Probability

Vittoria Silvestri

Department of Mathematics, University of Rome La Sapienza.

Contents

Preface

These lecture notes are for the course Probability at ACSAI, University of Rome La Sapienza, given in Fall 2022 jointly with Prof. Mauro Piccioni. The content is closely based on the following lecture notes, all available online:

- James Norris: http://www.statslab.cam.ac.uk/∼james/Lectures/p.pdf
- Douglas Kennedy: http://trin-hosts.trin.cam.ac.uk/fellows/dpk10/IA/IAprob.html
- Richard Weber: http://www.statslab.cam.ac.uk/∼rrw1/stats/Sa4.pdf

Please notify silvestri@mat.uniroma1.it for comments and corrections.

Additional suggested readings:

- D. Bertsekas, J. Tsitsiklis: *Introduction to probability* (2008) Athena Scientific (available online at MIT Open Course Ware).
- J. Blitzstein, J. Hwang, *Introduction to probability*, Taylor and Francis (available online at Stat 110: Probability).
- Sheldon M Ross: *A first course in Probability* (2019) Boston, MA: Pearson.

Note: the parts marked with (*) are optional, and can be skipped with no harm.

CHAPTER 1

Discrete probability

1. Introduction

This course concerns the study of experiments with *random* outcomes, such as throwing a die, tossing a coin or blindly drawing a card from a deck. Say that the set of possible outcomes is

$$\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}.$$

We call Ω *sample space*, while its elements are called *outcomes*. A subset $A \subseteq \Omega$ is called an *event*.

Example 1.1 (Tossing a coin). Toss a coin: the sample space is $\Omega = \{H, T\}$ where H stands for head and T stands for tail. Examples of events are:

$$\begin{aligned} \{H\} &\qquad \text{(the outcome is } H) \\ \{T\} &\qquad \text{(the outcome is } T) \\ \{H, T\} &\qquad \text{(the outcome } H \text{ or } T) \end{aligned}$$

Example 1.2 (Throwing a die). Toss a normal six-faced die: the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Examples of events are:

$$\begin{aligned} \{5\} &\qquad \text{(the outcome is 5)} \\ \{2, 4, 6\} &\qquad \text{(the outcome is even)} \\ \{3, 6\} &\qquad \text{(the outcome is divisible by 3)} \end{aligned}$$

Example 1.3 (Drawing a card). Draw a card from a standard deck: Ω is the set of all possible cards, so that $|\Omega| = 52$. Examples of events are:

$$\begin{aligned} &A_1 = \{\text{the card is a Jack}\}, \qquad |A_1| = 4 \\ &A_2 = \{\text{the card is Diamonds}\}, \qquad |A_2| = 13 \\ &A_3 = \{\text{the card is not the Queen of Spades}\}, \qquad |A_3| = 51. \end{aligned}$$

EXAMPLE 1.4 (Picking a natural number). Pick any natural number: the sample space is $\Omega = \mathbb{N}$. Examples of events are:

$$\begin{aligned}\{\text{the number is at most } 5\} &= \{1, 2, 3, 4, 5\} \\ \{\text{the number is even}\} &= \{2, 4, 6, 8 \dots\} \\ \{\text{the number is not } 7\} &= \mathbb{N} \setminus \{7\}.\end{aligned}$$

EXAMPLE 1.5 (Picking a real number). Pick any number in the closed interval $[0, 1]$: the sample space is $\Omega = [0, 1]$. Examples of events are:

$$\begin{aligned}\{x : x < 1/3\} &= [0, 1/3) \\ \{x : x \neq 0.7\} &= [0, 1] \setminus \{0.7\} = [0, 0.7) \cup (0.7, 1] \\ \{x : x = 2^{-n} \text{ for some } n \in \mathbb{N}\} &= \{1, 1/2, 1/4, 1/8 \dots\}.\end{aligned}$$

Recall that a set Ω is said to be *countable* if there exists a bijection between Ω and a subset of $\mathbb{N}$. A set is said to be *uncountable* if it is not countable. Thus, for example, any finite set is countable, the set of even (or odd) integers is countable, but $\mathbb{R}$ is uncountable. Note that the sample space Ω is *finite* in the first two examples, infinite but countable in the third example, and uncountable in the last example.

REMARK 1.6. *For the first part of the course we will restrict to* countable *sample spaces, thus excluding Example 1.5 above.*

We can now give a general definition.

DEFINITION 1.7. *Let Ω be any countable set (not necessarily finite) and denote by $\mathcal{F}$ the family of all subsets of Ω. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a* probability measure *if*

- $\mathbb{P}(\Omega) = 1$,
- *for any sequence of disjoint events $(A_n)_{n \geq 1}$ it holds*

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

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space.

REMARK 1.8 (*). *Note that the family of sets $\mathcal{F}$ satisfies:*

- $\Omega \in \mathcal{F}$,
- *if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,*
- *for every sequence $(A_n)_{n \geq 1}$ in $\mathcal{F}$, it holds $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.*

For this reason $\mathcal{F}$ is called a sigma algebra*. When the state space Ω is uncountable one cannot define the probability measure $\mathbb{P}$ on all subsets of Ω: in that case $\mathcal{F}$ will be strictly smaller than the set of all subsets of Ω, and $\mathbb{P}$ will only be defined on $\mathcal{F}$. We do not discuss the technical issues concerning uncountable state spaces in this course.*

We think of $\mathcal{F}$ as the collection of observable events. If $A \in \mathcal{F}$, then $\mathbb{P}(A)$ is the probability of the event A. In some probability models, such as the one in Example 1.5, the probability of each individual outcome is 0. This is one reason why we need to specify probabilities of events rather than outcomes.

1.1. Equally likely outcomes. The simplest case is that of a finite sample space Ω and equally likely outcomes:

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} \qquad \forall \omega \in \Omega.$$

Note that by definition of probability measure this implies that

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \qquad \forall A \in \mathcal{F}.$$

In words, the probability of an event A is the number of outcomes in A, divided by the total number of possible outcomes. Moreover

- $\mathbb{P}(\emptyset) = 0$,
- if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$,
- if $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

To check that $\mathbb{P}$ is a probability measure, note that $\mathbb{P}(\Omega) = |\Omega|/|\Omega| = 1$, and for disjoint events $(A_k)_{k=1}^n$ it holds

$$\mathbb{P}\left(\bigcup_{k=1}^n A_k\right) = \frac{|A_1 \cup A_2 \cup \ldots \cup A_n|}{|\Omega|} = \sum_{k=1}^n \frac{|A_k|}{|\Omega|} = \sum_{k=1}^n \mathbb{P}(A_k),$$

as wanted.

Example 1.9. When throwing a fair die there are 6 possible outcomes, all equally likely. Then

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \quad \mathbb{P}(\{i\}) = 1/6 \ \text{ for } i = 1 \ldots 6.$$

So $\mathbb{P}(\text{even outcome}) = \mathbb{P}(\{2, 4, 6\}) = 1/2$, while $\mathbb{P}(\text{outcome} \leq 5) = \mathbb{P}(\{1, 2, 3, 4, 5\}) = 5/6$.

EXAMPLE 1.10. Pick a card from a deck of 52. Then Ω is the set of cards, and $|\Omega| = 52$. Example of events are:

$$A = \{\text{pick the queen of hearts}\}$$
$$B = \{\text{pick a Diamond or a Jack}\}$$
$$C = \{\text{don't pick an ace}\}$$

and $\mathbb{P}(A) = 1/52$, $\mathbb{P}(B) = 16/52$, $\mathbb{P}(C) = 48/52$.

EXAMPLE 1.11. Pick *two* cards from a deck of 52. Then Ω is the set of all *ordered pairs* of cards, and examples of events are:

$$A = \{\text{pick an ace, then a 2}\}$$
$$B = \{\text{pick two diamonds}\}$$
$$C = \{\text{the first card is a jack}\}.$$

Can you write down the probability of the above events?

2. Exercises

Exercise 1. Take $\Omega = \{a, b, c\}$. Write down the collection $\mathcal{F}$ of all subsets of Ω (that is, the collection of all possible events). Repeat the exercise with $\Omega = \{a, b, c, d\}$. Check that in both cases $|\mathcal{F}| = 2^{|\Omega|}$.

Exercise 2. Let Ω be a sample space, and A be an event. Show that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Exercise 3. What is $\mathcal{F}$ if $\Omega = \emptyset$?

Exercise 4. Take equally likely outcomes on a finite set Ω. Show that if A and B are events with $A \subseteq B$, then $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$ (note that this is always non-negative).

Exercise 5. Roll a die. What is the probability of seeing a number which is at most 4? And at most 6?

Exercise 6. Roll two dice. What is the sample space Ω? Check that $|\Omega| = 36$. Write down the probability of the following events:

- $A = \{(1, 5)\}$
- $B = \{\text{the first die gives } 2\}$
- $C = \{\text{the second die does not give } 4\}$
- $D = \{\text{both dice give an even number}\}$
- $E = \Omega \setminus \{(1, 5)\}$

Exercise 7. Draw two cards from a deck of 52. What is the sample space Ω? Compute $|\Omega|$. Write down the probability of the following events:

- $A = \{$the first card is an ace$\}$
- $B = \{$the second card is not a jack$\}$
- $C = \{$both cards are hearts$\}$
- $D = \{$the first card is a diamond, the second is hearts$\}$
- $E = \{$the first card is a diamond, the second is a 3$\}$

Exercise 8. You have three books: Maths Stats and Physics. Throw them on a shelf, assuming that all possible orderings are equally likely. What is the sample space Ω? Compute $|\Omega|$. Now pick the book on top. What is the probability that you picked the Maths book?

Exercise 9. Can you answer the questions of Exercise 8 in the case of n books?

Exercise 10. Out of a deck of 5 cards, you pick 3 and arrange them from left to right on the table. What is the sample space Ω, and what is its cardinality? How about picking 3 cards from a deck of n cards (assume $n \geq 3$)?

Exercise 11. Again, pick 3 cards from a deck of 5, but now you keep them in your hands and don't care about their order. What is Ω, and what is $|\Omega|$? Can you answer the same question if you pick 3 cards from a deck of n?

Exercise 12. Pick 4 cards from a deck of n sequentially, reinserting each card in the deck before picking the next one (drawing with repetition). What is Ω and what is $|\Omega|$? Compute the probability that the chosen cards are identical, and the probability that they are all different.

3. Combinatorial analysis

We have seen that, when working with equally likely outcomes on a finite set, in order to compute the probability of events one needs to count the number of elements of Ω with a given property. We now take a systematic look at some counting methods.

3.1. Multiplication rule. Take N finite sets $\Omega_1, \Omega_2 \ldots \Omega_N$ (some of which might coincide), with cardinalities $|\Omega_k| = n_k$. We imagine to pick one element from each set: how many possible ways do we have to do so? Clearly, we have n_1 choices for the first element. Now, for each choice of the first element, we have n_2 choices for the second, so that

$$|\Omega_1 \times \Omega_2| = n_1 n_2.$$

Once the first two element, we have n_3 choices for the third, and so on, giving

$$|\Omega_1 \times \Omega_2 \times \ldots \times \Omega_N| = n_1 n_2 \cdots n_N.$$

We refer to this as the *multiplication rule.*

EXAMPLE 3.1. A restaurant offers 6 starters, 7 main courses and 5 desserts. The number of possible three–course meals is then $6 \times 7 \times 5 = 210$.

EXAMPLE 3.2. Throw 3 dice. Then the number of outcomes given by an even number, followed by a 6 and then by a number smaller than 4 is $3 \times 1 \times 4 = 12$.

EXAMPLE 3.3 (The number of subsets). Suppose a set $\Omega = \{\omega_1, \omega_2 \ldots \omega_n\}$ has n elements. How many subsets does Ω have? We proceed as follows. To each subset A of Ω we can associate a sequence of 0's and 1's of length n so that the i^{th} number is 1 if ω_i is in A, and 0 otherwise. Thus if, say, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ then

$$\begin{aligned} A_1 = \{\omega_1\} &\mapsto 1,0,0,0 \\ A_2 = \{\omega_1, \omega_3, \omega_4\} &\mapsto 1,0,1,1 \\ A_3 = \emptyset &\mapsto 0,0,0,0. \end{aligned}$$

This defines a bijection between the subsets of Ω and the strings of 0's and 1's of length n. Thus we have to count the number of such strings. Since for each element we have 2 choices (either 0 or 1), there are 2^n strings. This shows that a set of n elements has 2^n subsets. Note that this also counts the cardinality of the sigma algebra $\mathcal{F}$ associated to Ω, as well as the number of functions from a set of n elements to $\{0, 1\}$.

3.2. Permutations. How many possible orderings of n elements are there? Label the elements $\{1, 2 \ldots n\}$. A *permutation* is a bijection from $\{1, 2 \ldots n\}$ to itself, i.e. an ordering of the elements. We may obtain all permutations by subsequently choosing the image of element 1, then the image of element 2 and so on. We have n choices for the image of 1, then $n - 1$ choices for the image of 2, $n - 2$ choices for the image of 3 until we have only one choice for the image of n. Thus the total number of choices is, by the multiplication rule,

$$n! = n(n-1)(n-2)\cdots 1.$$

Thus there are $n!$ different orderings, or permutations, of n elements. Equivalently, there are $n!$ different bijections from any two sets of n elements.

EXAMPLE 3.4. There are $4! = 24$ possible ways of arranging 4 books on a shelf.

EXAMPLE 3.5. There are $52!$ possible orderings of a standard deck of cards.

3.3. Subsets. How many ways are there to choose k elements from a set of n elements?

3.3.1. *With ordering.* We have n choices for the first element, $n-1$ choices for the second element and so on, ending with $n-k+1$ choices for the k^{th} element. Thus there are

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!} \tag{3.1}$$

ways to choose k ordered elements from n. An alternative way to obtain the above formula is the following: to pick k ordered elements from n, first pick a permutation of the n elements ($n!$ choices), then forget all elements but the first k. Since for each choice of the first k elements there are $(n-k)!$ permutations starting with those k elements, we again obtain (3.1).

3.3.2. *Without ordering.* To choose k unordered elements from n, we could first choose k *ordered* elements, and then forget about the order. Recall that there are $n!/(n-k)!$ possible ways to choose k ordered elements from n. Moreover, any given k elements can be ordered in $k!$ possible ways. Thus there are

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

possible ways to choose k unordered elements from n.

More generally, suppose we have integers $n_1, n_2 \ldots n_k$ with $n_1 + n_2 + \cdots + n_k = n$. Then we have

$$\binom{n}{n_1 \ldots n_k} = \frac{n!}{n_1! \ldots n_k!}$$

possible ways to partition n elements in k subsets of cardinalities $n_1, \ldots n_k$.

Example 3.6. Imagine to have a box containing n balls. Then there are:

- $\frac{n!}{(n-k)!}$ ordered ways to pick k balls,
- $\binom{n}{k}$ unordered ways to pick k balls,
- $\binom{n}{n_1\, n_2\, n_3}$ to subdivide the balls into 3 unordered groups of n_1, n_2 and n_3 balls each.

3.4. Subsets with repetitions. How many ways are there to choose k elements from a set of n elements, allowing repetitions?

3.4.1. *With ordering.* We have n choices for the first element, n choices for the second element and so on. Thus there are

$$n^k = n \times n \times \cdots \times n$$

possible ways to choose k ordered elements from n, allowing repetitions.

3.4.2. *Without ordering.* Suppose we want to choose k elements from n, allowing repetitions but discarding the order. How many ways do we have to do so? Note that naïvely dividing n^k by $k!$ doesn't give the right answer, since there may be repetitions. Instead, we count as follows. Label the n elements $\{1, 2 \dots n\}$, and for each element draw a $*$ each time it is picked.

$$\begin{array}{c|c|c|c|c} 1 & 2 & 3 & \dots & n \\ ** & * & & \dots & *** \end{array}$$

Note that there are k $*$'s and $n-1$ vertical lines. Now delete the numbers:

$$* * | * | | \dots | * ** \tag{3.2}$$

The above diagram uniquely identifies an unordered set of (possibly repeated) k elements. Thus we simply have to count how many such diagrams there are. The only restriction is that there must be $n-1$ vertical lines and k $*$'s. Since there are $n+k-1$ locations, we can fix such a diagram by assigning the positions of the $*$'s, which can be done in

$$\binom{n+k-1}{k}$$

ways. This therefore counts the number of unordered subsets of k elements from n, without ordering. (This goes under the name of *sticks and stars* argument).

