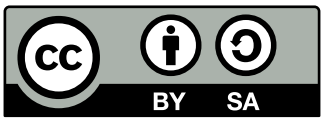
ST120 – Introduction to Probability

Lecture Notes

Department of Statistics
University of Warwick

Autumn 2022

Compiled on: 10th December 2022



Attribution

These notes were prepared by Leonardo T. Rolla and Tessy Papavasileiou in 2022. Although most of the material was written from scratch, big parts of it were taken from lecture notes for ST112, written by Daniel Valesin, and for ST115, produced by Giuseppe Cannizzaro. Those parts were nevertheless modified and adapted, and any mistakes found in the current material should be attributed to the current authors.

Contents

1 Uniform probability spaces

1.1 Probability encoded as sets and functions

Consider the following experiment: a box has 4 red balls, 6 blue balls and 10 green balls. We pick a ball *at random*. What is the *probability of* that ball being red? We are taught in school that this should be the number of red balls over the total number of balls, so $\frac{4}{20} = 0.2 = 20\%$ and this is indeed true under certain assumptions. To understand the assumptions we are implicitly making when doing this computation, we ask the following questions:

- What is a probability as a mathematical object?
- What other questions could we ask without changing the experiment?
- Would the answer be the same if some balls are harder to grasp (e.g. they are of difference sizes)?

Note that when we ask about a probability, we need to determine the event whose probability we are interested in – while the probability of a specific event (e.g. 'the ball is red') is a number in $[0, 1]$, *the probability on its own, is a map that attaches to each event a number.*

There are three possible outcomes for this experiment: red, blue and green. We call the set of all possible outcomes *sample space*, usually denoted by Ω. As a mathematical object, Ω is any non-empty set – in this case, $\Omega = \{red, blue, green\}$

We can, however, ask other questions as well. For example, we can ask for the probability that 'the ball is either blue or green' (which would have been equivalent to 'ball is not red'). In words, an event is a statement that you can tell whether it is true or not, after seeing the outcome of the experiment. In this case, all possible events are

- 'The ball is none of the three colours or any other colour' – mathematically, this will be denoted by the empty set $\emptyset$, since it contains none of the possible outcomes.
- 'The ball is red' – denoted by $\{red\}$
- 'The ball is blue' – denoted by $\{blue\}$
- 'The ball is green' – denoted by $\{green\}$
- 'The ball is either red or blue' – denoted by $\{red, blue\}$
- 'The ball is either red or green' – denoted by $\{red, green\}$

- 'The ball is either blue or green' – denoted by $\{blue, green\}$
- 'The ball is any of red, blue or green' – denoted by $\{red, blue, green\} = \Omega$.

From this exhaustive list, it is clear that all events are subset of Ω and in fact, in this case at least, all subsets of Ω are events. We call the collection of all events the *event space* (or, more formally, σ-algebra), usually denoted by $\mathcal{F}$. As a mathematical object, this is a collection of subsets of the sample space – we will later see that this collection has to satisfy certain properties but for now, we assume that it includes all subsets, so $\mathcal{F} = \mathcal{P}(\Omega)$, where $\mathcal{P}$ denotes the power-set (collection of all subsets).

We have already argued that the probability, usually denoted by $\mathbb{P}$, is a map from the event space to numbers $[0, 1]$ – we denote it by $\mathbb{P} : \mathcal{F} \to [0, 1]$. Using our intuition, however, we expect that

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(\{red\}) = 0.2$
- $\mathbb{P}(\{blue\}) = 0.3$
- $\mathbb{P}(\{green\}) = 0.5$
- $\mathbb{P}(\{red, blue\}) = 0.5$
- $\mathbb{P}(\{red, green\}) = 0.7$
- $\mathbb{P}(\{blue, green\}) = 0.8$
- $\mathbb{P}(\{red, blue, green\}) = 1$.

What implicit assumptions are we making when doing these computations, based on our intuition?

- The probability of the event $\emptyset$ that includes no outcomes should be 0.
- The probability of the event Ω that includes all outcomes should be 1.
- When an event can be broken down to the union of two disjoint events, then its probability should be the sum of the two probabilities, e.g. $\mathbb{P}(\{red, blue\}) = \mathbb{P}(\{red\}) + \mathbb{P}(\{blue\})$.

These are fundamental properties that a probability map should have.

In conclusion, the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ of sample space Ω, event space $\mathcal{F}$ and probability $\mathbb{P}$ form what is called a *probability space*.

1.2 Uniform probability space

In the previous example, the way we postulated the probability of each colour had an implicit assumption: that all balls have the same chance to be chosen. This is a correct assumption if the balls have the same weight, size, texture, etc.

What if instead coloured red, green or blue, the balls were numbered from 1 to 20? Under the same assumption (that the balls have the same weight, size, texture, etc.), we would have $\Omega = \{red_1, \ldots, red_4, blue_1, \ldots, blue_6, green_1, \ldots, green_{10}\}$ and $\mathbb{P}$ would be so that every ball has the same chance to be picked, that is, every element $\omega \in \Omega$ would have the same chance. Then, the event 'ball is red' would correspond to the event $A = \{red_1, \ldots, red_4\}$ and we would intuitively expect that the probability of getting a red ball would be equal to $\frac{4}{20}$, where 4 is the number of elements in A (and red balls) and 20 is the number of all possible outcomes.

When each outcome is equally likely, we have a uniform probability space.

