Table 3.1.*

Likewise, the concept of independence carries over to distributions.

Definition 3.17 (Independence of random variables) *Two random variables X, Y are independent (denoted by $X \perp Y$) if $P_{XY} = P_X P_Y$, i.e. $\forall x \in \text{supp}(X), \forall y \in \text{supp}(Y) : P(X = x, Y = y) = P(X = x)P(Y = y)$.*

As with events, independence is equivalent to $P_{X|Y} = P_X$ (and to $P_{Y|X} = P_Y$). Moreover, independence makes it much easier to calculate

the expectation and variance of functions of jointly distributed random variables[1]. For independent X and Y, we have

$$\mathbb{E}[XY] = \sum_{x \in \text{supp}(X)} \sum_{y \in \text{supp}(Y)} P(X = x, Y = y)xy \tag{3.23}$$

$$= \sum_{x \in \text{supp}(X)} \sum_{y \in \text{supp}(Y)} P(X = x)P(Y = y)xy \tag{3.24}$$

$$= \sum_{x \in \text{supp}(X)} P(X = x)x \sum_{y \in \text{supp}(Y)} P(Y = y)y \tag{3.25}$$

$$= \mathbb{E}[X]\mathbb{E}[Y] \tag{3.26}$$

Similarly, it does not really make sense to talk about the variance of two random variables (at least not as one single value). However, we can again compute the variance of functions. For addition, independence again makes our lives much easier.

Exercise 3.18 *Show that for independent RVs X and Y, we have that $var(X + Y) = var(X) + var(Y)$. Give an example of dependent X and Y with $var(X + Y) \neq var(X) + var(Y)$.*

3.4 Three Important Distributions

We will finish this chapter by introducing two extremely useful discrete distributions. They are very often used in computer science to model the probability of bit strings and in linguistics and natural language processing to model the probability of sentences.

To get us started, let us consider coin tosses. When you toss a fair coin, the probability that it lands on heads is 0.5 and so is the probability that it lands on tails. Let 1 stand for heads and 0 for tails in a random variable X modelling coin tosses. We can formalize this distribution as

$$P(X = x) = 0.5^x \times 0.5^{1-x} = 0.5 \text{ for } x \in \{0, 1\} \ .$$

Notice though that not all coins are fair and that in general we want to allow for probabilities that are not equal. We will therefore introduce a parameter θ that regulates the probability that we observe heads. In the above equation, we set $\theta = 0.5$. What happens if we set $\theta = 0.3$?

$$P(X = x) = 0.3^x \times 0.7^{1-x} = \begin{cases} 0.3 & \text{if } x = 1\,, \\ 0.7 & \text{if } x = 0\,. \end{cases} \tag{3.27}$$

[1]Notice that the joint random variable (X, Y) is not a *real-valued* random variable anymore. Hence $\mathbb{E}[X, Y]$ and $var(X, Y)$ are not defined. Rather, we can define a new (real-valued) random variable as $f(X, Y)$ for a function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. and compute $\mathbb{E}[f(X, Y)]$ and $var(f(X, Y))$.

Where did we get the 0.7 from? Well, we know that the coin has to land on either heads or tails. Thus, if the probability to land on heads is 0.3 the probability to land on tails has to be 0.7. Moreover, we chose 1 as a stand-in for landing on heads. As we would expect, $X = 1$ will return the 0.3 probability and $X = 0$ will return the probability of 0.7.