Example 3.7. Imagine to again have a box with n balls, but now each time a ball is picked, it is put back in the box (so that it can be picked again). There are:

- n^k ordered ways to pick k balls, and
- $\binom{n+k-1}{k}$ unordered ways to pick k balls.

3.5. Recap of formulas.

- Permutations of n elements: $n!$.
- Choose k elements from n, no repetitions:
  - with ordering: $\dfrac{n!}{(n-k)!}$,
  - without ordering: $\dbinom{n}{k}$.
- Partition a set of n elements into k subsets of cardinalities $n_1, n_2 \dots n_k$: $\dbinom{n}{n_1\, n_2 \cdots n_k}$.
- Choose k elements from n, with repetitions:
  - with ordering: n^k,
  - without ordering: $\dbinom{n-1+k}{k}$.

4. Exercises

Exercise 1. How many functions $f : \{1, 2, 3, 4, 5\} \to \{0, 1\}$ are there? How many functions $g : \{0, 1\} \to \{1, 2, 3, 4, 5\}$? Explain.

Exercise 2. How many possible ways of assigning 5 toys to 5 children are there? How about 4 toys to 5 children (one child is left with no toy)? And 3 toys to 5 children?

Exercise 3. In how many different ways can you pick 3 cards from a deck of 7? Answer the same question if each card is re-inserted in the deck after being picked.

Exercise 4. Toss 3 dice. How many possible outcomes are there?

Exercise 5. You pick 6 (distinct) numbers out of 90 to bet on in the national lottery. What is the probability that the winning combination matches your choice (taking the order into account)? What is the probability that the winning combination matches your choice of numbers, possibly with a different order?

Exercise 6. A monkey types a 4 letters word using letters from the alphabet {Q,U,I,Z} (it could use the same letter multiple times). What is the probability that it types the word QUIZ?

Exercise 7. What is the probability that a uniformly chosen function $f : \{0, 1\} \to \{1, 2, 3, 4, 5\}$ is not constant? How about $f : \{1, 2 \ldots k\} \to \{1, 2 \ldots n\}$?

Exercise 8. [The birthday problem] If there are n people in the room, what is the probability that at least two people have the same birthday?

Exercise 9. How many ways are there to divide a deck of 52 cards into two decks of 26 each? How many ways are there to do so, ensuring that each of the two decks contains exactly 13 black and 13 red cards?

Exercise 10. How many increasing functions $f : \{1, 2 \ldots k\} \to \{1, 2 \ldots n\}$ are there? Assume $k \leq n$.

Exercise 11 (*). How many non-decreasing functions $f : \{1, 2 \ldots k\} \to \{1, 2 \ldots n\}$ are there?

Exercise 12. Four mice are chosen (without replacement) from a litter, two of which are white. The probability that both white mice are chosen is twice the probability that neither is chosen. How many mice are there in the litter?

Exercise 13. Suppose that n (non-identical) balls are tossed at random into n boxes. What is the probability that no box is empty? What is the probability that all balls end up in the same box?

Exercise 14 (*). [Ordered partitions] An ordered partition of k of size n is a sequence $(k_1, k_2 \dots k_n)$ of non-negative integers such that $k_1 + \cdots + k_n = k$. How many ordered partitions of k of size n are there?

5. Properties of Probability measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and recall that $\mathbb{P} : \mathcal{F} \to [0,1]$ has the property that $\mathbb{P}(\Omega) = 1$ and for any sequence $(A_n)_{n\geq 1}$ of *disjoint* events in $\mathcal{F}$ it holds

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}(A_n).$$

Then we have the following:

- $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{F}$,
- if $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$,
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, since $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$,
- $\mathbb{P}(\emptyset) = 0$,
- if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$, since
$$\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A).$$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

5.1. Continuity of probability measures. For a non-decreasing sequence of events $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ we have

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Indeed, we can define a new sequence $(B_n)_{n\geq 1}$ as

$$B_1 = A_1, \quad B_2 = A_2 \setminus A_1, \quad B_3 = A_3 \setminus (A_1 \cup A_2) \quad \dots \quad B_n = A_n \setminus (A_1 \cup \cdots A_{n-1}) \quad \dots$$

Then the B_n's are *disjoint*, and so

$$\mathbb{P}(A_n) = \mathbb{P}(B_1 \cup \cdots \cup B_n) = \sum_{k=1}^{n} \mathbb{P}(B_k).$$

Taking the limit as $n \to \infty$ both sides, we get

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \mathbb{P}\Big(\bigcup_{n\geq 1} B_n\Big) = \mathbb{P}\Big(\bigcup_{n\geq 1} A_n\Big),$$

as wanted. Similarly, for a non-increasing sequence of events $B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$ we have

$$\mathbb{P}\left(\bigcap_{n\geq 1} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

EXAMPLE 5.1. Take $\Omega = \mathbb{N}$ and let $A_n = \{1, 2 \ldots n\}$ for $n \geq 1$. Then we have $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ and

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \mathbb{P}(\Omega) = 1.$$

If, on the other hand, we set $A_n = A_5$ for all $n \geq 5$, then we find

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \mathbb{P}(A_5).$$

EXAMPLE 5.2. Take $\Omega = \mathbb{N}$ and let $B_n = \{n, n+1, n+2 \ldots\}$ for $n \geq 1$. Then we have $B_1 \supseteq B_2 \supseteq B_3 \cdots$ and

$$\lim_{n\to\infty} \mathbb{P}(B_n) = \mathbb{P}\left(\bigcap_{n\geq 1} B_n\right) = \mathbb{P}(\emptyset) = 0.$$

If, on the other hand, we set $B_n = B_3$ for all $n \geq 3$, then we find

$$\lim_{n\to\infty} \mathbb{P}(B_n) = \mathbb{P}\left(\bigcap_{n\geq 1} B_n\right) = \mathbb{P}(B_3).$$

5.2. Subadditivity of probability measures. For any events $A_1, A_2 \ldots A_n$ we have

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \mathbb{P}(A_k). \tag{5.1}$$

To see this, define the events

$$B_1 = A_1, \quad B_2 = A_2 \setminus A_1, \quad B_3 = A_3 \setminus (A_1 \cup A_2) \quad \ldots \quad B_n = A_n \setminus (A_1 \cup \cdots A_{n-1}).$$

Then the B_k's are disjoint, and $\mathbb{P}(B_k) \leq \mathbb{P}(A_k)$ for all $k = 1 \ldots n$, from which

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_k\right) = \sum_{k=1}^{n} \mathbb{P}(B_k) \leq \sum_{k=1}^{n} \mathbb{P}(A_k).$$

The same proof shows that this also holds for infinite sequences of events $(A_n)_{n\geq 1}$

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mathbb{P}(A_n).$$

The above inequalities show that probability measures are *subadditive*, and are often referred to as *Boole's inequality*.

EXAMPLE 5.3. We have already seen that $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, which is an example of (5.1) with $n = 2$.

EXAMPLE 5.4. Throw a fair die, so that $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then

$$\mathbb{P}(\{1, 2, 3\} \cup \{2, 4\}) = \mathbb{P}(\{1, 2, 3, 4\}) = \frac{4}{6},$$

while

$$\mathbb{P}(\{1, 2, 3\}) + \mathbb{P}(\{2, 4\}) = \frac{3}{6} + \frac{2}{6} = \frac{5}{6},$$

so indeed $\mathbb{P}(\{1, 2, 3\} \cup \{2, 4\}) \leq \mathbb{P}(\{1, 2, 3\}) + \mathbb{P}(\{2, 4\})$.

EXAMPLE 5.5. Toss a fair coin 3 times. Write H for head and T for tail. Then

$$\mathbb{P}\big(\text{exactly one head } \cup \{(T, H, H)\} \cup \{(H, T, T)\}\big) =$$
$$= \mathbb{P}(\{(H, T, T), (T, H, T), (T, T, H), (T, H, H)\}) = \frac{4}{8} = \frac{1}{2},$$

while

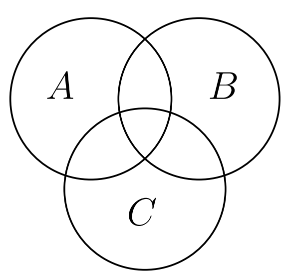
$$\mathbb{P}\big(\text{exactly one head}) + \mathbb{P}(T, H, H) + \mathbb{P}(H, T, T) = \frac{3}{8} + \frac{1}{8} + \frac{1}{8} = \frac{5}{8}.$$

5.3. Inclusion-exclusion formula. We have already seen that for any two events A, B

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

How does this generalise to an arbitrary number of events? For 3 events we have

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



In general, for n events $A_1, A_2 \ldots A_n$ we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}).$$

You are not required to remember the general formula for this course, but you should learn it for $n = 2, 3$.

Example 5.6. Pick a card from a standard deck of 52. Define the events

$$\begin{aligned} A &= \{\text{the card is diamonds}\},\\ B &= \{\text{the card is even}\},\\ C &= \{\text{the cards is at least 10.}\} \end{aligned}$$

Then

$$\begin{aligned} &\mathbb{P}(A) = \frac{13}{52}, \quad \mathbb{P}(B) = \frac{24}{52}, \quad \mathbb{P}(C) = \frac{16}{52},\\ &\mathbb{P}(A\cap B) = \frac{6}{52}, \quad \mathbb{P}(A\cap C) = \frac{4}{52}, \quad \mathbb{P}(B\cap C) = \frac{8}{52},\\ &\mathbb{P}(A\cap B\cap C) = \frac{2}{52}, \end{aligned}$$

which gives

$$\mathbb{P}(A\cup B\cup C) = \frac{13+24+16-6-4-8+2}{52} = \frac{37}{52}.$$

5.4. Limsup and liminf of events. Let $(A_n)_{n\ge1}$ be a sequence of events. We define

$$\begin{aligned} \limsup_{n\to\infty} A_n &:= \{A_n \text{ infinitely often}\} = \bigcap_{n\ge1}\bigcup_{m\ge n} A_m,\\ \liminf_{n\to\infty} A_n &:= \{A_n \text{ eventually}\} = \bigcup_{n\ge1}\bigcap_{m\ge n} A_m. \end{aligned}$$

In words:

- $\{A_n$ infinitely often$\}$ is the event that A_n holds for infinitely many $n \ge 1$, that is the event that for all $n \ge 1$ there exists $m \ge n$ such that A_m holds.
- $\{A_n$ eventually$\}$ is the event that all the A_n's hold for n large enough, that is the event that there exists $n \ge 1$ such that for all $m \ge n$ the event A_m holds.

Example 5.7. Toss a coin infinitely many times, and let $A_n = \{\text{head on the } n^{th} \text{ toss}\}$ for $n \ge 1$. Then

$$\begin{aligned} \limsup_{n\to\infty} A_n &:= \{A_n \text{ infinitely often}\} = \{\text{infinitely many heads}\},\\ \liminf_{n\to\infty} A_n &:= \{A_n \text{ eventually}\} = \{\text{only heads from some point onwards}\}. \end{aligned}$$

We will see later on that $\mathbb{P}(\text{heads infinitely often}) = 1$ while $\mathbb{P}(\text{heads eventually}) = 0$.

6. Exercises

Exercise 1. Let A, B, C be three events. Express in symbols the following events:

(i) only A occurs,
(ii) at least one event occurs,
(iii) exactly one event occurs,
(iv) no event occurs.

Exercise 2. Show that for any two sets A, B it holds $(A \cup B)^c = A^c \cap B^c$.

Exercise 3. Show that, for any three events A, B, C

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

How many numbers in $\{1, 2 \dots 500\}$ are not divisible by 7 but are divisible by 3 or 5?

Exercise 4. Let $(A_n)_{n \geq 1}$ be an increasing sequence of events

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$$

and assume that

$$\lim_{n \to \infty} \mathbb{P}(A_n) = p.$$

Explain why we can conclude that

$$\mathbb{P}\left(\bigcup_{n \geq 1} A_n \right) = p,$$

and

$$\mathbb{P}\left(\bigcap_{n \geq 1} A_n^c \right) = 1 - p.$$

Exercise 5. For three events A, B, C

(i) Express $\mathbb{P}(A \cup B)$ using the inclusion/exclusion formula.
(ii) Express $\mathbb{P}(A \cup B \cup C)$ using the inclusion/exclusion formula.

Exercise 6. Pick a card from a standard deck. Define the events

$$A = \{\text{the card is heart}\}$$

$$B = \{\text{the card is } \leq 5\}$$

$$C = \{\text{the card is multiple of } 3\}.$$

Find $\mathbb{P}(A \cup B)$ and $\mathbb{P}(A \cup B \cup C)$.

Exercise 7. A company has 5 employees: 3 women and 2 men.

(i) How many different committees of 3 employees can be formed?
(ii) What is the probability that a committee of 3 employees is only made of women?

Suppose each week a committee of 3 employees is formed.

(iii) What is the probability that in 2 months (8 weeks) the committee has been only made of women exactly once?
(iv) What is the probability that in 2 months (8 weeks) the committee has been only made of women exactly k times?

Exercise 8. A committee of size r is chosen at random from a set of n people. Calculate the probability that m given people will be in the committee.

Exercise 9. What is the probability that a non-decreasing function $f : \{1 \dots k\} \to \{1 \dots n\}$ is increasing?

Exercise 10. Show that

$$\left\{ \limsup_{n\to\infty} A_n \right\}^c = \liminf_{n\to\infty} A_n^c.$$

Exercise 11. Take an event A and define $A_n = A$ for all $n \geq 1$ (the events are all the same). Then show that $\limsup_n A_n = \liminf_n A_n = A$.

Exercise 12. Show that if the sequence of events $(A_n)_{n\geq 1}$ is non-decreasing, then

$$\limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = \bigcup_{n\geq 1} A_n,$$

while if the sequence$(A_n)_{n\geq 1}$ is non-increasing then

$$\limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = \bigcap_{n\geq 1} A_n.$$

7. Independence

We will now discuss what is arguably the most important concept in probability theory, namely *independence.*

DEFINITION 7.1. *Two events A, B are said to be* independent *if*

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

EXAMPLE 7.2. Throw two fair dice, and let

$$A = \{\text{the first number is even}\}, \qquad B = \{\text{the second number is } \leq 2\}.$$

Then $\mathbb{P}(A) = \frac{3}{6}$, $\mathbb{P}(B) = \frac{2}{6}$ and $\mathbb{P}(A \cap B) = \frac{6}{36}$, so $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ and the events A and B are independent.

EXAMPLE 7.3. Pick a card from a standard deck. Let

$$A = \{\text{the card is heart}\}, \qquad B = \{\text{the card is at most 5}\}.$$

Then $\mathbb{P}(A) = \frac{13}{52}$, $\mathbb{P}(B) = \frac{20}{52}$ and $\mathbb{P}(A \cap B) = \frac{5}{52}$, so $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ and therefore A and B are independent.

EXAMPLE 7.4. Toss two fair dice. Let

$$A = \{\text{the sum of the two numbers is 6}\}, \qquad B = \{\text{the first number is 4}\}.$$

Then $\mathbb{P}(A) = \frac{5}{36}$, $\mathbb{P}(B) = \frac{1}{6}$ and $\mathbb{P}(A \cap B) = \frac{1}{36}$. Thus $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$ and the events are not independent. Intuitively, the probability of getting 6 for the sum depends on the first outcome, since if we were to get 6 at the first toss then it would be impossible to obtain 6 for the sum, while if the first toss gives a number ≤ 5 then we have a positive probability of getting 6 for the sum.

