

# A Primer in Discrete Probability

Dominik Scheder

## Abstract

## 1 Finite Probability Spaces

Here is a random experiment all of us have done in our lives: rolling a die. There are six possible outcomes, 1,2,3,4,5, and 6. Assuming that the die is

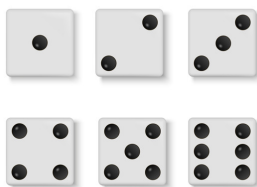


Figure 1: The six sides of a die.

“fair”, each outcome is equally likely, and thus has probability  $1/6$ . This is a toy example of a *probability space*.

**Definition 1.** A finite probability space is a finite set  $S$  together with a function  $P : S \rightarrow \mathbb{R}$  such that  $P(s) \geq 0$  for every  $s \in S$  and  $\sum_{s \in S} P(s) = 1$ .

In our die-rolling example,  $S$  would be the set  $\{1, 2, 3, 4, 5, 6\}$  and  $P$  is the function that maps every element to  $1/6$ . Suppose we roll two dice, a white and a red one. We get another probability space. This probability space has 36 elements, all possible combinations of rolling two dice. We could write down  $S$  explicitly, as  $S = \{(1, 1), (1, 2), (1, 3), \dots, (6, 4), (6, 5), (6, 6)\}$ , if we wanted to.

















































						
						
						
						
						
						
						

Figure 2: Rolling two dice.

## 1.1 Events

Sometimes, at least one of the two rolled dice shows a 5 or higher. Sometimes not, and then both dice are 4 or lower. What is the probability of this happening? Let  $\mathcal{E}$  denote the event that at least one of the two dice is 5 or higher. What kind of object is  $\mathcal{E}$ , mathematically speaking? Formally,  $\mathcal{E}$  is a subset of the probability space:  $\mathcal{E} \subseteq S$ . Indeed,

$$\begin{aligned} \mathcal{E} = \{ & (1, 5), (1, 6), (2, 5), (2, 6), (3, 5), (3, 6), (4, 5), (4, 6), \\ & (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ & (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \} . \end{aligned}$$

So what is the probability of  $\mathcal{E}$ ? There are 36 possible outcomes, i.e.,  $|S| = 36$ ; for 20 of them  $\mathcal{E}$  happens, i.e.,  $|\mathcal{E}| = 20$ . The total probability of all those elements in  $\mathcal{E}$  is

$$P(\mathcal{E}) = \sum_{s \in S} P(s) = \frac{|\mathcal{E}|}{|S|} = \frac{20}{36} = \frac{5}{9} .$$

**Definition 2.** Let  $S$  be a finite probability space. An event is a subset  $\mathcal{E} \subseteq S$  and its probability is  $P(\mathcal{E}) := \sum_{s \in S} P(s)$ .

## 1.2 Independence

Let us come up with a different, more elegant way of computing  $P(\mathcal{E})$ . We define two events  $A$  and  $B$  as follows:  $A$  is the event that the first die is 4 or lower;  $B$  is the event that the second die is 4 or lower. Let us observe three simple facts:

1.  $P(A) = 4/6$ .
2.  $P(B) = 4/6$ .
3.  $\mathcal{E} = S \setminus (A \cap B)$ .

How does this help us to compute  $P(\mathcal{E})$ ? We observe that  $A$  and  $B$  are independent events.

**Definition 3.** Let  $A, B \subseteq S$  be two events. We say  $A$  and  $B$  are independent if  $P(A \cap B) = P(A) \cdot P(B)$ .

You are encouraged to check that this indeed holds for our two events  $A$  and  $B$ . Intuitively,  $A$  talks only about the first die, and  $B$  only about the second, and the two rolls are independent. We can compute:

$$P(\mathcal{E}) = 1 - P(A \cap B) = 1 - P(A) \cdot P(B) = 1 - \frac{4}{6} \cdot \frac{4}{6} = \frac{5}{9}.$$

Suppose we roll a die  $n$  times. Let  $\mathcal{E}$  be the event that we see at least one 6. What is  $P(\mathcal{E})$ ? For  $i = 1, \dots, n$ , let  $A_i$  denote the event that the  $i^{\text{th}}$  die is 5 or lower. Clearly,  $P(A_i) = 5/6$  and  $\mathcal{E} = S \setminus (A_1 \cap \dots \cap A_n)$ . Observe that each  $A_i$  “talks about a different die”, so they are all independent.