Definition 1.1. A *uniform probability space* is defined as the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where

- Ω (the *sample space*) is a non-empty finite set of all possible outcomes of the experiment;
- $\mathcal{F}$ (the *event space*) is the collection of all events, given by the power-set $\mathcal{P}(\Omega)$ of Ω;
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a map from the event space to $[0, 1]$, satisfying

  - $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1\,,$ (1.1a)
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)\,,$ for every $A, B \in \mathcal{F}$ such that $A \cap B = \emptyset$ (finite additivity) (1.1b)
  - $\mathbb{P}(\{\omega\}) = \mathbb{P}(\{\tilde{\omega}\})$ for all $\omega, \tilde{\omega} \in \Omega$ (uniform) (1.1c)

As a result of the uniform assumption, computing the probability of any event comes down to computing the cardinality of the event. Remember that an event is a set (a subset of the sample space Ω) – the cardinality of a set is the number of its elements. For a set $A \subseteq \Omega$, it is denoted by $|A|$.

To formally define cardinality, we first need to define one-to-one correspondence between two sets: given two sets A, B we say that they are in a one-to-one

correspondence if there exists a bijective map between them i.e. a function $f : A \to B$ that is both injective and surjective.

Definition 1.2. A set A has cardinality $n \in \mathbb{N}$ if it is in a one-to-one correspondence with $\{1, 2, \ldots, n\}$ and A has cardinality 0 if $A = \emptyset$.

Proposition 1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a uniform probability space. Then for all $\omega \in \Omega$*

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}\,, \tag{1.2}$$

and for all $A \subseteq \Omega$ $(A \in \mathcal{F})$

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}\,. \tag{1.3}$$

Proof. Since $\forall\, \omega_1, \omega_2 \in \Omega$, $\mathbb{P}(\{\omega_1\}) = \mathbb{P}(\{\omega_2\})$, let $p \in [0, 1]$ s.t.

$$p := \mathbb{P}(\{\omega\}) \quad \forall\, \omega \in \Omega\,.$$

Since $\mathbb{P}$ is a probability measure

$$\begin{aligned} 1 = \mathbb{P}(\Omega) &= \sum_{\omega \in \Omega} \mathbb{P}(\{w\})] \quad \text{by (1.1b)} \\ &= \sum_{\omega \in \Omega} p = p \sum_{\omega \in \Omega} 1 = p|\Omega|\,. \end{aligned}$$

Therefore

$$p = \frac{1}{|\Omega|}\,.$$

showing (1.2). We see that (1.3) follows from (1.1b), as

$$\begin{aligned} \mathbb{P}(A) &= \sum_{\omega \in A} \mathbb{P}(\{\omega\}), \quad \text{by (1.1b)} \\ &= \sum_{\omega \in A} p = p \sum_{\omega \in A} 1 = p|A| = \frac{|A|}{|\Omega|}\,. \end{aligned}$$

□

Exercise 1.1. Consider an urn with 50 balls numbered 1 to 50. Assume that they are drawn uniformly at random. After defining a suitable probability space, determine the probability that the first ball drawn shows a number divisible by 12.

Solution: Define $(\Omega, \mathcal{F}, \mathbb{P})$ as follows: $\Omega = \{1, 2, 3, \ldots, 50\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}$ the uniform probability measure. Then the event in question is

$$E = \{12, 24, 36, 48\}\,.$$

By Proposition 1.1 (1.3)

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{4}{50} = \frac{2}{25}\,.$$

2 How to count

2.1 Basics of combinatorics

In general, formula (1.3) says that, in order to compute probabilities, we need to count. This section digs deeper into the problem of counting.

Example 2.1. If there are 30 people in a room, what is the probability of at least two of them to have the same birthday? (Assume that no one is born on February 29 and that any day has the same chance of being anyone's birthday).

To answer the question in Example 2.1, we need to compute the cardinality of the set of all possible combinations of birthdays as well as the cardinality of the set of all possible combination of birthdays where at least two are the same. How do we do that? We will need to use the the fundamental counting principle, which allows us to compute cardinalities of large and complex sets, where explicit counting is not possible.

First we start by identifying the fundamental rules of counting:

Correspondence Rule If A and B are in a one-to-one correspondence then $|A| = |B|$.

Addition Rule If $A_1, \ldots, A_n$ are pairwise disjoint subsets of some set then

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i|\,.$$

Fundamental Counting Principle Suppose that the elements of a finite set E can be determined in k successive steps, with n_1 possible choices for

step 1, n_2 possible choices for step 2, ..., n_k possible choices for step k. Suppose also that different choices lead to different elements. Then

$$|E| = n_1 \cdot n_2 \cdot \cdots \cdot n_k \,.$$

The set of all combinations of birthdays of 30 people appearing in Example 2.1 is an example of a set of ordered k-tuples of elements of a given set. More generally, this is defined as follows:

Definition 2.1. Let A be a finite set of cardinality $n \in \mathbb{N}$. A *sequence of length* $k \in \mathbb{N}$ *of elements of* A is an ordered k-tuple $(a_1, \ldots, a_k)$ s.t. $a_i \in A$, $i = 1, 2, \ldots k$. We denote by $S_{n,k}(A)$ the set of all sequences of length k of elements of A.

By "ordered" we mean that the order of the sequence matters, for example: $(a_1, a_2) \neq (a_2, a_1)$. Note also that repetitions of elements are allowed, for example $(1, 1)$ is a sequence of length 2 of elements of $\{1\}$.