If, on the other hand, we replace A with

$$A' = \{\text{the sum of the two numbers is 7}\},$$

then $\mathbb{P}(A') = \frac{6}{36}$, $\mathbb{P}(B) = \frac{1}{6}$ and $\mathbb{P}(A' \cap B) = \frac{1}{36}$. Thus $\mathbb{P}(A' \cap B) = \mathbb{P}(A')\mathbb{P}(B)$ and the events are independent. That is, knowing that the first toss gave 4 doesn't give us any information on the probability that the sum is equal to 7.

An important property of independence is the following: *if A is independent of B, then A is independent of B^c.* Indeed, we have

$$\begin{aligned}
\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) && \text{(by independence of } A, B) \\
&= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\
&= \mathbb{P}(A)\mathbb{P}(B^c) && (\text{since } \mathbb{P}(B^c) = 1 - \mathbb{P}(B))
\end{aligned}$$

which shows that A and B^c are independent.

EXAMPLE 7.5. Let A, B be defined as in Example 7.2, so that

$$B^c = \{\text{the second number is } \geq 3\}.$$

Then $\mathbb{P}(B^c) = 1 - \mathbb{P}(B) = \frac{4}{6}$, so $\mathbb{P}(A)\mathbb{P}(B^c) = \frac{12}{36} = \mathbb{P}(A \cap B^c)$, so A and B^c are independent.

EXAMPLE 7.6. Let A, B be defined as in Example 7.3, so that

$$B^c = \{\text{the card is at least } 6\}.$$

Then $\mathbb{P}(B^c) = 1 - \mathbb{P}(B) = \frac{32}{52}$, so $\mathbb{P}(A)\mathbb{P}(B^c) = \frac{8}{52} = \mathbb{P}(A \cap B^c)$, so A and B^c are independent.

We can also define independence for more than two events.

DEFINITION 7.7. *We say that the events $A_1, A_2 \dots A_n$ are independent if for any $k \geq 2$ and any collection of distinct indices $1 \leq i_1 < \cdots < i_k \leq n$ we have*

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Note that we require all possible subsets of the n events to be independent.

EXAMPLE 7.8. Toss 3 fair coins. Let

$$A = \{\text{first coin H}\}, \qquad B = \{\text{second coin H}\}, \qquad C = \{\text{third coin T}\}.$$

Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$ and

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B) \\
\mathbb{P}(B \cap C) &= \frac{1}{4} = \mathbb{P}(B)\mathbb{P}(C) \\
\mathbb{P}(A \cap C) &= \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C) \\
\mathbb{P}(A \cap B \cap C) &= \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C),
\end{aligned}$$

so A, B, C are independent.

At this point the reader may wonder whether independence of n events is equivalent to *pairwise independence*, that is the requirement that any two events are independent. The answer is no: independence is stronger than pairwise independence. Here is an example.

EXAMPLE 7.9. Toss 2 fair coins, so that

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}.$$

Let

$$\begin{aligned}
A &= \{\text{the first coin is T}\} = \{(T, H), (T, T)\}, \\
B &= \{\text{the second coin is T}\} = \{(H, T), (T, T)\}, \\
C &= \{\text{get exactly one T}\} = \{(T, H), (H, T)\}.
\end{aligned}$$

Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$, $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = 1/4$ so the events are pairwise independent. On the other hand,

$$\mathbb{P}(A \cap B \cap C) = 0,$$

so the events are *not* independent.

8. Conditional probability

Closely related to the notion of independence is that of *conditional probability.*

DEFINITION 8.1. *Let A, B be two events with $\mathbb{P}(B) > 0$. Then the* conditional probability of A given B *is*

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

We interpret $\mathbb{P}(A|B)$ as the probability that the event A occurs, when it is known that the event B occurs.

EXAMPLE 8.2. Toss a fair die. Let $A_1 = \{3\}$ and $B = \{\text{even}\} = \{2, 4, 6\}$. Then

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(\{3 \text{ and even}\})}{\mathbb{P}(\text{even})} = 0.$$

This is intuitive: if we know that the outcome is even, the probability that the outcome is 3 is zero.

If instead we take $A_2 = \{2, 3, 4, 5, 6\}$ then

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(\{2, 4, 6\})}{\mathbb{P}(\text{even})} = 1.$$

This is also intuitive: if we know that the outcome is even, then we are certain that it is at least 2.

Finally, take $A_3 = \{4, 5, 6\}$, for which

$$\mathbb{P}(A_3|B) = \frac{\mathbb{P}(\{4, 6\})}{\mathbb{P}(\text{even})} = \frac{2}{3}.$$

An important property to note is that if A and B are independent, then $\mathbb{P}(A|B) = \mathbb{P}(A)$. That is, when two events are independent, knowing that one occurs does not affect the probability of the other. This follows from

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

where we have used that A and B are independent in the second equality.

9. Exercises

Exercise 1. Toss two fair dice. Determine whether the following events are independent:

(i) $A = \{\text{the first die gives } 4\}$, $B = \{\text{the maximum outcome is } \leq 3\}$,
(ii) $A = \{\text{the first die gives } 4\}$, $B = \{\text{the maximum outcome is } \leq 5\}$,
(iii) $A = \{\text{the first die gives } 5\}$, $B = \{\text{the sum of the outcomes is } 6\}$,
(iv) $A = \{\text{the first die gives } 6\}$, $B = \{\text{the sum of the outcomes is } 7\}$.

Exercise 2. Toss a fair coin 3 times. Let

$$A = \{\text{at most two heads}\},$$
$$B = \{\text{the first coin toss gives head}\},$$
$$C = \{\text{all tails}\}.$$

(i) Are A, B independent?
(ii) Are A, B, C independent?
(iii) Are $A \setminus C$ and B independent?

Exercise 3. In New York it rains with probability 2/5. If it rains, Claire takes the umbrella with probability 9/10, and otherwise she takes the umbrella with probability 1/5. Given that Claire has taken the umbrella, what is the probability that it is raining?

Exercise 4. Let A and B be independent. Show that A^c and B^c are independent.

Exercise 5. Let A, B be two events with $\mathbb{P}(B) > 0$. Define $\mathbb{P}(A|B)$. Show that

(a) $\mathbb{P}(B|B) = 1$,
(b) if $A \subseteq B$ then $\mathbb{P}(A|B) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$,
(c) if $A \subsetneq B$ then $\mathbb{P}(A|B) < 1$.

Exercise 6. Give an example of two events for which $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$. Can you find two events A, B of positive probability such that $\mathbb{P}(A|B)\mathbb{P}(B) \neq \mathbb{P}(B|A)\mathbb{P}(A)$? Explain.

Exercise 7. Draw a card from a standard deck. What is the probability that the card is heart? What is the probability that the card is even, given that it is heart? What is the probability that the card is even, given that it is ≤ 5? What is the probability that the card is ≤ 5, given that it is even?

Exercise 8. Let A, B be two events such that $0 < \mathbb{P}(A) < 1$ and $0 < \mathbb{P}(B) < 1$. We say that B attracts A if $\mathbb{P}(A|B) > \mathbb{P}(A)$. Show that if B attracts A then A attracts B.

Exercise 9. What is the probability that a function $f : \{1, 2 \dots k\} \to \{1, 2 \dots n\}$ is constant, given that it is non-decreasing? What is the probability that $f : \{1, 2 \dots k\} \to \{1, 2 \dots n\}$ is non-decreasing, given that it is constant?

Exercise 10. You have four books: mathematics, physics, chemistry, biology. You arrange the books on a shelf uniformly at random. Let A denote the event that the mathematics book is in first position, and B the event that the biology book is in first position. Are A and B independent? How about A and the event $B' = \{\text{the biology book is in second position}\}$? Given that the mathematics book is in first position, what is the probability that the biology book is in second position?

10. The law of total probability

From the definition of conditional probability we see that for any two events A, B with $\mathbb{P}(B) > 0$ we have

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

This shows that the probability of two events occurring simultaneously can be broken up into calculating successive probabilities: first the probability that B occurs, and then the probability that A occurs given that B has occurred.

Clearly by replacing B with B^c in the above formula we also have

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

But then, since B and B^c are disjoint,

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

This has an important generalisation, called *law of total probability.*

Theorem 10.1 (Law of total probability). *Let $(B_n)_{n \geq 1}$ be a sequence of* disjoint *events of positive probability, whose union is the sample space Ω. Then for all events A*

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A|B_n)\mathbb{P}(B_n).$$

Proof. We know that, by assumption,

$$\bigcup_{n \geq 1} B_n = \Omega.$$

This gives

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega) = \mathbb{P}\left(A \cap \left(\bigcup_{n\geq 1} B_n\right)\right) = \mathbb{P}\left(\bigcup_{n\geq 1}(A \cap B_n)\right) \\ &= \sum_{n\geq 1} \mathbb{P}(A \cap B_n) = \sum_{n\geq 1} \mathbb{P}(A|B_n)\mathbb{P}(B_n).\end{aligned}$$

□

REMARK 10.2. *In the statement of the law of total probabilities we can also drop the assumption $\mathbb{P}(B_n) > 0$, provided we interpret $\mathbb{P}(A|B_n)\mathbb{P}(B_n) = 0$ if $\mathbb{P}(B_n) = 0$. It follows that we can also take $(B_n)_n$ to be a finite collection of events (simply set $B_n = \emptyset$ from some finite index onwards).*

EXAMPLE 10.3. An urn contains 10 black balls and 5 red balls. We draw two balls from the urn without replacement. What is the probability that the second ball drawn is black? Let

$$A = \{\text{the second ball is black}\}, \qquad B = \{\text{the first ball is black}\}.$$

Then

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) = \frac{10}{15} \cdot \frac{9}{14} + \frac{5}{15} \cdot \frac{10}{14} = \frac{2}{3}.$$

EXAMPLE 10.4. Toss a fair coin. If it comes up head, then toss one fair die, otherwise toss 2 fair dice. Call M the maximum number obtained from the dice, and let

$$A = \{M \leq 4\}.$$

What is the probability of A? We look at the two cases: either the coin came head and we tossed one die, or the coin came tail and we tossed two dice. Thus let

$$B = \{\text{the coin gives H}\}.$$

By the law of total probabilities we have

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) \\ &= \mathbb{P}(M \leq 4|H)\mathbb{P}(H) + \mathbb{P}(M \leq 4|T)\mathbb{P}(T) \\ &= \frac{1}{2} \cdot \frac{4}{6} + \frac{1}{2} \cdot \frac{4 \cdot 4}{6 \cdot 6} = \frac{5}{9}.\end{aligned}$$

EXAMPLE 10.5. Roll a fair dice: if you get k then toss a coin with probability $k/6$ of heads. What is the probability that the coin gives head? We have that

$$B_k = \{\text{the die gives } k\}, \qquad k = 1, 2, 3, 4, 5, 6$$

form a partition of Ω, since they are disjoint and their union is Ω. Moreover $\mathbb{P}(B_k) = 1/6$ for all k since the die is fair. Thus

$$\mathbb{P}(\text{head}) = \sum_{k=1}^{6} \mathbb{P}(A|B_k)\mathbb{P}(B_k) = \sum_{k=1}^{6} \frac{k}{6} \cdot \frac{1}{6} = \frac{1}{36} \sum_{k=1}^{6} k = 21/36.$$

11. Bayes' theorem

Note that if A and B are two events of positive probability we have, by definition of conditional probability,

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

Often, one of the conditional probabilities is easier to compute than the other. Bayes' theorem tells us how to switch between the two.

THEOREM 11.1 (Bayes' theorem). *For any two events A, B both of positive probability we have*

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

PROOF. We have

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

□

EXAMPLE 11.2. Going back to Example 10.3, suppose we are told that the second ball is black. What is the probability that the first ball was black? Applying Bayes' theorem we find

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{\frac{9}{14} \cdot \frac{10}{15}}{\frac{2}{3}} = \frac{9}{14}.$$

EXAMPLE 11.3. Going back to Example 10.4, suppose we are told that $M \leq 2$. What is the probability that the coin came up head? Recall that $B = \{\text{the coin gives H}\}$, and let $C = \{M \leq 2\}$. By Bayes' theorem, we have

$$\mathbb{P}(B|C) = \frac{\mathbb{P}(C|B)\mathbb{P}(B)}{\mathbb{P}(C)} = \frac{\mathbb{P}(C|B)\mathbb{P}(B)}{\mathbb{P}(C|B)\mathbb{P}(B) + \mathbb{P}(C|B^c)\mathbb{P}(B^c)} = \frac{\frac{2}{6} \cdot \frac{1}{2}}{\frac{2}{6} \cdot \frac{1}{2} + \frac{2 \cdot 2}{6 \cdot 6} \cdot \frac{1}{2}} = \frac{3}{4}.$$

12. Exercises

Exercise 1. Toss a fair coin. If you get head, toss another fair coin. What is the probability of seeing two heads?

Exercise 2. Iterate the experiment in the previous exercise until the first tail (that is, toss a fair coin and, at each step, if you get head toss it again). What is the probability of seeing n heads?

Exercise 3. Alice and Bob have one fair die each, and they play the following game. At each round they simultaneously toss their dice: if none of them gets a 6 they move to the next round, and otherwise whoever gets a 6 wins and the game ends. Note that there could be a tie, if they both get a 6.

(i) What is the probability that the game ends in the first round?
(ii) What is the probability that the game lasts exactly 2 rounds?

Exercise 4. This exercise generalises Example 10.3. An urn contains b black balls and r red balls. We draw two balls from the urn without replacement. What is the probability that the second ball drawn is black?

Exercise 5. Toss a fair coin twice. If get (H, H) then roll 1 die, if (T, T) roll 2 dice and otherwise roll 3 dice. Let M denote the maximum number obtained from the dice. What is the probability of the event $A = \{M \leq 4\}$?

Exercise 6. Toss a biased coin with probability p of heads. If head, then Alice tosses 4 fair coins, whereas if tail she tosses 6 fair coins.

(i) Compute the probability that Alice sees 5 heads.
(ii) Compute the probability that Alice sees 2 heads.
(iii) Compute the probability that the initial biased coin gave head, given that Alice sees 2 heads.
(iv) Determine for which value of p the probability in (iii) equals $1/2$.

Exercise 7. An urn contains two coins of type A and one coin of type B. When a type A coin is flipped, it comes up head with probability $1/4$, while when a type B coin is flipped it comes up head with probability $3/4$. A coin is randomly chosen from the urn and flipped. What is the probability to get head? Given that the coin gave head, what is the probability that it was a coin of type A?

Exercise 8. Students of a given class get grades A, B, C, D with probability $\frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{2}{8}$ respectively. If they misread the question, they usually do worse, getting grades A, B, C, D with probability $\frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}$ respectively. Each student misreads the question with probability $\frac{2}{3}$. What is the probability that:

(a) a student who read the question correctly gets A?
(b) a student who got B has read the question correctly?
(c) a student who got C has misread the question?
(d) a student who misread the question got C?
(e) a student gets D?

Exercise 9. Independent trials are performed, each with probability p of success. Let P_n be the probability that n trials result in an even number of successes. Show that

$$P_n = \frac{1}{2}(1 + (1-2p)^n).$$

13. Some natural probability distributions

Up to this point in the course we have only worked with equally likely outcomes. On the other hand, many other choices of probability measures are possible. We will next see some some of them, which arise naturally.

DEFINITION 13.1. *A probability distribution, or simply* distribution*, is a probability measure on some* $(\Omega, \mathcal{F})$*. We say that a distribution is* discrete *if* Ω *is a discrete set.*

For discrete Ω we always take the σ-algebra $\mathcal{F}$ to be the set of all subsets of Ω. In particular, it contains all sets $\{\omega\}$, $\omega \in \Omega$, made of a single outcome. Thus, if $\mathbb{P}$ is a probability distribution, we can write

$$p_\omega = \mathbb{P}(\{\omega\}).$$

Note that $p_\omega \in [0,1]$ for all $\omega \in \Omega$, and

$$\sum_{\omega \in \Omega} p_\omega = \mathbb{P}(\Omega) = 1.$$

Moreover, $\mathbb{P}$ is uniquely determined by the collection $(p_\omega)_{\omega\in\Omega}$, that we call *weights*. We think of the weight p_ω as the probability of seeing the outcome $\{\omega\}$, when sampling according to $\mathbb{P}$.