**Definition 4.** Let  $A_1, \dots, A_n \subseteq S$  be events in a probability space  $S$ . We say  $A_1, \dots, A_n$  are independent if

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i) \tag{1}$$

holds for all sets  $I \subseteq [n]$ .

Intuitively, since  $A_i$  talks about a different die, they should be independent. You are invited to verify this more formally. We compute

$$P(\mathcal{E}) = 1 - P(A_1 \cap \dots \cap A_n) = 1 - \left(\frac{5}{6}\right)^n.$$

**Note:** Often we write  $\Pr[\mathcal{E}]$  instead of  $P(\mathcal{E})$ , if  $P$  is clear from the context. The reason is that the reader will immediately recognize  $\Pr[\cdot]$  as a probability, whereas  $P(\cdot)$  can have many meanings in mathematics.

### 1.3 Example: Bit Strings Without 11

Consider the set

$$A_n := \{x \in \{0,1\}^n \mid x \text{ does not contain the substring } 11\}$$

We have seen in the lecture that  $|A_n| = F_{n+2}$ , the  $(n+2)^{\text{nd}}$  Fibonacci number. Let  $P_n$  be the uniform distribution over  $A_n$ . That is,  $P_n(x) = \frac{1}{|A_n|}$  for every  $x \in A_n$ . So  $(A_n, P_n)$  form a probability space.

**Exercise 1.1.** Let  $Q_n^i$  be the event that  $x_i = 1$ .

1. Give an explicit formula for  $\Pr[Q_n^i]$ . Recall the explicit formula for the Fibonacci numbers.
2. Give a simpler, “asymptotic” formula for  $\Pr[Q_n^i]$ , if  $n$  is large and  $i$  is neither close to 0 nor to 1. For example, take some  $0 < \delta < 1$  and compute the limit

$$q := \lim_{n \rightarrow \infty} \Pr[Q_n^i], \text{ where } i = \lfloor \delta n \rfloor$$

This is to say, “each bit of  $x$  not too close to the left or right is 1 with probability close to  $q$ ”.

3. Show that the events  $Q_n^i$  and  $Q_n^j$  are not independent.
4. Show that they are “asymptotically independent” if 1,  $i$ ,  $j$ , and  $n$  are sufficiently far apart. More precisely, show that for any  $0 < \alpha < \beta < 1$ , it holds that

$$\lim_{n \rightarrow \infty} \Pr[Q_n^i \cap Q_n^j] = \lim_{n \rightarrow \infty} \Pr[Q_n^i] \cdot \Pr[Q_n^j],$$

where  $i = \lfloor \alpha n \rfloor$  and  $j = \lfloor \beta n \rfloor$ .

## 2 Random Variables

Assume again that we roll two dice, and let  $X$  denote the value of the higher one. Obviously,  $X$  can take on values 1, 2, 3, 4, 5, 6, but higher values will be more likely. Formally, a random variable is a function  $S \rightarrow \mathbb{R}$ . In this concrete example, an element of  $S$  has the form  $(s_1, s_2)$ , and the random variable  $X$  is the function

$$X : S \rightarrow \mathbb{R}, \quad (a_1, a_2) \mapsto \max(a_1, a_2).$$













						
	1	2	3	4	5	6
	2	2	3	4	5	6
	3	3	3	4	5	6
	4	4	4	4	5	6
	5	5	5	5	5	6
	6	6	6	6	6	6

Figure 3: The maximum value of two dice.

What are random variables good for? First of all, they simplify notation. Our first event  $\mathcal{E}$  that at least one die has value 5 or larger, can be conveniently written as “ $X \geq 5$ ”. Its complement, that both are 4 or smaller, as “ $X \leq 4$ ”. Thus,

$$P(X \geq 5) = \frac{5}{9} .$$

Second, we can talk about “average outcomes”.

## 2.1 Expectation

What is the “average value” of  $X$ ? The possible outcomes of  $X$  are 1, 2, 3, 4, 5, 6, but larger values are more likely. In particular,

$i$	1	2	3	4	5	6
$P(X = i)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

Thus, the average value of  $X$  is

$$\frac{1}{36} \times 1 + \frac{3}{36} \times 2 + \frac{5}{36} \times 3 + \frac{7}{36} \times 4 + \frac{9}{36} \times 5 + \frac{11}{36} \times 6 = \frac{161}{36} = 4.7222\dots .$$

**Definition 5.** Let  $(S, P)$  be a finite probability space and  $X : S \rightarrow \mathbb{R}$  a random variable. The expectation of  $X$ , denoted  $\mathbb{E}[X]$ , is

$$\mathbb{E}[X] := \sum_{s \in S} P(s)X(s) .$$

Let  $X, Y : S \rightarrow \mathbb{R}$  be two random variables. Then  $X + Y$  a random variable, too.