Proposition 2.1. *Let A be a finite set of cardinality $n \in \mathbb{N}$. The set $S_{n,k}(A)$ of all sequences of length $k \in \mathbb{N}$ of elements of A has cardinality n^k, i.e.*

$$|S_{n,k}(A)| = n^k$$

Proof. To construct an arbitrary element $(a_1, a_2, \ldots, a_k)$ of $S_{n,k}(A)$, we perform the following steps

(a) choose the first value a_1. There are $n_1 = |A| = n$ ways to do this.
(b) choose the second value a_2. There are $n_2 = |A| = n$ ways to do this.

$$\vdots$$

k. I choose the k^{th} value a_k. There are $n_k = |A| = n$ ways to do this.

Thus we find

$$|S_{n,k}(A)| = n_1 \cdot n_2 \cdot \cdots \cdot n_k = \underbrace{n \cdot n \cdot \cdots \cdot n}_{k \text{ times}} = n^k \,.$$

□

To complete the Example 2.1, we also need to compute the cardinality of the

set of all possible birthdays where at least two are the same. It is easier and equivalent (why?) to compute the cardinality of the set where no two birthdays are the same. As there cannot be any repetition, this is an example of an 'ordering of length 30 of elements of $\{1, \ldots, 365\}$. More generally, we define orderings of length k as follows:

Definition 2.2. Let A be a finite set of cardinality $n \in \mathbb{N}$ and let $k \in \mathbb{N}$ such that $k \leqslant n$. An *ordering of length k of elements of A* is a sequence of length k of elements of A with no repetitions. We denote the set of orderings of length k of elements of A by $O_{n,k}(A)$. Thus we have

$$O_{n,k}(A) = \{(a_1, \ldots, a_k) : a_i \in A \, \forall i = 1, \ldots, k, \quad a_i \neq a_j \ \forall i \neq j\}.$$

Proposition 2.2. *Let A be a finite set of cardinality $n \in \mathbb{N}$ and $k \leqslant n$. Then*

$$|O_{n,k}(A)| = n(n-1)\ldots(n-k+1).$$

Proof. We determine an element of $O_{n,k}(A)$ by the following steps:

(a) We choose a_1. There are $n_1 = |A| = n$ choices for this.
(b) We choose a_2 such that $a_2 \neq a_1$. There are $n_2 = n - 1$ choices for this.
(c) We choose a_3 such that $a_3 \neq a_2$ and $a_3 \neq a_1$. There are $n_3 = n - 2$ choices for this.

$\vdots$

k. We choose a_k such that $a_k \neq a_i \ \forall \, i = 1, 2, \ldots, k-1$. There are $n_k = n - (k-1)$ choices for this.

By the fundamental counting principle

$$|O_{n,k}(A)| = n_1 n_2 \ldots n_k = n(n-1)\ldots(n-k+1).$$

□

Example 2.2 (continued). We can now answer the question of Example 2.1. First, we compute the probability that no two people have the same birthday, corresponding to event B. As discussed above, B is an ordering of length 30

($k = 30$ from a set of cardinality 365 ($n = 365$), so

$$\mathbb{P}(B) = \frac{|B|}{|\Omega|} = \frac{365 \times \cdots \times (365 - 30 + 1)}{365^{30}} \approx 0.29$$

The event that at least two people have the same birthday is the complement of event B, so the probability that at least two out of the 30 people have the same birthday will be close to 71%.

Remarks. We read $n!$ as n factorial. The following hold:

- $n!$ is the product of the first n natural numbers and by assumption $0! = 1$.
- $n!$ is the number of ways in which we can order the elements of a set of cardinality n or equivalently the number of ways to put the elements of a set of cardinality n in a row.
- $n(n-1)\dots(n-k+1) = \frac{n!}{(n-k)!}$ is the number of ways to put k elements of a set of cardinality n in a row.

Example 2.3. Now we consider a slightly different question to that of Example 2.1: what is the probability that *exactly* two people in the room with the same birthday? To construct such an example, we would need to

1. Choose the two people that have the same birthday.
2. Choose a day for their birthday.
3. Choose a day for everyone else's birthday, so that no other birthdays are the same.

In how many ways can we choose the two people that have the same birthday? We need to choose two numbers from $C = \{1, \dots, 30\}$ – this will be a sequence of length 2 with no repetition, but what is different to what we had before is that the order doesn't matter. Whether it is $(1, 2)$ or $(2, 1)$, it is still the same pair of people with the same birthday! To correct for that, we need to divide by all possible ways we can order the two elements – each such way will be a sequence of length 2 with no repetition, but now we are choosing from a set of just two points, so the set of all possibilities is $O_{2,2}$. This gives

$$\frac{|O_{30,2}|}{|O_{2,2}|} = \frac{30 \times 29}{2!} = \frac{30!}{28!2!}.$$

We denote this by $\binom{30}{2}$. Now, back to our question: we have computed the number of ways we can pick the two people with the same birthday. There are

are 365 ways to choose their birthday. For their remaining 28 people, there will be $364 \times \cdots \times (365 - 28)$ ways of picking their birthdays since they all need to be different. So, the total number of ways of selecting an outcome in the event 'exactly two people have the same birthday' is

$$\frac{30!}{28!2!} 365 \times 364 \times \cdots \times (365 - 28).$$

Finally, to compute the probability of exactly two people having the same birthday, we need to divide by the cardinality of all possible birthday combinations given by 365^{30}, which gives approximately 0.38 or 38%.

Choosing two people from a set of 30 is an example of a combination of 2 elemnts of $\{1, \ldots, 30\}$. More generally, we can ask for the number of combinations of k elements of a finite set A.

