

# MATH103: Probability

Lecturer: Dr John Haslegrave  
Email: `j.haslegrave@lancaster.ac.uk`

2025–26



# Contents

<b>1</b>	<b>Random events</b>	<b>2</b>
1.1	Events and the sample space . . . . .	3
1.2	The Discrete Uniform Law . . . . .	5
1.3	Operations on events . . . . .	10
<b>2</b>	<b>The axiomatic approach</b>	<b>12</b>
2.1	The axioms of probability . . . . .	12
2.2	Consequences of the axioms . . . . .	14
2.3	Conditional probability . . . . .	18
2.4	Bayes' theorem . . . . .	21
2.5	Independent events . . . . .	22
2.6	Summary . . . . .	27
<b>3</b>	<b>Discrete random variables</b>	<b>28</b>
3.1	Definition . . . . .	28
3.2	Probability mass functions . . . . .	31
3.3	The probability of an event . . . . .	33
3.4	Expectation . . . . .	36
3.5	Variance . . . . .	40
3.6	Chebyshev's inequality (not examinable) . . . . .	43
<b>4</b>	<b>Models for discrete random variables</b>	<b>44</b>
4.1	Useful mathematical identities proved elsewhere . . . . .	44
4.2	Discrete uniform random variables . . . . .	45
4.3	Bernoulli random variables . . . . .	46
4.4	Binomial random variables . . . . .	48
4.5	Geometric random variables . . . . .	53
4.6	Poisson random variables . . . . .	57
4.7	Negative binomial random variables (not examinable) . . . . .	60
4.8	Summary . . . . .	61
<b>5</b>	<b>More than one random variable</b>	<b>62</b>
5.1	Joint probability mass functions . . . . .	62
5.2	Independence . . . . .	64
5.3	The weak law of large numbers . . . . .	66
<b>6</b>	<b>Continuous random variables</b>	<b>69</b>

6.1	Introduction to continuous variables . . . . .	69
6.2	The cumulative distribution function . . . . .	69
6.3	The probability density function . . . . .	71
6.4	Expectation and variance . . . . .	75
6.5	Quantiles . . . . .	77
6.6	Transformations of random variables . . . . .	79
6.7	Jointly distributed continuous random variables . . . . .	80
<b>7</b>	<b>Models for continuous random variables</b>	<b>81</b>
7.1	The uniform distribution . . . . .	81
7.2	The exponential distribution . . . . .	83
7.3	The gamma distribution . . . . .	87
7.4	The normal distribution . . . . .	89

**Additional Reading:** Two standard references at this level are:

- S. Ross, **A First Course in Probability**, 5th Edition (2003). Macmillan: New York.
- G. Grimmett and D. Welsh, **Probability: An Introduction** (1986). Oxford University Press.

However, any basic book on probability should be helpful. Do not hesitate to browse the library and look for a book whose exposition style appeals to you.

There are many helpful on-line resources, such as the Khan Academy:

<https://www.khanacademy.org/math/statistics-probability/probability-library>.

You may also find Wikipedia <http://www.wikipedia.org> helpful, although please be aware that Wikipedia content is user-contributed and so is not always reliable.

# Chapter 1

## Random events

Much of what we do is based on the belief that the future is largely unpredictable. Very few people would play games such as roulette or the lottery, or buy and sell insurance policies if the outcomes were known in advance. Probability is the study of chance, and attempts to express ideas of uncertainty quantitatively and qualitatively. In many applications the concepts are easy and the results are consistent with intuition. However, in some cases it is more difficult and following our intuition can lead to contradictions. It is for this reason that we need a formal consistent mathematical theory of probability.

The mathematical study of probability began with a correspondence between Blaise Pascal and Pierre de Fermat in 1654. At the time, gamblers in France often had their games of chance interrupted by the authorities. Fermat wrote most of his mathematics in letters to other mathematicians, and corresponded with Pascal about the following question.

**Suppose a game between two equally skilled players is interrupted. Given the scores of the players at the time of interruption, and the number of points needed to win the game, what is a fair way to divide the stakes?**

In order to answer this, we need to be able to work out the likelihood of something (a player winning) happening next, given what has happened so far.

Here is a simplified, but still surprisingly subtle, example of this sort of question, that we shall return to later in the course.

Three indistinguishable purses each contain two coins. One purse contains two gold coins, another contains two silver coins and the third contains one gold coin and one silver coin.

GG

GS

SS

A purse is selected at random and a coin is selected from that purse at random. It turns out to be gold. What is the probability that the other coin in that purse is also gold?

Here intuition might lead different people to different answers, at most one of which can be correct.

Is it  $1/3$  because only one of the three purses contained two gold coins?

Is it  $1/2$  because the first gold coin must have been selected from one of the first two purses, and the other coin is also gold in one of these?

Could it even be  $2/3$  because two of the purses have two coins of the same type?

Thus we need to build up a theory of probability in order to be sure of finding the correct value. We will start with a simple motivating example that is appropriate for some common cases, then extending this via axioms for more general cases.

**Exercise 1.1.** You choose one card at random from a standard deck. What is the probability that it is an ace?

**Solution.** There are 4 aces and 52 cards in total, so the probability is  $4/52 = 1/13$ .

What we're assuming here is called the **discrete uniform law**, which is a special case of a **probability**. We need a few definitions to set this up in general.

## 1.1 Events and the sample space

In the example above, we perform an “experiment”, by choosing a card from the deck. This experiment can have many possible outcomes. But what we were interested in, “an ace”, is not actually a possible outcome. Instead it is a type of outcome. We want to be able to talk about the “probability” of such things. For this we use the language of **sets**, which occurs throughout mathematics, and should be covered elsewhere. The facts about sets that we need for this module is summarised in the supplementary notes available on Moodle.

### The sample space

The set of all possible outcomes of an experiment is  $\Omega$  and is known as the **sample space**.

A particular outcome  $\omega \in \Omega$  is a **sample point**.

$\Omega$  and  $\omega$  are the (capital and lower-case) Greek letter “Omega”.

When the experiment takes place, exactly one of the outcomes  $\omega \in \Omega$  occurs (i.e. happens).

In this case the sample space  $\Omega$  is the set of all 52 possible cards. A sample point is a single card, e.g.  $\diamond 7$ .

Different experiments have different sample spaces.

**Example 1.2.** a. A die is thrown so

$$\Omega = \{ 1, 2, 3, 4, 5, 6 \}.$$

b. Three coins are thrown so

$$\Omega = \{ \text{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT} \}.$$

c. A die and a coin are thrown so

$$\Omega = \{ H1, H2, \dots, H6, T1, \dots, T6 \}.$$

Sometimes it is helpful to represent the sample space diagrammatically.

**Example 1.3.** A diagram of the sample space when a coin is tossed and a die is rolled:

H	•	•	•	•	•	•
T	•	•	•	•	•	•
	1	2	3	4	5	6

Each point represents a possible outcome in the sample space.

**Example 1.4.** The sample space for drawing a single card from a deck:

♠	•	•	•	•	•	•	•	•	•	•	•	•	•
♥	•	•	•	•	•	•	•	•	•	•	•	•	•
♦	•	•	•	•	•	•	•	•	•	•	•	•	•
♣	•	•	•	•	•	•	•	•	•	•	•	•	•
	A	2	3	4	5	6	7	8	9	10	J	Q	K

For more complicated experiments, the sample space can be impractical to write out in full. For example, there are over 2.5 million different 5-card poker hands. Sample spaces can even be infinite.

**Example 1.5.** A coin is thrown until a tail is observed so that

$$\Omega = \{ T, HT, HHT, HHHT, \dots \}.$$

**Example 1.6.** The lifetime of a computer component is measured so that

$$\Omega = [0, \infty)$$

## Events

An **event**  $A$  is a subset of the possible outcomes contained in the sample space  $\Omega$ , which we write as  $A \subseteq \Omega$ .

An event  $A \subseteq \Omega$  **occurs** if, when the experiment is performed, the outcome  $\omega$  satisfies  $\omega \in A$ .

In our running example, the event was  $A = \{\spadesuit A, \heartsuit A, \diamondsuit A, \clubsuit A\}$ . If the result of the experiment was the sample point  $\diamondsuit 7$ , then  $A$  did not occur.

There are some special events which exist for any sample space.

- The whole sample space  $\Omega$  is an event; when the experiment is performed, the resulting sample point must be in  $\omega \in \Omega$ , so the event  $\Omega$  is certain to occur.
- The empty set  $\emptyset = \{\}$  is an event with no sample points;  $\omega \notin \emptyset$  for any  $\omega$ , so  $\emptyset$  is an event which will never occur.
- For any sample point  $\omega$  the set  $\{\omega\}$  is an event; this event occurs if and only if the outcome of the experiment is  $\omega$ .

**Exercise 1.7.** Illustrate the sample space when two dice are thrown and indicate the event that the sum is 5.

**Solution.**

6	•	•	•	•	•	•
5	•	•	•	•	•	•
4	*	•	•	•	•	•
3	•	*	•	•	•	•
2	•	•	*	•	•	•
1	•	•	•	*	•	•
	1	2	3	4	5	6

## 1.2 The Discrete Uniform Law

The Discrete Uniform law represents our intuitive notion of probability. It applies to situations where the set  $\Omega$  is finite,  $\Omega = \{\omega_1, \dots, \omega_n\}$ , and each of the  $n$  sample points is **equally likely**. The probability of the event  $A$  is given by

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

where  $|A|$  is the number of sample points in  $A$  and  $|\Omega|$  is the number of sample points in  $\Omega$ . Note that  $\mathbb{P}(\{\omega_i\}) = 1/|\Omega|$  for all  $i = 1, \dots, n$ , so that all sample points have equal probability associated to them.

**Exercise 1.8.** Consider the experiment of rolling two fair dice. What is  $|\Omega|$ ? If  $A$  is the event that the sum is 5, what is  $\mathbb{P}(A)$ ?

**Solution.** From the diagram above,  $|\Omega| = 6 \times 6 = 36$ . Also  $|A| = 4$  so  $\mathbb{P}(A) = 4/36 = 1/9$ .

In order to apply the discrete uniform law, we need to be able to work out how many sample points are in the full sample space, and how many sample points are in events of interest, even when writing them all down and counting them is impractical. Techniques for doing this are studied in their own right as a mathematical topic called **enumerative combinatorics**, which appears in the A-level syllabus and in MATH112.

Combinatorics plays an important part in probability, statistics, physics, actuarial science, operations research, and other fields. So that both the mathematics majors and non-majors are equipped with

basic skills in this arena, we will introduce some basic counting principles.

**Exercise 1.9.** What is the size of the sample space for the following experiments?

- Rolling two dice;
- rolling a die and picking a card;
- rolling two dice and picking a card.

**Solution.** a. We have already seen this is  $6 \times 6 = 36$ .

b. Similarly we can draw a diagram with 6 rows and 52 columns, so  $6 \times 52 = 312$ .

c. Listing all the outcomes of a as the rows of our table, we get  $36 \times 52 = 1872$ .

These examples illustrate the following multiplication principle.

Suppose an experiment is conducted in  $r$  stages, such that

- stage  $i$  has  $n_i$  possible outcomes;
- any outcome is possible, irrespective of previous outcomes;
- the order of outcomes matters.

Then there are  $n_1 n_2 \cdots n_r$  possible outcomes of the experiment.

In particular, if we have the same  $n$  possible outcomes at each stage, and we choose a sequence of  $r$  outcomes **with replacement** (i.e. repetitions are allowed), there are  $n^r$  possibilities.

**Exercise 1.10.** Use the multiplication principle to show that a finite set  $S$  has  $2^{|S|}$  subsets.

**Solution.** Suppose  $|S| = r$ , and write  $S = \{s_1, s_2, \dots, s_r\}$ . We can choose our subset  $T \subseteq S$  in  $r$  stages: for each  $i = 1, 2, \dots, r$  we decide whether  $s_i$  is in or out. Since there are 2 possibilities at each stage we have  $2^r$  possible outcomes.

We can extend this principle to cases where the possible outcomes at each stage do depend on previous stages, so long as the number is determined.

Suppose an experiment is conducted in  $r$  stages, such that

- the possible outcomes at each stage do depend on previous outcomes; but
- the number of possible outcomes at stage  $i$  that are consistent with the previous outcomes is always  $n_i$ ; and
- the order of outcomes matters.

Then there are  $n_1 n_2 \cdots n_r$  possible outcomes of the experiment.

In particular, if we have the same  $n$  possible outcomes at each stage, and we choose a sequence of  $r \leq n$  outcomes **without replacement** (i.e. repetitions are **not** allowed), there are  $n(n-1) \cdots (n-r+1)$  possibilities.

A special case is when  $n = r$ . Here, since there are no repetitions, we must have exactly one of every possible outcome in some order, i.e. we are putting  $n$  objects in order. The number of ways to do this is  $n(n-1)(n-2) \cdots 3 \times 2 \times 1$ , which we denote  $n!$  (" $n$  factorial"). We define  $0! = 1$ .

When  $r < n$  then  $n(n-1) \cdots (n-r+1)$  can be expressed in terms of factorials: it is  $n!$  with the factors  $(n-r)(n-r-1) \cdots 1$  missed off, or in other words  $n!/(n-r)!$ .

**Exercise 1.11.** Alice, Bob and Charlie queue for coffee in a random order. Illustrate the sample space, and the event  $E$  that Alice is at the front of the queue. Assuming the discrete uniform law, what is the probability of this event? What if there are  $n$  people?

**Solution.** The sample space  $\Omega = \{ABC, ACB, BAC, BCA, CAB, CBA\}$  and the event is  $E = \{ABC, ACB\}$ . So  $\mathbb{P}(E) = 2/6 = 1/3$ .

If there are  $n$  people then  $|\Omega| = n!$ , and  $|E| = (n-1)!$  since we need to order the other  $n-1$  people. So  $\mathbb{P}(E) = (n-1)!/n! = 1/n$ .

Next we turn to cases where order doesn't matter. How many ways are there to choose a committee of  $r$  people from  $n \geq r$ ?

We can actually deduce this from the previous rule.

Let  $x$  be the number of ways. Note that we can choose a sequence of  $r$  people, without replacement, by first choosing a committee, then ordering it. Since there are  $r!$  ways to order each committee, there are  $x \cdot r!$  ways to choose a sequence. But we know how many sequences there are, and we get

$$x \cdot r! = n!/(n-r)!$$

The number of ways to choose a committee of  $r$  people from  $n$  is given by the **Binomial coefficient**

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

for  $r = 0, 1, \dots, n$  and  $n \in \mathbb{N}_0$ . We read  $\binom{n}{r}$  as “ $n$  choose  $r$ ”.

The binomial coefficient  $\binom{n}{r}$  also comes up in the binomial theorem as the coefficient of  $x^r y^{n-r}$  in the expansion of  $(x+y)^n$ , hence the name.

A useful observation is the following. We can equivalently choose the people who are **not** going to be on the committee, and there must therefore be the same number of ways to choose either group, i.e.

$$\binom{n}{r} = \binom{n}{n-r}.$$

We can also see this directly from the formula.

Remember that we defined  $0!$  to be 1, so  $\binom{n}{0} = \binom{n}{n} = 1$ .

**Example 1.12.** We can select 4 people from 6 in  $\binom{6}{4} = \frac{6!}{4!2!} = 15$  ways.

**Exercise 1.13.** A coin is thrown five times. Heads are shown on exactly three throws. How many different sequences of heads and tails are there that have three heads? List them.

**Solution.** Choosing three of the five coin tosses to show H can be done in  $\binom{5}{3} = 10$  different ways. They are

HHHTT, HHTHT, HTHHT, THHHT, HHTTH,  
HTHTH, THHTH, HTTHH, THTHH, TTHHH.

**Exercise 1.14.** What is the probability of throwing exactly three heads when you toss a fair coin five times?

**Solution.**  $\Omega$  consists of all the sequences of H and T of length 5.

Each coin toss can be one of two outcomes, H or T.

So there are  $2 \times 2 \times 2 \times 2 \times 2 = 2^5 = 32$  different sequences of H and T.

This is the size of the sample space  $\Omega$ .

We have already seen that the event we care about has size  $\binom{5}{3} = 10$ .

So  $\mathbb{P}(\text{three heads}) = 10/32$ .

**Exercise 1.15.** Find the number of ways of choosing two digits (from 0 to 9) and four letters from the English alphabet (26 letters). How many ways don't include a vowel? Assuming the discrete uniform law, what is the probability that a random choice doesn't include a vowel?

**Solution.** Answer: there are  $\binom{10}{2}$  ways to choose the digits and  $\binom{26}{4}$  to choose the letters, so  $\binom{10}{2} \times \binom{26}{4}$ .

Without vowels we get  $\binom{10}{2} \times \binom{21}{4}$ . So the probability of no vowels is

$$\frac{\binom{10}{2} \times \binom{21}{4}}{\binom{10}{2} \times \binom{26}{4}} = \frac{21 \times 20 \times 19 \times 18}{26 \times 25 \times 24 \times 23}.$$

**Exercise 1.16.** Assume that a valid licence plate is formed by six symbols, two of which are digits and four of which are letters from the English alphabet, all chosen **without** repetition. How many valid license plates are there? How many don't include a vowel? Assuming the discrete uniform law, what is the probability that a random licence plate doesn't include a vowel?

**Note:** implicit in this problem is the fact that the **order matters**. For example, JA93TZ is one such licence plate, 9ZJ3AT is another – even though they use the same symbols, these are different licence plates. Frequently, the main difficulty in combinatorics is not applying the three principles, but rather working out which assumptions to make about whether order matters or repetitions are allowed.

**Solution.** Answer: by the previous problem, we have  $\binom{10}{2} \times \binom{26}{4}$  ways of selecting our 6 symbols as prescribed. Since all 6 are distinct, we have  $6!$  ways of arranging them to form a licence plate. So, there are  $\binom{10}{2} \times \binom{26}{4} \times 6!$  valid license plates. Similarly there are  $\binom{10}{2} \times \binom{21}{4} \times 6!$  without vowels. So the probability of no vowels is

$$\frac{\binom{10}{2} \times \binom{21}{4} \times 6!}{\binom{10}{2} \times \binom{26}{4} \times 6!} = \frac{21 \times 20 \times 19 \times 18}{26 \times 25 \times 24 \times 23}.$$

**Exercise 1.17.** Assume that a valid post code is of the format “letter letter digit digit letter letter”, but now repetitions **are** allowed. How many valid post codes are there? How many don't include a vowel? Assuming the discrete uniform law, what is the probability that a random postcode doesn't include a vowel?

**Solution.** Answer: We have 26 ways of choosing the first character. Regardless of our choice, we have 26 ways of choosing the second, 10 ways of choosing the third, and so on. So there are  $26^4 \times 10^2$  valid post codes.

Without vowels, we instead get  $21^4 \times 10^2$ . So the probability of not having a vowel is  $\frac{21^4}{26^4}$ .

## The birthday problem

**Exercise 1.18.** Would you bet that at least two people in your workshop had exactly the same birthday? How large must a workshop be to make the probability of finding two people with the same birthday at least 0.50?

For 2 people:

Ignoring leap years, there are  $365 \times 365$  possible pairs of birthdays, of which  $365 \times 364$  are different. Assuming all combinations are equally likely, the probability two birthdays are **different** is

$$\frac{365 \times 364}{365 \times 365} = \frac{364}{365}.$$

For  $n$  people:

There are  $365^n$  possible combinations of birthdays of which  $365 \times 364 \times \cdots \times (365 - n + 1)$  are all different. So the probability all their birthdays are different is

$$\frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}.$$

If  $n = 22$ , this probability is 0.52. If  $n = 23$ , this probability is 0.49.

## 1.3 Operations on events

Two related operations on events are intersection corresponding to “and”, and union corresponding to “or”.

The **union**  $A \cup B$  of the events  $A$  and  $B$  is the set of outcomes  $\omega$  that are in  $A$  **or** in  $B$  or in both.

The **intersection**  $A \cap B$  is the set of outcomes  $\omega$  that are in  $A$  **and** in  $B$ .

The events  $A$  and  $B$  are **mutually exclusive** or **disjoint** if they have no outcomes in common; that is  $A \cap B$  is the impossible event or equivalently  $A \cap B = \emptyset$ .

**Example 1.19.** Let  $A$  be the event that a person has normal diastolic blood pressure (DBP) reading  $\{\text{DBP} < 90\}$ , and let  $B$  be the event that person has borderline DBP readings  $\{90 \leq \text{DBP} \leq 95\}$ . What are the events  $A \cup B$  and  $A \cap B$ ?

The event  $A \cup B = \{\text{DBP} \leq 95\}$ , and  $A \cap B = \emptyset$ , i.e. the impossible event.

**Exercise 1.20.** A fair die is rolled: let  $A$  denote the event that the number shown is even, and let  $B$  be the event that it is prime. What are  $A \cup B$  and  $A \cap B$ ? Assuming the discrete uniform law, what are  $\mathbb{P}(A \cup B)$  and  $\mathbb{P}(A \cap B)$ ?

**Solution.** We have  $A = \{2, 4, 6\}$  and  $B = \{2, 3, 5\}$  so  $A \cup B = \{2, 3, 4, 5, 6\}$  and  $A \cap B = \{2\}$ . Since  $|\Omega| = 6$ , assuming the discrete uniform law we have  $\mathbb{P}(A \cup B) = 5/6$  and  $\mathbb{P}(A \cap B) = 1/6$ .

The **complementary** event to  $A$  is the event  $A^c$  consisting of those outcomes that are in  $\Omega$  but are not in  $A$ .

Note that  $A \cup A^c = \Omega$  and  $A \cap A^c = \emptyset$ .

**Example 1.21.** If  $A$  is the event that the die shows an even face then  $A^c = \{1, 3, 5\}$  and is the event that the die shows an odd face.

A **partition** of the sample space splits the sample space into disjoint subsets.

**Example 1.22.** The score on a die can be partitioned according to whether it is even or odd:  $\{1, 2, 3, 4, 5, 6\} = \{2, 4, 6\} \cup \{1, 3, 5\}$ .

Write  $A_1 = \{2, 4, 6\}$  and  $A_2 = \{1, 3, 5\}$ . Then  $A_1$  and  $A_2$  are mutually exclusive and exhaustive.

We may express the sample space as  $\Omega = A_1 \cup A_2$  where  $A_1 \cap A_2 = \emptyset$ .

More generally, the  $k$  sets  $A_1, A_2, \dots, A_k$  form a **partition** of the set  $B$  if the sets  $A_1, A_2, \dots, A_k$  are **mutually exclusive** and **exhaustive**, so that

$$A_i \cap A_j = \emptyset,$$

for all  $i \neq j$  and

$$B = A_1 \cup A_2 \cup \dots \cup A_k.$$

We can also define the difference of two events.

The **difference**  $A \setminus B$  of the events  $A$  and  $B$  is the set of outcomes  $\omega$  that are in  $A$  but not in  $B$ .

Note that there is no simple relationship between  $A \setminus B$  and  $B \setminus A$ .

We can rewrite this notation using the complement:  $A \setminus B = A \cap B^c$ .

The complementary event is a special case of a difference:  $A^c = \Omega \setminus A$ .

**Exercise 1.23.** Let  $\Omega$  be a finite sample space, and let  $A$  and  $B$  be disjoint events. What are  $\mathbb{P}(A \cup B)$  and  $\mathbb{P}(A^c)$  in terms of  $\mathbb{P}(A)$  and  $\mathbb{P}(B)$ ?