This leads to a general formulation where we do not specify the parameter θ but leave it to be filled in. The resulting distribution, known as **Bernoulli distribution** (with parameter θ), is shown in Equation (3.28).

$$P(X = x) = \theta^x \times (1 - \theta)^{1-x} = \begin{cases} \theta & \text{if } x = 1\,, \\ 1 - \theta & \text{if } x = 0\,. \end{cases} \tag{3.28}$$

The Bernoulli is defined for one coin flip, or more generally for the drawing of a coloured ball from a (very large) urn containing balls of exactly two colours (in proportion θ and $1 - \theta$). What if we draw balls repeatedly with replacement? We will get a sequence of coloured balls. Say we draw n balls and repeat this procedure five times. Chances are that the five sequences will look rather different. Thus, we can define a probability distribution over sequences of coloured balls of length n. This procedure can be described as a simple generalization of the Bernoulli. Whereas in the Bernoulli we made one draw we make n draws. This yields a sequence x in which each position $x_i, 1 \leq i \leq n$ is either 0 or 1.

$$P(X = x) = \theta^{\sum_i x_i} \times (1 - \theta)^{n - \sum_i x_i} \tag{3.29}$$

The adjustment is rather small. We basically only introduce an extra parameter, namely the number of draws n. However, there are two conditions for this distribution that we have not mentioned yet. First and foremost, the repeated Bernoulli draws are assumed to be independent. That is, for each draw in the sequence it should not matter which colours you have drawn so far and which ones you are still going to draw (as you are drawing with replacement). More formally, if we encode each draw in the sequence as a random variable, we postulate that these random variables are independent of each other.

The second point is that we cannot assign a probability to any sequence that contains more balls of one colour than there are balls in total. This is to say that we require that $0 \leq \sum_{i=1}^{n} x_i \leq n$ which is a reasonable restriction.

It is also worth mentioning that in general the two values that the distributions in (3.28) and (3.29) range over are generally referred to as success and failure where the success is usually encoded as 1 and the failure as 0.

Exercise 3.19 *Use the distribution in Equation (3.29) with $n = 10$ and $\theta = 0.8$ to determine the probability of the sequence (0,0,1,1,0,1,1,1,0,1).*

Another interesting observation is that, since the value at each position in the sequence is independent of all others, all sequences with the same number of successes and failures will have the same probability. At this point it is natural to ask for the probability to get *any* sequence with that amount of successes and failures. Say we are dealing with sequences of length n and wonder about the probability of obtaining a sequence with $0 \leq k \leq n$ successes. Then we simply need to sum up the probabilities of all sequences that contain k successes. But how many such sequences are there? In Chapter 1, we talked about how to choose a subset of k elements out of n. This is exactly the problem we are facing here. We want to know how many ways there are to choose k out of n positions that we interpret as successes. This counting is done by the binomial co-efficient $\binom{n}{k}$.

We can generalize the distribution from Equation (3.29) to a distribution that gives the probability of obtaining any sequence with k successes. This distribution is known as the **binomial distribution** (with parameters n and θ).

$$P(X = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{3.30}$$

The binomial is of crucial importance in many fields. For example in computer science, it is used to compute the probability of bit strings. Other applications include the assessment of failure rates. Say you run a company and a customer asks you to supply k items within a week. The capacity of your company allows you to produce at most $n > k$ items during that week. At the same time you know that each item that you produce has a probability of being faulty. Obviously, you cannot sell faulty items to your customer. The Binomial distribution allows you to compute the probability that you will be able to meet the customer's demand. Based on this calculation, you can evaluate whether or not it is reasonable to accept the customer's order.

The third important distribution that we get to know today is the **multinomial distribution**. It is basically a generalization of the binomial. A RV that is distributed according to the multinomial distribution can take on finitely many values. Let us say that there are m such values. Furthermore let us assume that each value occurs c_i times in a sequence of length n where $1 \leq i \leq n$ and $\sum_{i=1}^{m} c_i = n$. Then the multinomial with parameters n and $\theta_1, \ldots, \theta_n$ is given as

$$P(X_1 = c_1, \ldots, X_m = c_m) = \frac{n!}{\prod_{i=1}^{m} c_i!} \prod_{i=1}^{m} \theta_i^{c_i} \tag{3.31}$$

A further condition on the multinomial is that $\sum_{i=1}^{m} \theta_i = 1$. Just as with the binomial, we can in principle infer (the last) θ_m and c_m if we know the

value of the other θ_i and c_i, $1 \leq i \leq (m-1)$. However, this would clutter notation unnecessarily and we hence do not do it here. It is noteworthy, though, that because of this fact, the multinomial has $m-1$ and not m θ-parameters (since θ_m is determined by the other θ_i).