Definition 2.3. Let A be a finite set of cardinality $n \in \mathbb{N}$. A *combination of k elements of A* is a subset of A with k elements. We denote by $C_{n,k}(A)$ the set of combinations of k elements of A.

Proposition 2.3. *Let A be a finite set of cardinality $n \in \mathbb{N}$ and $k \leqslant n$. Then*

$$|C_{n,k}(A)| = \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Proof. Notice that an ordering of length k of elements of A can be obtained *uniquely* by the following steps

(a) Choose a combination of k elements of A. There are $n_1 = |C_{n,k}(A)|$ choices for this.
(b) Choose a permutation of these elements. By Corollary 2.1 there are $n_2 = k!$ choices for this.

By the fundamental counting principle,

$$|O_{n,k}(A)| = |C_{n,k}(A)| \cdot k!.$$

The above equation can be solved for the only unknown term, giving

$$|C_{n,k}(A)| = \frac{|O_{n,k}(A)|}{k!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

This concludes the proof. □

The number of orderings of length n when the cardinality of the set is also n is an important spacial case or orderings:

Definition 2.4. Let A be a finite set of cardinality $n \in \mathbb{N}$. An ordering of length n of elements of A is called a *permutation* of A.

Corollary 2.1. *Let A be a finite set of cardinality $n \in \mathbb{N}$. Then the number of permutations of the elements of A is $n(n-1)(n-2)\dots 1 = n!$.*

Proof. It suffices to take $k = n$ in Proposition 2.2. □

Exercise 2.1. A fair die is tossed 8 times. How many different outcomes can we have that contain the outcome 2 exactly three times and the outcome 3 exactly five?

Solution. Each possible outcome is completely determined by specifying which tosses result to a 2 – the remaining will be 3. For example, outcome $(2,3,2,2,3,3,3,3)$ is determined by the subset $\{1,3,4\}$ of set $\{1,2,3,4,5,6,7,8\}$, with the first corresponding to the positions of 2 and the second being the numbering of all tosses. Thus, to compute the number of possible outcomes, it is sufficient to compute the number of subsets of size 3 from a set of cardinality 8. This is exactly $C_{8,3}$, which is equal to 56. □

Example 2.4. Let's take exercise 2.1 a bit further: how would we compute the number of outcomes of 8 tosses of a die that contain exactly three 2's, three 4's and two 5's? As before, an outcome will be completely determined by specifying the positions of two out of the three possibilities, e.g. 2 and 4. For example, outcome $(2,4,5,2,2,5,4,4)$ corresponds to sets $A_2 = \{1,4,5\}$ and $A_4 = \{2,7,8\}$ where A_2 is the set of positions of outcome 2 and similarly for A_4. It follows that $A_5 = \{3,6\}$, as these are the only two positions left.

So, to answer the question, we need to count in how many ways we can pick a subset of $\{1,2,3,4,5,6,7,8\}$ of length 3 and then another subset of length 3 from the remaining 5 elements. We have $\binom{8}{3}$ choices for the first set and $\binom{5}{3}$ choices for the second. Applying the fundamental counting principle, we get

$$\binom{8}{3}\binom{5}{3} = \frac{8!}{5!3!}\frac{5!}{3!2!} = \frac{8!}{3!3!2!}.$$

The above is an example of counting the ways we can split a set into a fixed number of subsets. To define this formally, we first need to define a partition:

Definition 2.5. Let A be a set of cardinality $n \in \mathbb{N}$ and $r \in \mathbb{N}$ such that $r \leqslant n$. A *partition of A into r subsets* is a family $\{A_1, \ldots, A_r\}$ of subsets of A such that

(a) Every subset in the family is non-empty: $A_i \neq \emptyset \ \forall\, i = 1, 2, \ldots, r$.
(b) The subsets in the family are pairwise disjoint: $A_i \cap A_j = \emptyset \ \forall\, i \neq j$, $i, j = 1, 2, \ldots, r$.
(c) The union of all the subsets in the family is equal to A: $\bigcup_{i=1}^{r} A_i = A$.

Proposition 2.4. *Let A be a finite set with $|A| = n$. Let $r \in \mathbb{N}$, $r \leqslant n$. Then the number of partitions of A into r subsets $\{A_1, \ldots, A_r\}$ such that $|A_1| = k_1$, $|A_2| = k_2 \ldots$, $|A_r| = k_r$ ($k_1 + k_2 + \cdots + k_r = n$, $1 \leqslant k_i \leqslant n$) is given by*

$$\frac{n!}{k_1! k_2! \ldots k_r!}\,.$$

Proof. Every partition of A satisfying the assumptions can be uniquely determined via the following steps:

(a) Choose $A_1 \subseteq A$ such that $|A_1| = k_1$. There are $\binom{n}{k_1}$ choices for this step.
(b) Choose $A_2 \subseteq A$ such that $|A_2| = k_2$ and $A_1 \cap A_2 = \emptyset$ (which implies that $A_2 \subseteq A \setminus A_1$). There are $\binom{n-k_1}{k_2}$ choices for this step.
(c) Choose $A_3 \subseteq A$ such that $|A_3| = k_3$, $A_3 \cap A_2 = \emptyset$ and $A_3 \cap A_1 = \emptyset$ (which implies that $A_3 \subseteq A \setminus (A_1 \cup A_2)$). There are $\binom{n-k_1-k_2}{k_3}$ choices for this step.

$\vdots$

r. Finally choose the remaining set $A_r \subseteq A$ such that $|A_r| = k_r$ and $A_i \cap A_j = \emptyset$ for all $i = 1, 2, \ldots, r-1$ (we see that $A_r \subseteq A \setminus (A_1 \cup \cdots \cup A_r)$). There are $\binom{n-(k_1+k_2+\cdots+k_{r-1})}{k_r}$ choices for this step.