**Solution.** Since  $A$  and  $B$  are disjoint, we have  $|A \cup B| = |A| + |B|$ . Thus

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

Similarly,  $|A^c| = |\Omega| - |A|$ , so

$$\mathbb{P}(A^c) = \frac{|\Omega| - |A|}{|\Omega|} = 1 - \mathbb{P}(A).$$

## Chapter 2

# The axiomatic approach

A serious limitation of the discrete uniform law is that it only applies in situations where the sample space is finite and all outcomes are equiprobable. While this might be useful for drawing cards, rolling dice, or pulling balls from urns, it offers no method for dealing with outcomes with unequal probabilities, or where the sample space may be infinite.

Even if we are in a simple situation where the discrete uniform law appears plausible, are we sure that it is a safe assumption to make? How can we allow for the possibility that the dice, coins, roulette wheel, etc., might **not** be fair?

In this chapter we will build a rigorous framework for a general mathematical theory of probability. We start by assuming that each event has some “probability”, which doesn’t have to be anything in particular, but is subject to certain intuitive conditions.

### 2.1 The axioms of probability

Let  $\Omega$  be a sample space. The **probability**  $\mathbb{P}$  is a real-valued function defined on subsets of  $\Omega$  that satisfies the following three properties.

$\mathbb{P} : \{\text{events}\} \rightarrow \mathbb{R}$   
           $\uparrow$                    $\uparrow$   
      defined on      takes values in

**Axiom 1 (positivity)**  $\mathbb{P}(A) \geq 0$  for all  $A \subseteq \Omega$ .

**Axiom 2 (normalisation)**  $\mathbb{P}(\Omega) = 1$ .

**Axiom 3 (countable additivity)** If  $A_1$  and  $A_2$  are disjoint events, then

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

More generally, if  $A_1, A_2, A_3, \dots \subset \Omega$  are events that are pairwise disjoint (that is,  $A_i \cap A_j = \emptyset$  whenever  $i \neq j$ ), then

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

By the first part of Axiom 3, arguing by induction we deduce that whenever  $A_1, \dots, A_n$  are pairwise disjoint, we have

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i) \quad (\text{for any } n \in \mathbb{N}).$$

This part of the Axiom 3 is called **finite additivity** of  $\mathbb{P}$ .

However, suppose we now have a countable collection of pairwise disjoint sets  $A_1, A_2, \dots$ . (For example, suppose we're tossing a coin and let  $A_k$  be the event that we observe the first T on the  $i$ th toss, as in Example 1.5.) Unfortunately, the inductive argument cannot give us that  $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ . It is therefore necessary to include this statement as part of our third axiom. These two statements combined constitute the **countable additivity** of  $\mathbb{P}$ .

The number  $\mathbb{P}(A)$  is called the probability of the event  $A$  and can be thought of as a measure of the likelihood that  $A$  occurs.

The whole theory of probability relies on these axioms. Subject only to these axioms the probability  $\mathbb{P}$  is otherwise unspecified. We will soon see many examples of particular probability laws  $\mathbb{P}$  that play an important role in practice.

**Example 2.1.** Suppose the sample space  $\Omega$  contains four outcomes,  $\Omega = \{1, 2, 3, 4\}$ . Assuming  $\mathbb{P}$  satisfies Axiom 3, which of the following are valid probability distributions?

- i.  $\mathbb{P}(\{1\}) = 1/2, \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = 1/6$ .  $\mathbb{P}(\{x\}) \geq 0$  for all  $x$ , so Axiom 1 is satisfied.

$\mathbb{P}(\Omega) = \mathbb{P}(\{1\} \cup \{2\} \cup \{3\} \cup \{4\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) + \mathbb{P}(\{3\}) + \mathbb{P}(\{4\}) = 1$ , so Axiom 2 is satisfied.

So  $\mathbb{P}$  is a probability.

- ii.  $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = 1/2$ .

$\mathbb{P}(\Omega) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) + \mathbb{P}(\{3\}) + \mathbb{P}(\{4\}) = 2$ , which violates axiom 2.

- iii.  $\mathbb{P}(\{1\}) = -0.2$ , and  $\mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = 0.4$ .

$\mathbb{P}(\{1\}) < 0$ , which violates axiom 1.

**Exercise 2.2.** Show that the Discrete Uniform law, defined in Section 1.2 on page 5, is a probability.

**Solution.** For the discrete uniform law, we have  $\mathbb{P}(A) = |A|/|\Omega|$  for every event  $A$ . This is a function from  $\mathcal{P}(\Omega)$  to  $\mathbb{R}$ .

**Axiom 1** is satisfied since  $|A| \geq 0$  and  $|\Omega| > 0$ .

**Axiom 2** is satisfied since  $\mathbb{P}(\Omega) = \frac{|\Omega|}{|\Omega|} = 1$ .

**Axiom 3** is satisfied since if  $A \cap B = \emptyset$  then  $|A \cup B| = |A| + |B|$ , and so

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \frac{|A| + |B|}{|\Omega|} \\
&= \frac{|A|}{|\Omega|} + \frac{|B|}{|\Omega|} \\
&= \mathbb{P}(A) + \mathbb{P}(B).
\end{aligned}$$

**Example 2.3.** A fair coin is tossed twice so

$$\Omega = \{HH, HT, TH, TT\}.$$

Since the coin is fair, we may assume all sample points are equally likely:

$$\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}).$$

Now

$$\begin{aligned}
\mathbb{P}(\{HH\}) + \mathbb{P}(\{HT\}) + \mathbb{P}(\{TH\}) + \mathbb{P}(\{TT\}) &= \mathbb{P}(\{HH, HT, TH, TT\}) \quad \text{by axiom 3} \\
&= \mathbb{P}(\Omega) \\
&= 1 \quad \text{by axiom 2.}
\end{aligned}$$

Therefore the probability of each outcome is  $1/4$ .

We can deduce other probabilities from this; for example

$$\begin{aligned}
\mathbb{P}(\text{exactly one T}) &= \mathbb{P}(\{HT, TH\}) \\
&= \mathbb{P}(\{HT\}) + \mathbb{P}(\{TH\}) \quad \text{by axiom 3} \\
&= 1/4 + 1/4 \\
&= 1/2.
\end{aligned}$$

## 2.2 Consequences of the axioms

The statements contained in this section can be derived by mathematical deduction from the axioms. They are consequences of the axioms and while the axioms themselves are accepted on faith, any statement that follows from these needs to be justified.

**Theorem 2.4 (Monotonicity).** If  $A \subset B$ , then  $\mathbb{P}(A) \leq \mathbb{P}(B)$ .

*Proof.* • Observe that  $(B \cap A^c)$  and  $A$  are disjoint and, furthermore, that  $B = (B \cap A^c) \cup A$ . (Note: it helps to draw the Venn diagram.)

- Therefore, by Axiom 3,  $\mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(A)$ .
- But, by Axiom 1,  $\mathbb{P}(B \cap A^c) \geq 0$ .
- Therefore,  $\mathbb{P}(B) \geq \mathbb{P}(A)$ .

□

**Consequence of monotonicity:** Since every event  $A$  is a subset of  $\Omega$ , we have  $\mathbb{P}(A) \leq \mathbb{P}(\Omega)$ . So, in conjunction with Axiom 1, we have that

$$0 \leq \mathbb{P}(A) \leq 1.$$

**Theorem 2.5** (The law of **complementary events**).

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

*Proof.* •  $\Omega = A \cup A^c$  [either the event  $A$  or the event  $A^c$  must happen, they are exhaustive];

- Using axiom 2 we have  $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c)$ .
- But  $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$  by axiom 3, as  $A$  and  $A^c$  are disjoint.
- So  $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$ .
- Since  $\mathbb{P}(A)$  is finite (in fact, as previously shown, it is between 0 and 1), we can subtract it from both sides. We therefore obtain that  $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ , as desired.

□

**Exercise 2.6.** A fair coin is thrown three times. What is the probability that at least one head occurs?

**Solution.** Let  $A$  be the event that at least one head occurs.

Then  $A^c$  is the event no heads occur.

$$\mathbb{P}(A^c) = \mathbb{P}(\text{no H}) = \mathbb{P}(\{\text{TTT}\}) = 1/8.$$

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - 1/8 = 7/8.$$

**Exercise 2.7.** Show that  $\mathbb{P}(\emptyset) = 0$ .

**Solution.** By the law of complementary events and axiom 2,

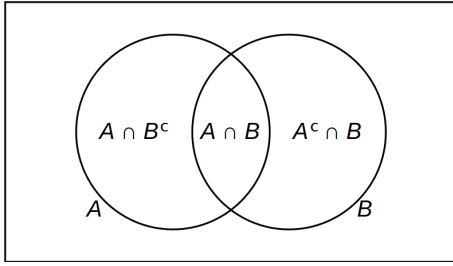
$$\begin{aligned} \mathbb{P}(\emptyset) &= 1 - \mathbb{P}(\emptyset^c) \\ &= 1 - \mathbb{P}(\Omega) \\ &= 1 - 1 = 0. \end{aligned}$$

The law of complementary events is a useful way to consider single properties. However it provides no mechanism for dealing with two events simultaneously. The first useful law for combining knowledge about more than one event is the following.

**Theorem 2.8** (The **partition law**).

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c).$$

*Proof.* Consider the Venn diagram:



We can see that  $A = (A \cap B) \cup (A \cap B^c)$ .

Note also that  $(A \cap B) \cap (A \cap B^c) = \emptyset$ , so  $A \cap B$  and  $A \cap B^c$  are exclusive.

Using axiom 3,

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c).$$

□

The ‘partition law’ can be formulated in a more general way, which we call the Total Probability Theorem. This more general version will be useful in several numerical examples later on.

**Theorem 2.9** (The Total Probability Theorem). Consider events  $B_1, B_2, \dots, B_n$  that are pairwise disjoint (that is,  $B_i \cap B_j = \emptyset$  whenever  $i \neq j$ ) and are such that  $\Omega = \cup_{i=1}^n B_i$ . Then, for any event  $A \subset \Omega$ , we have

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \cap B_i).$$

More generally, if  $\Omega = \cup_{i=1}^{\infty} B_i$  with  $B_i \cap B_j = \emptyset$  whenever  $i \neq j$ , we have

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i).$$

**Exercise 2.10.** If a randomly selected token is

- round and red with probability 0.1,
- round and blue with probability 0.2,
- square and red with probability 0.3, and
- square and blue with probability 0.4,

and there are no other possibilities, find the probability that the selected token is square, and the probability that it is blue.

**Solution.** Let  $A$  denote the event that the selected token is square and  $B$  the event that it is blue.

$$\begin{aligned} \mathbb{P}(\text{square}) &= \mathbb{P}(A) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \\ &= 0.4 + 0.3 = 0.7. \end{aligned}$$

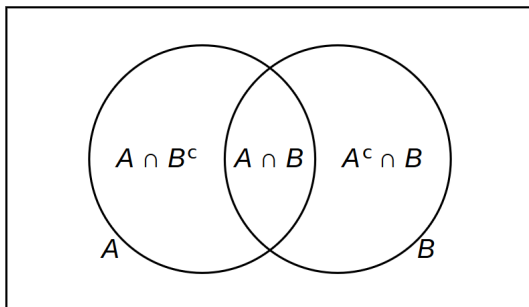
$$\begin{aligned}\mathbb{P}(\text{blue}) &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) \\ &= 0.4 + 0.2 = 0.6.\end{aligned}$$

If two events  $A$  and  $B$  are disjoint, we know from the additivity axiom that  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ . The addition law gives a general rule for any pair of events.

**Theorem 2.11** (The **addition law**, also known as the inclusion-exclusion formula).

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

*Proof.* Consider the Venn diagram:



We can see that

$$A \cup B = A \cup (B \cap A^c).$$

As  $A$  and  $B \cap A^c$  are disjoint, we may apply axiom 3:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c).$$

But by the partition law,

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c).$$

Combining these equations gives

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

□

**Exercise 2.12.** A fair die is thrown twice. Let the event  $A$  denote an even number on the first throw, and let  $B$  denote an even number on the second. Find the probability of having at least one even number.

A diagrammatic representation of the sample space makes this clearer:

second		2	4	6	1	3	5
first	5	b	b	b	*	*	*
	3	b	b	b	*	*	*
	1	b	b	b	*	*	*
	6	ab	ab	ab	a	a	a
	4	ab	ab	ab	a	a	a
	2	ab	ab	ab	a	a	a

Each point is equally probable.

**Solution.**

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(B \cap A) \\ &= 18/36 + 18/36 - 9/36 \\ &= 3/4.\end{aligned}$$

The laws of probability may be illustrated by a two-way table.

	$B$	$B^c$	total	law
$A$	$\mathbb{P}(A \cap B)$	$\mathbb{P}(A \cap B^c)$	$\mathbb{P}(A)$	partition
$A^c$	$\mathbb{P}(A^c \cap B)$	$\mathbb{P}(A^c \cap B^c)$	$\mathbb{P}(A^c)$	partition
total	$\mathbb{P}(B)$	$\mathbb{P}(B^c)$	1	complementary events

**Exercise 2.13.** Show that  $\mathbb{P}(A^c \cap B^c) = 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A \cap B)$ .

[Hint:  $A^c \cap B^c = (A \cup B)^c$ .]

**Solution.**

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

## 2.3 Conditional probability

The probability of an event may change depending not just on the experiment itself but on other information that we have. Conditional probability forms a framework in which this additional information can be incorporated.

Suppose we have two events,  $A$  and  $B$ , and we know that  $B$  has occurred. The question is, what does this tell us about whether  $A$  occurred?

We resort to extracting intuition from “empirical probability”: suppose we carry out the experiment a large number of times,  $n$ . We might expect that  $B$  occurs approximately  $n\mathbb{P}(B)$  times, and  $A$  and  $B$  occur together on approximately  $n\mathbb{P}(A \cap B)$  trials. (In fact this intuition is correct, in a way that we will explore later in the course.)

In other words, it seems reasonable to expect that  $A$  occurs on a proportion of approximately  $\mathbb{P}(A \cap B)/\mathbb{P}(B)$  of the trials in which  $B$  occurs.

This motivates the following definition.

If  $A$  and  $B$  are two events with  $\mathbb{P}(B) > 0$ , then the **conditional probability** of  $A$  given  $B$  is written as  $\mathbb{P}(A | B)$  and defined to be

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

Note the following immediate consequence of this definition:

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

**Exercise 2.14.** If a fair die is thrown and the face shows a number  $X$ .

- Suppose that  $X \neq 2$ . Find the probability that  $X$  is prime.
- Find the probability that  $X \neq 2$  given it is prime.

**Solution.** Write  $A = \{2, 3, 5\}$  and  $B = \{1, 3, 4, 5, 6\}$ . Then

$$\begin{aligned} \mathbb{P}(X \text{ prime} | X \neq 2) &= \mathbb{P}(A | B) \\ &= \mathbb{P}(A \cap B) / \mathbb{P}(B) \\ &= \mathbb{P}(\{3, 5\}) / \mathbb{P}(\{1, 3, 4, 5, 6\}) \\ &= \frac{2/6}{5/6} = 2/5. \end{aligned}$$

Similarly we have

$$\begin{aligned} \mathbb{P}(X \neq 2 | X \text{ prime}) &= \mathbb{P}(A \cap B) / \mathbb{P}(A) \\ &= \frac{2/6}{3/6} = 2/3. \end{aligned}$$

**Practice question 2.15.** A bag contains 3 blue, 5 white and 2 red marbles. A marble is selected at random; it turns out to be blue. Find the probability that the next marble selected (without replacing the first) is also blue.

**Exercise 2.16.** Three indistinguishable purses each contain two coins. One purse contains two gold coins, another contains two silver coins and the third contains one gold coin and a silver coin.

GG

GS

SS

A purse is selected at random, then at random a coin is selected from it. The selected coin turns out to be gold. Find the probability that the other coin in the purse is also gold.

**Solution.**

$$\begin{aligned} \mathbb{P}(G \text{ left} | G \text{ selected}) &= \mathbb{P}(G \text{ selected} \cap G \text{ left}) / \mathbb{P}(G \text{ selected}) \\ &= \mathbb{P}(GG) / \mathbb{P}(G \text{ selected}) \\ &= \frac{1/3}{1/2} = 2/3. \end{aligned}$$

**Exercise 2.17.** Does  $\mathbb{P}(A | B)$  satisfy  $0 \leq \mathbb{P}(A | B) \leq 1$ ?

**Solution.** Yes. First:  $A \cap B \subseteq B$  hence  $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$ . Then, since  $\mathbb{P}(B) > 0$ ,

$$\mathbb{P}(A | B) = \mathbb{P}(A \cap B) / \mathbb{P}(B) \leq \mathbb{P}(B) / \mathbb{P}(B) = 1.$$

Since  $\mathbb{P}(A \cap B) \geq 0$  by Axiom 1 and  $\mathbb{P}(B) > 0$  for the conditional probability to be defined, we have  $\mathbb{P}(A | B) \geq 0$ .

The partition law can be rephrased as the law of total probability, which is an extremely useful way to break down considerations about real life events.

**Theorem 2.18** (the **law of total probability**).

$$\mathbb{P}(A) = \mathbb{P}(A | B)\mathbb{P}(B) + \mathbb{P}(A | B^c)\mathbb{P}(B^c).$$

*Proof.* By the partition law, then the definition of conditional probability,

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \\ &= \mathbb{P}(A | B)\mathbb{P}(B) + \mathbb{P}(A | B^c)\mathbb{P}(B^c). \end{aligned}$$

□

**Example 2.19.** In a population of children, 60% are vaccinated against whooping cough. The probabilities of contracting whooping cough are 1/1000 if the child is vaccinated and 1/100 if not. Find the probability that a child selected at random will contract whooping cough.

Let  $W$  denote whooping cough and  $V$  denote vaccination. Then

$$\begin{aligned} \mathbb{P}(W) &= \mathbb{P}(W | V)\mathbb{P}(V) + \mathbb{P}(W | V^c)\mathbb{P}(V^c) \\ &= \frac{1}{1000}0.6 + \frac{1}{100}0.4 = \frac{4.6}{1000} = 0.0046. \end{aligned}$$

**Exercise 2.20.** Note that  $\{B, B^c\}$  is a partition of the sample space  $\Omega$ . Conjecture and prove a generalisation of the law of total probability to find  $\mathbb{P}(A)$  from  $\mathbb{P}(A | B_1)$ ,  $\mathbb{P}(A | B_2)$  and  $\mathbb{P}(A | B_3)$  when  $\{B_1, B_2, B_3\}$  are a partition of the sample space.

**Solution.**

$$\begin{aligned} A &= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \\ \mathbb{P}(A) &= \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \mathbb{P}(A \cap B_3) \\ &= \mathbb{P}(A | B_1)\mathbb{P}(B_1) + \mathbb{P}(A | B_2)\mathbb{P}(B_2) + \mathbb{P}(A | B_3)\mathbb{P}(B_3). \end{aligned}$$

**Exercise 2.21.** A test for a disease gives positive results 90% of the time when a disease is present, and 10% of the time when the disease is absent. It is known that 1% of the population have the disease. Of those that receive a positive test result, 80% of patients receive treatment. For a randomly selected member of the population, what is the probability of receiving treatment?

**Solution.** Let  $C$  be the event “has disease”:  $\mathbb{P}(C) = 0.01$ , and  $\mathbb{P}(C^c) = 0.99$ .

Let  $B$  be the event “positive test result”. We are told  $\mathbb{P}(B | C) = 0.9$  and  $\mathbb{P}(B | C^c) = 0.1$ .

Let  $A$  be the event “receives treatment”. We know  $\mathbb{P}(A | B) = 0.8$ . Presumably also  $\mathbb{P}(A | B^c) = 0$ .

We can now calculate

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A | B)\mathbb{P}(B) + \mathbb{P}(A | B^c)\mathbb{P}(B^c) \\ &= 0.8\mathbb{P}(B) + 0\mathbb{P}(B^c) \\ &= 0.8\mathbb{P}(B).\end{aligned}$$

Note that there's no need to find  $\mathbb{P}(B^c)$ . On the other hand we need

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B | C)\mathbb{P}(C) + \mathbb{P}(B | C^c)\mathbb{P}(C^c) \\ &= 0.9 \times 0.01 + 0.1 \times 0.99 = 0.108.\end{aligned}$$

It therefore follows that

$$\mathbb{P}(A) = 0.8 \times 0.108 = 0.0864.$$

## 2.4 Bayes' theorem

Thomas Bayes (1701–1761), a clergyman and amateur statistician, was ignored by his contemporaries but has had a profound effect on modern statistical thinking.

Often we care about the probability of  $A$  given  $B$  but information is given about the probability of  $B$  given  $A$ . Bayes' theorem provides the basis for transforming this information.

**Theorem 2.22** (Bayes' theorem). If  $A$  and  $B$  are events in the sample space with  $\mathbb{P}(A), \mathbb{P}(B) > 0$  then

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

*Proof.*

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

and

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

Therefore

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

The result follows by rearranging. □

Another way to express this theorem, using the law of total probability, is

$$\mathbb{P}(B | A) = \frac{\mathbb{P}(A | B)\mathbb{P}(B)}{\mathbb{P}(A | B)\mathbb{P}(B) + \mathbb{P}(A | B^c)\mathbb{P}(B^c)}.$$

To evaluate the right hand side we need to know the probabilities  $\mathbb{P}(A | B)$ ,  $\mathbb{P}(B)$ ,  $\mathbb{P}(A | B^c)$  and  $\mathbb{P}(B^c)$ .

When more than two possibilities are present, as when  $\{B_1, B_2, \dots, B_k\}$  form a partition of the sample space  $\Omega$ , Bayes' formula extends to

$$\begin{aligned}\mathbb{P}(B_i | A) &= \frac{\mathbb{P}(A | B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A | B_i)\mathbb{P}(B_i)}{\sum_{j=1}^k \mathbb{P}(A | B_j)\mathbb{P}(B_j)}.\end{aligned}$$

**Practice question 2.23.** For the whooping cough exercise above (Example 2.19 on p20), find the probability that a child is vaccinated given the occurrence of whooping cough.

**Exercise 2.24.** Return to the disease example (Example 2.21 on p21). If you receive a positive test result, what is the probability that you have the disease?

**Solution.** Recall that  $C$  is the event “have disease”, with  $\mathbb{P}(C) = 0.01$  and  $\mathbb{P}(C^c) = 0.99$ . Also that  $B$  is the event “positive test result”, with  $\mathbb{P}(B | C) = 0.9$  and  $\mathbb{P}(B | C^c) = 0.1$ . Using Bayes' theorem, we see that

$$\begin{aligned}\mathbb{P}(C | B) &= \frac{\mathbb{P}(B | C)\mathbb{P}(C)}{\mathbb{P}(B | C)\mathbb{P}(C) + \mathbb{P}(B | C^c)\mathbb{P}(C^c)} \\ &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} \\ &= 0.083.\end{aligned}$$

## 2.5 Independent events

In the previous section we saw that  $\mathbb{P}(A | B)$ , the conditional probability of  $A$  given  $B$ , where  $\mathbb{P}(B) > 0$ , was in general not equal to  $\mathbb{P}(A)$ , the unconditional probability of  $A$ . In the special case when

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

we say that  $A$  is **independent** of  $B$ . Independence means that knowing that the event  $B$  has occurred does not change the chance that  $A$  will occur.

Note that, using the definitions of conditional probability,  $\mathbb{P}(A \cap B) = \mathbb{P}(A | B)\mathbb{P}(B)$  so if  $A$  is independent of  $B$ , then  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ . We therefore get the following definition of independence.

$A$  and  $B$  are **independent** events if and only if

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

**Exercise 2.25.** If two coins are thrown and the four possible outcomes are equally likely, show that the events “head on first coin” and “head on second coin” are independent.