**Lemma 6** (Linearity of Expectation).  $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ .

*Proof.* Let us write down the definition of  $\mathbb{E}[X + Y]$  and perform some simple manipulations:

$$\begin{aligned}\mathbb{E}[X + Y] &= \sum_{s \in S} P(s) (X(s) + Y(s)) \\ &= \sum_{s \in S} (P(s)X(s) + P(s)Y(s)) \\ &= \sum_{s \in S} P(s)X(s) + \sum_{s \in S} P(s)Y(s) \\ &= \mathbb{E}[X] + \mathbb{E}[Y] .\end{aligned}$$

□

Note that the proof only uses the laws of associativity and commutativity. It is not required that  $X, Y$  be independent. Wait: what does it mean that “ $X$  and  $Y$  are independent”? We have defined independence only for *events*, not *random variables*. Well, let’s fix that.

**Definition 7.** Two random variables  $X, Y$  are independent if for all  $a, b \in S$ , the two events  $[X = a]$  and  $[Y = b]$  are independent. That is, if  $P(X = a \wedge Y = b) = P(X = a) \cdot P(Y = b)$ . In general,  $n$  random variables  $X_1, \dots, X_n$  are independent if for all  $a_1, \dots, a_n \in S$ , the events  $[X_1 = a_1], \dots, [X_n = a_n]$  are independent.

**Exercise 2.1.** Show that  $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$  if  $X, Y$  are independent.

## 2.2 Using Random Variables

Let us a coin. For head, let’s write a 1, for tail, a 0. Let’s do this  $n$  times, which produces a string of length  $n$ . This is our probability space:

$$S = \{0, 1\}^n .$$

Every string has the same probability  $2^{-n}$ . Let  $X$  be the number of times “11” appears in the string. For example, in 11011110 it appears 4 times,

starting at positions 1, 4, 5, 6. What is  $\mathbb{E}[X]$ ?

$s$	$X(s)$
000	0
001	0
010	0
011	1
100	0
101	0
110	1
111	2

Alright, so for  $n = 3$  we get  $\mathbb{E}[X] = \frac{1}{8}(1 + 1 + 2) = \frac{1}{2}$ . Can we get an exact formula for general  $n$ ? Think about the following: what is  $\Pr[X = j]$ ? In other words, how many bit strings of length  $n$  have exactly  $j$  occurrences of 11? Try to find a formula for this. I think it will be extremely messy. Fortunately, there is a powerful tool called *indicator variables*. Define  $X_1, \dots, X_{n-1}$  as follows:  $X_i$  is 1 if  $s_i = s_{i+1} = 1$ ; that is, if the pattern 11 occurs in  $s$  at position  $i$ . Note that  $X_i$  is a random variable. We often use the notation

$$X_i := [s_i = s_{i+1} = 1] .$$

The expression “[statement]” evaluates to 1 if “statement” is true and to 0 otherwise. This is an extremely convenient notation yet strangely not completely standard in mathematics. With this notation we observe that  $X = X_1 + X_2 + \dots + X_{n-1}$ . Indeed, every  $X_i$  indicates whether 11 occurs at position  $i$  (thus *indicator variable*), so their sum equals the number of occurrences. We compute the expectation of  $X_i$ :

$$\mathbb{E}[X_i] = 1 \cdot \Pr[X_i = 1] + 0 \cdot \Pr[X_i = 0] = \Pr[X_i = 1] = \Pr[s_i = s_{i+1} = 1] = \frac{1}{4} .$$

Note that  $X_i, X_{i+1}$  are *not* independent. Still, using linearity of expectation we compute

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \dots + \mathbb{E}[X_{n-1}] = \frac{n-1}{4} .$$