It is easy to see that the binomial is in fact just the special case $m = 2$ of the multinomial. Because of its general importance it is usually considered a separate distribution though.

Let us finish this section by introducing some useful notation. When a RV is distributed according to some distribution, one often uses the tilde($\sim$) to express this fact. If, for example, we have a random variable X that is distributed according to a binomial with $n = 100$ and $\theta = 0.5$, many authors will write

$$X \sim binom(100, 0.5)$$

Now that we know what it means for a random variable to be distributed according to some distribution, we can also clarify a question that was brought up in the beginning. What does sampling *uniformly at random* mean? It just means that we are sampling from a distribution that is the uniform distribution (all values in the support have the same probability). The *random* comes from the fact that we are sampling the values of a random variable. Just saying that some quantity is sampled at random is not enough. You should always add which distribution underlies that randomness!

3.5 The Negative Binomial Distribution*

Another interesting distribution that is well worth looking at is the negative binomial distribution. Suppose we do repeated independent Bernoulli trials and we wonder how many trials it will take until we have observed $k > 0$ successes. To answer this question, we would like to assign a probability to each integer $t \geq k$ which can be interpreted as the probability that we will take t trials until we have obtained the desired number of successes. Notice that for each t, we can infer the number of failures as $f = t - k$.

To approach this problem, let us start out from the simplest case where we are only waiting for one success. In that case, our sequence of trials ends in a success, which is also the only success in the sequence. Thus the sequence contains $f = t - 1$ failures. The success probability of a Bernoulli trial is θ, as usual. Then the associated probability distribution for T with parameters θ and $k = 1$ is

$$P(T = t) = \theta^1(1-\theta)^{t-1} \tag{3.32}$$

Notice that if $k = 1$ there is only one sequence of size t that ends in a success. If we let $k > 1$, there are more sequences of length y that end in a success and contain k successes in total. The probability for any such

sequence can easily be calculated by generalising Equation (3.32).

$$P(T = t) = \theta^k (1 - \theta)^{t-k} \tag{3.33}$$

As with the Binomial, we now have to add up the probabilities of all sequences of length t that contain exactly k successes *and end in a success*. The last condition is what separates the negative Binomial from the Binomial distribution. We first observe that the position of the last success is fixed. This means we are left with $k - 1$ which we can assign to different positions. In total there are $t-1$ positions that we can assign. Thus we have to choose $k - 1$ out of $t - 1$ position over which to distribute the remaining $k - 1$ successes. Thus, we conclude that there are $\binom{t-1}{k-1}$ sequences of length t that contain k successes, one of which occurs in the last position of the sequence. This is allows us to define the negative Binomial distribution with parameters θ and k as

$$P(T = t) = \binom{t-1}{k-1} \theta^k (1 - \theta)^{t-k} \tag{3.34}$$

where $t \in \mathbb{N}, t > k$.

Many authors use the notation $X \sim nbinom(\theta, k)$ to state that the RV X is distributed according to a negative Binomial distribution.

Chapter 4

Bayes' rule and its applications

4.1 The chain rule

This chapter is going to focus on how to re-write joint and conditional probabilities. When we turn to statistics later on, it will turn out that it is often hard to define a joint distribution over many variables. Likewise, it can be hard to calculate the probability distribution of a RV X conditioned on a RV Y but it may be much easier to find the distribution of Y conditioned on X. In this chapter we are essentially trying to find simpler expressions for distributions that may be hard to compute.