**Solution.**

$$\begin{aligned}\Omega &= \{HH, HT, TH, TT\} \\ \mathbb{P}(\text{H on first}) &= \mathbb{P}(\{HH, HT\}) = \frac{2}{4} = \frac{1}{2} \\ \mathbb{P}(\text{H on second}) &= \mathbb{P}(\{HH, TH\}) = \frac{2}{4} = \frac{1}{2} \\ \mathbb{P}(\text{H on both}) &= \mathbb{P}(\{HH\}) = \frac{1}{4}.\end{aligned}$$

We observe

$$\mathbb{P}(\text{H on both}) = \mathbb{P}(\text{H on first})\mathbb{P}(\text{H on second})$$

so they are independent.

Similar calculations show the independence of any pair of events with one referring to the first coin and the other to the second coin.

**Practice question 2.26.** Suppose that the coin is biased and that the probability of a head occurring on any throw is  $\theta$ , which can be any number between 0 and 1. Use independence to determine the probabilities of the four outcomes when throwing the coin twice.

**Example 2.27.** We roll a fair 4-sided die twice. The sample space is  $\Omega = \{(i, j) \mid i = 1, 2, 3, 4, j = 1, 2, 3, 4\}$ , with  $\mathbb{P}(\{(i, j)\}) = 1/16$  for each  $(i, j) \in \Omega$ .

Let  $A_1$  be the event that the 1st throw lands a 3 and  $A_2$  the event that the 2nd throw lands a 1. Clearly,

$$A_1 = \{(3, 1), (3, 2), (3, 3), (3, 4)\}, \quad A_2 = \{(1, 1), (2, 1), (3, 1), (4, 1)\}, \quad A_1 \cap A_2 = \{(3, 1)\}.$$

Hence

$$\mathbb{P}(A_1 \cap A_2) = \frac{1}{16} = \frac{1}{4} \times \frac{1}{4} = \mathbb{P}(A_1)\mathbb{P}(A_2).$$

Therefore,  $A_1$  and  $A_2$  are independent.

Now let  $B_1$  be the event that the first throw is 1 and  $B_2$  the event that the sum of the two throws is 2. That is,

$$B_1 = \{(1, 1), (1, 2), (1, 3), (1, 4)\}, \quad B_2 = \{(1, 1)\}, \quad B_1 \cap B_2 = \{(1, 1)\}.$$

Hence,

$$\mathbb{P}(B_1 \cap B_2) = \frac{1}{16} \neq \mathbb{P}(B_1)\mathbb{P}(B_2) = \frac{1}{4} \times \frac{1}{16}.$$

Therefore,  $B_1$  and  $B_2$  are not independent.

Finally, let  $B_1$  be again the event that the 1st throw lands a 1 and  $C_2$  the event that the sum of the two throws is 5. That is,

$$B_1 = \{(1, 1), (1, 2), (1, 3), (1, 4)\}, \quad C_2 = \{(1, 4), (4, 1), (2, 3), (3, 2)\}, \quad B_1 \cap C_2 = \{(1, 4)\}.$$

Hence,

$$\mathbb{P}(B_1 \cap C_2) = \frac{1}{16} = \frac{1}{4} \times \frac{1}{4} = \mathbb{P}(B_1)\mathbb{P}(C_2).$$

Therefore, the two events are independent! (So, even though knowing the outcome of the first throw gives us information about the sum, it does not change the probability of observing a sum of 5.)

**Exercise 2.28.** Suppose that the probability of mothers being hypertensive (having high blood pressure) is 0.1 and that for fathers is 0.2. Find the probability of a child's parents both being hypertensive, assuming both events are independent.

**Solution.**

$$\begin{aligned} \mathbb{P}(\text{both H}) &= \mathbb{P}(\text{mother H})\mathbb{P}(\text{father H}) \\ &= 0.1 \times 0.2 = 0.02. \end{aligned}$$

We might expect these two events to be independent if hypertension was caused by genetic factors, but not independent if hypertension was caused by environmental factors.

**Exercise 2.29.** If  $\mathbb{P}(A) = 0.2$  and  $\mathbb{P}(B) = 0.3$  find  $\mathbb{P}(A \cap B)$  if

- i.  $A$  and  $B$  are independent,
- ii.  $A$  and  $B$  are exclusive.

**Solution.** i.  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = 0.2 \times 0.3 = 0.06$ .

ii.  $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$ .

**Exercise 2.30.** Consider a mother and child's blood pressures. Let  $A = \{\text{mother's DBP} \geq 95\}$  and  $B = \{\text{child's DBP} \geq 80\}$ . Suppose we know that  $\mathbb{P}(A) = 0.1$ ,  $\mathbb{P}(B) = 0.2$  and  $\mathbb{P}(A \cap B) = 0.05$ . Are  $A$  and  $B$  independent?

**Solution.**

$$A \text{ indep } B \iff \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

But  $0.05 \neq 0.1 \times 0.2$  so they are not independent.

**Exercise 2.31.** Suppose that eye colours are only brown or green. Suppose also that there is a simple genetic coding for this: if both your eye colour alleles are G then you have green eyes, otherwise you have brown eyes. Suppose each allele is G with probability 0.1 (and you may assume independence). What is the probability that a random population member has green eyes? What is the probability that two randomly selected (unrelated) people both have green eyes?

**Solution.** For an individual to have green eyes, both alleles must be G. By independence,

$$\mathbb{P}(GG) = \frac{1}{10} \times \frac{1}{10} = \frac{1}{100}.$$

For two independent individuals, again by independence, we calculate

$$\mathbb{P}(\text{both green}) = \mathbb{P}(GG)\mathbb{P}(GG) = \frac{1}{100} \times \frac{1}{100} = \frac{1}{10000}.$$

Genes are passed from parent to child – each parent passes one of their two alleles, selected uniformly at random, to the child. If both parents have green eyes, what is the eye colour of the child?

**Solution.** Both parents are GG, so both pass a G to the child, so the child is GG and has green eyes.

If we don't know the parents' eye colour, but do know the eye colour of a sibling, this can give us useful information. What can we say about the probability of a brother and sister both having green eyes?

**Solution.** Suppose the brother has green eyes. Then each parent must have at least one G allele.

Therefore each parent passes a G allele to the sister with probability at least 1/2. Hence

$$\mathbb{P}(\text{sister GG} \mid \text{brother GG}) \geq \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

Now

$$\begin{aligned}\mathbb{P}(\text{both GG}) &= \mathbb{P}(\text{brother GG})\mathbb{P}(\text{sister GG} \mid \text{brother GG}) \\ &\geq \frac{1}{100} \times \frac{1}{4} \\ &= \frac{1}{400}.\end{aligned}$$

Note that if the population prevalence of G is much smaller than 1/10, then  $\mathbb{P}(GG)$  decreases significantly, but the bound on  $\mathbb{P}(\text{sister GG} \mid \text{brother GG})$  is unchanged.

Misunderstanding this kind of dependence can have tragic consequences. There have been several cases in which mothers were wrongfully convicted of murdering their children based on statistically illiterate testimony from doctors about the unlikelihood of multiple cases of sudden infant death (SIDS) in the same family. These doctors made the error of treating two siblings suffering SIDS as independent events. Essentially this is like confusing the independent case and the sibling case above, except that the figures for SIDS roughly correspond to taking  $\mathbb{P}(G) \approx \frac{1}{92}$ .

Calculate the probability for two independent people, and approximate the probability for two siblings, to have green eyes based on this value. The extremely low value we get in the independent case was used as evidence to convict (which can itself lead to a different error known as the Prosecutor's Fallacy), whereas the real probability will be closer to the value we get for siblings.

**Example 2.32.** If  $A \subset B$ , can  $A$  and  $B$  be independent?

Answer: observe that  $\mathbb{P}(A \cap B) = \mathbb{P}(A) \neq \mathbb{P}(A)\mathbb{P}(B)$  unless  $\mathbb{P}(A) = 0$  or  $\mathbb{P}(B) = 1$ .

**Example 2.33.** If  $A \cap B = \emptyset$ , can  $A$  and  $B$  be independent?

Answer: observe that  $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)$  unless  $\mathbb{P}(A) = 0$  or  $\mathbb{P}(B) = 0$ .

### Independence for multiple events

We have seen several examples concerning the independence (or otherwise) of two events. Often we have more than two events; what does it mean for multiple events to be independent?

Three events,  $A$ ,  $B$  and  $C$ , are **independent** if and only if all the following are satisfied:

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B), & \mathbb{P}(B \cap C) &= \mathbb{P}(B)\mathbb{P}(C), & \mathbb{P}(A \cap C) &= \mathbb{P}(A)\mathbb{P}(C), \\ \mathbb{P}(A \cap B \cap C) &= \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)\end{aligned}$$

It is important to know that the last statement does not follow from the others. That is, even if  $A$  and  $B$  are independent,  $B$  and  $C$  are independent, and  $A$  and  $C$  are independent, it is still possible that  $A$ ,  $B$  and  $C$  are **not** independent.

**Exercise 2.34.** I roll two standard fair dice, one red and one blue. Let  $A$  be the event that the red die shows an odd number,  $B$  the event that the blue die shows an odd number, and  $C$  be the event that the total is odd.

- a. Are  $A$  and  $C$  independent? What about  $B$  and  $C$ ?
- b. Are  $A$ ,  $B$  and  $C$  independent?

**Solution.** a. We have  $\mathbb{P}(A) = 3/6 = 1/2$ ,  $\mathbb{P}(C) = 18/36 = 1/2$  and  $\mathbb{P}(A \cap C) = 9/36 = 1/4$ , so  $A$  and  $C$  are independent. Similarly  $B$  and  $C$  are independent.

- b. No:  $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$  but  $\mathbb{P}(A \cap B \cap C) = 0$ , since if both dice show odd numbers, the total must be even.

## 2.6 Summary

We conclude with a summary of how the various notions introduced in this chapter are related.

Sample space: $\Omega$	a set.
Events (subsets of $\Omega$ ):	subsets of the sample space.
Probability law $\mathbb{P}(\cdot)$ :	a measure of “chance” (likelihood) that an event occurs.
Conditional probability $\mathbb{P}(\cdot \mid A)$ :	as above, but takes into account the fact that we already know that $A$ has occurred.
Independence of $A$ and $B$ :	concerns events $A$ and $B$ , but depends on the choice of $\mathbb{P}$ .

Since whether  $A$  and  $B$  are independent depends not only on the choice of  $A$  and  $B$  but also on  $\mathbb{P}$ , independent events may become dependent when conditioned on another event, or vice versa.

For example, let  $A$  be the event that person 1 has a rare genetic disease and  $B$  the event that person 2 has a rare genetic disease. When these two people are selected at random from the general population, we expect  $A$  and  $B$  to be independent. However, as we have discovered, things can change significantly if we discover that the two randomly selected people are in fact siblings.

**Practice question 2.35.** In a standard deck of cards, the one-eyed cards are the King of Diamonds, the Jack of Spades and the Jack of Hearts. A card is selected at random. Let  $E$  be the event that it is a one-eyed card,  $R$  be the event that it is a red card, and  $J$  be the event that it is a jack. Work out  $\mathbb{P}(E)$ ,  $\mathbb{P}(R)$  and  $\mathbb{P}(E \cap R)$ . Are  $E$  and  $R$  independent? Now work out the conditional probabilities  $\mathbb{P}(E \mid J)$ ,  $\mathbb{P}(R \mid J)$  and  $\mathbb{P}(E \cap R \mid J)$ . Are  $E$  and  $R$  independent given  $J$ ?

## Chapter 3

# Discrete random variables

We are not always interested in an experiment itself, but rather in some consequence of its random outcome. For example, in a football match, we may be interested the total number of goals that team  $A$  or team  $B$  scored, but not really concerned by how the game played out. Random variables give us a way to think about these consequences in those situations when they take real values.

### 3.1 Definition

A **random variable**  $X$  is a **function**  $X: \Omega \rightarrow \mathbb{R}$ .  $X$  associates each outcome  $\omega$  in the sample space  $\Omega$  with a unique real number  $X(\omega)$ .

**Example 3.1.** Suppose  $\Omega = \{(i, j) : i, j \in \{1, \dots, 6\}\}$  is the sample space resulting from rolling two dice. We can define natural random variables by

- $X((i, j)) = i + j$ , the sum of the values on the dice,
- $Y((i, j)) = \max\{i, j\}$ , the bigger of the two values on the dice.

Every time the experiment is conducted exactly one value of the random variable is observed; this is called a **realisation** of the random variable.

The range of values taken by the random variable  $X$  defined on  $\Omega$ , that is  $\{X(\omega) : \omega \in \Omega\}$ , is known as the **induced sample space** for  $X$  and is sometimes written as  $\mathcal{S}$ .

In this chapter we shall focus on **discrete random variables** – that is functions where  $\mathcal{S}$  is finite or countable e.g.  $\mathcal{S} = \mathbb{Z}$ . We will move on to **continuous random variables** – functions where  $\mathcal{S}$  is uncountable e.g.  $\mathcal{S} = \mathbb{R}$  – in Chapter 6.

Random variables are important as:

- they are the result of most real experiments
- additional structure imposed by the number system, such as ordering, enables further development of the ideas in the previous chapters.

### Introductory examples of random variables

Discrete random variables arise in a variety of ways: From experiments

- with a natural integer valued outcome
  - the number of buses to stop in the hour,
  - the number of goals in a football match.
- with a continuous outcome which is recorded on an integer scale
  - heights, ages, salaries
- with non-integer outcomes to which numerical values are assigned
  - a coin is tossed with outcome H or T, converted to 1 and 0 respectively.
  - disease stage coded on a numerical scale (e.g. 1, 2, ..., 5).

**Exercise 3.2.** A coin is tossed three times. The sample space is

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

Define a random variable giving the number of heads thrown. What is the induced sample space?

**Solution.** Define  $X(\omega) = \#H$  in  $\omega$ .

Explicitly,

$$\begin{aligned} X(TTT) &= 0 \\ X(HTT) &= X(THT) = X(TTH) = 1 \\ X(HHT) &= X(HTH) = X(THH) = 2 \\ X(HHH) &= 3. \end{aligned}$$

The induced sample space for  $X$  is  $\mathcal{S} = \{0, 1, 2, 3\}$ .

**Example 3.3.** Suppose we decide to record the number of children born in the local maternity ward tomorrow. Any outcome is a non-negative integer, so a suitable sample space is  $\Omega = \{0, 1, 2, \dots\}$ . The random variable is  $X(\omega) = \omega$ , giving the number of children born.

Suppose that  $A \subseteq \mathcal{S}$  is a subset of the induced sample space. Then we can define the event

$$\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\} \subseteq \Omega.$$

In the special case  $A = \{a\}$  we can write  $\{X = a\}$  for the event  $\{X \in A\}$ .

In the three-coins example above, we have that

$$\{X = 2\} = \{HHT, HTH, THH\}.$$

The right hand side of these equations is an event in  $\Omega$  and hence has a probability assigned to it. This induces a probability on the induced sample space, given by

$$\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$

**Exercise 3.4.** Suppose our sample space consists of the outcomes of throwing a fair die, and suppose we gamble on the outcome:

- lose £1 if the outcome is 1, 2 or 3;
- win nothing if the outcome is 4;
- win £2 if the outcome is 5 or 6.

Define  $X$  to be the random variable giving our profit. Find the induced sample space for  $X$ , and evaluate the probabilities on the induced sample space.

**Solution.**  $\Omega = \{1, 2, 3, 4, 5, 6\}$  and

$$\begin{aligned} X(1) &= X(2) = X(3) = -1, \\ X(4) &= 0, \\ X(5) &= X(6) = 2. \end{aligned}$$

The induced sample space for  $X$  is  $\mathcal{S} = \{-1, 0, 2\}$ . Now

$$\begin{aligned} \mathbb{P}(X = -1) &= \mathbb{P}(\{1, 2, 3\}) = \frac{1}{2}, \\ \mathbb{P}(X = 0) &= \mathbb{P}(\{4\}) = \frac{1}{6}, \\ \mathbb{P}(X = 2) &= \mathbb{P}(\{5, 6\}) = \frac{1}{3}. \end{aligned}$$

The probability associated to the other subsets can be obtained using axiom 3 e.g.

$$\mathbb{P}(X \in \{-1, 0\}) = \mathbb{P}(X = -1) + \mathbb{P}(X = 0) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}.$$

To summarise:

- The outcomes in the **sample space**,  $\Omega$ , of the probability experiment may or may not be numerically valued.
- A **random variable**  $X$  is a function that associates a unique real number with each outcome in the sample space,  $\Omega$ .
- A random variable is **not** a number. It is neither random, nor a variable. It is a **function**.
- The set of values taken by random variable  $X$  defined on  $\Omega$ , is known as the **induced sample space** for  $X$  and is sometimes written as  $\mathcal{S}$ .
- The event  $\{X = x\} = \{\omega : X(\omega) = x\}$  and this correspondence induces a probability distribution on  $\mathcal{S}$ .

## 3.2 Probability mass functions

We cannot predict the value of a discrete random variable  $X$  exactly, but we can state the values it could take and attach probabilities to these values.

The **probability mass function** (p.m.f.)  $p_X$  of a discrete random variable  $X$  is defined by

$$p_X(x) = \mathbb{P}(X = x)$$

for all  $x \in \mathcal{S}$ .

**Lemma 3.5.** If  $p_X$  is a probability mass function then it satisfies the conditions

$$p_X(x) \geq 0 \quad \forall x \quad \text{and} \quad \sum_{x \in \mathcal{S}} p_X(x) = 1.$$

*Proof.* Note that as  $p_X(x) = \mathbb{P}(X = x)$  is a probability, by axiom 1,

$$p_X(x) \geq 0 \quad \forall x.$$

Write  $\mathcal{S} = \{x_1, x_2, \dots\}$ . (An analogous argument will hold for the finite case  $\mathcal{S} = \{x_1, x_2, \dots, x_n\}$ .) As  $\Omega = \{X \in \mathcal{S}\}$ , we have

$$\begin{aligned} 1 &= \mathbb{P}(\Omega) && \text{by Axiom 2} \\ &= \mathbb{P}((X = x_1) \cup (X = x_2) \cup (X = x_3) \cup \dots) \\ &= \mathbb{P}(X = x_1) + \mathbb{P}(X = x_2) + \mathbb{P}(X = x_3) + \dots && \text{by Axiom 3} \\ &= p_X(x_1) + p_X(x_2) + p_X(x_3) + \dots \end{aligned}$$

□

In most cases we consider, we will have  $\mathcal{S} \subseteq \mathbb{N}_0$ , in which case we can write the sum above as  $\sum_{x=0}^{\infty} p_X(x) = 1$ , with the convention that if  $x \notin \mathcal{S}$  then  $p_X(x) = 0$ .

In theory, any function  $p : \mathbb{N}_0 \rightarrow \mathbb{R}$  satisfying  $p(x) \geq 0$  for all  $x$  and  $\sum_{x=0}^{\infty} p(x) = 1$  is a p.m.f. of some random variable. The next few exercises give some natural examples. In the next chapter we extend these to ones which correspond to random variables of interest.

**Example 3.6.** A random variable  $X$  which has the same outcome  $k$  every time the experiment is undertaken is a constant. The p.m.f. for this random variable is

$$p_X(x) = \begin{cases} 1 & \text{if } x = k; \\ 0 & \text{if } x \neq k. \end{cases}$$

This clearly satisfies the two conditions.

**Exercise 3.7.** Find the p.m.f. of the number of heads in three tosses of a fair coin. The sample space is  $\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$

**Solution.** The event equivalence gives

$$\begin{aligned}\{X = 0\} &= \{TTT\}; \\ \{X = 1\} &= \{HTT, THT, TTH\}; \\ \{X = 2\} &= \{HHT, HTH, THH\}; \\ \{X = 3\} &= \{HHH\}.\end{aligned}$$

So, by the Discrete Uniform Law,

$$\begin{aligned}p_X(0) &= \mathbb{P}(X = 0) = 1/8 \\ p_X(1) &= \mathbb{P}(X = 1) = 3/8 \\ p_X(2) &= \mathbb{P}(X = 2) = 3/8 \\ p_X(3) &= \mathbb{P}(X = 3) = 1/8, \\ p_X(x) &= 0 \quad \text{for } x \notin \{0, 1, 2, 3\}.\end{aligned}$$

Note:  $p_X(x) \geq 0$  for all  $x$  and  $\sum_{x=0}^{\infty} p_X(x) = p_X(0) + p_X(1) + p_X(2) + p_X(3) = 1$ .

**Exercise 3.8.** Suppose  $\Omega = \{a, b, c, d\}$  and each outcome occurs with equal probability. The random variable  $X$  is defined by  $X(a) = 2, X(b) = 4, X(c) = 3, X(d) = 2$ . Write down the induced sample space and p.m.f. of  $X$ .

**Solution.** The induced sample space is  $\mathcal{S} = \{2, 3, 4\}$ . By the Discrete Uniform Law,

$$\begin{aligned}p_X(2) &= \mathbb{P}(\{a, d\}) = \frac{1}{2}, \\ p_X(3) &= \mathbb{P}(\{c\}) = \frac{1}{4}, \\ p_X(4) &= \mathbb{P}(\{b\}) = \frac{1}{4}.\end{aligned}$$

**Exercise 3.9.** If a p.m.f. is specified by  $p_X(x) = c$  for  $x = 0, 1, \dots, m$  and  $p_X(x) = 0$  otherwise, where  $c$  is constant, then determine the value of  $c$ .

**Solution.**

$$1 = \sum_{x=0}^m p_X(x) = (m+1)c.$$

Therefore  $c = 1/(m+1)$ . Note that the condition  $p_X(x) \geq 0$  is also satisfied.

**Exercise 3.10.** If a p.m.f. is specified by  $p_X(x) = cx$  for  $x = 1, 2, 3, 4$  and  $p_X(x) = 0$  otherwise, where  $c$  is constant, then determine the value of  $c$ .

**Solution.**

$$\begin{aligned} 1 &= p_X(1) + p_X(2) + p_X(3) + p_X(4) \\ &= c + 2c + 3c + 4c = 10c. \end{aligned}$$

Therefore  $c = 1/10$ .

### 3.3 The probability of an event

If we are interested in evaluating the probability of some event occurring for a random variable, this can easily be obtained from the p.m.f.

**Lemma 3.11.** Let  $E \subseteq \mathcal{S}$  be an event in the induced sample space. The probability of  $E$  is given by

$$\mathbb{P}(X \in E) = \sum_{x \in E} p_X(x).$$

*Proof.* Write  $E = \{r_1, \dots, r_k\} \subseteq \mathcal{S}$ . Then

$$\begin{aligned} \mathbb{P}(X \in E) &= \mathbb{P}(\{X = x_1\} \cup \{X = x_2\} \cup \dots \cup \{X = x_k\}) \\ &= \mathbb{P}(X = x_1) + \mathbb{P}(X = x_2) + \dots + \mathbb{P}(X = x_k) \\ &= p_X(x_1) + p_X(x_2) + \dots + p_X(x_k) \\ &= \sum_{x \in E} p_X(x). \end{aligned}$$

An analogous argument holds in the case of a countably infinite set  $E = \{x_1, x_2, \dots\} \subseteq \mathcal{S}$ . □

**Exercise 3.12.** The length of stay in hospital after surgery is modelled as a random variable  $X$ . The following table gives the p.m.f. for  $X$ .

Days stayed	$x$	4	5	6	7	8	9	10	total
Probability	$p_X(x)$	0.038	0.114	0.430	0.300	0.080	0.030	0.008	1

Find the probability of being in hospital for:

- at most 6 days;
- between 5 and 7 days inclusive;
- at least 7 days.

**Solution.** a. at most 6 days:  $\mathbb{P}(X \leq 6) = \mathbb{P}(X = 4) + \mathbb{P}(X = 5) + \mathbb{P}(X = 6) = 0.582$

b. between 5 and 7:  $\mathbb{P}(5 \leq X \leq 7) = \mathbb{P}(X = 5) + \mathbb{P}(X = 6) + \mathbb{P}(X = 7) = 0.844$

c. at least 7:  $\mathbb{P}(X \geq 7) = 1 - \mathbb{P}(X \leq 6) = 1 - 0.582 = 0.418$ .