### 3 Tossing a Coin Infinitely Often

Imagine we toss a fair coin until we see three heads in a row. We kind of have an intuitive understanding of the randomness involved, but what is the

underlying probability space, formally? Well, there is no upper bound on how often we have to toss, so our probability space is

$$S := \{0, 1\}^\infty = \{(x_1 x_2 x_3 \dots), x_i \in \{0, 1\}\} ,$$

the set of all infinite bit sequences. This is a new situation: an infinite probability space, actually *uncountably infinite*. How should we define the probability function  $P : S \rightarrow \mathbb{R}$ ? For example, what is  $P(000\dots)$ , the probability we toss only zeroes? One thing should be clear: if we only talk about the first  $n$  tosses, they should just be uniformly distributed over  $\{0, 1\}^n$ . So for example it should hold that

$$\Pr[\text{the first } n \text{ tosses are } 0] = 2^{-n} .$$

So clearly  $P(000\dots) \leq 2^{-n}$ . Indeed, tossing only zeroes means in particular that the first  $n$  tosses are 0. Note that  $P(000\dots) \leq 2^{-n}$  holds for every number  $n \in \mathbb{N}$ . Thus, we can only conclude that  $P(000\dots) = 0$ . The same argument works for any other infinite sequence:  $P(x) = 0$  for every  $x \in S$ . Now this seems counter-intuitive. If we toss the coin infinitely many times, *something* must come up, but every concrete string has probability 0.

The way out of this uncomfortable situation requires a bit of formal machinery, which we do not want to introduce now. Let us stick to the following observation: events that only talk about the first  $n$  bits have a well-defined probability, since we can fall back on the discrete space  $\{0, 1\}^n$  with the uniform distribution. For example,

$$\Pr[\text{the first } n \text{ tosses do not contain the pattern } 11] = \frac{|A_n|}{2^n} = \frac{F_{n+2}}{2^n} .$$

How about the event  $\mathcal{E}$  that 11 appears earlier than 10? Certainly, this event does not only “talk about the first  $n$  bits”—there is no upper bound on how often we have to toss the coin until 11 or 10 appears. There is a nice way out: define  $\mathcal{E}_n$  to be the event that 11 appears first after  $n$  coin tosses but 10 does not appear in  $x_1 x_2 \dots x_n$ . A moment of thought shows that

$$\mathcal{E}_n = [x_1 = x_2 = \dots = x_{n-2} = 0, x_{n-1} = x_n = 1] .$$

So obviously  $\mathcal{E}_n$  talks only about the first  $n$  bits and thus  $P(\mathcal{E}_n)$  is defined; in this case, it is  $2^{-n}$ . Now  $\mathcal{E}$  happens if and only if one of  $\mathcal{E}_n$  happens:

$$\mathcal{E} = \bigcup_{n \geq 2} \mathcal{E}_n ,$$



and none of the events  $\mathcal{E}_2, \mathcal{E}_3, \dots$  can happen simultaneously. Therefore it is reasonable that

$$\Pr[\mathcal{E}] = \sum_{n \geq 2} \Pr[\mathcal{E}_n] .$$

If you continue thinking along these lines, starting with simple events—those only talking about the first  $n$  bits—and defining simple rules how to construct more complicated events, you will arrive at the formal definition of a probability space in the infinite case (the formal term is sigma-algebra, or  $\sigma$ -algebra; look it up if you want to learn more).

### 3.1 When Does the First 1 Appear?

Now that we have (informally) introduced how to talk about probabilities on infinite spaces, let us put it to use. Again, consider the space  $S = \{0, 1\}^\infty$  above. Let  $T_1$  be the time when the first 1 appears. We want to know  $\mathbb{E}[T_1]$ . In words: on average, how often do we have to toss a coin until we see heads for the first time? There are several ways to compute  $\mathbb{E}[T_1]$ , all rather instructive. Let us actually consider a more general setting in which the coin is biased and every toss comes up 1 with probability  $p$ . In this space, the event “ $x$  starts with 1001” has the probability  $p \cdot (1 - p) \cdot (1 - p) \cdot p$ . We observe that

$$\Pr[T_1 = n] = \Pr[x_1 = x_2 = \dots = x_{n-1} = 0, x_n = 1] = p \cdot (1 - p)^{n-1} ,$$

and therefore

$$\mathbb{E}[T_1] = \sum_{n \geq 1} n \cdot \Pr[T_1 = n] = \sum_{n \geq 1} np(1 - p)^{n-1} .$$

**Lemma 8.**  $\mathbb{E}[T_1] = 1/p$ .

I once had an argument with a (non-mathematician) friend who claimed that this does not need a proof because it is obvious; he thought it was the very definition of probability that this would hold. He was wrong, of course... We will give three different ways of proving this lemma.