The first general method for simplifying a joint distribution is known as the **chain rule**. For completeness' sake, we are going to formulate the chain rule first for events and then for random variables.

Theorem 4.1 ***(Chain rule)*** *The joint probability of events $E_1, \ldots, E_n$ can be factorised as*

$$\mathbb{P}(E_1, \ldots, E_n) = \mathbb{P}(E_1) \times \mathbb{P}(E_2|E_1) \times \ldots \times \mathbb{P}(E_n|E_1, \ldots, E_{n-1})$$

Recall from Definition 2.6 the notation $\mathbb{P}(E_1, E_2) = \mathbb{P}(E_1 \cap E_2)$ for denoting the probability that both events E_1 and E_2 occur. Also remember that we use the abbreviation $E_1^n := E_1, \ldots, E_n$; so for the case of events, we have $\mathbb{P}(E_1^n) = \mathbb{P}(\bigcap_{i=1}^n E_i)$. There are a couple of things to note about the chain rule: First of all, the numbering of the events is arbitrary. That means that it does not matter in which order we decompose the joint probability. We could just as well start with any E_i for $1 \leq i \leq n$. Second we used the word *factorise*. This simply means that we decompose any expression (in this case a joint probability) into a product. Products are nice in that we can arrange them in any order that we like (i.e. they commute). Moreover, products make a lot of calculations easier, as we will see later.

Let us go ahead and actually prove the chain rule.

Proof of Theorem 4.1 We are going to do so inductively and choose $\mathbb{P}(E_1, E_2)$ as our base case. Then we simply employ the definition of conditional probability to get

$$\mathbb{P}(E_1, E_2) = \mathbb{P}(E_1) \times \frac{\mathbb{P}(E_1, E_2)}{\mathbb{P}(E_1)} = \mathbb{P}(E_1) \times \mathbb{P}(E_2|E_1) \tag{4.1}$$

Let us assume that the chain rule holds for events $E_1, \ldots, E_{n-1}$. We will abbreviate them as E_1^{n-1}. Then we get

$$\mathbb{P}(E_1^{n-1}, E_n) = \mathbb{P}(E_1^{n-1}) \times \frac{\mathbb{P}(E_1^{n-1}, E_n)}{\mathbb{P}(E_1^{n-1})} = \mathbb{P}(E_1^{n-1}) \times \mathbb{P}(E_n|E_1^{n-1}) \tag{4.2}$$

Since $\mathbb{P}(E_1^{n-1})$ factorises according to the chain rule by our induction hypothesis, we have completed the proof. □

The chain rule can make our lives even simpler if we have independent events. Assume we want to compute the joint probability of 3 events E_1, E_2, E_3 and we also know that $E_1 \perp E_2$. In this case our factorisation becomes (4.3) where the first equality follows from the chain rule and the second equality follows from independence between E_1 and E_2.

$$\begin{aligned} \mathbb{P}(E_1, E_2, E_3) &= \mathbb{P}(E_1) \times \mathbb{P}(E_2|E_1) \times \mathbb{P}(E_3|E_1, E_2) \\ &= \mathbb{P}(E_1) \times \mathbb{P}(E_2) \times \mathbb{P}(E_3|E_1, E_2) \end{aligned} \tag{4.3}$$

We can now state the chain rule for random variables. There are two ways you can go about proving it. Either you calculate the probability of a specific setting of the variables or you just do the proof based on the distributions of the RVs. So in the first case you would have to prove that

$$\begin{aligned} \forall x_1, \ldots, x_n : &P(X_1 = x_1, \ldots, X_n = x_n) \\ &= P(X_1 = x_1) \times \ldots \times P(X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \end{aligned}$$

whereas in the second case you would simply prove that

$$P_{X_1^n} = \sum_{i=1}^{n} P_{X_i|X_1^{i-1}}$$

Incidentally, we also introduce a very short notation for the chain rule above. Note that it is not quite correct, since if $i = 1$ we would be conditioning on X_0. That is not to bad however, since we can always define ourselves a constant variable X_0 that does not affect the distribution. Moreover, this notation is really just meant to be convenient, so you should just accept it as is when you encounter it in papers.