**Exercise 3.13.** Find the probability of an odd number of heads in 3 tosses of a fair coin.

**Solution.** Let  $X$  be the number of heads. The p.m.f. of  $X$  is

$$p_X(x) = \begin{cases} 1/8 & \text{for } x = 0, 3 \\ 3/8 & \text{for } x = 1, 2 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore  $\mathbb{P}(\text{odd number Hs}) = \mathbb{P}(X \in \{1, 3\}) = \sum_{x=1,3} p_X(x) = 3/8 + 1/8 = 1/2$ .

A specific family of events in which we are often interested, particularly for continuous random variables, is  $\{x : x \leq m\}$  for different values of  $m$ ; we can equivalently write this event as  $X \leq m$ . These events are useful because for any event  $E \subseteq \mathcal{S}$  we can calculate  $\mathbb{P}(E)$  from the probabilities of events of the type  $\{x : x \leq m\}$ . For example, let  $E = \{4, 5, 6\}$ . Then

$$\{x : x \leq 6\} = \{x : x \leq 3\} \cup E,$$

a disjoint union. Therefore, by Axiom 3,

$$\mathbb{P}(E) = \mathbb{P}(X \leq 6) - \mathbb{P}(X \leq 3).$$

The **cumulative distribution function** or c.d.f. of a random variable  $X$  is a function  $F_X : \mathbb{X} \rightarrow \mathbb{X}$  given by

$$F_X(m) = \mathbb{P}(X \leq m).$$

For a discrete random variable  $X$  with induced sample space  $\mathcal{S} \subseteq \mathbb{N}_0$ , the cumulative distribution function is given by

$$F_X(m) = \mathbb{P}(X \leq m) = \sum_{x=0}^{\lfloor m \rfloor} p_X(x).$$

**Exercise 3.14.** What is the cumulative distribution function for a random variable  $X$  whose p.m.f. is specified by  $p_X(x) = x/10$  for  $x = 1, 2, 3, 4$ ?

**Solution.** Let us first compute some values. We have:

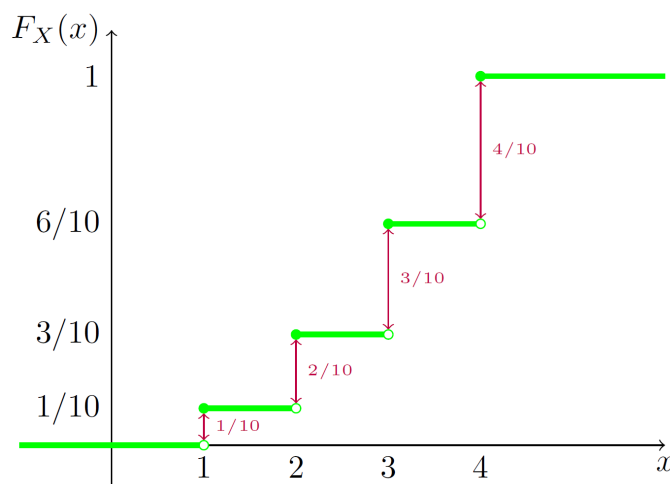
$$\begin{aligned} F_X(0) &= 0 \\ F_X(1) &= p_X(1) = 1/10 \\ F_X(2) &= p_X(1) + p_X(2) = 3/10 \\ F_X(3) &= p_X(1) + p_X(2) + p_X(3) = 6/10 = 3/5 \\ F_X(4) &= p_X(1) + p_X(2) + p_X(3) + p_X(4) = 1 \\ F_X(m) &= 1 \text{ for } m > 4. \end{aligned}$$

Of course, we're not done, as we need to specify  $F_X(x)$  for any real number  $x$  (since the c.d.f. is a function  $\mathbb{R} \rightarrow \mathbb{R}$ ). For this, we notice that for any  $x < 1$ , we have  $F_X(x) = 0$  since  $X$  as given in the problem cannot produce any value that is less than 1. Similarly, for any  $x \in [1, 2)$ , we have  $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X = 1) = 1/10$  since  $X$  as given in the problem is not allowed to take on non-integer values. Continuing in this vein, we obtain the full expression for  $F_X$ , namely:

$$F_X(x) = \begin{cases} 0, & x < 1 \\ 1/10, & x \in [1, 2) \\ 3/10, & x \in [2, 3) \\ 6/10, & x \in [3, 4) \\ 1, & x \geq 4 \end{cases}$$

Now we're done.

Though we weren't asked to do this, it is instructive to also plot  $F_X$ . It has the following graph:



In this case, the c.d.f. starts at the value 0, proceeds in jumps whose heights are the values of the p.m.f., and ends at the value 1. Is this an accident?

More generally, a cumulative distribution function  $F_X$  will have the following properties:

- $F_X$  is non-decreasing, meaning that if  $x_1 < x_2$ , we must have  $F_X(x_1) \leq F_X(x_2)$ .
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$ .
- $\lim_{x \rightarrow \infty} F_X(x) = 1$ .
- For any  $x_0 \in \mathbb{R}$ ,  $\mathbb{P}(X = x_0) = F_X(x_0) - \lim_{x \rightarrow x_0^-} F_X(x)$ . (Here,  $\lim_{x \rightarrow x_0^-} F_X(x)$  denotes the limit of  $F_X(x)$  as  $x$  approaches  $x_0$  from the left.)

These properties will be satisfied by any cumulative distribution function, including in the case of continuous random variables, and even more generally. You may wish to use these properties to check that your c.d.f. (on a coursework, on an exam) was indeed computed correctly. (For instance, if your c.d.f. starts decreasing, you must have made a mistake.)

### 3.4 Expectation

Expectation is a simple measure to calculate the average value taken by a random variable.

Suppose the outcome of an experiment is the random variable  $X$ . If the experiment is repeated, we observe outcomes  $x_1, x_2, \dots$ . The mean observed value of  $X$  is

$$\frac{x_1 + x_2 + \dots + x_n}{n}.$$

Let  $n_x$  be the number of times that  $x$  is observed in the  $n$  experiments. Then

$$x_1 + x_2 + \dots + x_n = \sum_{x \in \mathcal{S}} x n_x.$$

We motivated probability with the notion (as yet unproved) that

$$\frac{n_x}{n} \rightarrow \mathbb{P}(X = x) = p_X(x)$$

as  $n \rightarrow \infty$ . If this holds then

$$\begin{aligned} \frac{x_1 + x_2 + \dots + x_n}{n} &= \frac{\sum_{x \in \mathcal{S}} x n_x}{n} \\ &= \sum_{x \in \mathcal{S}} x \frac{n_x}{n} \\ &\rightarrow \sum_{x \in \mathcal{S}} x p_X(x). \end{aligned}$$

This motivates the following definition:

The **expected value** or **expectation** of a discrete random variable  $R$  is written  $\mathbb{E}[X]$  and defined by

$$\mathbb{E}[X] = \sum_{x \in \mathcal{S}} x p_X(x).$$

The expectation is also known as the **mean** or the **first moment** of a random variable.

An equivalent expression of the expectation that can be useful if the p.m.f. of a random variable is not known is the following:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\{\omega\}).$$

The equivalence between these two expressions can be shown using the way the p.m.f. is defined, and is left as an exercise.

**Example 3.15.** Recall the gambling exercise above, where the random variable  $X$ , profit, is defined by

$$X(\omega) = \begin{cases} -1 & \text{if } \omega = 1, 2, 3 \\ 0 & \text{if } \omega = 4 \\ 2 & \text{if } \omega = 5, 6 \end{cases}$$

from the throw of a fair die. The induced sample space for  $X$  is  $\mathcal{S} = \{-1, 0, 2\}$ . The p.m.f. of  $X$  is  $p_X(-1) = 3/6$ ,  $p_X(0) = 1/6$ ,  $p_X(2) = 2/6$ . Let us find the expected profit, using both definitions above.

Using the first definition:

$$\begin{aligned}\mathbb{E}[X] &= -1 \times p_X(-1) + 0 \times p_X(0) + 2 \times p_X(2) \\ &= -1 \times 3/6 + 0 \times 1/6 + 2 \times 2/6 \\ &= 1/6.\end{aligned}$$

Using the second definition:

$$\begin{aligned}\mathbb{E}[X] &= -1 \times 1/6 + -1 \times 1/6 + -1 \times 1/6 + 0 \times 1/6 + 2 \times 1/6 + 2 \times 1/6 \\ &= 1/6.\end{aligned}$$

Both give the same result of  $\mathcal{L}1/6$ .

**Exercise 3.16.** Find the expected number of heads in three tosses of a fair coin.

**Solution.** We first calculate the p.m.f.:

$$\begin{aligned}X(TTT) &= 0, p_X(0) = 1/8, \\ X(HTT) &= X(THT) = X(TTH) = 1, p_X(1) = 3/8, \\ X(HHT) &= X(HTH) = X(THH) = 2, p_X(2) = 3/8, \\ X(HHH) &= 3, p_X(3) = 1/8,\end{aligned}$$

Then we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=0}^3 x p_X(x) \\ &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} \\ &= 12/8 = 1.5.\end{aligned}$$

Note  $\mathbb{P}(\{X = \mathbb{E}[X]\}) = 0$ , which may seem surprising. The expected value of a random variable is **not** the value that you expect to obtain.

A similar calculation gives the expected number of tails is 1.5. This accords with intuition: we expect **on average** the same number of heads and tails for a fair coin.

**Exercise 3.17.** Find the expected value of the score on a fair die.

**Solution.** The discrete uniform law gives  $p_X(1) = p_X(2) = \dots = p_X(6) = 1/6$ . Let  $X$  represent the score, so that

$$\mathbb{E}[X] = \sum_{x=1}^6 x p_X(x) = \frac{1}{6}(1 + 2 + \dots + 6) = 3.5.$$

**Example 3.18.** For any event  $A \subseteq \Omega$ , the **indicator function** of  $A$  is the function  $I_A : \Omega \rightarrow \{0, 1\}$  such that  $I_A(\omega) = 1$  if  $\omega \in A$ , and  $I_A(\omega) = 0$  otherwise. Since  $I_A$  is a real-valued function on  $\Omega$ , it is a random variable. What is its expected value?

First find the p.m.f. of  $I_A$ .

$$\begin{aligned} p_I(1) &= \mathbb{P}(I_A = 1) = \mathbb{P}(\{\omega : \omega \in A\}) = \mathbb{P}(A) \\ p_I(0) &= \mathbb{P}(I_A = 0) = \mathbb{P}(\{\omega : \omega \in A^c\}) = 1 - \mathbb{P}(A). \end{aligned}$$

Therefore

$$\mathbb{E}[I_A] = 0 \times p_I(0) + 1 \times p_I(1) = \mathbb{P}(A).$$

If  $g : \mathbb{R} \rightarrow \mathbb{R}$  is any real valued function, then  $g(X) = g \circ X$  is a function from  $\Omega \rightarrow \mathbb{R}$ , so is also a random variable. We can therefore work out its expected value:

Using the second definition of expectation:

$$\begin{aligned} \mathbb{E}[g(X)] &= \sum_{\omega \in \Omega} g(X(\omega)) \mathbb{P}(\{\omega\}) \\ &= \sum_{x \in \mathcal{S}} \sum_{\omega: X(\omega)=x} g(X(\omega)) \mathbb{P}(\{\omega\}) \\ &= \sum_{x \in \mathcal{S}} g(x) \sum_{\omega: X(\omega)=x} \mathbb{P}(\{\omega\}) \\ &= \sum_{x \in \mathcal{S}} g(x) p_X(x), \end{aligned}$$

where the last line follows from the definition of a p.m.f.

We have proved the following:

**Lemma 3.19.** The expected value of a function  $g$  of a discrete random variable  $X$  is

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{S}} g(x) p_X(x).$$

**Exercise 3.20.** If  $p_X(x) = \frac{1}{3}$  for  $x = 0, 1, 2$ , find  $\mathbb{E}[X^2]$ .

**Solution.** The function here is  $g(x) = x^2$ . So

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_{x=0}^2 x^2 p_X(x) \\ &= 0^2 p_X(0) + 1^2 p_X(1) + 2^2 p_X(2) \\ &= \frac{1}{3}(0^2 + 1^2 + 2^2) = 5/3.\end{aligned}$$

**Exercise 3.21.** Find  $\mathbb{E}[X^3 + 2X]$  if  $p_X(1) = 3/4$  and  $p_X(2) = 1/4$ .

**Solution.**

$$\mathbb{E}(X^3 + 2X) = (1^3 + 2 \times 1)\frac{3}{4} + (2^3 + 2 \times 2)\frac{1}{4} = \frac{21}{4}.$$

Expectation obeys two important rules of linearity. For arbitrary functions  $g$  and  $h$ , and constant  $c$ :

$$\begin{aligned}\mathbb{E}[g(X) + h(X)] &= \mathbb{E}[g(X)] + \mathbb{E}[h(X)] \\ \mathbb{E}[cg(X)] &= c\mathbb{E}[g(X)].\end{aligned}$$

A special case is that  $\mathbb{E}[c] = c$ .

These results can be verified using the definition of the expectation of a function. We show how to obtain the first identity; the others are obtained similarly.

$$\begin{aligned}\mathbb{E}[g(X) + h(X)] &= \sum_{x \in \mathcal{S}} (g(x) + h(x))p_X(x) \\ &= \sum_{x \in \mathcal{S}} g(x)p_X(x) + \sum_{x \in \mathcal{S}} h(x)p_X(x) \\ &= \mathbb{E}[g(X)] + \mathbb{E}[h(X)]\end{aligned}$$

**Exercise 3.22.** Find  $\mathbb{E}[X]$  if it is known that  $\mathbb{E}[X(X - 1)] = 4$  and  $\mathbb{E}[X^2] = 3$ .

**Solution.**

$$\begin{aligned}4 = \mathbb{E}[X(X - 1)] &= \mathbb{E}[X^2 - X] \\ &= \mathbb{E}[X^2 + (-1)X] \\ &= \mathbb{E}[X^2] + \mathbb{E}[(-1)X] \\ &= \mathbb{E}[X^2] + (-1)\mathbb{E}[X] \\ &= 3 - \mathbb{E}[X]\end{aligned}$$

Therefore  $\mathbb{E}[X] = -1$ .

### 3.5 Variance

Expectation is a weighted average of the possible values of a random variable, and consequently is a measure of the location of the p.m.f.

The spread, or dispersion, of a random variable is usually measured by either the variance or the standard deviation, two closely related concepts that we define below.

The **variance** of the random variable  $X$ , written  $\text{Var}(X)$ , is defined as

$$\text{Var}(X) = \mathbb{E}((X - \mathbb{E}[X])^2),$$

The **standard deviation** of  $X$ , written  $\text{s.d.}(X)$ , is defined as the square root of the variance:

$$\text{s.d.}(X) = \sqrt{\text{Var}(X)}.$$

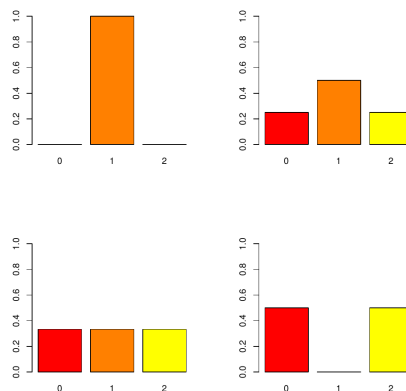
The variance is the expectation of the function of the random variable  $g(X) = (X - m)^2$ , where  $m = \mathbb{E}(X)$  is a real number.

The standard deviation quantifies how far away from the mean we can expect a random variable to typically be. To get a feel for this, in practice (and in theory that appears in later courses),<sup>1</sup> many random variables we deal with are within  $\pm 2$  standard deviations of the mean approximately 95% of the time. However, not every random variable has this property, as we shall see at the end of this chapter.

**Exercise 3.23.** Suppose that four random variables  $X_1$ ,  $X_2$ ,  $X_3$  and  $X_4$  on  $\mathcal{S} = \{0, 1, 2\}$  have p.m.f.

	0	1	2
$p_1(x)$	0	1	0
$p_2(x)$	1/4	1/2	1/4
$p_3(x)$	1/3	1/3	1/3
$p_4(x)$	1/2	0	1/2

respectively. These are plotted below.



<sup>1</sup>"In theory there's no difference between theory and practice, but in practice there is." – attributed to Yogi Berra

Note for each p.m.f. the sum of the probabilities is 1. The expectations are the same so that

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mathbb{E}[X_3] = \mathbb{E}[X_4] = 1.$$

Find the variances.

**Solution.** The different variances are

$$\begin{array}{llll} (0-1)^2 \times 0 & +(1-1)^2 \times 1 & +(2-1)^2 \times 0 & = 0, \\ (0-1)^2 \times 1/4 & +(1-1)^2 \times 1/2 & +(2-1)^2 \times 1/4 & = 1/2, \\ (0-1)^2 \times 1/3 & +(1-1)^2 \times 1/3 & +(2-1)^2 \times 1/3 & = 2/3, \\ (0-1)^2 \times 1/2 & +(1-1)^2 \times 0 & +(2-1)^2 \times 1/2 & = 1. \end{array}$$

We see that  $\text{Var}[X_1] < \text{Var}[X_2] < \text{Var}[X_3] < \text{Var}[X_4]$ . This agrees with our intuition of how dispersed each variable is (based on looking at the barplots).

This formulation of the variance is inconvenient for calculation, so we next derive an equivalent alternative form which is often easier to use in practice. Writing  $\mathbb{E}[X]$  as  $m$ , a constant, we have

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2], \\ &= \mathbb{E}[(X - m)^2], \\ &= \mathbb{E}[X^2 - 2Xm + m^2], \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[mX] + \mathbb{E}[m^2], \quad \text{linearity} \\ &= \mathbb{E}[X^2] - 2m\mathbb{E}[X] + m^2, \quad \text{linearity} \\ &= \mathbb{E}[X^2] - 2mm + m^2, \\ &= \mathbb{E}[X^2] - m^2, \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \end{aligned}$$

**Exercise 3.24.** Find the variance of a random number  $X$  uniformly distributed on the integers  $1, 2, \dots, 6$ .

**Solution.** From Exercise 3.17 we know  $\mathbb{E}[X] = \frac{7}{2}$ .

$$\begin{aligned} \mathbb{E}[X^2] &= \frac{1}{6}1^2 + \frac{1}{6}2^2 + \dots + \frac{1}{6}6^2 \\ &= \frac{1}{6} \times \frac{6 \times 7 \times 13}{6} = \frac{91}{6}. \end{aligned}$$

Therefore

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 \\ &= \frac{182}{12} - \frac{147}{12} = \frac{35}{12}. \end{aligned}$$

As with linearity for expectation there is an important result for the variance of linear transformations of a random variable  $X$ . Suppose  $a$  and  $b$  are constants. We will determine  $\text{Var}(aX + b)$  in terms of  $\text{Var}(X)$ .

$$\begin{aligned}
 \text{Var}(aX + b) &= \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2] \\
 &= \mathbb{E}[(aX + b - (a\mathbb{E}[X] + b))^2] && \text{linearity of } \mathbb{E} \\
 &= \mathbb{E}[(aX - a\mathbb{E}[X])^2] \\
 &= \mathbb{E}[a^2(X - \mathbb{E}[X])^2] \\
 &= a^2\mathbb{E}[(X - \mathbb{E}[X])^2] && \text{linearity of } \mathbb{E} \\
 &= a^2 \text{Var}(X).
 \end{aligned}$$

This result shows how variance is affected by linear transformations:

$$\begin{aligned}
 \text{Var}(X + b) &= \text{Var}(X) \\
 \text{Var}(aX) &= a^2 \text{Var}(X).
 \end{aligned}$$

**Exercise 3.25.** For a random variable  $X$ ,  $\mathbb{E}[X] = 3$  and  $\text{Var}(X) = 2$ . Find

- i.  $\mathbb{E}[2X] = 2\mathbb{E}[X] = 6$ .
- ii.  $\mathbb{E}[-2X + 6] = -2\mathbb{E}[X] + 6 = 0$ .
- iii.  $\text{Var}(2X) = 2^2 \text{Var}(X) = 8$ .
- iv.  $\text{Var}(-2X + 6) = (-2)^2 \text{Var}(X) = 8$ .

In summary

	$  \begin{aligned}  \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] \\  &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2.  \end{aligned}  $
It satisfies	$\text{Var}(aX + b) = a^2 \text{Var}(X).$

### 3.6 Chebyshev's inequality (not examinable)

A first step to understanding why the variance matters is given by **Chebyshev's inequality**. Let  $X$  be **any** random variable. Suppose  $\mathbb{E}(X) = m$  and  $\text{Var}(X) = \sigma^2$ . Let  $c > 0$  be any constant: we will find a bound on the probability

$$\mathbb{P}(|X - m| \geq c\sigma)$$

that  $X$  is at least  $c$  standard deviations away from its expected value.

Let  $A$  be the event that  $|X - m| \geq c\sigma$ , and let  $I_A$  be the indicator of  $A$ . Recall that

$$\mathbb{E}[I_A] = 1 \times \mathbb{P}(I_A = 1) + 0 \times \mathbb{P}(I_A = 0) = \mathbb{P}(A).$$

Also define the function  $g(x) = (x - m)^2 / (c\sigma)^2$ , and notice that  $g(x) \geq 0$  for all  $x$ , and  $g(x) \geq 1$  whenever  $|X - m| \geq c\sigma$ , i.e. whenever  $A$  occurs.

So if  $A$  does not occur, then  $I_A = 0 \leq g(X)$ .

And if  $A$  does occur, then  $I_A = 1 \leq g(X)$ .

So  $I_A \leq g(X)$  and it follows that

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{E}[I_A] \leq \mathbb{E}[g(X)] \\ &= \mathbb{E}[(X - m)^2 / (c\sigma)^2] \\ &= \frac{1}{c^2\sigma^2} \mathbb{E}[(X - m)^2] = \frac{1}{c^2}.\end{aligned}$$

Consequently,

$$\mathbb{P}(|X - m| \geq c\sigma) \leq \frac{1}{c^2}.$$

This is Chebyshev's inequality. It is the strongest inequality that is true for **every** random variable.

Suppose  $c > 1$  and  $X$  is a random variable with p.m.f.  $p_X(0) = \frac{1}{2c^2}$ ,  $p_X(1) = 1 - \frac{1}{c^2}$  and  $p_X(2) = \frac{1}{2c^2}$ . Then we can calculate that  $\mathbb{E}[X] = 1$  and s.d.( $X$ ) =  $1/c$ . But

$$\mathbb{P}(|X - 1| \geq c \text{ s.d.}(X)) = \mathbb{P}(|X - 1| \geq 1) = p_X(0) + p_X(2) = \frac{1}{c^2},$$

which is exactly the bound from Chebyshev's inequality.

## Chapter 4

# Models for discrete random variables

From a mathematical viewpoint any sequence of numbers,  $p(x)$ , satisfying

$$p(x) \geq 0 \text{ for all } x \text{ and } \sum_{x=0}^{\infty} p(x) = 1,$$

is a valid p.m.f. corresponding to some random variable. However, we are interested mainly in those random variables which arise from experiments of practical relevance, and thus form plausible models for future statistical modelling.

In this chapter we examine some examples of random variables that result from experiments with well-defined physical mechanisms.

### 4.1 Useful mathematical identities proved elsewhere

The following identities are helpful in the subsequent developments.