*Proof 1, Using Calculus.* Define  $F(p) := \sum_{n \geq 0} (1 - p)^n$ . This is a geometric series, so we easily compute  $F(p) = \frac{1}{p}$ . Let us compute  $F'(p)$ . On the one hand, this is

$$F'(p) = \left( \frac{1}{p} \right)' = \frac{-1}{p^2} .$$

Of course, this would require a proof, too, but I assume we have all seen such a proof. On the other hand, differentiating the sum gives

$$F'(p) = \left( \sum_{n \geq 0} (1-p)^n \right) = - \sum_{n \geq 0} n(1-p)^{n-1} = - \sum_{n \geq 1} n(1-p)^{n-1}$$

So  $\sum_{n \geq 1} n(1-p)^{n-1} = \frac{1}{p^2}$  and

$$\sum_{n \geq 1} np(1-p)^{n-1} = p \cdot \frac{1}{p^2} = \frac{1}{p} .$$

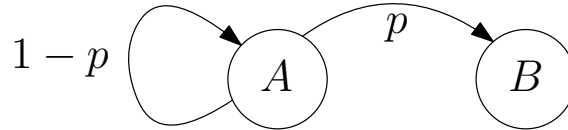
□

*Proof 2, by shifting the sum.*

$$\begin{aligned} \mathbb{E}[T_1] &= \sum_{n=1}^{\infty} np(1-p)^{n-1} = \sum_{k=0}^{\infty} (k+1)p(1-p)^k \\ &= (1-p) \sum_{k=0}^{\infty} kp(1-p)^{k-1} + \sum_{k=0}^{\infty} p(1-p)^k \\ &= (1-p) \sum_{k=1}^{\infty} kp(1-p)^{k-1} + \sum_{n=1}^{\infty} p(1-p)^{n-1} \\ &= (1-p)\mathbb{E}[T_1] + \sum_{n=1}^{\infty} \Pr[T_1 = n] \\ &= (1-p)\mathbb{E}[T_1] + 1 , \end{aligned}$$

from which it follows that  $\mathbb{E}[T_1] = 1/p$ . It remains to justify the final equality: isn't  $\sum_{n=1}^{\infty} \Pr[T_1 = n]$  obviously 1? After all,  $T_1$  has to be *some* natural number. Well, it might be that we *never* see a 1, and in that case  $T_1$  would be infinite, or undefined. Thus,  $\sum_{n=1}^{\infty} \Pr[T_1 = n]$  equals 1 minus the probability that we toss only zeroes; the latter probability is 0, as we have already seen. □

*Proof 3, by drawing a picture.* Consider the following small automaton:



We start in state  $A$ . We toss a coin; if we toss a 0 (with probability  $1 - p$ ) we stay in state  $A$ ; from there, it will take us, on expectation,  $\mathbb{E}[T_1]$  more coin tosses. If we toss a 1 (probability  $p$ ) we enter state  $B$  and stop. Thus, we can write the following equation:

$$\mathbb{E}[T_1] = 1 + (1 - p)\mathbb{E}[T_1] + p \cdot 0 .$$

The 1 appears since we have to toss at least one coin. Then, with probability  $1 - p$ , we stay at  $A$ , thus have to “start over”; with probability  $p$  we move to  $B$ , and no further tosses are required. Solving the above equation immediately yields  $\mathbb{E}[T_1] = 1/p$ .  $\square$

In general, for some finite bit string  $z \in \{0, 1\}^*$ , let  $T_z$  denote the number of tosses until  $z$  appears the first time. For example, if we toss 0010110, then  $T_{10} = 4$  and  $T_{110} = 7$ .

**Exercise 3.1.** Consider an unbiased coin, i.e., 0 and 1 come up with probability  $p = 1/2$  each. Compute  $\mathbb{E}[T_{11}]$  and  $\mathbb{E}[T_{10}]$ . You can choose any of the three proof methods above (but two of them won’t be fun).

You might have noticed that  $\mathbb{E}[T_{10}] < \mathbb{E}[T_{11}]$ , i.e., 10 appears earlier than 11, on average.

**Exercise 3.2.** Let  $\mathcal{E}$  be the event that 10 appears earlier than 11. What is  $\Pr[\mathcal{E}]$ ?

**Exercise 3.3.** Toss two coins repeatedly, producing two sequences  $x_1x_2 \dots$  and  $y_1y_2 \dots$ . We stop once we see 10 in the first sequence or 11 in the second sequence. Formally, we toss the two coins

$$T := \min(T_{10}(x), T_{11}(y))$$

times.

1. What is  $\mathbb{E}[T]$ ?
2. What are the probabilities that 10 appears in  $x$  (a) before 11 appears in  $y$ , (b) at the same time as 11, (c) later than 11?

**Hint.** I think the only way to solve this exercise without going crazy is by applying proof method 3 from above, drawing a small automaton; or in this case, not so extremely small anymore.