Exercise 4.2 *Prove the chain rule for random variables. The proof is totally analogous to the one give for events.*

Exercise 4.3 *Let X_0 be a constant RV, i.e. there exists $c \in \mathbb{R}$ such that $P(X_0 = c) = 1$. Prove that X_0 is independent of any set of other random variables $X_1, \ldots, X_n$.*

4.2 Bayes' rule

In this section we are going to prove **Bayes' rule**. The rule follows directly from the chain rule. The proof is really simple and thus of no great interest in and by itself. The consequences of Bayes' rule are huge however. It will basically allow us to invert a conditional probability distribution. You may rightfully ask: what's the deal? Well, as we said in the beginning, it may be hard to compute a conditional distribution in one direction but much easier to compute it in the other direction. On top of that, Bayes' rule opens up a whole range of new possibilities. We will discuss those as we proceed in this chapter.

Theorem 4.4 ***(Bayes' rule)*** *The probability distribution of a random variable X given a random variable Y can be computed as*

$$P_{X|Y} = \frac{P_{Y|X} P_X}{P_Y}$$

And here comes the proof:

$$P_{X|Y} = \frac{P_{XY}}{P_Y} = \frac{P_{Y|X} P_X}{P_Y}\,. \qquad \square \tag{4.4}$$

That was the proof! Considering how simple it was, it will be surprising to see what kind of benefits we can get out of Bayes' rule. To get us started, let us introduce some terminology. In particular, each of the terms in Bayes' rule has a specific name. You should really learn these names by heart as they crop up all over the place.

$$\textit{posterior} = \frac{\textit{likelihood} \times \textit{prior}}{\textit{marginal likelihood}}$$

The posterior is what we get after we have completed the computation. However, its name is related to the prior. The prior is just the probability that we would place on $P(X = x)$ *a priori*. Therefore P_X is also known as the prior distribution. When we divide the product of likelihood and prior by the marginal likelihood we get a new distribution over X that is conditioned on Y. This is the distribution that we place on X *a posteriori*,

i.e. after having taken into account information about X that we may get from knowing the value of Y. The marginal likelihood of Y is simply needed to normalize the expression to a probability distribution (i.e. to make sure that it sums to one). Why is it called marginal likelihood? The reason for this is how you can compute it. Recall that when we are given a joint distribution P_{XY}, we can obtain the distribution P_Y by simply marginalizing over X.

$$P(Y = y) = \sum_{x \in \text{supp}(X)} P(X = x, Y = y) \tag{4.5}$$

In addition to that, the chain rule allows us to factorise the joint probability. Thus we get

$$P(Y = y) = \sum_{x \in \text{supp}(X)} P(Y = y | X = x) \times P(X = x) \tag{4.6}$$

If you think that this looks an awful lot like the enumerator of Bayes' rule then you are exactly on the right track. Essentially, we are just summing over all possible denominators (with respect to X). Let us make this more concrete with an example. Assume that we are given two coins. One of them is fair, meaning that it is equally probable to come up heads or tails. The other coin is biased towards tails and we happen to know that its probability to come up heads is only 0.3. Which coin is flipped is captured by a random variable X that takes on the value 0 if the fair coin is used and the value 1 if the biased coin is used. We have no idea which coin is going to be tossed, it could be either one. Therefore we set our prior to $P(X = 0) = P(X = 1) = 0.5$.