**Arithmetic progression:**

$$\sum_{i=0}^n i = n(n+1)/2.$$

**Sum of squares:**

$$\sum_{i=0}^n i^2 = n(n+1)(2n+1)/6.$$

**Exponential series:**

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

Note also that:

$$\begin{aligned}\exp(x+y) &= \exp(x) \exp(y) \\ \exp(-x) &= 1/\exp(x)\end{aligned}$$

## Geometric type sums

### Partial geometric sum:

$$S_m = 1 + x + x^2 + \dots + x^m = \frac{1 - x^{m+1}}{1 - x}$$

### Geometric sum:

$$S_\infty = 1 + x + x^2 + \dots = \frac{1}{1 - x} \quad (4.1)$$

for  $|x| < 1$ .

**Weighted geometric sum:** Using the result for geometric sums further results can be derived. Assuming  $|x| < 1$ , we can differentiate (4.1) and then multiply by  $x$  to obtain

$$\sum_{i=0}^{\infty} i x^i = x + 2x^2 + 3x^3 + \dots = \frac{x}{(1 - x)^2}.$$

Applying a similar process to the equation above gives

$$\sum_{i=0}^{\infty} i^2 x^i = x + 4x^2 + 9x^3 + \dots = \frac{x + x^2}{(1 - x)^3}.$$

## Binomial expansion

For any positive integer  $n$

$$\begin{aligned} (p + q)^n &= \binom{n}{0} p^n q^0 + \binom{n}{1} p^{n-1} q^1 + \dots + \binom{n}{i} p^{n-i} q^i + \dots + \binom{n}{n} p^0 q^n \\ &= \sum_{i=0}^n \binom{n}{i} p^{n-i} q^i. \end{aligned}$$

## 4.2 Discrete uniform random variables

Consider an experiment where the sample space is  $\{a, a + 1, \dots, b\}$  for some integers  $a, b$  with  $0 \leq a < b$ , and the random variable  $X$  corresponds to an outcome being picked at random from the sample space. Each outcome is **equally likely**.

We write this as  $X \sim \text{Uniform}(a, b)$ .

Examples include:

- the score on a fair die, where  $a = 1$  and  $b = 6$ ;
- the date of someone's birthday, given that they were born in January (assuming all birthdays are equally likely), where  $a = 1$  and  $b = 31$ .

**Exercise 4.1.** Write down the p.m.f. of a discrete uniform random variable.

**Solution.** For  $x = a, a + 1, \dots, b$ ,

$$p_X(x) = \frac{1}{b - a + 1}.$$

Otherwise  $p_X(x) = 0$ .

**Example 4.2.** Calculate the expectation and variance of a discrete uniform random variable.

**Solution.**

$$\begin{aligned}
 \mathbb{E}[X] &= \sum_{x=0}^{\infty} x p_X(x) \\
 &= \sum_{x=a}^b x \frac{1}{b-a+1} \\
 &= \frac{1}{b-a+1} \left( \sum_{x=0}^b r - \sum_{x=0}^{a-1} x \right) \\
 &= \frac{1}{b-a+1} \times \frac{b(b+1) - a(a-1)}{2} \\
 &= \frac{b+a}{2}.
 \end{aligned}$$

To find the variance we need  $\mathbb{E}[X^2]$ .

$$\begin{aligned}
 \mathbb{E}[X^2] &= \sum_{x=0}^m x^2 \frac{1}{b-a+1} \\
 &= \frac{1}{b-a+1} \left( \sum_{x=0}^b x^2 - \sum_{x=0}^{a-1} x^2 \right) \\
 &= \frac{1}{b-a+1} \times \frac{b(b+1)(2b+1) - (a-1)a(2a-1)}{6} \\
 &= \frac{(b-a)(2b-2a+1)}{6}.
 \end{aligned}$$

Now

$$\begin{aligned}
 \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \frac{(b-a)(2b-2a+1)}{6} - \frac{(b-a)^2}{4} \\
 &= \frac{(b-a)(b-a+2)}{12}.
 \end{aligned}$$

### 4.3 Bernoulli random variables

Jacob Bernoulli (1654–1705) was a member of a prolific scientific family: at least twelve members contributed to some branch of mathematics or physics and at least five to probability. Jacob and his brother John were great rivals and would only communicate in print arguing over the correctness of each other's mathematical proofs. Jacob gives his name to the simplest family of non-uniform random variables.

Consider an experiment where the sample space is  $\{0, 1\}$  and the probability of a 1 is  $\theta$  ( $0 \leq \theta \leq 1$ ). A random variable  $X$  with such a p.m.f. is termed a **Bernoulli random variable** with **parameter**  $\theta$ . Examples include:

- number of heads on a single toss of a biased coin;
- the number of positive results on a single COVID test;
- 1 if a candidate passes an exam and 0 otherwise;
- 1 if the next baby is a girl, 0 otherwise (here  $\theta \approx 0.49$ ).

Here outcomes of 1 and 0 are sometimes called “success” and “failure” respectively – although, as in the COVID example above, “failure” can be the better outcome! Note that all of these examples give an indicator function for some event, and indicator functions of events are always Bernoulli random variables.

For Bernoulli random variables

$$\mathbb{P}(X = 1) = \theta, \quad \mathbb{P}(X = 0) = 1 - \theta,$$

and  $\mathbb{P}(X = x) = 0$  otherwise.

The **p.m.f. of a Bernoulli random variable**  $X$  is

$$p_X(i) = \begin{cases} (1 - \theta) & \text{if } i = 0 \\ \theta & \text{if } i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

We write  $X \sim \text{Bernoulli}(\theta)$ .

**Exercise 4.3.** Find the expectation and variance of a Bernoulli random variable.

**Solution.**

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x=0}^{\infty} x p_X(x) \\ &= 0 \times (1 - \theta) + 1 \times \theta \\ &= \theta. \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X^2] &= \sum_{x=0}^{\infty} x^2 p_X(x) \\ &= 0^2 \times (1 - \theta) + 1^2 \times \theta \\ &= \theta. \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \theta - \theta^2 = \theta(1 - \theta). \end{aligned}$$

Notice if  $\theta = 0$  or  $\theta = 1$  the variance is 0, whereas the maximum possible variance is 0.25 when  $\theta = 0.5$ . Is this logical?

## 4.4 Binomial random variables

Consider an experiment in which  $n$  independent Bernoulli trials are carried out, each with probability of success being  $\theta$ . Let  $X$  be the random variable reporting the number of successes in these  $n$  trials. The induced sample space is  $\{0, 1, \dots, n\}$ . The random variable  $X$  is a **binomial random variable** with **parameters**  $n$  and  $\theta$ . We write  $X \sim \text{Bin}(n, \theta)$ .

Examples include:

- the number of heads in  $n$  tosses of a biased coin;
- the number of positive COVID tests when  $n$  people are tested;
- the number of girls among the next  $n$  babies born.

The derivation is a little more complex here so first consider the  $n = 3$  case with S and F denoting success and failure respectively. The sample space for the experiment is

$$\Omega = \{SSS, SSF, SFS, FSS, SFF, FSF, FFS, FFF\}.$$

The random variable of interest,  $X$ , is the number of successes.

**Exercise 4.4.** Find  $\mathbb{P}(X = x)$  for  $x = 0, 1, 2, 3$ .

**Solution.** Previously, when  $\theta = 0.5$ , we used the discrete uniform law to derive the p.m.f. This is not possible with an arbitrary  $\theta$ . Instead we need to use independence to calculate the probabilities of the sample points. This results in the following calculations:

$$\begin{aligned} p_X(0) &= \mathbb{P}(\{FFF\}) \\ &= (1 - \theta)(1 - \theta)(1 - \theta) \\ &= (1 - \theta)^3, \\ p_X(1) &= \mathbb{P}(\{SFF\}) + \mathbb{P}(\{FSF\}) + \mathbb{P}(\{FFS\}) \\ &= 3\theta(1 - \theta)^2, \\ p_X(2) &= \mathbb{P}(\{SSF\}) + \mathbb{P}(\{SFS\}) + \mathbb{P}(\{FSS\}) \\ &= 3\theta^2(1 - \theta), \\ p_X(3) &= \mathbb{P}(\{SSS\}) \\ &= \theta^3. \end{aligned}$$

with  $p_X(x) = 0$  for other values of  $x$ .

We can summarise these results as

$$p_X(x) = \binom{3}{x} \theta^x (1 - \theta)^{3-x}$$

for  $x = 0, 1, 2, 3$ .

The more general form for the p.m.f. is as follows:

**Lemma 4.5** (The p.m.f. of a binomial random variable). The p.m.f. of a binomial random variable  $X \sim \text{Bin}(n, \theta)$  is

$$p_X(r) = \binom{n}{r} \theta^r (1 - \theta)^{n-r}$$

for  $r = 0, 1, 2, \dots, n$ , with  $p_X(r) = 0$  otherwise, where  $0 \leq \theta \leq 1$ .

*Proof.* • For any sample point with  $r$  successes and  $n - r$  failures, the probability of the event consisting solely of that sample point is  $\theta^r(1 - \theta)^{n-r}$ , by independence.

• There are  $\binom{n}{r}$  sample points with  $r$  successes and  $n - r$  failures (choose  $r$  of the  $n$  trials to be the successes).

• Hence  $\mathbb{P}(X = r) = \sum_{\omega: X(\omega)=r} \mathbb{P}(\{\omega\}) = \sum_{\omega: X(\omega)=r} \theta^r(1 - \theta)^{n-r} = \binom{n}{r} \theta^r(1 - \theta)^{n-r}$ .

□

**Exercise 4.6.** Show that  $\sum_{r=0}^m p_X(r) = 1$

**Solution.** By the binomial theorem,

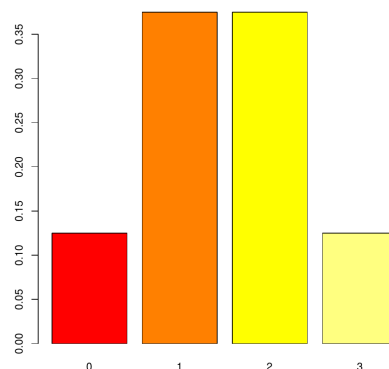
$$\sum_{r=0}^m \binom{m}{r} a^r b^{m-r} = (a + b)^m.$$

Putting  $a = \theta$  and  $b = 1 - \theta$  gives the result.

The software package R can evaluate p.m.f.s from many standard probability models, including the binomial.

**Example 4.7.** The random variable  $X \sim \text{Bin}(3, 0.5)$ . Use R to evaluate and plot the p.m.f. of  $X$ . Repeat with  $\theta = 0.4$ .

```
dbinom(0:3,size=3,prob=0.5)
dbinom(0:3,size=3,prob=0.4)
# Note how the probabilities change.
p = dbinom(0:3,size=3,prob=0.5)
barplot(p, names.arg=c(0:3))
```



**Exercise 4.8.** Find the probability of rolling a fair die and finding

- i. two sixes in four rolls,
- ii. two sixes in five rolls,
- iii. at least two sixes in four rolls.

**Solution.** i. Let  $X$  be the number of sixes in four rolls, so  $X \sim \text{Bin}(4, 1/6)$ .

$$\begin{aligned}\mathbb{P}(X = 2) &= \binom{4}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2 \\ &= 25/216 \approx 0.116.\end{aligned}$$

ii. Let  $Y$  be the number of sixes in five rolls, so  $Y \sim \text{Bin}(5, 1/6)$ .

$$\begin{aligned}\mathbb{P}(Y = 2) &= \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 \\ &\approx 0.161.\end{aligned}$$

iii. With  $X$  as in part i,

$$\begin{aligned}\mathbb{P}(X \geq 2) &= 1 - \mathbb{P}(X < 2) \\ &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) \\ &= 1 - \binom{4}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^4 - \binom{4}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^3 \\ &= 19/144 = 0.132.\end{aligned}$$

We could calculate these in R using the following commands:

```
dbinom(2,size=4,prob=1/6)
dbinom(2,size=5,prob=1/6)
1-dbinom(0,size=4,prob=1/6)-dbinom(1,size=4,prob=1/6)
```

**Exercise 4.9.** There are two families each with three children. Suppose each child is independently a girl with probability  $1/2$ . Find the probability that the families have the same number of girls.

R hint: `sum( dbinom(0:3, size=3, prob=1/2)^2 )`

**Solution.** Let  $X_i$  be the number of girls in family  $i$ , for  $i = 1, 2$ .

By independence, we have  $X_1, X_2 \sim \text{Bin}(3, 0.5)$ . Therefore

$$\begin{aligned}\mathbb{P}(X_1 = X_2) &= \sum_{r=0}^3 \mathbb{P}(X_1 = r, X_2 = r) \quad \text{additivity} \\ &= \sum_{r=0}^3 \mathbb{P}(X_1 = r) \mathbb{P}(X_2 = r) \quad \text{independence} \\ &= \sum_{r=0}^3 \left[ \binom{3}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{3-r} \right]^2\end{aligned}$$

$$\begin{aligned}
&= \left(\frac{1}{2}\right)^6 \sum_{r=0}^3 \binom{3}{r}^2 \\
&= \frac{1}{64} (1^2 + 3^2 + 3^2 + 1^2) \\
&= 20/64 = 0.3125.
\end{aligned}$$

## Expectation and variance

Using the definitions and algebraic manipulation gives

For a binomial random variable  $X \sim \text{Bin}(n, \theta)$

$$\begin{aligned}
\mathbb{E}[X] &= n\theta \\
\text{Var}(X) &= n\theta(1 - \theta).
\end{aligned}$$

The general proof is given in a worksheet solution. We instead consider a special case:

**Example 4.10.** If  $X \sim \text{Bin}(3, \theta)$  show that  $\mathbb{E}[X] = 3\theta$ .

**Solution.**

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{r=0}^{\infty} r p_X(r) \\
&= \sum_{r=0}^3 r p_X(r) \\
&= 0 + \sum_{r=1}^3 r p_X(r)
\end{aligned}$$

Now substitute for the p.m.f. and simplify

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{r=1}^3 r \binom{3}{r} \theta^r (1 - \theta)^{3-r} \\
&= \sum_{r=1}^3 r \frac{3!}{r! (3-r)!} \theta^r (1 - \theta)^{3-r} \\
&= 3\theta \sum_{r=1}^3 \frac{2!}{(r-1)! (3-r)!} \theta^{r-1} (1 - \theta)^{3-r}
\end{aligned}$$

Now put  $s = r - 1$  and use the binomial theorem

$$\begin{aligned}
\mathbb{E}[X] &= 3\theta \sum_{s=0}^2 \frac{2!}{(s)! (2-s)!} \theta^s (1 - \theta)^{2-s} \\
&= 3\theta (\theta + (1 - \theta)^2) \\
&= 3\theta.
\end{aligned}$$

As  $p_Y(s)$  is the p.m.f. of  $Y \sim \text{Bin}(2, \theta)$ .

**Practice question 4.11.** The popular website Dialtome shows its users adverts when they log in, and gains revenue when users click on adverts, so its owners need to predict how many times this happens. Suppose that 900 million users log in any given day, and each user independently clicks an advert with probability 0.01. What is the expected number of clicks in a day, and the variance of this number?

**Exercise 4.12.** Suppose an experiment is carried out  $n$  times, let  $A$  be an event associated with the experiment, and let  $\theta$  be the probability of the event  $A$ .

Let  $X$  count the number of times that event  $A$  occurs in the  $n$  experiments.  $X$  is therefore a  $\text{Bin}(n, \theta)$  random variable.

Calculate the expectation and variance of  $X/n$ , the proportion of times that  $A$  occurs. Use Chebyshev's inequality to say something about how close  $X/n$  is to  $\theta$  for large  $n$ .

**Solution.** We know that

$$\begin{aligned}\mathbb{E}[X/n] &= \frac{1}{n} \mathbb{E}[X] \\ &= \frac{1}{n} n\theta \\ &= \theta \\ \text{Var}(X/n) &= \frac{1}{n^2} \text{Var}(X) \\ &= \frac{1}{n^2} n\theta(1 - \theta) \\ &= \frac{\theta(1 - \theta)}{n}\end{aligned}$$

Chebyshev's inequality tells us that

$$\mathbb{P}(|X/n - \mathbb{E}[X/n]| > c \text{ s.d.}(X/n)) < \frac{1}{c^2}.$$

Hence

$$\mathbb{P}(|X/n - \theta| > c\sqrt{\theta(1 - \theta)/n}) < \frac{1}{c^2},$$

and therefore

$$\mathbb{P}(|X/n - \theta| \leq c\sqrt{\theta(1 - \theta)/n}) \geq 1 - \frac{1}{c^2},$$

Taking, for example,  $c = 10$ , we see the probability  $X/n$  is within  $10\sqrt{\theta(1 - \theta)/n}$  of  $\theta$  is at least 0.99. Since  $10\sqrt{\theta(1 - \theta)/n} \rightarrow 0$  as  $n \rightarrow \infty$ , it follows that  $X/n$  is likely to be very close to  $\theta$  for large  $n$ .

This confirms that the axioms we set up support our intuitive beliefs: the proportion of times an event  $A$  will be close to  $\mathbb{P}(A)$  if the number of trials is large.

## 4.5 Geometric random variables

Consider an experiment based on independent Bernoulli trials, each with the probability of a success being  $\theta$ .

Now define the variable of interest,  $X$ , to be the number of trials up to but **not including** the first success.

Here the induced sample space is  $\mathcal{S} = \{0, 1, 2, \dots\}$ , and is infinite, corresponding to outcomes in the original sample space

$$\Omega = \{S, FS, FFS, FFFS, \dots\}.$$

If, for example, the sequence FFFFS occurs then we have  $X(\text{FFFS}) = 4$ .

Such a random variable is called a **geometric random variable** with parameter  $\theta$ .<sup>1</sup> Examples of geometric random variables include:

- the number of heads of a coin toss before the first tail;
- the number of boys born before the first girl;
- the number of black cars passed before a red car;
- the number of years to pass before your team wins the FA cup.

We write  $X \sim \text{Geometric}(\theta)$ .

**Exercise 4.13.** Use the independence of the Bernoulli random variables to derive the p.m.f. of the geometric random variable.

**Hint:**  $X = 4$  corresponds to the sample point FFFFS.

**Solution.**

$$\begin{aligned} p_X(r) &= \mathbb{P}(X = r) = \mathbb{P}(\underbrace{\{FF \cdots FF\}_r S}) \\ &= \mathbb{P}(F)\mathbb{P}(F) \cdots \mathbb{P}(F)\mathbb{P}(S) && \text{independence} \\ &= (1 - \theta)^r \theta && \text{for } r = 0, 1, 2, \dots \end{aligned}$$

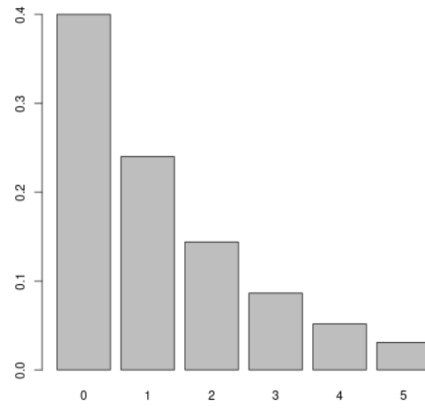
**Exercise 4.14.** The random variable  $X \sim \text{Geometric}(0.3)$ . Use R to evaluate and plot the p.m.f. of  $X$  for  $r = 0, 1, 2, \dots, 5$ , with the commands

```
dgeom(0:5,prob=0.3)
dgeom(0:5,prob=0.4)
# Note how the probabilities change
barplot( dgeom(0:5,prob=0.4),names.arg=c(0:5) )
```

Repeat with  $\theta = 0.4$  and plot.

---

<sup>1</sup>Note that this is slightly different to the definition in some textbooks, which use  $X + 1$  instead of  $X$ .



**Exercise 4.15.** Verify that  $\sum_{r=0}^{\infty} p_X(r) = 1$  for the geometric p.m.f. This requires familiarity with the sum of a geometric series given at the start of this chapter.

**Solution.**

$$\begin{aligned}
 \sum_{r=0}^{\infty} p_X(r) &= \sum_{r=0}^{\infty} (1-\theta)^r \theta \\
 &= \theta \sum_{r=0}^{\infty} (1-\theta)^r \\
 &= \theta \frac{1}{1-(1-\theta)} = 1.
 \end{aligned}$$

**Exercise 4.16.** For a general  $X \sim \text{Geometric}(\theta)$ , find  $\mathbb{P}(X \geq r)$ .

**Solution.**

$$\begin{aligned}
 \mathbb{P}(X \geq r) &= \sum_{s=r}^{\infty} p_X(s) \\
 &= \sum_{s=r}^{\infty} (1-\theta)^s \theta \\
 &= (1-\theta)^r \theta \sum_{s=r}^{\infty} (1-\theta)^{s-r} \\
 &= (1-\theta)^r \theta \sum_{s'=0}^{\infty} (1-\theta)^{s'} \quad \text{setting } s' = s - r \\
 &= (1-\theta)^r \theta \frac{1}{1-(1-\theta)} \\
 &= (1-\theta)^r
 \end{aligned}$$

Note that this is simply the probability that the first  $r$  Bernoulli trials are all  $F$ .

**Example 4.17.** Find  $\mathbb{E}[X]$  and  $\text{Var}(X)$  for a geometric random variable.

We need to use the basic identities from Section 4.1 on page 44.

**Solution.**

$$\begin{aligned}
 \mathbb{E}[X] &= \sum_{r=0}^{\infty} r p_X(r) \\
 &= 0 + \sum_{r=1}^{\infty} r (1 - \theta)^r \theta \\
 &= (1 - \theta) \theta \sum_{r=1}^{\infty} r (1 - \theta)^{r-1} \\
 &= (1 - \theta) \theta [(1 - (1 - \theta))^{-2}] \\
 &= \frac{1 - \theta}{\theta}.
 \end{aligned}$$

Now to find  $\text{Var}(X)$  we begin by calculating  $\mathbb{E}[X^2]$ :

$$\begin{aligned}
 \mathbb{E}[X^2] &= \sum_{r=0}^{\infty} r^2 p_X(r) \\
 &= \sum_{r=0}^{\infty} r^2 (1 - \theta)^r \theta \\
 &= \theta \frac{(1 - \theta) + (1 - \theta)^2}{(1 - (1 - \theta))^3} \\
 &= \frac{2 - 3\theta + \theta^2}{\theta^2}
 \end{aligned}$$

We then substitute this to get

$$\begin{aligned}
 \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \frac{2 - 3\theta + \theta^2}{\theta^2} - \frac{1 - 2\theta + \theta^2}{\theta^2} \\
 &= \frac{1 - \theta}{\theta^2}.
 \end{aligned}$$

Note that if the Bernoulli probability  $\theta$  is very small then the expectation and variance are both very large – in the limit as  $\theta \rightarrow 0$  these go to infinity.

To summarise:

<p>For a geometric random variable <math>X \sim \text{Geometric}(\theta)</math></p> $p_X(r) = (1 - \theta)^r \theta \text{ for } r = 0, 1, 2, \dots$ $p_X(r) = 0 \text{ otherwise}$ $\mathbb{E}[X] = \frac{1 - \theta}{\theta}$ $\text{Var}(X) = \frac{1 - \theta}{\theta^2}$
---

**Exercise 4.18.** Alice and Bob play a series of chess matches. Alice's probability of winning a match is 0.4, Bob's probability of winning a match is 0.3, and the probability of a draw is 0.3.

- a. What is the probability that Alice wins three matches out of five?

Answer:  $\binom{5}{3}0.4^30.6^2$ .

- b. Find the p.m.f. of the number of matches won by Alice out of  $n$  matches played.

Answer: Let  $K$  be the random variable denoting the number of matches won by Alice out of the  $n$  matches played. Then  $K \sim \text{Bin}(n, 0.4)$ , so

$$p_K(k) = \begin{cases} \binom{n}{k}0.4^k0.6^{n-k} & \text{for } k = 0, 1, \dots, n \\ 0 & \text{otherwise.} \end{cases}$$

- c. Find the p.m.f. of the number of draws up until the first match won by either Alice or Bob.