We now list several natural probability distributions.

13.1. Bernoulli distribution.

Take $\Omega = \{0,1\}$ and define $\mathbb{P}$ to be the probability distribution given by the weights

$$p_1 = p, \qquad p_0 = 1-p.$$

Then $\mathbb{P}$ models the number of heads obtained in one biased coin toss (the coin gives head with probability p and tail with probability $1-p$). Such $\mathbb{P}$ is called *Bernoulli distribution of parameter* p, and it is denoted by *Bernoulli*(p).

13.2. Binomial distribution.

Fix an integer $N \geq 1$ and let $\Omega = \{0, 1, 2 \ldots N\}$. Define $\mathbb{P}$ to be the probability distribution given by the weights

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}, \qquad 0 \leq k \leq N.$$

Then $\mathbb{P}$ models the number of heads in N biased coin tosses (again, the coin gives head with probability p, and tail with probability $1-p$). Such $\mathbb{P}$ is called *Binomial distribution of parameters* N, p, and it is denoted by $Binomial(N,p)$. Note that

$$\mathbb{P}(\Omega) = \sum_{k=0}^{N} p_k = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} = (p + (1-p))^N = 1.$$

13.3. Geometric distribution. Let $\Omega = \{1, 2, 3\ldots\}$, and define $\mathbb{P}$ as the distribution given by the weights

$$p_k = (1-p)^{k-1} p, \qquad k \geq 1.$$

for all $k \geq 1$. Then μ models the number of biased coin tosses up to (and including) the first head. Such $\mathbb{P}$ is called *Geometric distribution of parameter* p, and it is denoted by $Geometric(p)$. Note that

$$\mathbb{P}(\Omega) = \sum_{k=1}^{\infty} p_k = p \sum_{k=1}^{\infty} (1-p)^{k-1} = p \sum_{k=0}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1.$$

13.4. Poisson distribution. Let $\Omega = \mathbb{N}$, and for a fixed parameter $\lambda \in (0, +\infty)$ let $\mathbb{P}$ be the probability distribution given by the weights

$$p_k = \frac{e^{-\lambda} \lambda^k}{k!}, \qquad k \geq 0.$$

Such $\mathbb{P}$ is called *Poisson distribution of parameter* λ, and it is denoted by $Poisson(\lambda)$. Note that

$$\mathbb{P}(\Omega) = \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

This distribution arises as the limit of a Binomial distribution with parameters $N, \frac{\lambda}{N}$ as $N \to \infty$. Indeed, if $p_k(N, \lambda/N)$ denote the weights of a $Binomial(N, \lambda/N)$ distribution, we have

$$\begin{aligned} p_k(N, \lambda/N) &= \binom{N}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \frac{N(N-1)\cdots(N-k+1)}{N^k} \left(1 - \frac{\lambda}{N}\right)^{N-k} \frac{\lambda^k}{k!} \longrightarrow \frac{e^{-\lambda}\lambda^k}{k!} \end{aligned}$$

as $N \to \infty$, since

$$\frac{N(N-1)\cdots(N-k+1)}{N^k} \to 1, \qquad \left(1 - \frac{\lambda}{N}\right)^N \to e^{-\lambda}, \qquad \left(1 - \frac{\lambda}{N}\right)^{-k} \to 1.$$

14. Exercises

Exercise 1. Let $\Omega = \{0, 1, 2\}$. How many events are there on Ω? Consider the probability distribution on Ω given by the weights $p_0 = 1/2$, $p_1 = 1/3$, $p_2 = 1/6$. Check that this is a valid probability distribution. Compute $\mathbb{P}(A)$ for all A events on Ω.

Exercise 2. An infinite sequence of independent trials is to be performed. Each trial results in a success with probability $p \in (0, 1)$, and a failure with probability $1 - p$. Determine the probability that:

(a) no success occurs in the first n trials,
(b) at least one success occurs in the first n trials,
(c) exactly k successes occur in the first n trials,
(d) the first success occurs exactly on the n^{th} trial.

Exercise 3. An infinite sequence of independent trials is to be performed. Each trial results in a success with probability $p \in (0, 1)$, and a failure with probability $1 - p$. What is the probability distribution of the number of trials until (and including) the first success? Write down the corresponding weights $(p_k)_{k \geq 1}$.

Exercise 4. A variant of the geometric distribution counts the number of tails until the first head, when we have a sequence of independent coin tosses with probability p of heads. Explain why in this case $\Omega = \{0, 1, 2, 3 \dots\}$ with

$$p_k = (1 - p)^k p, \qquad k \geq 0,$$

and check that this is a valid probability distribution.

Exercise 5. Each day of the week James flips a biased coin: if it comes up head (which happens with probability p), then he goes out for a walk. What is the distribution of the number of times James goes for a walk in one week? Write down the corresponding weights $(p_k)_{0 \leq k \leq 7}$.

Exercise 6. Prove the identity

$$(a + b)^N = \sum_{k=0}^{N} \binom{N}{k} a^k b^{N-k}$$

by reasoning as follows.

- Note that $(a + b)^N = (a + b) \cdot (a + b) \cdot (a + b) \cdots (a + b)$.
- Each term in the above product is a monomial of degree N, of the form $a^k b^{N-k}$ for some $0 \leq k \leq N$.
- For any fixed k between 0 and N, there are exactly $\binom{N}{k}$ monomials of the form $a^k b^{N-k}$.

Exercise 7. Let $\Omega = \{0, 1, 2 \ldots\}$ and consider the collection of non-negative numbers

$$q_k = e^{-\lambda k}, \qquad k \geq 0.$$

Show that these are normalizable to a probability distribution if and only if $\lambda \in (0, \infty)$. Assuming this is the case, compute C such that the collection $(p_k)_{k\geq 0}$ with $p_k = Cq_k$ is a probability distribution on Ω.

Exercise 8. On $\Omega = \{1, 2, 3 \ldots\}$ consider the collection of weights $(p_k)_{k\geq 1}$ with $p_k = 1/k$. Show that this is not normalizable to a probability distribution. Show that, on the other hand, the collection of square weights $(q_k)_{k\geq 1}$ with $q_k = p_k^2$ is normalizable to a probability distribution.

Exercise 9. Show that if $|\Omega| = +\infty$ then we cannot define a probability measure on Ω such that all outcomes are equally likely.

Exercise 10. Tom is undecided as to whether to take a Mathematics course or a Literature course. He estimates that the probability of getting an A would be $1/3$ in a Maths course, and $1/2$ in a Literature course. Suppose he chooses which course to take by tossing a fair coin: what is the probability that he gets an A? Suppose that instead he tosses a biased coin: if it comes up head, which happens with probability $p \in (0, 1)$, he takes the Maths course, and otherwise he takes the Literature course. What is the probability that Tom gets an A?

Exercise 11. At a certain stage of a criminal investigation the inspector is 60 per cent convinced of the guilt of a certain suspect. Suppose, however, that a new piece of evidence is uncovered: it is found that the criminal is left-handed. If the probability of being both left-handed and non-guilty is $2/25$, how certain of the guilt of the suspect should the inspector be if it turns out that the suspect is left-handed?

Exercise 12. Toss a biased coin, which gives head with probability $p \in (0, 1)$ and tail with probability $1 - p$. If it comes up head, roll two dice and record the sum. If tail, pick a card from a standard deck and record its number (between 1 and 13). What is the probability that the number you have recorded is 4? What is the probability that it is ≤ 2?

15. Random variables

It is often the case that when a random experiment is conducted we don't just want to record the outcome, but perhaps we are interested in some functions of the outcome.

DEFINITION 15.1. *A* random variable *on* $(\Omega, \mathcal{F})$ *taking values in a discrete set* S *is a function* $X : \Omega \to S$.

Typically $S \subseteq \mathbb{R}$ or $S \subseteq \mathbb{R}^k$, in which case we say that X is a real-valued random variable.

EXAMPLE 15.2. Toss a biased coin. Then $\Omega = \{H, T\}$, where H, T stands for head and tail. Define

$$X(H) = 1, \qquad X(T) = 0.$$

Then $X : \Omega \to \{0, 1\}$ counts the number of heads in the outcome.

EXAMPLE 15.3. Toss two biased coins, so that $\Omega = \{H, T\}^2$. Define

$$X(HH) = 2, \qquad X(TH) = X(HT) = 1, \qquad X(TT) = 0.$$

Then $X : \Omega \to \{0, 1, 2\}$ counts the number of heads in 2 coin tosses.

EXAMPLE 15.4. Roll two dice, so that $\Omega = \{1, 2, 3, 4, 5, 6\}^2$. For $(i, j) \in \Omega$, set

$$X(i, j) = i + j.$$

Then $X : \Omega \to \{2, 3 \dots 12\}$ records the sum of the two dice. We could also define the random variables

$$Y(i, j) = \max\{i, j\}, \qquad Z(i, j) = i.$$

Then $Y : \Omega \to \{1, 2, 3, 4, 5, 6\}$ records the maximum of the two dice, while $Z : \Omega \to \{1, 2, 3, 4, 5, 6\}$ records the outcome of the first die.

EXAMPLE 15.5. For a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $A \in \mathcal{F}$, define

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \text{for } \omega \in A, \\ 0, & \text{for } \omega \notin A. \end{cases}$$

Then $\mathbb{1}_A : \Omega \to \{0, 1\}$ tells us whether the outcome was in A or not. Note that

(i) $\mathbb{1}_{A^c} = 1 - \mathbb{1}_A$,
(ii) $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$,
(iii) $\mathbb{1}_{A \cup B} = 1 - (1 - \mathbb{1}_A)(1 - \mathbb{1}_B)$.

You should prove (iii) as an exercise.

For a subset $T \subseteq S$ we denote the set $\{\omega \in \Omega : X(\omega) \in T\}$ simply by $\{X \in T\}$. Since X takes values in a discrete S, we let

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

for all $x \in S$. The collection $(p_x)_{x \in S}$ is referred to as the *probability distribution of* X. If the probability distribution of X is, say, *Geometric*(p), then we say that X is a Geometric random variable of parameter p, and write $X \sim \mathit{Geometric}(p)$. Similarly, for the other distributions we have encountered write $X \sim \mathit{Bernoulli}(p)$, $X \sim \mathit{Binomial}(N, p)$, $X \sim \mathit{Poisson}(\lambda)$.

Given a random variable X taking values in $S \subset \mathbb{R}$, the function $F_X : \mathbb{R} \to [0,1]$ given by

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

is called *distribution function* of X. Note that F_X is piecewise constant, non-decreasing, right-continuous, and

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 1.$$

EXAMPLE 15.6. Toss a biased coin, which gives head with probability p, and define the random variable

$$X(H) = 1, \qquad X(T) = 0.$$

Then

$$F_X(x) = \begin{cases} 0, \text{ if } x < 0 \\ 1-p, \text{ if } x \in [0,1) \\ 1, \text{ if } x \geq 1. \end{cases}$$

Note that F_X is piecewise constant, non-decreasing, right-continuous, and its jumps are given by the weights of a Bernoulli distribution of parameter p.

Knowing the probability distribution function is equivalent to knowing the collection of weights $(p_x)_{x \in S}$ such that $p_x = \mathbb{P}(X = x)$, and hence it is equivalent to knowing the probability distribution of X.

DEFINITION 15.7 (Independent random variables). *Two random variables X, Y, taking values in S_X, S_Y respectively, are said to be* independent *if*

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

for all $(x, y) \in S_X \times S_Y$. In general, the random variables $X_1, X_2 \dots X_n$ taking values in $S_1, S_2 \dots S_n$ are said to be independent if

$$\mathbb{P}(X_1 = x_1, X_2 = x_2 \dots X_n = x_n) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2) \cdots \mathbb{P}(X_n = x_n)$$

for all $(x_1, x_2 \dots x_n) \in S_1 \times S_2 \times \cdots \times S_n$.

Note that if $X_1, X_2 \dots X_N$ are independent, then for any $k \leq N$ and any distinct indices $1 \leq i_1 < i_2 < \cdots < i_k \leq N$, the random variables $X_{i_1}, X_{i_2} \dots X_{i_k}$ are independent. You are asked to prove this as an exercise.

16. Exercises

Exercise 1. Let $X \sim \mathit{Bernoulli}(p)$. Determine the distribution of $Y = 1 - X$.

Exercise 2. Show that for any two sets A, B

$$\mathbb{1}_{A\cup B} = 1 - (1 - \mathbb{1}_A)(1 - \mathbb{1}_B).$$

Exercise 3. Show that if A, B are independent events then $\mathbb{1}_A$ and $\mathbb{1}_B$ are independent random variables. Compute the expectation of $\mathbb{1}_{A\cap B}$ and $\mathbb{1}_{A\cup B}$.

Exercise 4. Roll two dice, and let $\Omega = \{1, 2 \ldots 6\}^2$ be the associated sample space. Define the random variable $X : \Omega \to \{1, 2 \ldots 6\}$ by

$$X(i, j) = \max\{i, j\}$$

for all $(i, j) \in \Omega$. Write down the weights of the random variable X. Draw the graph of the associated distribution function F_X, and check that

$$\lim_{x\to-\infty} F_X(x) = 0, \qquad \lim_{x\to+\infty} F_X(x) = 1.$$

Exercise 5. Roll two dice, and let $\Omega = \{1, 2 \ldots 6\}^2 = \{(i, j) : 1 \le i, j \le 6\}$ be the associated sample space. Consider the event $A = \{i = 4\}$, and define the random variable

$$X(i, j) = \mathbb{1}_A(i, j) = \begin{cases} 1 \text{ if } i = 4, \\ 0 \text{ if } i \neq 4. \end{cases}$$

Define further the random variable

$$Y(i, j) = i + j.$$

Determine whether X and Y are independent.

Exercise 6. Let X be a random variable taking values in $\{0, 1, 2\}$ with probabilities $p_0 = 1/3$, $p_1 = 1/4$ and $p_2 = 5/12$. Draw the distribution function of X, namely $F_X(x) = \mathbb{P}(X \le x)$.

Exercise 7. Draw the distribution function of a geometric random variable of parameter $p = 1/2$.

Exercise 8. Let X, Y, Z be independent random variables taking values in S_X, S_Y, S_Z respectively. Show that X, Y, Z are pairwise independent.

Exercise 9. Generalise the above exercise by showing that if $X_1, X_2 \ldots X_N$ are independent, then for any $k \le N$ and any distinct indices $1 \le i_1 < i_2 < \cdots < i_k \le N$, the random variables $X_{i_1}, X_{i_2} \ldots X_{i_k}$ are independent.

Exercise 10. Roll a fair die repeatedly. Let N and M denote the number of tosses until the first 1 and the second 1 respectively.

(i) Compute $\mathbb{P}(M = 2)$. Are the events $\{N = 1\}$ and $\{M = 2\}$ independent?
(ii) Compute $\mathbb{P}(M = N + 1)$ (the probability of seeing the second 1 right after the first 1). Are the events $\{N = 1\}$ and $\{M - N = 1\}$ independent?
(iii) Are N and M independent? Are N and $M - N$ independent? Explain.
(iv) Compute $\mathbb{P}(N = k|M = 5)$ for $k = 1, 2, 3, 4$.

17. Expectation

From now on all the random variables we consider are assumed to take real values, unless otherwise specified. We say that a random variable X is non-negative if X takes values in $S \subseteq [0, \infty)$.

Definition 17.1 (Expectation of a random variable). *For a random variable $X : \Omega \to S$ we define the* expectation *(or* expected value*, or* mean value*) of X to be*

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

Thus the expectation of X is the average of the values taken by X, averaged with weights corresponding to the probabilities of the values.