We flip the chosen coin 10 times and obtain 8 heads. The number of heads obtained during the 10 tosses is going to be encoded by Y. Since all tosses are independent of each other, Y will follow a binomial distribution. For each of the two coins we also know the parameter of the binomial distribution. For the fair coin it is $\theta = 0.5$ and for the biased coin it is $\theta = 0.3$. Let us compute each of the enumerators separately.

$$P(Y = 8 | X = 0) \times P(X = 0) = \binom{10}{8} 0.5^8 (1 - 0.5)^2 \times 0.5 = 0.02195 \tag{4.7}$$

$$P(Y = 8 | X = 1) \times P(X = 1) = \binom{10}{8} 0.3^8 (1 - 0.3)^2 \times 0.5 = 0.0007 \tag{4.8}$$

Remember that $Y \sim binom(10, \theta)$ and that $\theta = 0.5$ if $X = 0$ and $\theta = 0.3$ if $X = 1$.

All that is left do is to compute the marginal likelihood of Y. Luckily for us, X only assumes two values, so we only need to add up (4.7) and (4.8).

$$\begin{aligned} P(Y = 8) =& P(Y = 8 | X = 0) \times P(X = 0) \\ &+ P(Y = 8 | X = 1) \times P(X = 1) = 0.02265 \end{aligned} \tag{4.9}$$

And finally we can apply Bayes' rule to compute the posterior probabilities of X.

$$P(X=0|Y=8) = \frac{P(Y=8|X=0) \times P(X=0)}{P(Y=8)} \tag{4.10}$$

$$= \frac{0.2195}{0.02265} = 0.969$$

$$P(X=1|Y=8) = \frac{P(Y=8|X=1) \times P(X=1)}{P(Y=8)} \tag{4.11}$$

$$= \frac{0.0005}{0.02265} = 0.031$$

(4.12)

There is a probability of 0.969 that the fair coin has been tossed when a sequence with eight heads is generated and only a probability of 0.031 that the biased coin was tossed. Obviously, the probability of the fair coin is much higher. But how much higher? We can take the ratio of the two probabilities. This gives us $^{0.969}/_{0.031} \approx 31$. We can conclude that the fair coin is 31 times more likely to have generated the sequence with 8 heads than the biased coin. But wait a second, can we maybe find this ratio somewhere else? It turns out that the ratio of the likelihoods is the same! That is $^{0.0439}/_{0.0014} \approx 31$.

We started out by assuming that both coins were equally likely to be used. However, we then observed a sequence of 10 tosses, 8 of which were heads and that made it 31 times more likely that the fair coin was used. What if the priors had not been equal? Actually, there is a more general story: While calculating the actual probabilities involves a lot of number crunching, just telling whether or not an observation will make one or the other event more likely is not too hard. [For the rest of this chapter, we assume that we only condition on events with non-zero probabilities such as $P(Y=y) > 0$ so that we are never dividing by 0].

$$\frac{P(X=x_1|Y=y)}{P(X=x_2|Y=y)} = \frac{\frac{P(Y=y|X=x_1)P(X=x_1)}{P(Y=y)}}{\frac{P(Y=y|X=x_2)P(X=x_2)}{P(Y=y)}}$$

$$= \frac{P(Y=y|X=x_1)P(X=x_1)}{P(Y=y|X=x_2)P(X=x_2)}$$

From the above equalities, we see that the ratio of the posterior probabilities is determined by the ratio of the likelihood times the prior. In our coin example, the priors were the same so it was only the likelihood that mattered. If the ratio of any of the above terms is greater than 1, the posterior will change in favour of $X = x_1$. If the ratio is smaller than 1 the

posterior changes in favour of $X = x_2$. If the ratio is exactly 1, the posterior stays unchanged.