Thus, by the fundamental counting principle the number of partitions of A into r subsets $\{A_1, A_2, \ldots, A_r\}$ such that $|A_1| = k_1, \ldots, |A_r| = k_r$ with $k_1 + k_2 + \cdots + k_r = n$ is

$$\binom{n}{k_1}\binom{n-k_1}{k_2}\cdots\binom{n-(k_1+\cdots+k_{r-1})}{k_r}$$

$$= \frac{n!}{k_1!(n-k_1)!}\frac{(n-k_1)!}{k_2!(n-k_1-k_2)!}\cdots\frac{(n-(k_1+\cdots+k_{r-1}))!}{k_r!(n-(k_1+\cdots+k_r))!}$$

$$= \frac{n!}{k_1!k_2!\ldots k_r!},$$

where the last equality comes by cancellation of the fractions and noticing that $(n-(k_1+\cdots+k_r))! = (n-n)! = 0! = 1$. □

2.2 Sampling

Picking a subset out of a larger set is also called sampling. Sampling allows us to use information about a small group in order to make an inference on the properties or preferences of a larger group. It is a fundamental tool in statistics. When the population is 'homogeneous' (every person in the group is likely to have the same properties or preferences), any sample picked at random (with a uniform probability) will be representative of the group - we still need to use advanced mathematical tools to quantify the uncertainty in our inference about a population, given the size of the sample, but this is the subject for a different module. What we would like to understand now is how the probability structure on the samples varies when the population is mixed. To understand the question, consider the following.

Example 2.5. The lecturer of ST120 wants to know to what extend students have understood the concept of a probability space. Given the size of the class, it is not possible to ask every single student. Instead, they want to sample a small group of students and ask them. How should they pick the students?

A practical solution is to talk to some of the students in the lecture hall. However, there is a bias - students that come to lectures are more likely to understand the concepts! So, the lecturer decides to pick at random a number of student ID numbers and email them and let's assume that all students reply. This would be a representative group, if the class was homogeneous. However, we know that the class consists of two groups of n_1 Mathematics students and n_2 Computer Science students. To understand the bias, the lecturer needs to compute the probability that the group ends up with k_1 Mathematics students and k_2 Computer Science students. What would that be?

There are $\binom{n_1}{k_1}$ ways to choose k_1 Mathematics students and $\binom{n_2}{k_2}$ to choose k_2 Computer Science students. So, the cardinality of the set of all groups of k_1 Mathematics students and k_2 Computer Science students will be

$$\binom{n_1}{k_1}\binom{n_2}{k_2}$$

The cardinality of the set of all groups of k_1+k_2 students will be $\binom{n_1+n_2}{k_1+k_2}$. So, the probability of picking a group with k_1 Mathematics students and k_2 Computer Science students is

$$\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_1+n_2}{k_1+k_2}}.$$

The next step would be to use this probability in order to remove the bias, but they will need to consult the lecturers of more advanced statistics modules on how to do this!

The above is an example of sampling from a population of size $n \in \mathbb{N}$, which has $n_1 \in \mathbb{N}$ ($n_1 \leqslant n$) individuals of type 1 and $n_2 = n - n_1$ individuals of type 2. We draw a sample of size $k < N$ from the whole population 'at random' *without replacement* (i.e. an individual cannot be picked twice). Then, the probability of the sample containing k_1 individuals of type 1 and $k_2 = k - k_1$ individuals of type 2 is given by

$$\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n}{k}}. \tag{2.1}$$

Exercise 2.2. Construct the uniform probability space corresponding to sampling without replacement from a population with two types of individuals and prove (2.1). In other words, give the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ such that, by proposition 1.1, all the events $\omega \in \Omega$ have the same probability. The event of interest ω is the set of all the combinations containing k_1 numbers from 1 to n_1 and $k_2 = k - k_1$ numbers from $n_1 + 1$ to n.

Example 2.6. Now consider the following problem: an environmental biologist is studying the weight of a specific species of fish. To do that, they sample from that fish population, weigh the fish and then release them back to the wild. Assume that there is no way of knowing that a fish has already been sampled. Also, assume that the population consists of only two types of fish - male and female - which have slightly different average weights and this needs to be taken into consideration when de-biasing.

If the whole population of fish is n, with n_1 male and $n_2 = n - n_1$ females and the sample size is k, what is the probability of having k_1 males and k_2 females in the sample?

The difference with this example and example 2.5 is that now individuals can be picked more than once. As a result, it is no more sufficient to only determine the positions of type 1 individuals in the sample but we also need to specify the individuals so that we keep track of the ones chosen repeatedly. So, we go through the following process:

- Choose the position of males (the remaining positions will be taken by females) – there are $\binom{k}{k_1}$ choices.
- Choose the males that are picked – there are $n_1^{k_1}$ choices.
- Choose the females that are picked – there are $n_2^{k_2}$ choices.