Example 17.2. If $X \sim Bernoulli(p)$ then

$$\mathbb{E}(X) = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = p.$$

Example 17.3. If $X \sim Binomial(N, p)$ then

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=0}^{N} k\binom{N}{k}p^k(1-p)^{N-k} = Np\sum_{k=1}^{N}\binom{N-1}{k-1}p^{k-1}(1-p)^{(N-1)-(k-1)} \\ &= Np\underbrace{\sum_{k'=0}^{N-1}\binom{N-1}{k}p^{k'}(1-p)^{N-1-k'}}_{1} = Np.\end{aligned}$$

Example 17.4. If $X \sim Geometric(p)$ then

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty}\underbrace{k(1-p)^{k-1}}_{-\frac{d}{dp}(1-p)^k} = -p\frac{d}{dp}\Big(\sum_{k=1}^{\infty}(1-p)^k\Big) \\ &= -p\frac{d}{dp}\Big(\frac{1}{1-(1-p)}\Big) = -p\Big(-\frac{1}{p^2}\Big) = \frac{1}{p}.\end{aligned}$$

EXAMPLE 17.5. If $X \sim Poisson(\lambda)$ then

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \underbrace{\sum_{k'=0}^{\infty} \frac{\lambda^{k'}}{k'!}}_{e^{\lambda}} = \lambda.$$

REMARK 17.6. *Note that, if X takes infinitely many values and it is non-negative, then the expectation of X might diverge to $+\infty$ (see Exercise 18 below). We remark that if X takes infinitely many values of both signs, then $\mathbb{E}(X)$ may be ill-defined (one would have to look at its positive and negative part separately). We do not address this latter issue in this course.*

We have defined the expectation of a random variable $X : \Omega \to S_X$ as

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

In what follows, it will be convenient to note the equivalent formula

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

The equivalence of the two formulas for $\mathbb{E}(X)$ follows from reindexing the summation according to the values of the random variable X:

$$\sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}) = \sum_{x \in S_X} \sum_{\omega \in \Omega : X(\omega) = x} X(\omega)\mathbb{P}(\{\omega\}) = \sum_{x \in S_X} x \underbrace{\sum_{\omega \in \Omega : X(\omega) = x} \mathbb{P}(\{\omega\})}_{\mathbb{P}(X = x)} = \sum_{x \in S_X} x \mathbb{P}(X = x).$$

Properties of the expectation. The expectation of a random variable $X : \Omega \to S$ satisfies the following properties, some of which we state without proof:

(1) If $X \geq 0$ then $\mathbb{E}(X) \geq 0$, and $\mathbb{E}(X) = 0$ if and only if $\mathbb{P}(X = 0) = 1$.
(2) If $c \in \mathbb{R}$ is a constant, then $\mathbb{E}(c) = c$ and $\mathbb{E}(cX) = c\,\mathbb{E}(X)$.
(3) For random variables X, Y, it holds $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

PROOF.

$$\begin{aligned} \mathbb{E}(X+Y) &= \sum_{\omega \in \Omega} (X+Y)(\omega)\mathbb{P}(\{\omega\}) = \sum_{\omega \in \Omega} (X(\omega) + Y(\omega))\mathbb{P}(\{\omega\}) \\ &= \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}) + \sum_{\omega \in \Omega} Y(\omega)\mathbb{P}(\{\omega\}) = \mathbb{E}(X) + \mathbb{E}(Y). \end{aligned}$$

□

The above properties generalise by induction, to give that for any constants $c_1, c_2 \ldots c_n$ and random variables $X_1, X_2 \ldots X_n$ it holds

$$\mathbb{E}\left(\sum_{k=1}^{n} c_k X_k\right) = \sum_{k=1}^{n} c_k \mathbb{E}(X_k).$$

Thus the expectation is *linear*.

(4) For any function $g : S \to S'$, $g(X) : \Omega \to S'$ is a random variable taking values in S', and

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

An important example is given by $g(x) = x^k$, and the corresponding expectation $\mathbb{E}(X^k)$ is called *the* k^{th} *moment* of the random variable X.

Proof.

$$\mathbb{E}(g(X)) = \sum_{\omega \in \Omega} g(X(\omega))\mathbb{P}(\{\omega\}) = \sum_{x \in S} \sum_{\omega \in \Omega : X(\omega) = x} g(X(\omega))\mathbb{P}(\{\omega\}) = \sum_{x \in S} g(x)\mathbb{P}(X = x).$$

□

(5) If $X \geq 0$ takes integer values, then

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k).$$

Proof. Exercise. □

Example 17.7. Let $X \sim \mathit{Geometric}(p)$. Then $\mathbb{P}(X \geq k) = (1-p)^{k-1}$, and

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \sum_{k=1}^{\infty} (1-p)^{k-1} = \sum_{k=0}^{\infty} (1-p)^k = \frac{1}{p}.$$

This is the quickest way to compute the expectation of a geometric random variable (compare it to the computation in Example 17.4).

(6) If X, Y are independent random variables, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. In general, if $X_1, X_2 \dots X_N$ are independent random variables,

$$\mathbb{E}(X_1 \cdot X_2 \cdots X_N) = \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) \cdots \mathbb{E}(X_N).$$

Proof. We prove it with two random variables.

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_{\omega \in \Omega} (XY)(\omega)\mathbb{P}(\{\omega\}) = \sum_{x \in S_X} \sum_{y \in S_Y} \sum_{\omega : X(\omega) = x, Y(\omega) = y} X(\omega)Y(\omega)\mathbb{P}(\{\omega\}) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} xy\mathbb{P}(X = x, Y = y) = \sum_{x \in S_X} \sum_{y \in S_Y} xy\mathbb{P}(X = x)\mathbb{P}(Y = y) \\
&= \left(\sum_{x \in S_X} x\mathbb{P}(X = x) \right) \cdot \left(\sum_{y \in S_Y} x\mathbb{P}(Y = y) \right) = \mathbb{E}(X)\mathbb{E}(Y).
\end{aligned}$$

Note that, crucially, we have used independence in the second line in order to say that $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$. □

(7) If $X : \Omega \to S_X$ and $Y : \Omega \to S_Y$ are two independent random variables, and $f : S_X \to S'_X$ and $g : S_Y \to S'_Y$ are two functions, then $f(X), g(Y)$ are independent random variables. This generalises to arbitrarily many random variables, to give that if $X_1, X_2 \dots X_N$ are independent, then $f_1(X_1), f_2(X_2) \dots f_N(X_N)$ are independent.

18. Exercises

Exercise 1. Let A be an event and define the random variable

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Show that $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$. If B is another event, show that $\mathbb{E}(\mathbb{1}_A \mathbb{1}_B) = \mathbb{P}(A \cap B)$. If, moreover, A and B are independent show that $\mathbb{E}(\mathbb{1}_A \mathbb{1}_B) = \mathbb{P}(A)\mathbb{P}(B)$, and deduce that $\mathbb{E}(\mathbb{1}_A \mathbb{1}_B) = \mathbb{E}(\mathbb{1}_A)\mathbb{E}(\mathbb{1}_B)$.

Exercise 2. Write a random word made of 10 characters by choosing a character uniformly at random 10 times independently, from an alphabet of 26 characters. Let X be the random variable that counts the number of A's in the resulting word. What is the distribution of X? Compute $\mathbb{E}(X)$, that is the average number of A's in a random word of length 10. How about repeating the experiment with a word of length N?

Exercise 3. Let X be a random variable taking values in the set $S_X = \{0, 1, 2, 3\}$ with probability distribution $(p_k)_{k=0}^3$ given by $p_0 = 1/4$, $p_1 = 1/8$, $p_2 = 1/8$, $p_3 = 1/2$. Compute $\mathbb{E}(X)$.

Exercise 3. Compute the average number of times you have to roll a fair die in order to get a 4.

Exercise 4. Draw a card from a standard deck N times and let X count the number of cards which are diamonds. What is the distribution of X? Compute $\mathbb{E}(X)$, that is the average number of diamonds in N cards drawn independently from a standard deck.

Exercise 5. Let $X : \Omega \to \mathbb{N}$ have probability distribution $(p_k)_{k \in \mathbb{N}}$ where

$$p_k = \mathbb{P}(X = k) = \frac{6}{(\pi k)^2}, \qquad k \geq 1.$$

Show that $\mathbb{E}(X) = +\infty$.

Exercise 6. Let $X : \Omega \to \mathbb{Z}$ have probability distribution $(p_k)_{k\in\mathbb{Z}}$ where

$$\begin{aligned} p_0 &= \mathbb{P}(X=0) = 0, \\ p_k &= \mathbb{P}(X=k) = \mathbb{P}(X=-k) = 2^{-|k|-1}, \qquad k \geq 1. \end{aligned}$$

Show that $\mathbb{E}(X) = 0$.

Exercise 7. Show that if $X \geq 0$ takes integer values, then

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k).$$

Exercise 8. Roll two dice, and let X, Y denote the outcomes of the first and second die respectively. Compute $\mathbb{E}(X)$ and $\mathbb{E}(Y)$. Show that X and Y are independent. Compute the expectation of the random variables $S = X + Y$ and $P = XY$.

Exercise 9. Generalise the above exercise to an arbitrary number of dice. Let $X_1, X_2 \ldots X_N$ denote the outcome of N consecutive rolls of a die, and define

$$S = \sum_{k=1}^{N} X_k, \qquad P = \prod_{k=1}^{N} X_k.$$

Compute $\mathbb{E}(S)$ and $\mathbb{E}(P)$.

Exercise 10. Roll two dice. Let

$$A = \{\text{the first die gives } 4\}, \qquad B = \{\text{the sum of the two dice is } 7\}.$$

Then we have seen that A, B are independent events. Define the indicator random variables

$$X = \mathbb{1}_A, \qquad Y = \mathbb{1}_B.$$

Show that X, Y are independent. Compute $\mathbb{E}(X), \mathbb{E}(Y), \mathbb{E}(X+Y), \mathbb{E}(XY)$. If, on the other hand, we introduce

$$C = \{\text{the sum of the two dice is } 3\}, \qquad Z = \mathbb{1}_C$$

then we have seen that A, C are *not* independent. Check that X, Z are not independent by showing that $\mathbb{E}(XZ) \neq \mathbb{E}(X)\mathbb{E}(Z)$.

Exercise 11. Let $X_1, X_2 \ldots X_N$ be i.i.d. Bernoulli(p) random variables, and set $S = X_1 + X_2 + \cdots + X_N$. Then S takes values in $\{0, 1 \ldots N\}$. Show that

$$\mathbb{P}(S = k) = \binom{N}{k} p^k (1-p)^{N-k},$$

and conclude that $S \sim$Binomial(N, p). Now use the linearity of the expectation to show that $\mathbb{E}(S) = Np$. This is by far the quickest way to compute the expected value of a Binomial random variable.

Exercise 12. Alice and Bob have one deck of cards each, and they play the following game. At each round they simultaneously pick a card from their decks: if none of them gets an ace they re-insert the card in the deck and move to the next round, and otherwise whoever gets an ace wins and the game ends. Note that there could be a tie, if they both get an ace. What is the expected duration of the game?

19. Variance and covariance

19.1. Variance. Once we know the mean of a random variable X, we may ask how much typically X deviates from it. This is measured by the variance.

DEFINITION 19.1. *For any random variable X with finite mean $\mathbb{E}(X)$, the* variance *of X is defined to be*

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

Thus $\mathrm{Var}(X)$ measures how much the distribution of X is concentrated around its mean: the smaller the variance, the more the distribution is concentrated around $\mathbb{E}(X)$.

Properties of the variance. The variance of a random variable $X : \Omega \to S$ satisfies the following properties:

(1) $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

PROOF. We have

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2 + (\mathbb{E}(X))^2 - 2X\mathbb{E}(X)] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$

□

(2) If $\mathbb{E}(X) = 0$ then $\mathrm{Var}(X) = \mathbb{E}(X^2)$.
(3) $\mathrm{Var}(X) \geq 0$, and $\mathrm{Var}(X) = 0$ if and only if $\mathbb{P}(X = c) = 1$ for some constant $c \in \mathbb{R}$.
(4) If $c \in \mathbb{R}$ then $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.
(5) If $c \in \mathbb{R}$ then $\mathrm{Var}(X + c) = \mathrm{Var}(X)$.
(6) If X, Y are independent then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. In general, if $X_1, X_2 \ldots X_N$ are independent then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_N) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_N).$$

EXAMPLE 19.2. If $X \sim$Bernoulli(p) then $\mathbb{E}(X) = p$ and $\mathrm{Var}(X) = p(1 - p)$. It follows that if $X_1, X_2 \ldots X_N$ are i.i.d. Bernoulli(p) and $S = X_1 + X_2 + \cdots + X_N$ then $S \sim$Binomial(N, p) and $\mathbb{E}(S) = Np$,

$$\mathrm{Var}(S) = \sum_{k=1}^{N} \mathrm{Var}(X_k) = Np(1 - p).$$

This is by far the quickest way to compute the variance of a Binomial random variable.

19.2. Covariance. Closely related to the concept of variance is that of covariance.

DEFINITION 19.3. *For any two random variables X, Y with finite mean, we define their* covariance *to be*

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].$$

19.3. Properties of the covariance. The covariance of two random variables X, Y satisfies the following properties:

(1) $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

PROOF. We have

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}[XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

□

(2) $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
(3) $\mathrm{Cov}(X, c) = 0$ for all $c \in \mathbb{R}$.
(4) For $c \in \mathbb{R}$, $\mathrm{Cov}(cX, Y) = c\,\mathrm{Cov}(X, Y)$ and $\mathrm{Cov}(X + c, Y) = \mathrm{Cov}(X, Y)$.
(5) For X, Y, Z random variables, $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y)$.

The above properties generalise, to give that the covariance is *bilinear*. That is, for any collections of random variables $X_1, X_2 \dots X_n$ and $Y_1, Y_2 \dots Y_n$, and constants $a_1, a_2 \dots a_n$ and $b_1, b_2 \dots b_n$, it holds

$$\mathrm{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^n b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \mathrm{Cov}(X_i, Y_j).$$

(6) For any two random variables with finite variance, it holds

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$

PROOF. We have

$$\begin{aligned}\mathrm{Var}(X + Y) &= \mathbb{E}[((X + Y) - \mathbb{E}(X + Y))^2] \\ &= \mathbb{E}[(X - \mathbb{E}(X) + Y - \mathbb{E}(Y))^2] \\ &= \mathbb{E}[(X - \mathbb{E}(X))^2] + \mathbb{E}[(Y - \mathbb{E}(Y))^2] + 2\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).\end{aligned}$$

□

(7) If X, Y are independent then $\mathrm{Cov}(X, Y) = 0$.

PROOF. We have

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0,$$

where the second equality follows from independence. □

Note that properties (6) and (7) together give that if X, Y are independent then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. We remark that the converse of property (7) is false: while X, Y independent implies $\mathrm{Cov}(X,Y) = 0$, the fact that $\mathrm{Cov}(X,Y) = 0$ does not imply that X, Y are independent. We show this in the following example.

EXAMPLE 19.4 (Zero covariance does not imply independence). Let

$$X = \begin{cases} 2, & \text{with probability } \frac{1}{4} \\ 1, & \text{with probability } \frac{1}{4} \\ -1, & \text{with probability } \frac{1}{4} \\ -2, & \text{with probability } \frac{1}{4} \end{cases}, \qquad Y = X^2.$$

Then $\mathbb{E}(X) = 0$ and $\mathbb{E}(XY) = \mathbb{E}(X^3) = 0$ by symmetry, so

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0.$$

However, X, Y are *not* independent, since

$$\mathbb{P}(X = 1, Y = 4) = 0,$$

while

$$\mathbb{P}(X = 1)\mathbb{P}(Y = 4) = \mathbb{P}(X = 1)\mathbb{P}(\{X = 2\} \cup \{X = -2\}) = \frac{1}{4}\frac{1}{2} > 0.$$

20. Exercises

Exercise 1. Toss two biased coin (giving head with probability p). Let

$$A = \{\text{the first coin gives head}\},$$
$$B = \{\text{the second coin gives head}\},$$
$$C = \{\text{at least one head}\}.$$

Let $X = \mathbb{1}_A, Y = \mathbb{1}_B, Z = \mathbb{1}_C$. Compute expectation and variance of X, Y, Z. Compute $\mathrm{Cov}(X,Y)$, $\mathrm{Cov}(X,Z)$. Are X, Z independent? Compute $\mathbb{E}(X+Y)$, $\mathbb{E}[(X+Y)^2]$, $\mathbb{E}[\log(X+2)]$, $\mathbb{E}(4YZ - X^2)$.