Notice that in general, although our observations may shift the posterior in favour of $X = x_2$, say, this shift does not necessarily imply that $P(X = x_2|Y = y)$ will be greater than $P(X = x_1|Y = y)$. The condition that $P(X = x_2|Y = y)$ is bigger than $P(X = x_1|Y = y)$ can be rewritten as follows

$$\begin{aligned}
P(X = x_1|Y = y) &< P(X = x_2|Y = y) && \Leftrightarrow \\
\frac{P(Y = y|X = x_1)P(X = x_1)}{P(Y = y)} &< \frac{P(Y = y|X = x_2)P(X = x_2)}{P(Y = y)} && \Leftrightarrow \\
P(Y = y|X = x_1)P(X = x_1) &< P(Y = y|X = x_2)P(X = x_2) && \Leftrightarrow \\
\frac{P(Y = y|X = x_1)}{P(Y = y|X = x_2)} &< \frac{P(X = x_2)}{P(X = x_1)}
\end{aligned}$$

The last line is of particular interest as it elucidates the relationship between the prior and the likelihood. Only if the likelihood ratio for $X = x_1$ over $X = x_2$ is smaller than the reversed prior ratio will the posterior probability of $X = x_2$ be greater than that of $X = x_1$. This means that if we have strongly asymmetric priors (like $P(X = x_1) = 0.9$ and $P(X = x_2) = 0.1$), the likelihood needs to discriminate very well between the two cases in order to tip the scale in favour of $X = x_2$. In that sense the prior and the likelihood can be seen as battling forces whose equilibrium gives us the posterior.

But enough theory about Bayes' rule, it is about time you apply it! To that end, we present you an exercise that is, in some variation, contained in virtually every textbook on probability theory, statistics or machine learning. Have fun with it!

Exercise 4.5 *A random person walks into the doctor's office to be tested for a particular disease. The disease can be fatal if not treated. However, successful treatment is possible if the disease is discovered early enough. It is commonly known that the disease occurs in 1 out of 1000 people of the country's population. The doctor will administer a test that with a probability of 99% returns a positive results if the patient does indeed have the disease. At the same time, the test also returns a positive result in 5% of the cases where the patient does not have the disease. After the test has been administered to the patient in question, it returns a positive result. What is the probability that the patient is infected with the disease?*
Proceed as follows:

1. *Write down a guess for what you think the probability might be (do not consider any math at this point).*

2. *Calculate that probability.*

3. *Check whether there is a considerable difference between your initial guess and the calculated probability. Go on to examine how the different factors have influenced the probability of the patient having the disease.*

Let us finish up this section with some more notation. In many applications of Bayes' rule we only want to know which outcome is the most likely, without worrying too much about the actual probabilities. Likewise, there is a range of situations where we just want to assign a score to outcomes and do not demand this score to be a probability. Throughout this chapter, we have repeatedly encountered the following phenomenon: In order to rank the values of an RV according to their probabilities, we do not necessarily need to compute the marginal likelihood since it cancels in all these comparisons anyway. Therefore, you will often see authors stating that

$$P(X = x|Y = y) \propto P(Y = y|X = x)P(X = x) \tag{4.13}$$

This equation reads as "the posterior is proportional to the product of the likelihood and the prior". In general, if we have two quantities a and b, then by $a \propto b$ we mean that there is some constant $C \in \mathbb{R} \setminus \{0\}$ such that $a = Cb$. Notice that the probability distribution is a function and hence we require C to be the same across the domain of that function (that is C should be the same for all values of X).

Exercise 4.6 *What is the value of C in Equation (4.13)?*

4.3 Naïve Bayes

In this section, we introduce a rather crude application of Bayes's rule which is surprisingly successful nonetheless. Assume that instead of one random variable we are observing a sequence of random variables. Thus our problem is the following:

$$P(Y = y|X_1^n = x_1^n) \propto P(X_1^n = x_1^n|Y = y) \times P(Y = y) \tag{4.14}$$

By the chain rule we can decompose the right-hand side into

$$\begin{aligned} P(Y = y|X_1^n = x_1^n) \propto & P(X_1 = x_1|Y = y) \times \dots \\ & \times P(X_n = x_n|Y = y, X_1^{n-1} = x_1^{n-1}) \times P(Y = y) \end{aligned}$$

We are now going to introduce the aforementioned crudeness into the model by assuming that all $X_1, \dots, X_n$ are conditionally independent given Y. Notice that this is just an assumption that we are making without

justification. In fact, it is very likely wrong. However, it makes our live much easier because we only have to deal with very simple terms of the form $P(X_i = x_i|Y = y)$. Because of the crudeness of our assumptions, this probabilistic model is known as **naïve Bayes** (sometimes also stupid Bayes).