So, the number of different samples with k_1 males and k_2 females will be

$$\binom{k}{k_1} n_1^{k_1} n_2^{k_2}.$$

Given that the total number of possible samples are n^k, the probability of picking k_1 males and k_2 females will be given by

$$\frac{\binom{k}{k_1} n_1^{k_1} n_2^{k_2}}{n^k} = \binom{k}{k_1} \left(\frac{n_1}{n}\right)^{k_1} \left(1 - \frac{n_1}{n}\right)^{k-k_1}. \tag{2.2}$$

The above is an example of sampling from a population of size $n \in \mathbb{N}$, which has $n_1 \in \mathbb{N}$ ($n_1 \leqslant n$) individuals of type 1 and $n_2 = n - n_1$ individuals of type 2. We draw a sample of size $k < N$ from the whole population 'at random' *with replacement* (i.e. an individual can be picked twice). Then, the probability of the sample containing k_1 individuals of type 1 and $k_2 = k - k_1$ individuals of type 2 is given by (2.2).

Exercise 2.3. Construct the uniform probability space corresponding to sampling with replacement from a population with two types of individuals and prove (2.2). Now the event of interest ω is the set of all the sequences containing k_1 numbers from 1 to n_1 and $k_2 = k - k_1$ numbers from $n_1 + 1$ to n.

Exercise 2.4. The sampling probabilities should not be sensitive to what type

was deemed '1' and that type was deemed '2'. Indeed,

$$\binom{k}{k_1}\left(\frac{n_1}{n}\right)^{k_1}\left(1-\frac{n_1}{n}\right)^{k-k_1} = \binom{k}{k_2}\left(\frac{n_2}{n}\right)^{k_2}\left(1-\frac{n_2}{n}\right)^{k-k_2}.$$

Assuming that $k_1 + k_2 = k$ and $n_1 + n_2 = n$, check the above identity.

3 Probability spaces

In section 1, we defined a uniform probability space as the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a finite set, $\mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfies properties (1.1a) - (1.1c). We will now generalise the concept of a probability space, to allow for

- Arbitrary sample spaces Ω.
- Event spaces that reflect partial information – it is not necessary or indeed sometimes possible that $\mathcal{F} = \mathcal{P}(\Omega)$.
- Probability defined on a general event space.

3.1 Sample space and event space

Definition 3.1. A *sample space* Ω is the set of all possible outcomes of a random process (or experiment), i.e. a process whose outcome cannot be determined in advance. It can be any set.

Example 3.1. What is the sample space corresponding to the following processes?

- The flip of a coin: $\Omega = \{H, T\}$.
- The roll of a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- The number of emails send by a @google.com address in a year: $\Omega = \mathbb{N}$.
- The weight of an apple: $\Omega = [0, 1]$.
- The position of a dart thrown onto a square board of size 1: $\Omega = [0, 1] \times [0, 1]$.
- The price of a Twitter stock in a year: $\Omega = \mathbb{R}$.
- The temperature fluctuations at Coventry in 2023: in this case, the sample space is a whole function, mapping time t to a number in $[-50, 50]$.
- The state of the world in a year! In this case, the sample space cannot be described, but it exists as a concept and it can be partially observed through its interaction with processes that can be measured (e.g. it will

affect interest rates or number of hospital admissions in a certain day in the future, frequency of extreme weather events etc).

Remark 1. While Ω can be any set in theory, in ST120 we will only consider the cases where the cardinality of Ω is equal to n, for some $n \in \mathbb{N}$ (finite state space), $|\Omega| = |\mathbb{N}|$ (countable state space) or $|\Omega| = |\mathbb{R}|$ (uncountable or continuous state space – to include intervals or Cartesian products of $\mathbb{R}$ and its intervals).

As we have already seen in the case of the uniform probability space, events are subsets of Ω. More generally, an event is a subset of Ω when it is possible to say whether any given outcome belongs to the set (i.e. 'the event happened') or not, *given the information we have about the outcome* – note that we don't always have full information about the outcome and the event space reflects the information that we do have. The following example demonstrates exactly this property of the event space.

Example 3.2. Suppose that I roll a die ($\Omega = \{1, 2, 3, 4, 5, 6\}$) and I report the following information to two students: I tell James if the outcome is an even number or not and I tell Lily the quotient of $(\omega - 1)$ divided by two (i.e., I report 0 for $\{1, 2\}$, 1 for $\{3, 4\}$ and 2 for $\{5, 6\}$). What are the corresponding event spaces?

James only knows whether the outcome is odd or even, so he can only tell whether it belongs to $\{1, 3, 5\}$ or $\{2, 4, 6\}$. Since, by construction of Ω, all outcomes are in Ω, he can also tell that the outcome is in Ω and not in $\emptyset$ – note that both $\emptyset$ and Ω are subsets of Ω. So, the collection of events (i.e. the event space) corresponding to the information that James has is

$$\mathcal{F}_J = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}\,.$$

(Think about why James cannot tell with certainty whether the outcome is in any other subset).

Based on the information given to her, Lily will be able to tell whether the outcome is in $\{1, 2\}, \{3, 4\}$ or $\{5, 6\}$. She can also tell whether the outcome will be in $\{1, 2, 3, 4\}, \{1, 2, 5, 6\}$ or $\{3, 4, 5, 6\}$ (e.g. $\{1, 2, 3, 4\}$ corresponds to the number reported to Lily being 0 or 1) and she can tell, by default that the outcome will be in Ω and not in $\emptyset$ (the $\emptyset$ is defined as the complement of Ω with respect to Ω, so all points in Ω that are not in Ω, which is of course none! It is easier to

think about it as the event where the outcome is not in the sample space rather than the event when nothing happens). So, the event space corresponding to information given to Lily is

$$\mathcal{F}_L = \{\emptyset, \{1,2\}, \{3,4\}, \{5,6\}, \{1,2,3,4\}, \{1,2,5,6\}, \{3,4,5,6\}, \Omega\}\,.$$

Note that there is redundancy in the way information is encoded in the event space – information corresponds to knowing simultaneously the answer to 'event happened or not' for all events in the event space but knowing, for example, that both $\{1,2\}$ and $\{1,2,3,4\}$ have happened allows us to deduce the answer for everything else. What assumption are we making that allows us to say this?