Exercise 2. Let X be a random variable taking values $\{0, 1, 2, 3\}$ with probability p_0, p_1, p_2, p_3 respectively, with probability

$$p_0 = p_1 = p_2 = \frac{1}{6}, \qquad p_3 = \frac{1}{2}.$$

Compute $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathrm{Var}(X)$. Describe the random variable $Y = |X - 2|$ (that is, list the possible values of Y and the associated weights). Compute $\mathbb{E}(Y)$, $\mathbb{E}(Y^2)$, $\mathrm{Var}(Y)$ and $\mathrm{Cov}(X, Y)$. Are X, Y independent? Compute $\mathbb{E}(X + Y + 3)$, $\mathrm{Var}(X + Y - 2)$, $\mathbb{E}[(X - Y)X]$, $\mathbb{E}[(X + Y)(X - Y)]$.

Exercise 3. Draw two cards from a standard deck. Let

$$A = \{\text{the first card is heart}\}, \qquad B = \{\text{the second card is heart}\}.$$

Define $X = \mathbb{1}_A$, $Y = \mathbb{1}_B$. Are A, B independent? Are X, Y independent? Write down the weights of X and Y, and compute $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathrm{Cov}(X, Y)$, $\mathrm{Var}(X + Y)$.

Exercise 4. Let $X \sim \mathit{Geometric}(p)$. Compute $\mathbb{E}(X(X - 1))$. Deduce $\mathbb{E}(X^2)$ and $\mathrm{Var}(X)$.

Exercise 5. Let $X \sim \mathit{Poisson}(\lambda)$. Compute $\mathbb{E}(X(X - 1))$. Deduce $E(X^2)$ and $\mathrm{Var}(X)$.

Exercise 6. In a sequence of n independent trials the probability of a success at the i^{th} trial is p_i. Let N denote the total number of successes. Find $\mathbb{E}(N)$ and $\mathrm{Var}(N)$.

Exercise 7. Let X be a random variable with finite mean μ. For $x \in \mathbb{R}$ define the function

$$f(x) = \mathbb{E}[(X - x)^2].$$

Computing $f'(x)$, show that $f(x)$ is minimised at $x = \mu$, and compute its minimum value.

Exercise 8. Let X be a random variable with mean μ and variance σ^2. Compute $\mathbb{E}(X^2)$ and $\mathbb{E}[(X - 2)^2]$ in terms of μ, σ^2.

Exercise 9. Give an example of two random variables with equal mean but different variance. Give an example of two random variables with the same mean and variance, but different distribution.

Exercise 10. Let $X \sim \mathit{Poisson}(\lambda)$. Compute $\mathbb{E}(e^X)$. Compute $\mathbb{E}(e^{\alpha X})$ for all $\alpha \in \mathbb{R}$. Let X, Y be independent $\mathit{Poisson}(\lambda)$ random variables. For any $\alpha, \beta \in \mathbb{R}$ compute

$$\mathbb{E}(e^{\alpha X + \beta Y}).$$

Exercise 11. Roll a die N times. Denote by X the total number of $3's$ and by Y the total number of 6's. Are X and Y independent? Denote by Z the total number of even outcomes. Are X and Z independent? Compute expectation and variance of X, Y, Z. [Hint: write X, Y, Z as sums of independent indicator random variables.]

21. Joint and conditional distributions

21.1. Joint distribution.

Definition 21.1 (Joint distribution). *Given two random variables $X : \Omega \to S_X$ and $Y : \Omega \to S_Y$, their* joint distribution *is defined as the collection of probabilities*

$$\mathbb{P}(X = x, Y = y), \qquad x \in S_X,\ y \in S_Y.$$

Note that the joint distribution of X and Y is a probability distribution on $S_X \times S_Y$, so that one could think of (X, Y) as a random vector taking values in $S_X \times S_Y$.

Example 21.2. Let X, Y be independent Bernoulli random variables of parameter p, and define $Z = X + Y$. Then, if we interpret X, Y as recording the outcome of two independent coin tosses, Z denotes the total number of head. Here $S_X = S_Y = \{0, 1\}$ while $S_Z = \{0, 1, 2\}$. Moreover, by independence

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \qquad \text{for all } (x, y) \in S_X \times S_Y.$$

On the other hand, X and Z are not independent, and they have joint distribution

$$\begin{aligned} &\mathbb{P}(X = 0, Z = 0) = (1 - p)^2, \quad \mathbb{P}(X = 0, Z = 1) = (1 - p)p, \\ &\mathbb{P}(X = 1, Z = 1) = p(1 - p), \quad \mathbb{P}(X = 1, Z = 2) = p^2. \end{aligned}$$

Note that the weights of the joint distribution sum up to 1.

Remark 21.3. *The joint distribution of two* independent *random variables is the* product *of the marginal distributions.*

From the joint distribution, one can recover the *marginal distribution* of X and Y by mean of the total probability law:

$$\mathbb{P}(X = x) = \sum_{y \in S_Y} \mathbb{P}(X = x, Y = y) \qquad \text{for all } x \in S_X,$$

$$\mathbb{P}(Y = y) = \sum_{x \in S_X} \mathbb{P}(X = x, Y = y) \qquad \text{for all } y \in S_Y.$$

Example 21.4. Continuing Example 21.2, from the joint distribution of X, Z we can recover the marginal distribution of Z via

$$\begin{aligned}\mathbb{P}(Z=0) &= \mathbb{P}(Z=0, X=0)+\mathbb{P}(Z=0, X=1)=(1-p)^2,\\ \mathbb{P}(Z=1) &= \mathbb{P}(Z=1, X=0)+\mathbb{P}(Z=1, X=1)=2p(1-p),\\ \mathbb{P}(Z=2) &= \mathbb{P}(Z=2, X=0)+\mathbb{P}(Z=2, X=1)=p^2.\end{aligned}$$

This generalises to arbitrarily many random variables: for $X_1 \dots X_n$ random variables taking values in $S_1 \dots S_n$, their joint distribution is given by

$$\mathbb{P}(X_1=x_1, X_2=x_2, \dots X_n=x_n), \qquad x_1\in S_1,\dots,x_n\in S_n,$$

and we can recover the marginal distributions via

$$\mathbb{P}(X_i=x)=\sum_{x_1\dots x_{i-1},x_{i+1}\dots x_n}\mathbb{P}(X_1=x_1,\dots,X_i=x,\dots,X_n=x_n)$$

for all $x\in S_i$.

21.2. Conditional distribution.

Definition 21.5 (Conditional distribution). *For two random variables X, Y, the* conditional distribution *of X given the event $\{Y=y\}$ (which we assume to have positive probability) is the collection of probabilities*

$$\mathbb{P}(X=x|Y=y), \qquad x\in S_X,$$

where, clearly,

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

If we know the distribution of Y, then we can recover the marginal distribution of X via the total probability law:

$$\mathbb{P}(X=x)=\sum_{y\in S_Y}\mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y).$$

Example 21.6. Continuing Example 21.2, the conditional distribution of Z given $\{X=0\}$ is

$$\mathbb{P}(Z=0|X=0)=1-p, \quad \mathbb{P}(Z=1|X=0)=p, \quad \mathbb{P}(Z=2|X=0)=0.$$

Thus, conditional on the event $\{X=0\}$, Z has Bernoulli(p) distribution.

On the other hand, the conditional distribution of Z given $\{X=1\}$ is

$$\mathbb{P}(Z=0|X=1)=0, \quad \mathbb{P}(Z=1|X=1)=1-p, \quad \mathbb{P}(Z=2|X=1)=p,$$

which shows that $Z-1$ has Bernoulli(p) distribution. We can recover the marginal distribution of Z via the law of total probability

$$\begin{aligned}
\mathbb{P}(Z=0) &= \mathbb{P}(Z=0|X=0)\mathbb{P}(X=0)+\mathbb{P}(Z=0|X=1)\mathbb{P}(X=1) = (1-p)^2,\\
\mathbb{P}(Z=1) &= \mathbb{P}(Z=1|X=0)\mathbb{P}(X=0)+\mathbb{P}(Z=1|X=1)\mathbb{P}(X=1) = 2p(1-p),\\
\mathbb{P}(Z=2) &= \mathbb{P}(Z=2|X=0)\mathbb{P}(X=0)+\mathbb{P}(Z=2,X=1)\mathbb{P}(X=1) = p^2.
\end{aligned}$$

REMARK 21.7. *If X,Y are independent, then the conditional distribution of X given $\{Y=y\}$ coincides with the distribution of X, since $\mathbb{P}(X=x|Y=y)=\mathbb{P}(X=x)$ for all $(x,y)\in S_X\times S_Y$.*

22. Inequalities

It is often the case that, rather than computing the exact value of $\mathbb{E}(X)$, we are only interested in having a bound for it.

THEOREM 22.1 (Markov's inequality). *Let X be a non-negative random variable, and $\lambda\in(0,\infty)$. Then*

$$\mathbb{P}(X\geq\lambda)\leq\frac{\mathbb{E}(X)}{\lambda}.$$

THEOREM 22.2 (Chebyshev's inequality). *Let X be a random variable with finite mean $\mathbb{E}(X)$. Then, for $\lambda\in(0,\infty)$,*

$$\mathbb{P}(|X-\mathbb{E}(X)|\geq\lambda)\leq\frac{\mathrm{Var}(X)}{\lambda^2}.$$

PROOF. We have

$$\mathbb{P}(|X-\mathbb{E}(X)|\geq\lambda)=\mathbb{P}((X-\mathbb{E}(X))^2\geq\lambda^2)\leq\frac{\mathbb{E}[(X-\mathbb{E}(X))^2]}{\lambda^2}=\frac{\mathrm{Var}(X)}{\lambda^2},$$

where the inequality follows from Markov's inequality. □

We now discuss an application of the above inequalities. Toss a biased coin (giving head with probability p) repeatedly, and let $X_k=\mathbb{1}(\text{head on the }k^{th}\text{ coin toss})$ for $k=1,2\dots n$. Then $X_1,X_2\dots X_n$ are independent *Bernoulli*(p) random variables. Define

$$\bar{X}=\frac{1}{n}\sum_{k=1}^{n}X_k,$$

and note that $\bar{X}$ records the proportion of heads in n coin tosses. Then

$$\mathbb{E}(\bar{X})=\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(X_k)=p$$

and

$$\mathrm{Var}(\bar{X})=\frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}(X_k)=\frac{p(1-p)}{n}.$$

It therefore follows from Chebyshev's inequality that

$$\mathbb{P}(|\bar{X} - p| \geq 0.1) \leq \frac{\operatorname{Var}(\bar{X})}{(0.1)^2} = \frac{p(1-p)}{n \cdot 0.01}. \tag{22.1}$$

Note that the function $f(p) = p(1-p)$ on $[0,1]$ is maximised at $p = 1/2$, as can be checked by differentiating it, so that $f(p) \leq f(1/2) = 1/4$. Using this to bound the right hand side of (22.1) we get

$$\mathbb{P}(|\bar{X} - p| \geq 0.1) \leq \frac{p(1-p)}{n \cdot 0.01} \leq \frac{1}{4n \cdot 0.01} = \frac{100}{4n}.$$

This can be made as small as we want by taking n large enough.

Theorem 22.3 (Cauchy-Schwarz inequality). *For all random variables X, Y, it holds*

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

We do not prove this inequality. Note that if Y is constant, say $\mathbb{P}(Y = 1) = 1$, then the inequality gives

$$\mathbb{E}(|X|) \leq \sqrt{\mathbb{E}(X^2)}.$$

Note that for arbitrary random variables X, Y we have

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

by definition, from which

$$\begin{aligned} |\operatorname{Cov}(X, Y)| &\leq \mathbb{E}[|X - \mathbb{E}(X)| \cdot |Y - \mathbb{E}(Y)|] \\ &\leq \sqrt{\mathbb{E}[(X - \mathbb{E}(X))^2] \cdot \mathbb{E}[(Y - \mathbb{E}(Y))^2]} \\ &= \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}, \end{aligned}$$

where in the second inequality we have applied Cauchy–Schwarz to the random variables $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$. It follows that if we define the *correlation coefficient* of X and Y as

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

then $|\operatorname{Corr}(X, Y)| \leq 1$ and so

$$\operatorname{Corr}(X, Y) \in [-1, 1].$$

Note that $\operatorname{Corr}(X, Y) = 0$ if X and Y are independent, while $\operatorname{Corr}(X, Y) = 1$ if $X = Y$ and $\operatorname{Corr}(X, Y) = -1$ if $X = -Y$. The correlation coefficient is a standardized way of measuring the correlation between X and Y.

Theorem 22.4 (Jensen's inequality). *Let X be an integrable random variable and let $f : \mathbb{R} \to \mathbb{R}$ be a convex function (it suffices for f to be convex on the image of X, i.e. on an interval that contains S_X). Then*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

We do not prove this. To remember the direction of the inequality, take $f(x) = x^2$ and use that $\mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \mathrm{Var}(X) \geq 0$. Note that the function $f(x) = |x|$ is also convex, so by Jensen's inequality

$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|).$$

22.1. The Law of Large Numbers. As an important application of Chebyshev's inequality we discuss the (Weak) Law of Large Numbers.

THEOREM 22.5 (Weak law of large numbers). *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2, and define*

$$S_N = X_1 + X_2 + \cdots + X_N.$$

Then for all $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_N}{N} - \mu\right| > \varepsilon\right) \to 0$$

as $N \to \infty$.

PROOF. Note that $\mathbb{E}(S_N/N) = \mu$ and $\mathrm{Var}(S_N/N) = \sigma^2/N$. It then follows from Chebyshev's inequality that

$$\mathbb{P}\left(\left|\frac{S_N}{N} - \mu\right| > \varepsilon\right) = \mathbb{P}\left(\left|\frac{S_N}{N} - \mathbb{E}\left(\frac{S_N}{N}\right)\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(S_N/N)}{\varepsilon^2} = \frac{\sigma^2}{N\varepsilon^2} \to 0$$

as $N \to \infty$. □

23. Exercises

Exercise 1. Let A, B denote two events on the same sample space Ω, and define two random variables on this space by setting $X = \mathbb{1}(A)$ and $Y = \mathbb{1}(B)$. Compute the joint distribution of X and Y.

Exercise 2. Let X, Y have joint distribution given by

$$\mathbb{P}(X = 0, Y = 1) = 1/4$$
$$\mathbb{P}(X = 1, Y = 1) = 1/4$$
$$\mathbb{P}(X = 2, Y = 1) = 0$$
$$\mathbb{P}(X = 3, Y = 1) = 1/8$$
$$\mathbb{P}(X = 0, Y = 5) = 1/8$$
$$\mathbb{P}(X = 1, Y = 5) = 1/12$$
$$\mathbb{P}(X = 2, Y = 5) = 1/12$$
$$\mathbb{P}(X = 3, Y = 5) = 1/12$$

(you should check that the above weights sum up to 1). Compute the marginal distributions of X and Y. Compute $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$. Are X and Y independent?

Exercise 3. Let X, Y have joint distribution

$$\begin{aligned}
\mathbb{P}(X=0, Y=1) &= 1/3 \\
\mathbb{P}(X=0, Y=2) &= 1/6 \\
\mathbb{P}(X=0, Y=3) &= 1/12 \\
\mathbb{P}(X=1, Y=1) &= 1/12 \\
\mathbb{P}(X=1, Y=2) &= 1/6 \\
\mathbb{P}(X=1, Y=3) &= 1/6.
\end{aligned}$$

Compute the marginal distribution of X and the marginal distribution of Y. Compute $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$. Are X and Y independent?