Answer: let  $X$  be the number of draws up until the first match won by either Alice or Bob. For any match, the probability of a draw is 0.3, the probability of a win (by either Alice or Bob) is 0.7, and the outcomes of the matches are independent of one another. Therefore  $X \sim \text{Geometric}(0.7)$ , so

$$p_X(x) = \begin{cases} 0.3^x0.7 & \text{for } x = 0, 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

- d. If we know that Alice did not win a particular match, what is the probability that it was a draw?

Answer:  $0.3/(0.3 + 0.3) = 0.5$ .

- e. If Alice and Bob play until the first time Alice wins, find the p.m.f. of  $M$ , the number of matches that are played.

Answer: Notice that  $M - 1$  is the number of draws or wins by Bob up until the first match won by Alice, so  $M - 1 \sim \text{Geometric}(0.4)$ . Therefore

$$p_M(m) = \begin{cases} 0.6^{m-1}0.4 & \text{for } m = 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

(In some textbooks, this would be called a geometric distribution, but we will refer to it as a **shifted geometric**.)

- f. Find the p.m.f. of the number  $D$  of draws up until the first match won by Alice.

Answer: By the total probability theorem, for any  $d = 0, 1, \dots$ ,

$$p_D(d) = \mathbb{P}(D = d) = \sum_{m=1}^{\infty} \mathbb{P}(D = d \cap M = m).$$

In order to have  $M = m$  and  $D = d$  we need  $m \geq d + 1$ . Then there are  $\binom{m-1}{d}$  choices for which of the first  $m - 1$  matches are draws, and the probability of a specific sequence of  $d$  draws,  $m - 1 - d$  wins for Bob, and a final win for Alice, is  $0.3^{m-1}0.4$ , so giving

$$p_D(d) = \sum_{m=d+1}^{\infty} \binom{m-1}{d} 0.3^{m-1}0.4, \quad d = 0, 1, 2, \dots$$

## 4.6 Poisson random variables

Unlike the other random variables discussed here, we define the Poisson random variable directly through its p.m.f.:

Let  $\lambda$  be a positive real number. The p.m.f. of a **Poisson random variable**  $X$  with parameter  $\lambda$  is

$$p_X(r) = \frac{\lambda^r \exp(-\lambda)}{r!},$$

for  $r = 0, 1, 2, \dots$ , with  $p_X(r) = 0$  otherwise. We write  $X \sim \text{Pois}(\lambda)$ .

The Poisson random variable arises physically in two ways:

1. The number of events in a fixed time interval of a continuous time process where events occur at random at a given rate over time. (Covered in later courses.)
2. As an approximation for the number of successes when there are many trials but the probability of success is very rare.

Examples of quantities that may be modelled by Poisson random variables are:

- the number of raindrops to land on your head in a given time interval;
- the number of floods of a river in a year;
- the number of deaths due to typhoid over a year (assuming typhoid cases are independent);
- the number of hits on a website in a given period of time.

**Practice question 4.19.** Calculate the probabilities of 0, 1 and 2 deaths from typhoid in a year if the number  $X$  has a Poisson random variable with  $\lambda = 4.6$ .

R hint: `dpois(0:2,lambda=4.6)`

**Exercise 4.20.** Verify that  $p_X(r) = \frac{\lambda^r \exp(-\lambda)}{r!}$  is a proper probability mass function. You will need to use the definition of the exponential function through its series expansion.

**Solution.** For each  $r \geq 0$ ,

$$\begin{aligned} p_X(r) &= \frac{\lambda^r \exp(-\lambda)}{r!} \\ &\geq 0. \end{aligned}$$

We also need that the probabilities sum to 1.

$$\begin{aligned} \sum_{r=0}^{\infty} p_X(r) &= \sum_{r=0}^{\infty} \frac{\lambda^r \exp(-\lambda)}{r!} \\ &= \exp(-\lambda) \sum_{r=0}^{\infty} \frac{\lambda^r}{r!} \\ &= \exp(-\lambda) \exp(\lambda) \\ &= 1. \end{aligned}$$

**Exercise 4.21.** If  $X \sim \text{Pois}(\lambda)$  show that  $\mathbb{E}[X] = \lambda$ . The technique is similar to that in the above example.

**Solution.**

$$\begin{aligned}
 \mathbb{E}[X] &= \sum_{r=0}^{\infty} r p(r) \\
 &= \sum_{r=0}^{\infty} r \frac{\lambda^r \exp(-\lambda)}{r!} \\
 &= 0 + \sum_{r=1}^{\infty} r \frac{\lambda^r \exp(-\lambda)}{r!} \\
 &= \lambda \exp(-\lambda) \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} \\
 &= \lambda \exp(-\lambda) \exp(\lambda) \\
 &= \lambda.
 \end{aligned}$$

**Example 4.22.** Find the variance of a Poisson random variable by first computing  $\mathbb{E}[X(X-1)]$ .

**Solution.**

$$\begin{aligned}
 \mathbb{E}[X(X-1)] &= \sum_{r=0}^{\infty} r(r-1) p_X(r) \\
 &= 0 + 0 + \sum_{r=2}^{\infty} r(r-1) p_X(r) \\
 &= \sum_{r=2}^{\infty} r(r-1) \frac{\lambda^r \exp(-\lambda)}{r!} \\
 &= \lambda^2 \exp(-\lambda) \sum_{r=2}^{\infty} \frac{\lambda^{r-2}}{(r-2)!} \\
 &= \lambda^2 \exp(-\lambda) \exp(\lambda) = \lambda^2.
 \end{aligned}$$

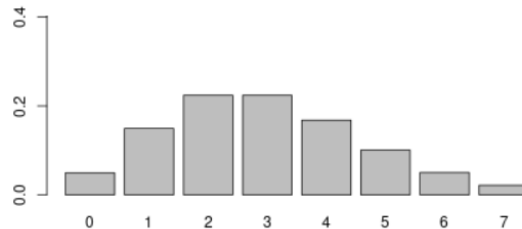
Hence

$$\begin{aligned}
 \text{Var}(X) &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 \\
 &= \lambda^2 + \lambda - (\lambda)^2 \\
 &= \lambda.
 \end{aligned}$$

Thus a key property of the Poisson p.m.f. is that the expectation and variance are both equal to  $\lambda$ .

**Example 4.23.** Use R to give a barplot of the Poisson p.m.f. on  $0, 1, \dots, 7$ , when  $\lambda = 0.5$  and when  $\lambda = 3$ .

```
par(mfrow=c(1,2))
barplot( dpois(0:7,lambda=0.5),names.arg=c(0:7),ylim=c(0,1) )
barplot( dpois(0:7,lambda=3),names.arg=c(0:7),ylim=c(0,1) )
```



### Poisson approximation to the binomial

One use of Poisson random variables is for rare events. Consider a binomial random variable  $X$  from  $n$  trials with the probability of success being  $\theta$ . Suppose  $n$  is very large, but  $\theta$  is very small.

Each event has a small probability of occurring, but when there are a large number of trials, the probability that one or more events occur is not negligible. Define  $\lambda = n\theta$ . We will make  $n$  large while keeping  $\lambda$  fixed (and so  $\theta = \lambda/n$ ).

We will use the following fact:

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = \exp(x)$$

for all  $x \in \mathbb{R}$ .

$$\begin{aligned} \mathbb{P}(X = r) &= \binom{n}{r} \theta^r (1 - \theta)^{n-r} \\ &= \frac{n!}{r!(n-r)!} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r} \\ &= \frac{\lambda^r}{r!} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\rightarrow \exp(-\lambda)} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-r}}_{\rightarrow 1} \underbrace{\frac{n!}{(n-r)!n^r}}_{\rightarrow 1} \\ &\rightarrow \frac{\lambda^r}{r!} \exp(-\lambda), \end{aligned}$$

where the limits are taken as  $n \rightarrow \infty$ .

Therefore a binomial p.m.f. with large  $n$  and correspondingly small  $\theta$  can be approximated by a Poisson p.m.f. with parameter  $\lambda = n\theta$ . For a good approximation we should have  $n \geq 100$  and  $\theta \leq .01$ .

**Exercise 4.24.** The number of cases of a rare disease is on average 3.67 per month. In the month after a festival there were 14 cases of the disease reported.

How unusual is this? Compute the  $p$ -value  $= \mathbb{P}(\text{observed value or worse})$ , under the assumption of natural Poisson variability.

It is  $\mathbb{P}(X \geq 14)$  and measures the worry in observing 14 cases.

R hint: `1-sum( dpois(0:13,lambda=3.67) )`

**Solution.** We could count the number of susceptible people, and estimate the probability that each catches the disease. However, since this is a rare event, we can instead model the number of cases each month as  $R \sim \text{Pois}(3.67)$ . This makes the calculations simpler, and also means that we don't need to know the population.

$$p_X(r) = \frac{3.67^r \exp(-3.67)}{r!}$$

for  $r = 0, 1, 2, \dots$ , so the  $p$ -value is

$$\mathbb{P}(X \geq 14) = 1 - \sum_{r=0}^{13} p(r) \approx 0.000031.$$

Very small and so worrying: the festival may have caused an outbreak.

**Exercise 4.25.** I believe that the number of mushrooms that grow on my lawn in any given week is a Poisson random variable. I have observed that no mushrooms grow with probability 0.1. What is the average number of mushrooms growing in a week?

**Solution.** The number  $M \sim \text{Pois}(\lambda)$ . We know  $\mathbb{P}(M = 0) = e^{-\lambda} = 0.1$  and so  $e^{\lambda} = 10$  giving  $\lambda = \log 10 \approx 2.3$ . This is also  $\mathbb{E}[M]$ .

## 4.7 Negative binomial random variables (not examinable)

Consider an experiment for which the random variable of interest,  $X$ , corresponds to the number of failures to occur before the  $k$ th success in a series of independent Bernoulli trials, each with probability of success being  $\theta$ .

Here the sample space when  $k = 2$  is  $\{\text{SS}, \text{FSS}, \text{SFS}, \text{FFSS}, \text{FSFS}, \text{SFFS}, \dots\}$ , and the induced sample space is  $\{0, 1, 2, \dots\}$ .

What is the probability that  $X = r$ ? This requires  $r$  failures and  $k$  successes in  $r + k$  trials, with the last trial being a success (or we would have stopped earlier). In other words, we need  $k - 1$  successes out of the first  $r + k - 1$  trials, followed by a success. This has probability

$$\binom{k+r-1}{k-1} (1-\theta)^r \theta^{k-1} \times \theta = \binom{k+r-1}{k-1} (1-\theta)^r \theta^k.$$

It is possible to show that

$$\mathbb{E}[X] = \frac{k(1-\theta)}{\theta}$$

$$\text{Var}(X) = \frac{k(1-\theta)}{\theta^2}.$$

Note that, setting  $k = 1$ , these results match what we already know about the geometric p.m.f.

## 4.8 Summary

There are many discrete probability models based on Bernoulli trials. In these cases

- the sample space  $\Omega$  is a set of sequences of S and F (successes and failures);
- the induced sample space  $\mathcal{S}$  is usually  $\{0, 1, 2, \dots\}$  or  $\{0, 1, \dots, n\}$  for some parameter  $n$ ;
- trials are independent, with the same success probability on each trial  $\mathbb{P}(S) = \theta$ .

A random variable  $X$  is a function from  $\Omega$  to  $\mathcal{S}$ , with p.m.f.  $p_X(r) = \mathbb{P}(\{X = r\})$ .

Natural constructions for  $X$  give the p.m.f. for Bernoulli, binomial, geometric random variables etc. The following table summarises the distributions we have looked at.

	$\mathcal{S}$	construction	$p_X(0)$
Bernoulli	$\{0, 1\}$	single trial	$1 - \theta$
Binomial	$\{0, 1, \dots, n\}$	successes in $n$ trials	$(1 - \theta)^n$
Geometric	$\{0, 1, \dots\}$	failures before first success	$\theta$
Poisson	$\{0, 1, \dots\}$	Limit of $\text{Bin}(n, \lambda/n)$ as $n \rightarrow \infty$	$\exp(-\lambda)$
Negative binomial	$\{0, 1, \dots\}$	failures before $k$ th success	$\theta^k$

**Exercise 4.26.** I sell items door-to-door. For each of the following random variables, express it in terms of a standard random variable. If the expectation is 10, what must the parameter(s) be?

- a. The number  $W$  of houses where someone is in out of the first 25 I try.

*Solution:* This is  $\text{Bin}(25, p)$ . Since  $\mathbb{E}[W] = 25p = 10$ ,  $p = 0.4$ .

- b. The number  $X$  of houses I try **before** the first one where someone buys something.

*Solution:* This is  $\text{Geometric}(\theta)$ . Since  $\mathbb{E}[X] = \frac{1-\theta}{\theta} = 10$ , we get  $1 - \theta = 10\theta$  so  $\theta = 1/11$ .

- c. The number  $Y$  of houses I try **including** the first one where someone buys something.

*Solution:* This is  $1 + \text{Geometric}(\theta)$ . Since  $\mathbb{E}[Y] = 1 + \frac{1-\theta}{\theta} = 10$ , we get  $1/\theta = 10$  so  $\theta = 1/10$ .

- d. The number  $Z$  of cars that pass in the first hour.

*Solution:* This is  $\text{Pois}(\lambda)$ , and  $\lambda = \mathbb{E}[Z] = 10$ .

# Chapter 5

## More than one random variable

Often we have more than one random quantity in an experiment. For example:

- the height and weight of a randomly sampled tree;
- the amount of precipitation and the height of a river;
- the outcomes of repeated trials in an experiment.

So far we have only introduced the machinery to deal with a single random variable for each experiment. In this chapter we will introduce some basic principles that enable us to deal with more than one random quantity at a time. In this module we only cover this for discrete random variables.

### 5.1 Joint probability mass functions

Recall that a random variable is simply a function from the sample space  $\Omega$  to the real numbers  $\mathbb{R}$ , mapping each elementary outcome  $\omega$  to a number. Formally, there is no reason not to define several such functions,  $X_1, X_2, \dots$  such that for each  $\omega$  we get a set of numbers  $X_1(\omega), X_2(\omega), \dots$

**Example 5.1.** Let  $\Omega$  be the set of all trees in a forest. An experiment consists of selecting a tree  $\omega$ . Let  $X_1$  report the height of a selected tree, and  $X_2$  report the weight of a selected tree. Then the reported numbers on carrying out an experiment are the measured height  $X_1(\omega)$  and weight  $X_2(\omega)$ .

We present some basic theory for discrete random variables. Let  $X$  and  $Y$  be discrete random variables defined on the same sample space  $\Omega$ . Their **joint probability mass function** is

$$p_{X,Y}(x, y) = \mathbb{P}(\{X = x\} \cap \{Y = y\}).$$

For simplicity we will assume our random variables only take non-negative integer values.

**Properties of  $p_{X,Y}(x, y)$ :**

1. For all  $x$  and  $y$ , we have  $0 \leq p_{X,Y}(x, y) \leq 1$ ,
2.  $\sum_{\text{all } x, y} p_{X,Y}(x, y) = 1$ ,
3.  $\mathbb{P}((X, Y) \in A) = \sum_{(x, y) \in A} p_{X,Y}(x, y)$ .

**Exercise 5.2.** The joint p.m.f. of  $X$  and  $Y$  is

$$p_{X,Y}(x, y) = (x + y)/18$$

for  $x, y = 0, 1, 2$ .

- Write out the joint probability table.
- Show this is a valid joint p.m.f.
- Evaluate (i)  $\mathbb{P}(X = 2)$ , (ii)  $\mathbb{P}(X = Y)$ , (iii)  $\mathbb{P}(X + Y \geq 2)$ .

**Solution.** (a)

		$y$		
		0	1	2
$x$	0	0	1/18	2/18
	1	1/18	2/18	3/18
	2	2/18	3/18	4/18

(b)  $p_{X,Y}(x, y) \geq 0$  for all  $x, y$ , and  $\sum_{\text{all } (x,y)} p_{X,Y}(x, y) = (0+1+2+1+2+3+2+3+4)/18 = 1$ .

(c) (i)  $\mathbb{P}(X = 2) = \mathbb{P}((X, Y) \in \{(2, 0), (2, 1), (2, 2)\}) = 2/18 + 3/18 + 4/18 = 9/18$ .

(ii)  $\mathbb{P}(X = Y) = \mathbb{P}((X, Y) \in \{(0, 0), (1, 1), (2, 2)\}) = 0/18 + 2/18 + 4/18 = 6/18$ .

(iii)  $\mathbb{P}(X + Y \geq 2) = \mathbb{P}((X, Y) \in \{(0, 2), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2)\}) = (2 + 2 + 3 + 2 + 3 + 4)/18 = 16/18$ .

Note that each of  $X$  and  $Y$  still have their own probability mass functions  $p_X$  and  $p_Y$ . In the context of jointly distributed random variables, these are called the **marginal probability mass functions**.

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}((X, Y) \in \{(x, 0), (x, 1), (x, 2), \dots\}) = \sum_{y=0}^{\infty} p_{X,Y}(x, y).$$

Similarly,

$$p_Y(y) = \sum_{x=0}^{\infty} p_{X,Y}(x, y).$$

**Exercise 5.3.** Bivariate random variables  $X$  and  $Y$  have joint p.m.f.

		$y$				$p_X(x)$
		0	1	2	3	
$x$	1	5/60	8/60	2/60	1/60	16/60
	2	12/60	7/60	3/60	2/60	24/60
	3	4/60	8/60	6/60	2/60	20/60
$p_Y(y)$		21/60	23/60	11/60	5/60	1

Fill in the marginal p.m.f.s in the final row and column.

## 5.2 Independence

Independence is the simplest form for **joint** behaviour of two (or more) random variables. Informally, two random variables  $X$  and  $Y$  are independent if knowing the value of one of them gives **no information** about the value of the other.

The outcomes of, say, rolls of two separate dice are independent in exactly this sense: knowing that the red die showed a particular value does not give us any information about the score of the blue die, or vice versa.

Two random variables  $X$  and  $Y$  are **independent** if the events  $\{X \in A\}$  and  $\{Y \in B\}$  are independent for all sets  $A$  and  $B$ , i.e.

$$\mathbb{P}(\{X \in A\} \cap \{Y \in B\}) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all sets  $A, B$ .

**Theorem 5.4.** Two discrete random variables  $X$  and  $Y$  with joint probability mass function  $p_{X,Y}$  are independent if and only if

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

for all  $x$  and  $y$ .

*Proof.* Let  $X$  and  $Y$  be independent, and let  $A = \{x\}$  and  $B = \{y\}$ . Then

$$\begin{aligned} p_{X,Y}(x, y) &= \mathbb{P}(\{X \in A\} \cap \{Y \in B\}) \\ &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \\ &= p_X(x)p_Y(y). \end{aligned}$$

Conversely, suppose  $p_{X,Y}(x, y) = p_X(x)p_Y(y)$  for all  $x, y$ . Then, for arbitrary sets  $A$  and  $B$ ,

$$\begin{aligned}
 \mathbb{P}(\{X \in A\} \cap \{Y \in B\}) &= \sum_{x \in A} \sum_{y \in B} p_{X,Y}(x, y) \\
 &= \sum_{x \in A} \sum_{y \in B} p_X(x)p_Y(y) \\
 &= \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y) \\
 &= \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \square
 \end{aligned}$$

If  $X$  and  $Y$  are discrete random variables, and  $p_Y(y) > 0$ , the **conditional probability mass functions** are defined as

$$\begin{aligned}
 p_{X|Y}(x | y) &= \frac{p_{X,Y}(x, y)}{p_Y(y)}, \\
 p_{Y|X}(y | x) &= \frac{p_{X,Y}(x, y)}{p_X(x)}.
 \end{aligned}$$

Thus  $p_{X|Y}(x | y) = \mathbb{P}(\{X = x\} \cap \{Y = y\}) / \mathbb{P}(Y = y) = \mathbb{P}(X = x | Y = y)$ .

**Exercise 5.5.** Show that if the discrete variables  $(X, Y)$  are independent then for all  $x, y$  we have

$$p_{X|Y}(x | y) = p_X(x).$$

**Solution.**

$$\begin{aligned}
 p_{X|Y}(x | y) &= \frac{p_{X,Y}(x, y)}{p_Y(y)} \\
 &= \frac{p_X(x)p_Y(y)}{p_Y(y)} \\
 &= p_X(x)
 \end{aligned}$$

These results conform with intuition as, when  $X$  and  $Y$  are independent, knowing the value of  $X$  should tell us nothing about  $Y$ .

The converse is also true: if the conditional distribution of  $X$  given  $Y = y$  does not depend on  $y$ , or equivalently, the conditional distribution of  $Y$  given  $X = x$  does not depend on  $x$ , then  $X$  and  $Y$  are independent.

**Exercise 5.6.** A fair coin is tossed. If it shows  $H$  a fair die is thrown, if  $T$  a biased die. The biased die makes even numbers twice as probable as odd numbers. Let  $X$  be the random variable with value 1 if the coin shows a head and 0 otherwise, and let  $Y$  be the score on the die. Find the joint p.m.f. of  $X$  and  $Y$ .

**Solution.** First, let's work out the probability distribution for the biased die. Each odd number has probability  $c$  and each even number has probability  $2c$  for some  $c$ . For this to be a probability distribution, we must have  $c + 2c + c + 2c + c + 2c = 1$ , so  $c = 1/9$ .

We know that the marginal p.m.f.  $p_X(x) = 1/2$  for  $x = 0, 1$ .

From the information given, we can work out the conditional p.m.f.

$x = 1$ :  $p_{Y|X}(y | 1) = 1/6$  for  $y = 1, 2, \dots, 6$ .

$x = 0$ :  $p_{Y|X}(y | 0) = 1/9$  for  $y = 1, 3, 5$  and  $p_{Y|X}(y | 0) = 2/9$  for  $y = 2, 4, 6$ .

Using  $p_{X,Y}(x, y) = p_{Y|X}(y | x) p_X(x)$  gives

	1	2	3	4	5	6
0	1/18	2/18	1/18	2/18	1/18	2/18
1	1/12	1/12	1/12	1/12	1/12	1/12

**Example 5.7.** For the joint p.m.f. in Example 5.3 obtain the conditional p.m.f. of  $X$  given  $Y = 2$ .

		Y			
		0	1	2	3
X	1			2/60	16/60
	2			3/60	24/60
	3			6/60	20/60
				11/60	

So

$$p_{X|Y}(x | 2) = p_{X,Y}(x, 2)/p_Y(2).$$

Thus  $x = 1$  with probability  $2/11$  and  $x = 2$  with probability  $3/11$  and  $x = 3$  with probability  $6/11$ .

We have seen that when  $X$  and  $Y$  are both discrete, they are independent if and only if their joint p.m.f. can be factorised as a product of the marginal p.m.f.s.

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

The result is needed for constructing **likelihood-based estimates** in statistics: often it is assumed that repeated experiments result in  $n$  independent observations of a random variable, and the joint density function of the observations is the product of the marginal densities.

## 5.3 The weak law of large numbers

Recall from Exercise 4.12 that if an experiment is repeated  $n$  times then, as  $n$  increases, the proportion of times an event  $A$  occurs converges to  $\mathbb{P}(A)$ . We will now prove a similar result concerning the average of several realisations of a random variable converging to the expected value. We start with a lemma which is proved in MATH230.