If A and B are events, then according to out intuition we expect the following to be events as well:

- $A \cap B$ (Both A and B have happened).
- $A \cup B$ (Either A or B has happened).
- A^c (A has not happened).
- $A \setminus B = A \cap B^c$ (A has happened but B has not).

Notation. When we write A^c, it is implicit that we are taking complement with a given sample space Ω. A more explicit notation is to write $\Omega \setminus A$.

So, we would like the event space to be closed under the operations of union, intersection, complement and difference (when we say that a set is closed under an operation, we mean that if we apply the operation to any elements in the set, the outcome is still in the set). Is that sufficient? Let us consider the following

Example 3.3. Consider the case where $\Omega = \mathbb{N}$ (i.e. any natural number can be the outcome of the random process that we consider) and suppose that we have sufficient information to say whether event $\{n\}$ happened or not, for any $n \in \mathbb{N}$. According to our intuition, if we can tell whether any outcome ω belongs to any set $\{n\}$ or not (what can you tell for ω when $\omega \in \{n\}$?), then we should also be able to tell whether $\omega \in \{2n | n \in \mathbb{N}\}$ or not, i.e. we should be able to tell whether ω is even. So, we would expect

$$\bigcup_{n=0}^{\infty} \{2n\} = \{2n | n \in \mathbb{N}\}$$

to be in the event space. We are now making the assumption that the event space is not just closed to unions but also to unions of countably (i.e infinite but with cardinality equal to $|\mathbb{N}|$) many sets.

Following our intuition, we define the event space as follows

Definition 3.2. Let Ω be the sample space and $\mathcal{F}$ be a collection of subsets of Ω. $\mathcal{F}$ is an *event space* (also called σ*-algebra*) if it satisfies

(*i*) $\Omega \in \mathcal{F}$.
(*ii*) if $A \subseteq \Omega$, $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ ($\mathcal{F}$ is closed under compliments).
(*iii*) if $\{A_n : n \in \mathbb{N}\}$ is such that $A_n \in \mathcal{F}\ \forall\, n$ then

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}\,.$$

($\mathcal{F}$ is closed under countable unions.)

Exercise 3.1. Let Ω be a non-empty set and $\mathcal{F} = \mathcal{P}(\Omega)$ i.e. the set of all subsets of Ω. Then $\mathcal{F}$ is an event space on Ω.

Solution. We need to show that $\mathcal{F} = \mathcal{P}(\Omega)$ satisfies the three properties in 3.2.

(*i*) $\Omega \subseteq \Omega$ and thus $\Omega \in \mathcal{P}(\Omega) = \mathcal{F}$.
(*ii*) Let $A \subseteq \Omega$, $A \in \mathcal{F}$. Then, $A^c = \Omega \setminus A \subseteq \Omega$. Thus, $A^c \in \mathcal{P}(\Omega) = \mathcal{F}$.
(*iii*) Let A_n be such that $A_n \subseteq \Omega$, $\forall n \in \mathbb{N}$. Then

$$\bigcup_{n=1}^{\infty} A_n \subseteq \bigcup_{n=1}^{\infty} \Omega = \Omega \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{P}(\Omega) = \mathcal{F}\,.$$

Exercise 3.2. Let $A \subseteq \Omega$ be a non-empty subset of Ω. Then

$$\{\emptyset, A, A^c, \Omega\}$$

is an event space on Ω.

Solution. We need to show that $\{\emptyset, A, A^c, \Omega\}$ satisfies the three properties in 3.2.

(*i*) $\Omega \in \mathcal{F}$ by definition.
(*ii*) We check that the property holds for each event: $\emptyset^c = \Omega \in \mathcal{F}$, $A^c \in \mathcal{F}$, $(A^c)^c = A \in \mathcal{F}$, $\Omega^c = \emptyset \in \mathcal{F}$.
(*iii*) Let $\{B_n : n \in \mathbb{N}\}$ be a sequence of subsets of Ω such that $B_n \in \mathcal{F}\ \forall\, n$. We go through all possibilities:

- If all the sets B_n are identical and equal B (so $B \in \mathcal{F}$), or there is a subsequence B_{n_k} of sets that are identical to B with the rest being all $\emptyset$, then $\bigcup_{n=1}^{\infty} B_n = B \in \mathcal{F}$.
- If at least one of the sets is the sample space (e.g. $B_i = \Omega$, for some $i \in \mathbb{N}$) then
$$\bigcup_{n=1}^{\infty} B_n = \Omega$$
and thus $\bigcup_{n=1}^{\infty} B_n \in \mathcal{F}$.
- if there is at least one set A and one set A^c, then
$$\Omega = A \cup A^c \subseteq \bigcup_{n=1}^{\infty} B_n \subseteq \Omega,$$
where the last relationship follows from the fact that all events are subsets of Ω. It follows that $\bigcup_{n=1}^{\infty} B_n = \Omega$ and thus $\bigcup_{n=1}^{\infty} B_n \in \mathcal{F}$.

Proposition 3.1. *Let $\mathcal{F}$ be a event space on Ω. Then*

(a) *$\mathcal{F}$ is closed under finite unions.*
(b) *$\mathcal{F}$ is closed under finite intersections.*
(c) *$\mathcal{F}$ is closed under countable intersections.*

Proof.