Exercise 4. Roll two dice, and let X, Y denote the outcomes of the first and second die respectively. Compute the joint distribution of X and Y. Compute the distribution of X given the event $\{Y = 2\}$.

Exercise 5. Let $X \sim \mathit{Geometric}(p)$. Compute the distribution of X conditional on the event $\{X \geq j\}$, for any j positive integer, that is the collection of weights

$$\mathbb{P}(X = k | X \geq j), \qquad k \geq 1.$$

Exercise 6. Toss n fair coins, and denote the outcomes by $X_1, X_2 \dots X_n$. Compute the joint distribution of $X_1, X_2 \dots X_n$. Use the law of total probability to compute the marginal distribution of X_1 from the joint distribution.

Exercise 7. Let X be a Binomial(N, p) random variable. Compute the joint distribution of X and X^2.

Exercise 8. Let K be a random variable taking values $\{0, 1, 2 \dots 7\}$ with probability $1/8$ each. Define

$$X = \cos\left(\frac{K\pi}{4}\right), \qquad Y = \sin\left(\frac{K\pi}{4}\right).$$

Show that $\mathrm{Cov}(X, Y) = 0$ but X, Y are not independent.

Exercise 9. Toss a biased coin repeatedly, and let X denote the number of tosses until (and including) the second head. Show that

$$\mathbb{P}(X = k) = (k-1)(1-p)^{k-2}p^2$$

for $k \geq 2$. Explain how X can be represented as the sum of two independent and identically distributed random variables. Use this representation to derive the mean and variance of X.

Exercise 10. In the same setting as the above exercise, let X denote the number of tosses until (and including) the a^{th} head, for $a \geq 2$. Show that in this case

$$\mathbb{E}(X) = \frac{a}{p}, \qquad \mathrm{Var}(X) = \frac{a(1-p)}{p^2}.$$

Exercise 11. Let $X_1, X_2 \dots X_n$ be independent and identically distributed random variables with finite mean μ and variance σ^2. Define the sample mean

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Use Chebyshev's inequality to show that

$$\mathbb{P}(|\bar{X} - \mu| > 2\sigma) \leq \frac{1}{4n}.$$

Use this bound to find n such that

$$\mathbb{P}(|\bar{X} - \mu| > 2\sigma) \leq 0.01.$$

Exercise 12. Toss a fair coin repeatedly. Let X denote the number of tosses until (and including) the first head, while Y denotes the number of tosses until (and including) the first tail. Show that

$$X \sim \textit{Geometric}(1/2), \qquad Y \sim \textit{Geometric}(1/2),$$

and X, Y are *not* independent. Use Cauchy–Schwarz to show that $\mathbb{E}(XY) \leq 6$ and deduce that $\mathrm{Cov}(X, Y) \leq 2$.

Exercise 13. Let X be a bounded random variable. Show that for every $p > 0$ and all $x > 0$

$$\mathbb{P}(|X| \geq x) \leq \mathbb{E}(|X|^p)x^{-p}.$$

Show further that for all $\beta \geq 0$

$$\mathbb{P}(X \geq x) \leq \mathbb{E}(e^{\beta X})e^{-\beta x}.$$

Exercise 14. Let X be a Poisson random variable of parameter λ. We have seen in the previous exercise that for all $\beta \geq 0$

$$\mathbb{P}(X \geq x) \leq \mathbb{E}(e^{\beta X})e^{-\beta x}.$$

By finding the value of β that minimises the right hand side, show that for all $x \geq \lambda$

$$\mathbb{P}(X \geq x) \leq \exp\{-x \log(x/\lambda) - \lambda + x\}.$$

Exercise 15(*). Let $x_1, x_2 \dots x_n$ be positive numbers, and let X be a random variable taking value x_k with probability $1/n$ for each $k = 1 \dots n$ (that is, X is uniformly distributed over the set of values $\{x_1, x_2 \dots x_n\}$). By applying Jensen's inequality to X show that

$$\left(\frac{1}{n}\sum_{k=1}^{n}\frac{1}{x_k}\right)^{-1} \leq \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Show that if $y_1, y_2 \dots y_n$ is any reordering of the values $x_1, x_2 \dots x_n$, then

$$\frac{1}{n}\sum_{k=1}^{n}\frac{y_k}{x_k} \geq 1.$$

CHAPTER 2

Continuous probability

1. Some natural continuous probability distributions

Up to this point we have only considered probability distributions on *discrete* sets, such as finite sets, $\mathbb{N}$ or $\mathbb{Z}$. We now turn our attention to *continuous* distributions.

Definition 1.1 (Probability density function). *A function* $f : \mathbb{R} \to \mathbb{R}$ *is a* probability density function *if*

$$f(x) \geq 0 \quad \textit{for all } x \in \mathbb{R}, \qquad \int_{-\infty}^{+\infty} f(x)dx = 1.$$

We will often write p.d.f. in place of probability density function for brevity. Given a p.d.f., we can define a probability measure $\mathbb{P}$ on $\mathbb{R}$ by setting

$$\mathbb{P}([a,b]) = \int_a^b f(x)dx$$

for all $a, b \in \mathbb{R}$. Note that

$$\mathbb{P}((-\infty, a]) = \int_{-\infty}^{a} f(x)dx, \qquad \mathbb{P}(\mathbb{R}) = \int_{-\infty}^{+\infty} f(x)dx = 1,$$

and $\mathbb{P}([a,b)) = \mathbb{P}((a,b]) = \mathbb{P}((a,b)) = \mathbb{P}([a,b])$. We interpret $f(x)$ as the *density of probability* at x (the continuous analogue of the weight of x), and $\mathbb{P}([a,b])$ as the probability of the interval $[a,b]$. For this course we restrict to piecewise continuous probability density functions.

Example 1.2. Take $f(x) = \frac{e^{-|x|}}{2}$. Then $f(x) \geq 0$ for all $x \in \mathbb{R}$, and

$$\int_{-\infty}^{+\infty} f(x)dx = 2\int_0^{\infty} \frac{e^{-x}}{2}dx = 1.$$

Thus f is a valid probability density function. If μ denotes the associated probability measure, we have

$$\mathbb{P}([0,\infty)) = \frac{1}{2}\int_0^{\infty} e^{-x}dx = \frac{1}{2}, \qquad \mathbb{P}([-1,1]) = \frac{1}{2}\int_{-1}^{1} e^{-|x|}dx = \int_0^1 e^{-x}dx = 1 - e^{-1}.$$

This tells us that if we choose a real number at random according to $\mathbb{P}$, the probability that this number is non-negative is $1/2$ (which was also clear as f is symmetric), while the probability that it has modulus at most 1 is $1 - e^{-1}$.

We now list some natural continuous probability distributions. Throughout, we use the following notation: for $A \subseteq \mathbb{R}$

$$\mathbb{1}_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A, \end{cases}$$

in analogy with the notation introduced for indicator random variables.

1.1. Uniform distribution. For $a, b \in \mathbb{R}$ with $a < b$, the density function

$$f(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x)$$

defines the *uniform distribution* on $[a, b]$. Note that

$$\int_{-\infty}^{+\infty} f(x)dx = \int_a^b f(x)dx = 1.$$

The uniform distribution gives to each interval contained in $[a, b]$ a probability proportional to its length.

1.2. Exponential distribution. For $\lambda \in (0, \infty)$, the density function

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

defines the *exponential distribution* of parameter λ. Note that

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = 1.$$

1.3. Cauchy distribution. The density function

$$f(x) = \frac{1}{\pi(1+x^2)}$$

defines the *Cauchy distribution*. Note that

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{2}{\pi} \int_0^{+\infty} \frac{1}{1+x^2} dx = \frac{2}{\pi} \arctan(x) \Big|_0^{+\infty} = 1.$$

1.4. Gaussian distribution. The density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

defines the *standard Gaussian distribution*, or *standard normal distribution*, written $N(0, 1)$. More generally, for $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$, the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

defines the *Gaussian distribution of mean* μ *and variance* σ^2, written $N(\mu, \sigma^2)$. To check that f integrates to 1 we can use Fubini's theorem and change variables into polar coordinates to obtain

$$\left(\int_{-\infty}^{+\infty} e^{-x^2/2}dx\right)^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-(x^2+y^2)/2}dxdy = \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}rdrd\theta = 2\pi.$$

2. Continuous random variables

A random variable $X : \Omega \to \mathbb{R}$ has probability density function f if

$$\mathbb{P}(X \in [a,b]) = \int_a^b f(x)dx$$

for all $a < b \in \mathbb{R}$. Note that $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$, in contrast with the discrete case.

We again define the distribution function of X by

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(y)dy.$$

Definition 2.1. *A random variable X is said to be* continuous *if the associated distribution function F_X is continuous.*

Note that by the fundamental theorem of calculus

$$F_X'(x) = f_X(x),$$

thus the probability density function is the derivative of the distribution function. Moreover,

$$\lim_{x\to-\infty} F_X(x) = 0, \qquad \lim_{x\to+\infty} F_X(x) = 1$$

and F_X is non-decreasing, since $F_X'(x) = f_X(x) \ge 0$.

Definition 2.2. *Let X be a continuous random variable with probability density function f and distribution function F. Then, provided*

$$\Big|\int_0^{+\infty} xf(x)dx\Big| < \infty, \qquad \text{or} \qquad \Big|\int_{-\infty}^{0} xf(x)dx\Big| < \infty,$$

we define

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

The above finiteness assumptions are needed to ensure that the integral defining $\mathbb{E}(X)$ is well defined.

Example 2.3. If $X \sim \mathit{Exponential}(\lambda)$ then we integrate by parts to get

$$\mathbb{E}(X) = \int_0^{\infty} x\lambda e^{-\lambda x}dx = -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}dx = \frac{1}{\lambda}.$$

EXAMPLE 2.4. If $X \sim \textit{Cauchy}$ then

$$\int_0^\infty \frac{x}{\pi(1+x^2)} dx = +\infty$$

and similarly

$$\int_{-\infty}^0 \frac{x}{\pi(1+x^2)} dx = -\infty,$$

so the expectation *is not defined.*

EXAMPLE 2.5. If $X \sim N(0,1)$ then

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi}} e^{-x^2/2} dx = 0$$

by symmetry.

The expectation satisfies the same properties as in the case of discrete random variables. In particular it is *linear*, and for a function $g : \mathbb{R} \to \mathbb{R}$ we have

$$\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx.$$

Moreover, if X is non-negative, we have the alternative formula

$$\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X \geq x) dx = \int_0^{+\infty} (1 - F(x)) dx.$$

EXAMPLE 2.6. If $X \sim \textit{Exponential}(\lambda)$ then $\mathbb{P}(X \geq x) = e^{-\lambda x}$, so we can use the above formula to compute

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geq x) dx = -\frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty = \frac{1}{\lambda}.$$

DEFINITION 2.7. *A random variable X is said to be* integrable *if*

$$\mathbb{E}(|X|) = \int_{-\infty}^{+\infty} |x| f(x) dx < \infty.$$

For X integrable random variable, we define the *variance* of X by

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \int (x - \mathbb{E}(X))^2 f(x) dx.$$

In analogy with the discrete case,

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \int x^2 f(x) dx - \left(\int x f(x) dx \right)^2.$$

EXAMPLE 2.8. If $X \sim N(0,1)$ then we integrate by parts to get

$$\mathrm{Var}(X) = \mathbb{E}(X^2) = \int_{-\infty}^{+\infty} \frac{x^2}{\sqrt{2\pi}} e^{-x^2/2} dx = -\frac{x}{\sqrt{2\pi}} e^{-x^2/2}\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = 1.$$

3. Exercises

Exercise 1. Let X be a random variable with p.d.f. $f(x) = ce^{-3x}\mathbb{1}_{[0,\infty)}(x)$. Find c. Find $\mathbb{P}(X \in [2,3])$, $\mathbb{P}(|X| \leq 1)$, $\mathbb{P}(X > 4)$.

Exercise 2. Compute mean and variance of X if $X \sim$Uniform$[a,b]$.

Exercise 3. If $X \sim \mathit{Exponential}(\lambda)$ compute the variance of X.

Exercise 4. Let X be a Uniform random variable in $[-1,1]$. Compute $\mathbb{E}(X^3)$ and $\mathbb{E}(X^4)$. Can you compute $\mathbb{E}(X^{2n+1})$ for $n \geq 0$?

Exercise 5. Let X be a continuous random variable with probability density function

$$f(x) = c(\alpha)x^{\alpha}\mathbf{1}_{[0,1]}(x),$$

for some $\alpha > 0$ and some constant $c(\alpha)$.

(i) Compute the constant $c(\alpha)$.
(ii) Compute $\mathbb{E}(X)$.
(iii) Determine α such that $\mathbb{E}(X) = 1/2$.
(iv) Assuming that $\alpha > 1$, compute $\mathbb{E}(1/X)$.

Exercise 6. The Gamma function is defined by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx.$$

Using integration by parts, show that for all k positive integers

$$\Gamma(k+1) = k\Gamma(k).$$

By comparison with the exponential distribution, explain why $\Gamma(1) = 1$. Use this to conclude that

$$\Gamma(k) = (k-1)!$$

for all k positive integers.

Exercise 7. [The Gamma distribution] For $\alpha, \lambda \in (0, +\infty)$, define

$$f(x) = \frac{\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}\mathbb{1}_{[0,+\infty)}(x)$$

where

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx.$$

Show that the f above defines a p.d.f.. This is the Gamma distribution of parameters α, λ. Note that by setting $\alpha = 1$ we recover the exponential distribution, of which the Gamma distribution is therefore a generalisation. Show that $\mathbb{E}(X) = \alpha/\gamma$.

Exercise 8. Let X be a random variable with p.d.f. $f(x) = cx\mathbb{1}_{[0,2]}(x)$. Find c. Find the distribution function of $Y = 2x$. Differentiate con compute the probability density function of Y.

Exercise 9. Let X be a continuous random variable with probability density function

$$f_X(x) = 3x^2\mathbf{1}_{[0,1]}(x).$$

(i) Write down the distribution function F_X of X.
(ii) Compute mean and variance of X.
(iii) Set $Y = 1 - X$. Compute the distribution function F_Y of Y, and hence the probability density function $f_Y(y)$.
(iv) Compute mean and variance of Y.

Exercise 10. A stick of unit length is cut in two random points U_1 and U_2 independent and uniformly distributed in $(0,1)$. Determine the probability $\mathbb{P}(\min(U_1, U_2) > x)$ for all $x \in \mathbb{R}$. Set $X = \min(U_1, U_2)$. Write down the distribution function F_X of X. Differentiate to compute the probability density function f_X of X.

Exercise 11. Let X be an exponential random variable of parameter λ. Show that for any $0 < s < t$ it holds

$$\mathbb{P}(X > t | X > s) = \mathbb{P}(X > t - s).$$

This is the *memoryless property* of the exponential distribution. Indeed, thinking of X as a random time (say the time of arrival of a bus), it tells us that given that $X > s$, the random time $X - s$ still has exponential distribution of parameter λ. In fact, this property characterises the exponential distribution, but we won't prove this.

Exercise 12. Show that if X is a standard Gaussian $X \sim N(0,1)$, then

$$\mathbb{E}(e^{\theta X}) = e^{\theta^2/2}.$$

Deduce that if $Y = \sigma X + \mu$ then

$$\mathbb{E}(e^{\theta Y}) = e^{\theta\mu + \frac{\theta^2\sigma^2}{2}}.$$

4. Transformations of one-dimensional random variables

Let X be a random variable taking values in $S \subseteq \mathbb{R}$ (that is, with probability density function f_X supported on S). We consider a function $g : S \to S'$ and define the random variable $Y = g(X)$. What is the probability density function f_Y of Y?