Definition 4.7 *A naïve Bayes model is a probabilistic model that assumes*

$$P_{Y|X_1^n} \propto P_Y P_{X_1|Y} P_{X_2|Y} \cdots P_{X_n|Y}$$

Once we know all the component distributions $P_{X_i|Y}$, calculating the result is pretty straightforward.

In order to illustrate how naïve Bayes works we are going to employ one of its showcase applications where it indeed had a lot of success in real life. The application we are talking about is text classification. The task is the following: you are given some documents and for each of the documents you have to assign a label signifying its class. What you consider a class depends on your actual application setting, but usually classes are broad categories, such as legal texts, medical texts etc. If you manage to succeed at this task, you can accomplish a lot of things automatically that required humans before. For example, you could tag online news with their relevant categories and people who are interested in a particular category will then have an easier time finding the news related to that category. Crucially, since you will write a computer program that does the classification for you, you will not need to read any of the texts yourself. This automation will obviously allow you to classify huge quantities of text in a very short amount of time.

Exercise 4.8 *A collection of text (or any other kind of data for that matter) is often called a* ***corpus****. Here we are going to use a toy corpus. The corpus just consists of two sentences and we assume that each sentence constitutes a document. The categories that you can label the documents with are finance (0), medicine (1) or law (2). You can find the corpus (the pmfs of the distributions) below. For simplicity, we are not going to distinguish between lower and upper case words (this is actually common practice). For better readability, we are also using the actual words instead of their numerical encodings as values for the random variables. Just remember that those words could also be represented as real random variables. To shorten notation, we will use pmfs. If the probability of a word given a category is not specified, take it to be 0.*

Your task is to classify these two documents correctly using a Naïve Bayes Model that conditions each word's probability on the document class. Please also report the posterior probability for the correct label.

The corpus:

- a fact has been revealed
- the doctor's judgement has not been reliable

The document category pmfs:

- $p(0) = 0.3$
- $p(1) = 0.2$
- $p(2) = 0.5$

The lexical distribution for document category finance (0):

$p(a|0) = 0.19 \quad p(fact|0) = 0.14 \quad p(has|0) = 0.13 \quad p(been|0) = 0.12$
$p(revealed|0) = 0.04 \quad p(the|0) = 0.21 \quad p(doctor's|0) = 0.03$
$p(judgement|0) = 0 \quad p(not|0) = 0.11 \quad p(reliable|0) = 0.03$

The lexical distribution for document category medicine (1):

$p(a|1) = 0.02 \quad p(fact|1) = 0.08 \quad p(has|1) = 0.13 \quad p(been|1) = 0.13$
$p(revealed|1) = 0.01 \quad p(the|1) = 0.18 \quad p(doctor's|1) = 0.06$
$p(judgement|1) = 0.14 \quad p(not|1) = 0.20 \quad p(reliable|1) = 0.05$

The lexical distribution for document category law (2):

$p(a|2) = 0.18 \quad p(fact|2) = 0.03 \quad p(has|2) = 0.05 \quad p(been|2) = 0.13$
$p(revealed|2) = 0.10 \quad p(the|2) = 0.14 \quad p(doctor's|2) = 0.06$
$p(judgement|2) = 0.07 \quad p(not|2) = 0.08 \quad p(reliable|2) = 0.16$

Further Reading

Here, we have only scratched the surface of what Bayes' rule allows us to do. To get a wider outlook on what else is possible, you can consult Kevin Murphy's webpage.