**Lemma 5.8.** Let  $X_1, X_2, \dots, X_n$  be jointly distributed random variables with finite expectation and variance. Then

- $\mathbb{E}(X_1 + X_2 + \dots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_n)$ , and
- if  $X_1, X_2, \dots, X_n$  are independent then

$$\text{Var}(X_1 + X_2 + \dots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n).$$

Now suppose that  $X_1, X_2, \dots, X_n$  are independent copies of a random variable  $X$ . For example, suppose we repeated an experiment  $n$  times, and  $X_i$  is the measured outcome on the  $i$ th experiment. This means that for each  $i$  we have

$$\begin{aligned}\mathbb{E}(X_i) &= \mathbb{E}(X) \\ \text{Var}(X_i) &= \text{Var}(X)\end{aligned}$$

If we want to report a value, scientists will usually measure it  $n$  times and report the average measured value. Let  $X_i$  be the measured value on the  $i$ th experiment. The average measured value is

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n).$$

Why do we do this?

Let's consider the properties of  $\bar{X}$ . For simplicity, write  $\mu$  for  $\mathbb{E}(X)$  and  $\sigma^2$  for  $\text{Var}(X)$ .

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) \\ &= \frac{1}{n}\mathbb{E}(X_1 + X_2 + \dots + X_n) && \text{by linearity of } \mathbb{E} \\ &= \frac{1}{n}\{\mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_n)\} && \text{by Lemma 5.8} \\ &= \frac{1}{n}\{\mathbb{E}(X) + \mathbb{E}(X) + \dots + \mathbb{E}(X)\} && \text{since } \mathbb{E}(X_i) = \mathbb{E}(X) \\ &= \frac{1}{n}\{n\mu\} \\ &= \mu\end{aligned}$$

So  $\bar{X}$  has expectation the quantity we wish to report, the true expected value of  $X$ . Of course, simply reporting the first measurement  $X_1$  would also have this expected value.

Consider now the variance of  $\bar{X}$ :

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) \\ &= \frac{1}{n^2}\text{Var}(X_1 + X_2 + \dots + X_n) && \text{by the calculation on p42} \\ &= \frac{1}{n^2}\{\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)\} && \text{by Lemma 5.8} \\ &= \frac{1}{n^2}\{n\sigma^2\} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

The variance of our reported quantity,  $\bar{X}$ , decreases as the number of measurements  $n$  increases.

We can use Chebychev's inequality (Section 3.6) to be more precise about this. Recall that for any random variable  $R$  with expected value  $m$  and standard deviation  $s$

$$\mathbb{P}(|R - m| > cs) \leq \frac{1}{c^2},$$

for any  $c > 0$ .

Hence for the random variable  $\bar{X}$  with expected value  $\mu$ , variance  $\sigma^2/n$  and hence standard deviation  $\sigma/\sqrt{n}$ , we have

$$\mathbb{P}(|\bar{X} - \mu| > \frac{c\sigma}{\sqrt{n}}) \leq \frac{1}{c^2}$$

By taking  $k = c/\sqrt{n}$ , we can rearrange this expression to

$$\mathbb{P}(|\bar{X} - \mu| > k\sigma) \leq \frac{1}{k^2n}.$$

We see that as  $n$  gets large, the probability that the sample average  $\bar{X}$  is more than distance  $k\sigma$  away from the expected value of the original random quantity  $X$  converges to 0.

Since  $k$  is arbitrary, in some sense we can say that  $\bar{X}$  **converges to  $\mu$** . This is called the **weak law of large numbers**. You will see various other forms of convergence of random variables in later courses.

One final thing to note: the standard deviation  $\sigma$  is exactly the right quantity for determining the appropriate scale for measuring distance here: the events are of the type “a random variable is more than  $k$  standard deviations away from the mean”.

## Chapter 6

# Continuous random variables

### 6.1 Introduction to continuous variables

Discrete random variables described the outcomes of experiments which were in a countable set (usually non-negative integer values). This covered models for the number of successes in a fixed number of trials, the number of floods of a river in a year, the number of children in a family until a girl is born.

Focusing only on discrete random variables is too restrictive for many situations. Examples include the nicotine levels in the blood plasma of smokers, the time intervals between floods of a river, and the waiting time for admissions to an intensive care unit. In each case the outcome of the experiment is a measurement on a continuous scale. This suggests we need to consider **continuous random variables**, that is variables whose set of possible values is uncountable (usually the real numbers, or real numbers in some fixed range).

To describe continuous random variables we need slightly different mathematical tools than we used for discrete random variables. For example, the discrete probability mass function  $p_X(x) = \mathbb{P}(X = x)$  does not work for a continuous random variable  $X$ , since  $\mathbb{P}(X = x) = 0$  for any specific value of  $x$ . We therefore focus on probabilities of events instead of probabilities of single outcomes. In particular we focus on events of the form

$$\{X \leq x\}$$

for fixed  $x$ , and consider these as  $x$  varies. For discrete random variables we defined (on page 34) the cumulative distribution function

$$F_X(x) = \mathbb{P}(X \leq x)$$

for all real numbers  $x$ .

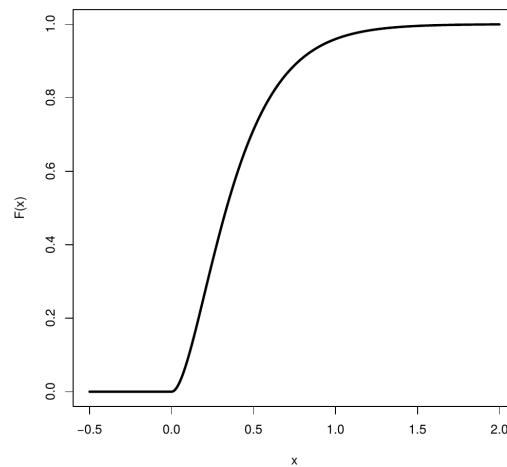
Recall that as  $x$  varies the function  $\mathbb{P}(X \leq x)$  jumps at the values that  $X$  could take, and the size of the jump is equal to the corresponding value of the p.m.f.

### 6.2 The cumulative distribution function

The **cumulative distribution function** (c.d.f.) of a (continuous or discrete) random variable,  $X$ , is defined, for all real values of  $x$ , by

$$F_X(x) = \mathbb{P}(X \leq x),$$

the probability that the random variable  $X$  takes a value less than or equal to  $x$ . When  $F_X$  is continuous, we have a **continuous random variable**. Below is an example of  $F_X$  for a continuous random variable.



### Properties of $F_X(x)$ :

1.  $0 \leq F_X(x) \leq 1$ , with  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  and  $\lim_{x \rightarrow \infty} F_X(x) = 1$ ,
2.  $F_X(x)$  is non-decreasing function of  $x$ . [Why?](#)

The distribution function is particularly useful for continuous random variables as we often want to know the probability of events that can be related by the laws of probability into probability statements about the event  $\{X \leq x\}$  for some  $x$ .

### Probabilities of intervals

Often the probability of the random variable  $X$  falling in the interval  $(a, b]$  is of interest for some real numbers  $a, b$  with  $a < b$ . This corresponds to the event  $\{a < X \leq b\}$ . By using the law of total probability  $\mathbb{P}(X \leq b) = \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b)$  so the probability of the interval event is

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a).$$

As  $\mathbb{P}(X = x) = 0$  for all  $x$ , for any continuous random variable  $X$

$$\mathbb{P}(X \leq x) = \mathbb{P}(X < x).$$

So  $\mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b)$ .

**Exercise 6.1.** Let  $X$  be a random variable with cumulative distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } 0 < x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

Obtain the following probabilities:

- a.  $\mathbb{P}(X \leq 0.5) = F_X(0.5) = 0.5$ ,
- b.  $\mathbb{P}(X > 0.5) = 1 - \mathbb{P}(X \leq 0.5) = 1 - F_X(0.5) = 0.5$ ,
- c.  $\mathbb{P}(X = 0.5) = 0$ ,
- d.  $\mathbb{P}(X < .9) = 0.9$ ,
- e.  $\mathbb{P}(0.5 < X \leq 0.9) = \mathbb{P}(X \leq .9) - \mathbb{P}(X \leq 0.5) = 0.9 - 0.5 = 0.4$ .

## 6.3 The probability density function

### Recap of definite integrals

We will frequently have to evaluate

$$F(x) := \int_{-\infty}^x f(s) \, ds.$$

For any particular value of  $x$  (e.g.  $x = 2$ ) this is the **definite integral** that you know. It is **NOT** the same as the **indefinite integral**, i.e.

$$F(x) \neq \int f(x) \, dx.$$

For example, suppose

$$f(x) = \begin{cases} 0 & \text{when } x < 1 \\ 1/x^2 & \text{otherwise.} \end{cases}$$

Then for  $x < 1$  we have  $F(x) = 0$ , and when  $x \geq 1$ ,

$$\int_{-\infty}^x f(s) \, ds = \int_1^x f(s) \, ds = \left[ -\frac{1}{s} \right]_1^x = 1 - 1/x.$$

Whereas, for  $x \geq 1$

$$\int f(x) \, dx = -1/x + c,$$

which gives a whole family of functions (including the one we want) depending on  $c$ .

**One of the most frequent single mistakes by students encountering continuous random variables for the first time is evaluating an indefinite integral with  $c = 0$  when they should have been evaluating a definite integral.**

### The probability density function

Recall (page 34) that the c.d.f.,  $F_R(r)$ , of a discrete random variable,  $R$ , is the sum of its p.m.f. up to the value  $r$ , i.e.  $\sum_{i=-\infty}^r p_R(i)$ .

Analogously the c.d.f.  $F_X(x)$  of a continuous random variable should correspond to an **integral**. We therefore define the **probability density function**, or p.d.f.,  $f_X(x)$ , of a continuous random variable  $X$ , to be

$$f_X(x) = \frac{d}{dx} F_X(x),$$

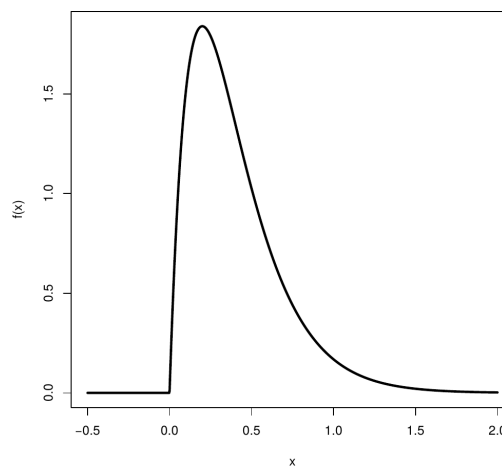
so that it satisfies

$$F_X(x) = \int_{-\infty}^x f_X(s) ds.$$

For example

$$P(X \leq 10) = F_X(10) = \int_{-\infty}^{10} f_X(s) ds.$$

Note that  $s$  is a **dummy variable**; one could use any letter for the integrand except  $x$ ,  $d$  or  $f$ . The following is the p.d.f. corresponding to the c.d.f. we saw earlier.



Notice that the p.d.f. is zero in regions where there are no outcomes, in this example for  $x \leq 0$ . Note also that it exceeds 1 in some places, so it cannot be interpreted as a probability despite having some properties (which we shall explore in what follows) that are very similar to those of the probability mass function.

### Properties of $f_X(x)$

1. Positivity:  $f_X(x) \geq 0$  for all  $x$ , [Why?](#)
2. Unit-integrability:  $\int_{-\infty}^{\infty} f_X(x) dx = 1$ . [Why?](#)

[Note that, unlike the c.d.f. of a continuous random variable, the p.d.f. is \*\*not\*\* necessarily continuous.](#)

What is the probability that an observation on a continuous random variable  $X$  lies in the interval  $(a, b]$ ?

$$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a) \\ &= \int_{-\infty}^b f_X(x) dx - \int_{-\infty}^a f_X(x) dx = \int_a^b f_X(x) dx. \end{aligned}$$

We see that  $\mathbb{P}(a < X \leq b)$  may be calculated as the area under the curve of  $f_X(x)$  between  $x = a$  and  $x = b$ .

### An illuminating idea

For some **very small** interval width  $\delta$ ,

$$\begin{aligned}\mathbb{P}(x < X \leq x + \delta) &= \int_x^{x+\delta} f_X(s) \, ds \\ &\approx f_X(x)\delta.\end{aligned}$$

Thus  $f_X(x)\delta$  can be thought of as (approximately) the probability that  $X$  is between  $x$  and  $x + \delta$ .

Compare this with the meaning of the p.m.f.  $p_R(r)$ .

**Example 6.2.** A random variable  $X$  has cumulative distribution function

$$F_X(x) = \begin{cases} 1 - \exp(-3x) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

Find the p.d.f. of  $X$ .

**Solution.**

$$f_X(x) = \frac{d}{dx} F_X(x) = \begin{cases} 3 \exp(-3x) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

**Exercise 6.3.** A triangular p.d.f.: a random variable  $X$  has p.d.f.

$$f_X(x) = \begin{cases} 2(1-x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Obtain the c.d.f.  $F_X(x)$ .

**Solution.** We **split the range** of  $x$ ,  $(-\infty, \infty)$  into sensible intervals.

- For  $x \in (-\infty, 0]$ ,

$$F_X(x) = \int_{-\infty}^x f_X(s) \, ds = \int_{-\infty}^x 0 \, ds = 0.$$

- For  $x \in (0, 1]$ ,

$$\begin{aligned}F_X(x) &= \int_{-\infty}^x f_X(s) \, ds \\ &= F_X(0) + \int_0^x 2(1-s) \, ds \\ &= 0 + [2s - s^2]_0^x \\ &= 2x - x^2\end{aligned}$$

- For  $x \in (1, \infty)$ ,

$$F_X(x) = \int_{-\infty}^x f_X(s) \, ds = F_X(1) + \int_1^x 0 \, ds = 1.$$

Hence

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 2x - x^2 & \text{for } 0 < x \leq 1 \\ 1 & \text{for } x > 1. \end{cases}$$

Now find  $\mathbb{P}(X < 0.5)$ ,  $\mathbb{P}(0.5 < X < 0.8)$  and  $\mathbb{P}(0.5 < X < 1.75)$ .

$$F(0.5) = 2(0.5) - (0.5)^2 = 0.75$$

$$F(0.8) - F(0.5) = 2(0.8) - (0.8)^2 - 0.75 = 1.6 - 0.64 - 0.75 = 0.21$$

$$F(1.75) - F(0.5) = 1 - 0.75 = 0.25.$$

Obtain  $\mathbb{P}(0.5 < X < 0.8)$  directly from the p.d.f.

$$\begin{aligned} \mathbb{P}(0.5 < X < 0.8) &= \int_{0.5}^{0.8} f_X(s) \, ds = \int_{0.5}^{0.8} 2(1-s) \, ds = [2s - s^2]_{0.5}^{0.8} \\ &= 2(0.8) - (0.8)^2 - 2(0.5) + (0.5)^2 = 0.21 \end{aligned}$$

**Example 6.4.** The lifetime, in years, that a computer functions before breaking down is a continuous random variable  $X$  with p.d.f. given by

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

The parameter  $\lambda$  depends on the type of computer. To set a time for a guarantee the company wants to know the time  $t$  for which with probability 0.9 the lifetime of the computer will exceed  $t$ .

**Solution.**

$$\begin{aligned} \mathbb{P}(X > t) &= \int_t^{\infty} f_X(x) \, dx \\ &= \int_t^{\infty} \lambda \exp(-\lambda x) \, dx \\ &= [-\exp(-\lambda x)]_t^{\infty} = \exp(-\lambda t), \end{aligned}$$

so that  $\mathbb{P}(X > t) = 0.9$  implies  $t = -\lambda^{-1} \log(0.9)$ .

The quantity that we have just found is called a quantile of the distribution; see Section 6.5.

## 6.4 Expectation and variance

All the information about the distribution of a continuous random variable  $X$  is contained in either the c.d.f.  $F_X(x)$  or the p.d.f.  $f_X(x)$ . However it is often helpful to summarise the main characteristics of the distribution in terms of a few values, analogously to the discrete case.

The expected value of a continuous random variable  $X$  can be thought of as the average of the different values that  $X$  may take, weighted according to their chance of occurrence.

The **expected value** of a continuous random variable  $X$  is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

This is very similar to the definition of expectation of a discrete random variable.

The differences are that the sum has become an integral, and the p.m.f. has become the p.d.f.

Similarly one can show that the expected value of a real-valued function  $g(X)$  of a continuous random variable  $X$  is

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

For example,

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx.$$

**Exercise 6.5.** For a random variable  $X$  with p.d.f. given by

$$f_X(x) = \begin{cases} 2x & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

find  $\mathbb{E}[X]$ ,  $\mathbb{E}[2X]$  and  $\mathbb{E}[X^2]$ .

**Solution.**

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \left[ \int_{-\infty}^0 + \int_0^1 + \int_1^{\infty} \right] x f_X(x) dx \\ &= 0 + \int_0^1 x 2x dx + 0 \\ &= \left[ \frac{2}{3} x^3 \right]_0^1 = \frac{2}{3}. \end{aligned}$$

Also

$$\begin{aligned} \mathbb{E}(2X) &= \int_0^1 2x 2x dx = \frac{4}{3} = 2\mathbb{E}(X) \\ \mathbb{E}(X^2) &= \int_0^1 x^2 2x dx = \left[ \frac{1}{2} x^4 \right]_0^1 = \frac{1}{2}, \end{aligned}$$

using the same method to split the range.

For continuous variables expectation has the same linearity properties that we found in Chapter 3 for discrete random variables.

For arbitrary functions  $g$  and  $h$  and constant  $c$ ,

$$\begin{aligned}\mathbb{E}[g(X) + h(X)] &= \mathbb{E}[g(X)] + \mathbb{E}[h(X)]; \\ \mathbb{E}[cg(X)] &= c\mathbb{E}[g(X)]; \\ \mathbb{E}[g(X) + c] &= \mathbb{E}[g(X)] + c.\end{aligned}$$

Derivation (of first property):

$$\begin{aligned}\mathbb{E}[g(X) + h(X)] &= \int_{-\infty}^{\infty} [g(x) + h(x)]f_X(x) \, dx \\ &= \int_{-\infty}^{\infty} g(x)f_X(x) \, dx + \int_{-\infty}^{\infty} h(x)f_X(x) \, dx \\ &= \mathbb{E}[g(X)] + \mathbb{E}[h(X)]\end{aligned}$$

The other properties may be shown similarly.

We can now define the variance in exactly the same way as for discrete random variables. Again, the standard deviation is the square root of this.

The **variance**, measuring of the spread or dispersion of a random variable about the expectation, for a continuous distribution is

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

As with discrete random variables, the easiest way to evaluate the variance is usually

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Note that the calculation of this formula on page 41 uses only the linearity properties of expectation given above, and so is just as valid for continuous random variables as it was for discrete random variables.

**Exercise 6.6.** For a random variable  $X$  with p.d.f. given by

$$f_X(x) = \begin{cases} 2x & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

find  $\text{Var}(X)$ . You might well want to use some results from Exercise 6.5.

**Solution.** From the previous exercise we know that  $\mathbb{E}[X] = \frac{2}{3}$  and  $\mathbb{E}[X^2] = \frac{1}{2}$ . Therefore

$$\begin{aligned}\text{Var } X &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \frac{1}{2} - \frac{4}{9} = \frac{1}{18}.\end{aligned}$$

**Warning:** Sometimes the expectation or variance do not exist. This occurs when the chance of obtaining very large values is too big.

If the expectation does not exist, the variance can't even be defined, but it is possible that the expectation exists but the variance does not.

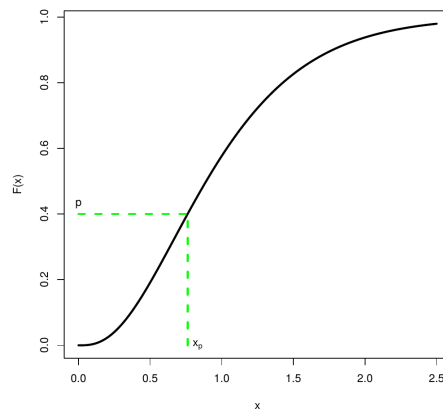
## 6.5 Quantiles

The c.d.f. tells us the probability that a continuous random variable does not exceed a specified value. Often we are interested in the reverse question: we specify a probability and want to know the value which is not exceeded with that probability.

Such values are termed **quantiles** with  $x_p$  the  $100p\%$  quantile defined by

$$F_X(x_p) = p.$$

They can be read off from the graph of the c.d.f.: the graph below shows the 40% quantile. (This is sometimes referred to as the “40th centile”.)



Certain quantiles are of special interest:

**Median:** the median is the middle of the distribution in the sense that the value of the random variable is equally likely to fall on either side of this value. The median is the 50% quantile,  $x_{0.5}$ , so that  $F(x_{0.5}) = 0.5$ . As a measure of location, the median has the advantage over the expectation of existing for all distributions. It also is less sensitive to small changes in the distribution of very large values. For this reason, the “average” income given in news stories almost always uses the median.

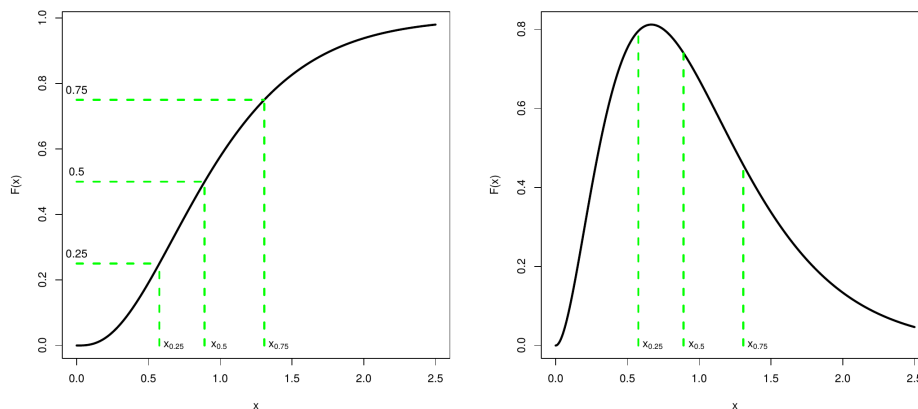
**Quartiles:** the quartiles split the distribution into four equally likely regions,  $x_{0.25}$  the lower quartile,  $x_{0.5}$  the median and  $x_{0.75}$  the upper quartile.

$$\mathbb{P}(X < x_{0.25}) = \mathbb{P}(x_{0.25} < X < x_{0.5}) = \mathbb{P}(x_{0.5} < X < x_{0.75}) = \mathbb{P}(X > x_{0.75}) = 0.25.$$

This is illustrated below for the c.d.f. and p.d.f. of the same random variable.

**Inter-quartile range:** the difference in values of quartiles provides a measure of the variability of a random variable (measured in the units of the variable) that does not require the evaluation of the standard deviation (which can be infinite). The inter-quartile range is

$$x_{0.75} - x_{0.25}.$$



**Example 6.7.** A possible model for the claim sizes received by an insurance company is a random variable with c.d.f.

$$F_X(x) = 1 - \exp(-\lambda x),$$

for some  $\lambda > 0$ . The company is legally obliged to pay the smallest 99% of claims without requiring re-insurance support. What claim size must the company be able to pay without re-insurance support?

**Solution.** The company must pay without support when a claim  $X$  is less than  $x_{0.99}$ , where

$$0.99 = \mathbb{P}(X \leq x_{0.99}) = F_X(x_{0.99})$$

$$1 - \exp(-\lambda x_{0.99}) = 0.99$$

$$\exp(-\lambda x_{0.99}) = 1 - 0.99$$

$$-\lambda x_{0.99} = \log(1 - 0.99)$$

$$x_{0.99} = -\frac{1}{\lambda} \log(0.01).$$

**Exercise 6.8.** It is considered suitable to model the annual maximum sea level by an extreme value distribution

$$F_X(x) = \exp[-\exp\{-(x - \alpha)/\beta\}],$$

for  $\beta > 0$ . The sea flood defence needs to be built to withstand a flood of the size which occurs in any year with probability 0.01 (i.e. once on average every 100 years). Evaluate the required height of the flood defence.