To answer this question, let us assume that g is strictly increasing, so that $g'(x) > 0$ for all x. Then we can look at the distribution function

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left(g(X) \leq y\right) = \mathbb{P}\left(X \leq g^{-1}(y)\right) = F_X(g^{-1}(y)).$$

Differentiating both sides, we find

$$f_Y(y) = \frac{d}{dy}F_X\left(g^{-1}(y)\right) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y).$$

A similar argument applies for the case of g decreasing.

Example 4.1. Let $X \sim Uniform[0,1]$. We may assume that X only takes values in $(0,1]$ since $\mathbb{P}(X=0)=0$. Define $Y = -\log X$, so that Y takes values in $[0,\infty)$. Then we have $S=(0,1]$, $S'=[0,\infty)$ and $g(x) = -\log x$ strictly decreasing. Then for $y \geq 0$

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(-\log X \leq y) = \mathbb{P}(X \geq e^{-y}) = 1 - e^{-y}.$$

By differentiating both sides with respect to y, we find

$$f_Y(y) = e^{-y},$$

so $Y \sim Exponential(1)$.

Example 4.2. Let $X \sim N(0,1)$ and for $\mu \in \mathbb{R}$ and $\sigma > 0$ define $Y = \sigma X + \mu$. Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\sigma X + \mu \leq y) = \mathbb{P}\Big(X \leq \frac{y-\mu}{\sigma}\Big) = F_X\Big(\frac{y-\mu}{\sigma}\Big),$$

from which, differentiating with respect to to y, we get

$$f_Y(y) = f_X\Big(\frac{y-\mu}{\sigma}\Big)\frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$

which shows that $Y \sim N(\mu,\sigma^2)$.

A similar argument shows that, conversely, if $X \sim N(\mu,\sigma^2)$ then $Y = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Note that the above example gives us a way to deduce the mean and variance of a $N(\mu,\sigma^2)$ from the ones of $N(0,1)$. Indeed, if $X \sim N(0,1)$ then $Y = \sigma X + \mu \sim N(\mu,\sigma^2)$, from which, using the linearity of expectation,

$$\mathbb{E}(Y) = \sigma\mathbb{E}(X) + \mu = \mu,$$

and

$$\mathrm{Var}(Y) = \sigma^2\mathrm{Var}(X) = \sigma^2.$$

5. Multivariate distributions

5.1. Joint distribution. We start by defining the joint probability density function of two real-valued random variables.

DEFINITION 5.1. *Two random variables* X, Y *are said to have joint probability density function* $f_{X,Y}$ *if*

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy$$

for all subsets $A \subseteq \mathbb{R}^2$. *Their joint distribution function is given by*

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v)dudv.$$

Properties of the joint probability density function $f_{X,Y}$

(i) $f_{X,Y}(x,y) \ge 0$,

(ii) $\displaystyle\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = 1$.

Note that $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ and $F_{X,Y} : \mathbb{R}^2 \to [0,1]$, and

$$\mathbb{P}((X,Y) \in (a,b) \times (c,d)) = \int_a^b \int_c^d f_{X,Y}(x,y)dxdy$$

for all $(a,b) \times (c,d) \subseteq \mathbb{R}^2$. Moreover,

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y),$$

so we can obtain the joint p.d.f. by differentiating the joint distribution function.

EXAMPLE 5.2. Let (X,Y) have joint p.d.f.

$$f_{X,Y}(x,y) = 6xy^2 \mathbb{1}_{[0,1]^2}(x,y).$$

We compute $\mathbb{P}((X,Y) \in A)$ where $A = \{(x,y) : x + y \ge 1\}$. We have

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy = \int_0^1 \int_{1-y}^1 6xy^2 dxdy = \int_0^1 (6y^3 - 3y^4)dy = \frac{9}{10}.$$

From the joint p.d.f. we can recover the marginals by integrating over one of the variables:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy, \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx.$$

EXAMPLE 5.3. Continuing with the above example, we find

$$f_X(x) = \int_0^1 6xy^2 dy = 2x \qquad \text{for } x \in [0,1],$$

and

$$f_Y(y) = \int_0^1 6xy^2 dx = 3y^2 \qquad \text{for } y \in [0,1].$$

Note that $f_{X,Y}(x,y) = f_X(x) f_Y(y)$. In this case we say that X and Y are *independent.*

DEFINITION 5.4. *Two random variables X, Y are said to be* independent *if*

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(Y \leq y)$$

for all $(x,y) \in \mathbb{R}^2$. Equivalently, they are independent if and only if their joint p.d.f. is given by the product of the marginals, that is

$$f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

EXAMPLE 5.5. Let X, Y have joint p.d.f.

$$f_{X,Y}(x,y) = \lambda e^{-\lambda x} \mathbb{1}_{[0,\infty)\times[0,1]}(x,y)$$

for $\lambda > 0$. Then

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,\infty)}(x), \qquad f_Y(y) = \mathbb{1}_{[0,1]}(y),$$

so $f_{X,Y} = f_X f_Y$ and X, Y are independent exponential random variables, with $X \sim exp(\lambda)$ and $Y \sim U[0,1]$.

All the above definitions generalise in the obvious way for the joint distribution of an arbitrary number of random variables. .

5.2. Covariance. Given two random variables X, Y and a function $g : \mathbb{R}^2 \to \mathbb{R}$, we have the formula

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y) dx dy.$$

In particular, taking $g(x,y) = xy$ we find

$$\mathbb{E}(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{X,Y}(x,y) dx dy.$$

DEFINITION 5.6 (Covariance). *Given two random variables X, Y with joint p.d.f. $f_{X,Y}$ we define their covariance as*

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Note that if X, Y are independent then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ and so

$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f_{X,Y}(x,y)dxdy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f_X(x)f_Y(y)dxdy \\ &= \left(\int_{-\infty}^{+\infty} x f_X(x)dx\right)\left(\int_{-\infty}^{+\infty} y f_Y(y)dy\right) = \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

which shows that $\mathrm{Cov}(X,Y) = 0$. The covariance satisfies all the properties that we have listed for the discrete case.

Example 5.7. Let X, Y be two random variables with joint p.d.f. $f_{X,Y}(x,y) = cxy$ for $0 \le x \le y^2 \le 1$ and $y \in [0,1]$. To ensure that $f_{X,Y}$ integrates to 1 we take $c = 12$. We find

$$\mathbb{E}(XY) = \int_0^1 \int_0^{y^2} 12x^2y^2 dxdy = 4\int_0^1 y^8 dy = \frac{4}{9},$$

and marginals

$$f_X(x) = \int_{\sqrt{x}}^1 12xydy = 6x(1-x), \qquad f_Y(y) = \int_0^{y^2} 12xydx = 6y^5,$$

from which $\mathbb{E}(X) = 1/2$ and $\mathbb{E}(Y) = 6/7$. It follows that $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{4}{9} - \frac{1}{2}\frac{6}{7} \neq 0$, and therefore X, Y are not independent.

6. Limit theorems

In this section we show how the Gaussian distribution arises in the limiting behaviour of sums of i.i.d. (independent and identically distributed) random variables. Recall the weak law of large numbers.

Theorem 6.1 (Weak law of large numbers). *Let $(X_k)_{k\ge 1}$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2, and define*

$$S_N = X_1 + X_2 + \cdots + X_N.$$

Then for all $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{S_N}{N} - \mu\right| > \varepsilon\right) \to 0$$

as $N \to \infty$.

We have already proved this result, as a corollary of Chebyshev's inequality. Once we know that $S_N/N \to \mu$ as $N \to \infty$, we could ask what is the error we make when approximating S_N by $N\mu$ for large N. This is the object of the next result, and it is where the Gaussian distribution comes in.

THEOREM 6.2 (Central Limit Theorem). *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2. Then for all $x \in \mathbb{R}$*

$$\mathbb{P}\left(\frac{S_N - N\mu}{\sigma\sqrt{N}} \leq x\right) \to \Phi(x)$$

as $N \to \infty$, where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

is the distribution function of a standard Gaussian random variable.

Thus the central limit theorem tells us that, for large N, we can approximate the sum S_N by a Gaussian random variable with mean $N\mu$ and variance $N\sigma^2$.

7. Exercises

Exercise 1. Let $X \sim \mathit{Exponential}(1)$. Find the p.d.f. of $Y = X^2$.

Exercise 2. Let $X \sim \mathit{Cauchy}$. Find the p.d.f. of $Y = \arctan X$.

Exercise 3. Conversely, show that if $X \sim U[-\pi/2, \pi/2]$ and $Y = \tan X$, then $Y \sim \mathit{Cauchy}$.

Exercise 4. Let $X \sim U[-1, 1]$. Find the p.d.f. of $Y = 2X + 5$ and $Z = X^3 + 1$.

Exercise 5. Let $X \sim U[0, 1]$, and let F be a strictly increasing distribution function. Show that the random variable $Y = F^{-1}(X)$ has distribution function F. *This is important for simulating random variables on a computer.*

Exercise 6. Show that if $X \sim N(\mu, \sigma^2)$ then $Y = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

Exercise 7. Let (X, Y) be uniformly distributed on $[0, 1]^2$, which is equivalent to saying that the joint p.d.f. of X, Y is given by

$$f_{X,Y}(x, y) = \mathbb{1}_{[0,1]^2}(x, y).$$

Show that $P(X < Y) = 1/2$ and $\mathbb{P}(X \leq 1/2, X + Y \leq 1) = 3/8$.

Exercise 8. Let X, Y be independent exponential random variables of parameter $\lambda > 0$. Compute $\mathbb{P}(X \geq z, Y \geq z)$. Deduce the distribution function $F_Z(z)$ of $Z = \min\{X, Y\}$. Differentiate to get the p.d.f. of Z: do you recognise it? Deduce $\mathbb{E}(Z)$ and $\mathrm{Var}(Z)$.

Exercise 9. Let X, Y have joint p.d.f. $f_{X,Y}(x, y) = c(x+2y)\mathbb{1}_{[0,1]^2}(x, y)$. Find c. Compute $\mathbb{P}(X < Y)$ and $\mathbb{P}(X + Y < 1/2)$. Find the marginals f_X and f_Y. Compute $\mathbb{E}(X)$, $\mathbb{E}(Y)$ and $\mathrm{Cov}(X, Y)$. Compute $\mathbb{E}(XY^2)$.

Exercise 10. Let X, Y have joint distribution function

$$F_{X,Y}(x, y) = 1 - e^{-x} - e^{-y} + e^{-(x+y)}$$

for $(x, y) \in [0, \infty)^2$. Find $\mathbb{P}(X < 4, Y < 2)$. Find $\mathbb{P}(X < 3)$. Differentiate to compute the joint p.d.f. $f_{X,Y}$ of X, Y. Check that $f_{X,Y}$ integrates to 1.

Exercise 11. Let X, Y have joint p.d.f. $f_{X,Y}(x, y) = cx^2 e^y \mathbb{1}_{\{0<x<y<1\}}(x, y)$. Find c. Find the marginals f_X and f_Y. Are X, Y independent? Compute $\mathbb{E}(XY)$, $\mathbb{E}(X^2 e^{-Y})$, $\mathbb{E}(X(1+4Y))$, $\mathbb{E}(e^{X+Y})$, $\mathbb{E}(e^{X-Y})$.

Exercise 12. The radius of a circle is exponentially distributed with parameter λ. Determine the probability density function of the perimeter of the circle and of the area of the circle.

Exercise 13. Let $Y \sim N(\mu, \sigma^2)$. Determine the p.d.f. of $X = e^Y$.

Exercise 14. Let X, Y have joint p.d.f. $f_{X,Y}(x, y) = \frac{1}{y} \mathbb{1}_{\{0<x<y<1\}}(x, y)$. Check that $f_{X,Y}$ integrates to 1. Find the marginal p.d.f.'s f_X and f_Y. Are X, Y independent? Compute $\mathbb{E}(XY)$, $\mathbb{E}(XY^k)$ for k positive integer.

APPENDIX A

Background on set theory

A *set* is a collection of elements, which can be finite or infinite. If A is a finite set, $|A|$ denotes the number of elements of A. The empty set is the set consisting of no element, and it is denoted by $\emptyset$. If A and B are two sets, then

- $A \cup B = \{x : x \in A \text{ or } x \in B\}$ (union).
- $A \cap B = \{x : x \in A \text{ and } B\}$ (intersection).
- $A \setminus B = \{x \in A \text{ but } x \notin B\}$ (difference).
- $A \subset B$ if all elements of A are also elements of B.
- If $A \cap B = \emptyset$ then we say that A and B are *disjoint*.

Note that the empty set is disjoint from every set, including itself. In fact, the empty set is the only set disjoint from itself. If Ω is a set, and $A \subseteq \Omega$, we define the *complement* of A in Ω as $A^c = \Omega \setminus A = \{x \in \Omega \text{ but } x \notin A\}$. Note that $A \cap A^c = \emptyset$ so A and A^c are disjoint.

EXAMPLE 0.1. Take $A = \{1, 2, 3\}$, $B = \{2, 4, 6, 8\}$. Then

$$A \cup B = \{1, 2, 3, 4, 6, 8\}, \quad A \cap B = \{2\}, \quad A \setminus B = \{1, 3\}, \quad B \setminus A = \{4, 6, 8\}.$$

If $C = \{6, 7, 8\}$ then

$$A \cup C = \{1, 2, 3, 6, 7, 8\}, \quad A \cap C = \emptyset, \quad A \setminus C = A, \quad A \cup B \cup C = \{1, 2, 3, 4, 6, 7, 8\}, \quad A \cap B \cap C = \emptyset.$$

Take $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Then $A, B, C \subseteq \Omega$. Their complements in Ω are

$$A^c = \{4, 5, 6, 7, 8, 9, 10\}, \quad B^c = \{1, 3, 5, 7, 9, 10\}, \quad C^c = \{1, 2, 3, 4, 5, 9, 10\},$$

We can also define infinite unions and intersections. For a sequence of sets $(A_n)_{n \in \mathbb{N}}$ set

$$\bigcup_{n \in \mathbb{N}} A_n = A_1 \cup A_2 \cup A_3 \cup \cdots = \{x : x \in A_n \text{ for some } n \in \mathbb{N}\},$$
$$\bigcap_{n \in \mathbb{N}} A_n = A_1 \cap A_2 \cap A_3 \cap \cdots = \{x : x \in A_n \text{ for all } n \in \mathbb{N}\}.$$

EXAMPLE 0.2. Take $A_n = \{n, n+1, n+2 \ldots\} = \{k \in \mathbb{N} : k \geq n\}$. Then

$$\bigcup_{n \in \mathbb{N}} A_n = \mathbb{N}, \qquad \bigcap_{n \in \mathbb{N}} A_n = \emptyset.$$

If on the other hand we take A_n as before but set $A_n = A_{10}$ for all $n \geq 10$ then

$$\bigcup_{n\in\mathbb{N}} A_n = \mathbb{N}, \qquad \bigcap_{n\in\mathbb{N}} A_n = A_{10} = \{10, 11, 12 \ldots\}.$$

Given two sets A, B, we define their product as

$$A \times B = \{(x, y) : x \in A, y \in B\}.$$

EXAMPLE 0.3. If $A = \{1, 2, 3\}$ and $B = \{\alpha, \beta\}$ then

$$\begin{aligned} A \times B &= \{(1, \alpha), (1, \beta), (2, \alpha), (2, \beta), (3, \alpha), (3, \beta)\}, \\ A \times A &= \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}, \\ B \times B &= \{(\alpha, \alpha), (\alpha, \beta), (\beta, \alpha), (\beta, \beta)\}. \end{aligned}$$

Note that the pairs are *ordered*, and that $|A \times B| = |A| \cdot |B|$. We can define the product of an arbitrary number of sets by setting

$$A_1 \times A_2 \times A_3 \cdots A_n = \{(x_1, x_2, x_3 \ldots x_n) : x_k \in A_k\}.$$

EXAMPLE 0.4. If $A = \{0, 1\}$, $B = \{a\}$ and $C = \{x, y\}$ then

$$A \times B \times C = \{(0, a, x), (0, a, y), (1, a, x), (1, a, y)\}.$$

Note that $|A \times B \times C| = 2 \cdot 1 \cdot 2 = 4$.