**Solution.** We seek  $x$  such that  $\mathbb{P}(X > x) = 0.01$ .

Equivalently  $F_X(x) = 0.99$

We thus solve

$$\exp[-\exp\{-(x - \alpha)/\beta\}] = 0.99$$

$$-(x - \alpha)/\beta = \log[-\log\{0.99\}]$$

Therefore

$$x = \alpha - \beta \log\{-\log(0.99)\}.$$

## 6.6 Transformations of random variables

Suppose we have a continuous random variable  $X$ , but actually want to know about a random variable  $Y = f(X)$  for some function  $f$  that is increasing (on the induced sample space).

For a discrete random variable, it is straightforward to write down the p.m.f. of  $Y$  based on the p.m.f. of  $X$ . For continuous random variables, we can't go directly from the p.d.f. of  $X$  to the p.d.f. of  $Y$ , but need to make use of the c.d.f.

**Example 6.9.** Suppose  $X$  is a random variable with p.d.f.  $f_X(x) = 1$  for  $0 \leq x \leq 1$  (and 0 otherwise). What is the p.d.f. of  $Y = \sqrt{X}$ ?

**Solution.** We first work out the c.d.f. of  $X$ .

$$F_X(x) = \int_{-\infty}^x f_X(z) dz = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases}$$

Next we find the c.d.f. of  $Y$ .

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq y^2) = F_X(y^2), \text{ so}$$

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ y^2 & \text{if } 0 \leq y \leq 1 \\ 1 & \text{if } y > 1. \end{cases}$$

Finally, we differentiate to get the p.d.f.

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} 2y & \text{if } 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

We can follow a similar process for more general increasing functions  $f$ . If  $f$  is decreasing, we have to consider the complementary event in the middle step.

**Example 6.10.** What is the p.d.f. of  $Z = 1/X$ ?

**Solution.** Note that we always have  $Z > 1$ , and for any  $z > 1$  we have  $\mathbb{P}(Z \leq z) = \mathbb{P}(X \geq 1/z) = 1 - F_X(1/z)$ , so

$$F_Z(z) = \begin{cases} 0 & \text{if } z < 1 \\ 1 - 1/z & \text{if } z \geq 1, \end{cases}$$

which gives

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \begin{cases} z^{-2} & \text{if } z \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

## 6.7 Jointly distributed continuous random variables

For two continuous random variables we can define independence in the same way as for discrete random variables.

Two random variables  $X$  and  $Y$  are **independent** if the events  $\{X \in A\}$  and  $\{Y \in B\}$  are independent for all sets  $A$  and  $B$ , i.e.

$$\mathbb{P}(\{X \in A\} \cap \{Y \in B\}) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all sets  $A, B$ .

It is beyond the scope of this module, but it is also possible to define a joint probability density function  $f_{X,Y}(x, y)$  for jointly distributed continuous random variables  $X$  and  $Y$ . Then we have the following equivalent of Theorem 5.4.

**Theorem 6.11.** Two continuous random variables  $X$  and  $Y$  are independent if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

for almost all  $x$  and  $y$ .

Everything we did in Section 5.3 to prove the weak law of large numbers for discrete random variables also works for continuous random variables.

## Chapter 7

# Models for continuous random variables

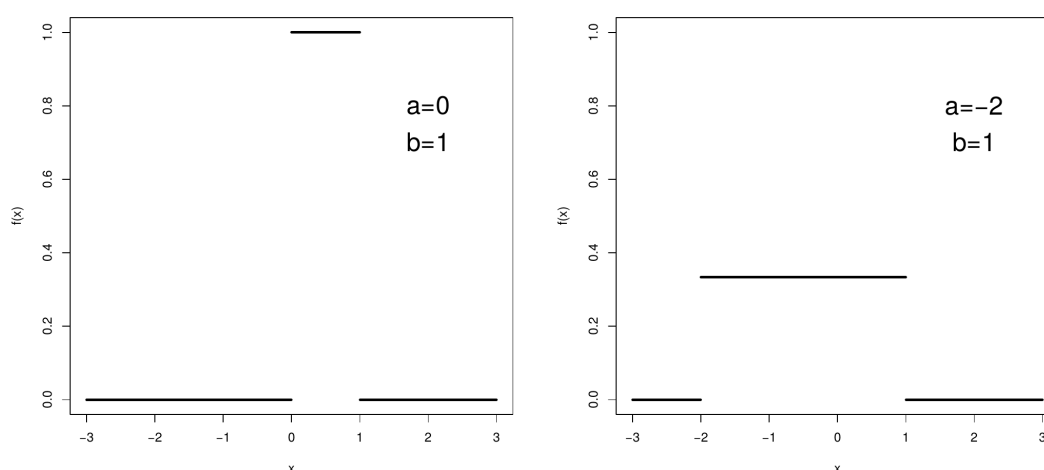
As with discrete random variables, there are a number of “standard” continuous random variables. In this chapter we give the details of some of the most important for subsequent study of probability and statistics.

### 7.1 The uniform distribution

A continuous random variable with a fixed range, for which all outcomes in that range have equal chance of occurring is said to be uniformly distributed. Specifically, a random variable  $X$  has a **uniform distribution** over the interval  $[a, b]$  if the p.d.f. is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

We write  $X \sim U(a, b)$ . This p.d.f. for two different sets of parameter values is illustrated below.



We find that for all  $x$  and  $x + c$  such that  $a \leq x < x + c \leq b$

$$\mathbb{P}(x < X \leq x + c) = c/(b - a),$$

so the probability of  $X$  falling in any interval of length  $c$  in the range  $(a, b)$  is the same for all  $x$ , i.e. independent of the position  $x$  and proportional to the interval length  $c$ .

Possible examples of uniform random variables are: completely random numbers between 0 and 1, the times of births in a 24 hour period, and the times of goals in a football match.

**Exercise 7.1.** Find the c.d.f., expected value and variance of the  $U(a, b)$  distribution.

**Solution.** The c.d.f. is given by

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(s) \, ds \\ &= \begin{cases} \int_{-\infty}^x 0 \, ds & \text{if } x \leq a, \\ \int_{-\infty}^a 0 \, ds + \int_a^x \frac{1}{b-a} \, ds & \text{if } a < x \leq b, \\ \int_{-\infty}^a 0 \, ds + \int_a^b \frac{1}{b-a} \, ds + \int_b^x 0 \, ds & \text{if } b < x, \end{cases} \\ &= \begin{cases} 0 & \text{if } x \leq a, \\ \frac{x-a}{b-a} & \text{if } a < x \leq b, \\ 1 & \text{if } b < x. \end{cases} \end{aligned}$$

The expectation is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_a^b \frac{x}{b-a} \, dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

To calculate the variance, we first calculate

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx = \int_a^b \frac{x^2}{b-a} \, dx = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}.$$

So the variance is

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{b^2 + ab + a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

These results seem logical as if all values in the interval  $(a, b)$  are equally likely then the expected value should be in the middle. Similarly the wider the interval, the more variable the outcomes, hence the larger the variance should be.

**Exercise 7.2.** Find the upper quartile of the random variable  $X \sim U(3, 5)$ .

**Solution.** Note that  $x_{0.75}$  lies in  $(3, 5)$ .

$$\begin{aligned} \mathbb{P}(X < x_{0.75}) &= 0.75 \\ \frac{x_{0.75} - 3}{5 - 3} &= 0.75 \\ x_{0.75} &= 4.5. \end{aligned}$$

**Exercise 7.3.** If  $X \sim U(0, 10)$ , use R to calculate the probability that (a)  $X < 3$ , (b)  $X > 6$ , (c)  $3 < X < 8$ , (d)  $8 < X < 10$ , (e)  $8 < X < 13$ .

```
punif(3,min=0,max=10)
1-punif(6,0,10)
punif(8,0,10)-punif(3,0,10)
punif(10,0,10)-punif(8,0,10)
punif(13,0,10)-punif(8,0,10)
```

**Example 7.4.** Suppose that you know the score in a football game was 1–0. You watch a recording. Assuming a uniform distribution for the time of goals (and no extra time, so the match lasts 90 minutes), what is the expected time until the goal? What is the probability it is in the first half?

**Solution.** Let  $X$  model the time to the goal.

If  $X \sim U(0, 90)$  then  $\mathbb{E}[X] = 45$  mins and  $F(45) = 0.5$ .

## 7.2 The exponential distribution

A random variable  $X$  has an **exponential distribution** with rate  $\beta$  if its p.d.f. is given by

$$f_X(x) = \begin{cases} C \exp(-\beta x) & x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where  $\beta > 0$  and  $C$  is a **normalising constant**. We write  $X \sim \text{Exp}(\beta)$ .

**Exercise 7.5.** For which value of  $C$  is this a valid p.d.f. (i.e. is non-negative and integrates to 1)?

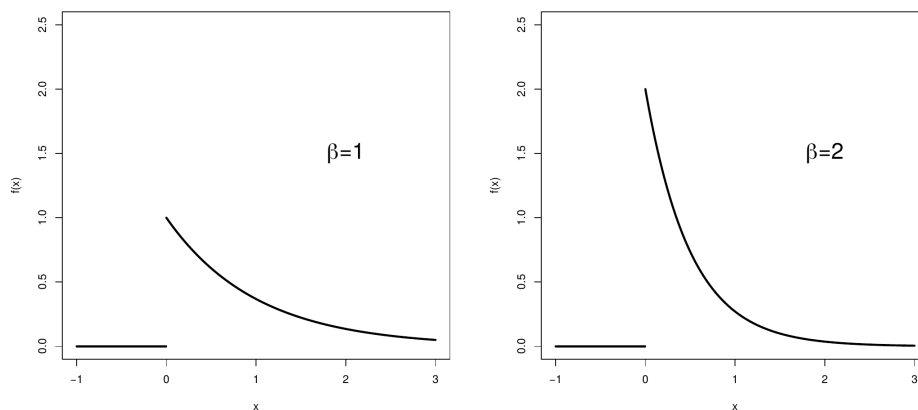
**Solution.** For  $f_X(x)$  to be non-negative, we need  $C \geq 0$ .

We also need

$$1 = \int_{-\infty}^{\infty} f_X(x) dx = \int_0^{\infty} C \exp(-\beta x) dx = \left[ \frac{C}{-\beta} \exp(-\beta x) \right]_0^{\infty} = \frac{C}{\beta}.$$

This means that  $C = \beta$ .

The p.d.f.s for two different values of  $\beta$  are shown below.



**Example 7.6.** Find the c.d.f. of the  $\text{Exp}(\beta)$  distribution.

**Solution.**

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(s) \, ds \\ &= \begin{cases} \int_{-\infty}^x 0 \, ds & \text{if } x \leq 0 \\ \int_{-\infty}^x 0 \, ds + \int_0^x \beta \exp(-\beta s) \, ds & \text{if } x > 0 \end{cases} \\ &= \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - \exp(-\beta x) & \text{if } x > 0. \end{cases} \end{aligned}$$

**Example 7.7.** The exponential distribution is often used to model waiting times. Suppose that the length of a phone call to a company, in minutes, is distributed as  $\text{Exp}(1/10)$ . If you phone and the number is busy then you are put on hold until the operator is free. If someone phones the company immediately before you, find the probability that you will have to wait (a) less than 10 minutes, (b) between 10 and 20 minutes, and (c) between 10 and 20 minutes once you have already waited for 10 minutes.

**Solution.** Let  $X$  be the time you have to wait, which equals the length of the call.  $X \sim \text{Exp}(1/10)$ .

(a)  $\mathbb{P}(X < 10) = F_X(10) = 1 - \exp(-\frac{1}{10} \times 10) = 1 - e^{-1}$ .

(b)  $\mathbb{P}(10 < X < 20) = F_X(20) - F_X(10) = e^{-1} - e^{-2}$ .

(c) We calculate as follows.

$$\begin{aligned} \mathbb{P}(10 < X < 20 \mid X > 10) &= \mathbb{P}(10 < X < 20 \cap X > 10) / \mathbb{P}(X > 10) \\ &= \mathbb{P}(10 < X < 20) / \mathbb{P}(X > 10) \\ &= \frac{e^{-1} - e^{-2}}{e^{-1}} = 1 - e^{-1}. \end{aligned}$$

## Lack of memory

A key property of the exponential distribution is its lack of memory property.

A random variable satisfies the **lack of memory property** if

$$\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s) \quad \text{for all } s, t > 0.$$

That is, the conditional probability that a variable exceeds  $s + t$ , given that it exceeds  $t$ , is independent of  $t$ .

If we interpret  $X$  as a waiting time to an event, this means that the probability that you have to wait a further time  $s$  is independent of how long you have waited already.

To show this result holds for  $X \sim \text{Exp}(\beta)$  recall that  $\mathbb{P}(X > x) = \exp(-\beta x)$  for all  $x > 0$ . Hence, for  $s > 0, t > 0$

$$\begin{aligned}\mathbb{P}(X > s + t \mid X > t) &= \frac{\mathbb{P}(\{X > s + t\} \cap \{X > t\})}{\mathbb{P}(X > t)} \\ &= \frac{\mathbb{P}(\{X > s + t\})}{\mathbb{P}(X > t)} \\ &= \frac{\exp\{-\beta(s + t)\}}{\exp(-\beta t)} = \exp(-\beta s) \\ &= \mathbb{P}(X > s).\end{aligned}$$

Note that we already observed this phenomenon in Example 7.7

Actually, the exponential distribution is the **only** continuous distribution with the lack of memory property. We will not prove that here.

## The Gamma function

The integral required to obtain the expected value and variance of an exponential random variable will occur several times in this chapter. We first define a slightly simplified form, which occurs in many areas of mathematics, and discover several of its properties.

The **Gamma function**,  $\Gamma(\alpha)$ , is defined as

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} \exp(-t) dt$$

Firstly we note that

$$\Gamma(1) = \int_0^{\infty} \exp(-t) dt = \left[ -\exp(-t) \right]_0^{\infty} = 1.$$

**Lemma 7.8.** For  $\alpha > 0$ ,

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha).$$

*Proof.* We use integration by parts:

$$\begin{aligned}\Gamma(\alpha + 1) &= \int_0^{\infty} t^{\alpha} \exp(-t) dt \\ &= \left[ t^{\alpha}(-1) \exp(-t) \right]_0^{\infty} + \int_0^{\infty} \alpha t^{\alpha-1} \exp(-t) dt \\ &= 0 - 0 + \alpha \Gamma(\alpha) \text{ for } \alpha > 0.\end{aligned}$$

□

Since  $\Gamma(1) = 1$ , we have that  $\Gamma(2) = 1$ ,  $\Gamma(3) = 2\Gamma(2) = 2$ ,  $\Gamma(4) = 3\Gamma(3) = 6, \dots$ . By induction we can easily see that for  $n$  a positive integer  $\Gamma(n) = (n - 1)!$ .

## Expectation and variance

The  **$r$ th moment** of a general random variable  $X$  is defined to be  $\mathbb{E}[X^r]$ . In the case of exponential random variables, we have that

$$\begin{aligned}\mathbb{E}[X^r] &= \int_{-\infty}^{\infty} x^r f_X(x) \, dx \\ &= \int_0^{\infty} x^r \beta \exp(-\beta x) \, dx \\ &= \beta \int_0^{\infty} x^r \exp(-\beta x) \, dx.\end{aligned}$$

We will need to evaluate integrals of this form many times, so the following will be useful.

**Lemma 7.9.**

$$\int_0^{\infty} x^{\alpha-1} \exp(-\beta x) \, dx = \frac{\Gamma(\alpha)}{\beta^\alpha}.$$

*Proof.* Substituting  $t = \beta x$  gives

$$\begin{aligned}\int_0^{\infty} x^{\alpha-1} \exp(-\beta x) \, dx &= \int_0^{\infty} \left(\frac{t}{\beta}\right)^{\alpha-1} \exp(-t) \frac{dt}{\beta} \\ &= \frac{1}{\beta^\alpha} \int_0^{\infty} t^{\alpha-1} \exp(-t) \, dt \\ &= \frac{\Gamma(\alpha)}{\beta^\alpha}.\end{aligned}$$

□

Using Lemma 7.9 with  $\alpha = r + 1$  we see that

$$\begin{aligned}\mathbb{E}[X^r] &= \beta \int_0^{\infty} x^r e^{-\beta x} \, dx \\ &= \beta \frac{\Gamma(r+1)}{\beta^{r+1}} \\ &= \frac{\Gamma(r+1)}{\beta^r}.\end{aligned}$$

For integer values of  $r$ , therefore,  $\mathbb{E}[X^r] = r!/\beta^r$ .

In particular the expectation and variance of an exponential random variable are

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{\beta} \\ \text{Var}(X) &= \frac{2}{\beta^2} - \frac{1}{\beta^2} = \frac{1}{\beta^2}.\end{aligned}$$

Hence the expectation and standard deviation are the same. Note that the expectation decreases with  $\beta$ ;  $\beta$  is the rate at which events occur, so the higher the rate of events the shorter the expected waiting time to the next event.

**Beware:** Some computer packages use a different parameterisation of exponential random variables, in which an  $\text{Exp}(m)$  random variable has expected value  $m$ .

## 7.3 The gamma distribution

A random variable  $X$  has a **gamma distribution** with **shape** parameter  $\alpha$  and **rate** parameter  $\beta$  if its p.d.f. is given by

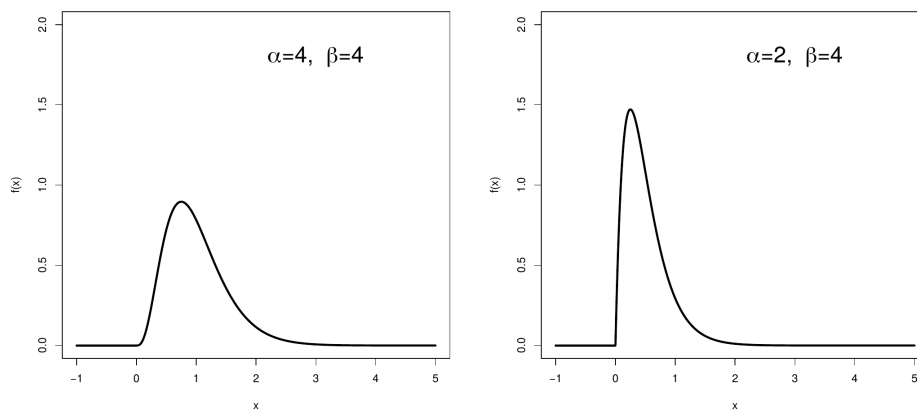
$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) & x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

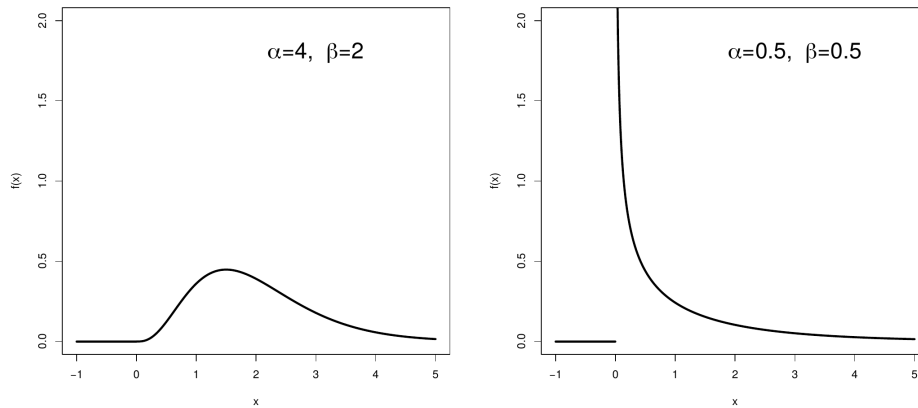
where  $\alpha > 0$  and  $\beta > 0$ . We write  $X \sim \text{Gamma}(\alpha, \beta)$ .

Lemma 7.9 shows directly that

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x) dx &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{\infty} x^{\alpha-1} \exp(-\beta x) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha)}{\beta^\alpha} \\ &= 1. \end{aligned}$$

Plotted below are the p.d.f.s of four different gamma distributions.





The family of gamma distributions provides a flexible class of p.d.f.s which may describe the distribution of a non-negative variable even when there is no strong model-based justification.

Note that when  $\alpha = 1$  the gamma distribution reduces to the exponential distribution. However, unlike the exponential distribution we cannot evaluate the c.d.f. in closed form for a general (non-integer) value of  $\alpha$ .

The statistical package R has built-in functions for evaluating the p.d.f., c.d.f. and inverse c.d.f. (for obtaining quantiles) for many common distributions including the gamma distribution.

```
> dgamma(4,shape=6,rate=1) # p.d.f. of Gamma(6,1) evaluated at x=4, i.e. f(4)
[1] 0.1562935
> dgamma(4,6,1)           # p.d.f. of Gamma(6,1) evaluated at x=4, i.e. f(4)
[1] 0.1562935
> pgamma(2,0.5,1)         # c.d.f. of Gamma(0.5,1) evaluated at x=2, i.e. P(X<2)
[1] 0.9544997
> qgamma(0.5,3,1)         # the median of the Gamma(3,1) distribution
[1] 2.67406
```

The  $r$ th moment of a gamma random variable is

$$\begin{aligned}
 \mathbb{E}[X^r] &= \int_{-\infty}^{\infty} x^r f_X(x) dx = \int_0^{\infty} x^r \beta^\alpha x^{\alpha-1} \exp(-\beta x) / \Gamma(\alpha) dx \\
 &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^{\infty} x^{r+\alpha-1} \exp(-\beta x) dx \\
 &= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(r+\alpha)}{\beta^{r+\alpha}} \\
 &= \frac{\Gamma(r+\alpha)}{\beta^r \Gamma(\alpha)}.
 \end{aligned}$$

where the penultimate line follows from Lemma 7.9.

An alternative to remembering or rederiving Lemma 7.9 is to use the **unit integrability** property of

the density (this trick can be useful for densities other than the gamma).

$$\begin{aligned}
 \mathbb{E}[X^r] &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{r+\alpha-1} \exp(-\beta x) \, dx \\
 &= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \frac{\Gamma(\alpha+r)}{\beta^{\alpha+r}} \times \int_0^\infty \frac{\beta^{\alpha+r}}{\Gamma(\alpha+r)} x^{r+\alpha-1} \exp(-\beta x) \, dx \\
 &= \frac{\Gamma(\alpha+r)}{\beta^r \Gamma(\alpha)} \times 1.
 \end{aligned}$$

Both approaches result in the same formula for  $\mathbb{E}[X^r]$ . We can now derive the expectation and variance.

$$\begin{aligned}
 \mathbb{E}[X] &= \frac{\Gamma(\alpha+1)}{\beta \Gamma(\alpha)} \\
 &= \frac{\alpha}{\beta}, \\
 \mathbb{E}[X^2] &= \frac{\Gamma(\alpha+2)}{\beta^2 \Gamma(\alpha)} \\
 &= \frac{(\alpha+1)\alpha}{\beta^2}, \\
 \text{Var}(X) &= \frac{(\alpha+1)\alpha}{\beta^2} - \frac{\alpha^2}{\beta^2} \\
 &= \frac{\alpha}{\beta^2}.
 \end{aligned}$$

## 7.4 The normal distribution

A random variable  $X$  has a **normal distribution**, also known as a **Gaussian distribution**, with parameters  $\mu$  and  $\sigma^2$  if its p.d.f. is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

for  $x \in \mathbb{R}$ . We write  $X \sim N(\mu, \sigma^2)$ .

This is the most commonly used continuous random variable, cropping up all over the place, for good theoretical reasons. We will not study it in detail here; this will be done in MATH104.