(a) Let $A_1, \ldots, A_n \in \mathcal{F}$. Set $A_j = \emptyset \in \mathcal{F} \; \forall \; j > n$, thus $A_n \in \mathcal{F} \, \forall \, n \geqslant 1$. Since the empty sets will not contribute to the union, we can show that
$$\bigcup_{j=1}^{n} A_j = \bigcup_{j=1}^{\infty} A_j \in \mathcal{F},$$
as $\mathcal{F}$ is closed under countable unions.

(b) Let $A_1, \ldots, A_n \in \mathcal{F}$. We want to show that $\bigcap_{j=1}^{n} A_j \in \mathcal{F}$. By De Morgan's law (show that every element of one set needs to belong to the other set as well)
$$\bigcap_{j=1}^{n} A_j = \left(\bigcup_{j=1}^{n} A_j^c \right)^c.$$
Since the event space $\mathcal{F}$ is closed under taking complemets, $A_j^c \in \mathcal{F}$, for every $j = 1, \ldots, n$. Since it is closed under finite unions (statement 1 of

the proposition, shown above),

$$\bigcup_{j=1}^{n} A_j^c \in \mathcal{F}$$

By taking complement once more, it follows that $\bigcap_{j=1}^{n} A_j \in \mathcal{F}$.

(c) The proof is similar to that of statement 2 above, noting that De Morgan's law also holds for countable unions and intersections. That is, we can write

$$\bigcap_{j=1}^{\infty} A_j = \left(\bigcup_{j=1}^{\infty} A_j^c \right)^c .$$

□

3.2 Probability

The last element of the probability space triplet is the probability measure $\mathbb{P}$. In definition 1.1 of the uniform probability space, we defined the probability measure as a map from the event space to $[0, 1]$, such that the probability of the event Ω ('the outcome is in the sample space) is 1 and for two disjoint events A, B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ – the later property can be generalised by induction to finite additivity: if $A_1, \ldots, A_n$ are disjoint events, then

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

Is this sufficient for infinite probability spaces? Let us consider the following

Example 3.4. Let $\Omega = \mathbb{N}^* = \{1, 2, \ldots\}$ be the positive natural numbers and $\mathcal{F} = \mathcal{P}(\Omega)$. Suppose that $\mathbb{P}(\{n\}) = \frac{1}{2^n}$, for every $n \geqslant 1$. What would we expect the event $\{2n | n \geqslant 1\}$ ('the outcome is an even number') to be?

Intuitively, what we would do is to sum up the probabilities corresponding to the outcome being even, i.e.

$$\mathbb{P}(\{2n | n \geqslant 1\}) = \sum_{n=1}^{\infty} \mathbb{P}(\{2n\}) = \sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \sum_{n=1}^{\infty} \frac{1}{4^n} = \frac{1}{3}.$$

(Note that the event $\{n\}$ corresponds to 'the outcome is n'). The computation

above cannot be justified, unless we extend the property of finite additivity to also hold for countable unions of disjoint events. Indeed, this is what we do!

Definition 3.3 (Probability measure)**.** Given a sample space Ω and an event space $\mathcal{F}$, a function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a *probability measure* if it satisfies

(*i*) $\mathbb{P}(B) \in [0, 1]$ for every $B \in \mathcal{F}$;
(*ii*) $\mathbb{P}(\Omega) = 1$;
(*iii*) (Countable additivity) For every $A_n \in \mathcal{F},\ n \geqslant 1$ disjoint events (i.e. for all $m, n \geqslant$ such that $m \neq n$, $A_m \cap A_n = \emptyset$,

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

We now give the definition of an abstract probability space.

Definition 3.4. A *probability space* is defined as the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where

- Ω (the *sample space*) is the set of all possible outcomes of the experiment (we always assume that it is not empty);
- $\mathcal{F}$ is an event space of subsets of Ω.
- $\mathbb{P}$ is a probability measure on $\mathcal{F}$.

Proposition 3.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then, $\mathbb{P}$ has the following properties*

(a) *If $A, B \in \mathcal{F}$ such that $A \subseteq B$, then*

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

Note that $B - A = B \cap A^c$ and is to be interpreted as 'all elements of B that are not in A.

(b) *For every $A \in \mathcal{F}$,*

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

(c) $\mathbb{P}(\emptyset) = 0$.

Proof.

(a) We write B as the union of the set will those elements in B that are not

in A and those that are, so $B = (B - A) \cup A$. By finite additivity

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

which proves the claim.

(b) Using the above property,

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

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

□

We can use countable additivity to compute the probability of a union of disjoint events. How can we compute the probability of any union of events? The following proposition gives us a way to do this.

Proposition 3.3 (Inclusion-Exclusion Formula)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then, for any finite collection $A_1, \ldots, A_n$ of events in $\mathcal{F}$, we have*

$$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}).$$

Remark 2. Formula 3.3 above uses concise notation and is not straight forward to interpret. To understand it better, let us consider some specific cases.

$n = 2$

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k=1}^{2} A_k\right) &= \sum_{k=1}^{2} (-1)^{k-1} \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant 2} \mathbb{P}(A_1 \cap \cdots \cap A_k) \\
&= \sum_{1 \leqslant i \leqslant 2} \mathbb{P}(A_i) - \sum_{1 \leqslant i_1 < i_2 \leqslant 2} \mathbb{P}(A_{i_1} \cap A_{i_2}) \\
&= \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2).
\end{aligned}$$

$n = 3$

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k=1}^{3} A_k\right) &= \sum_{k=1}^{3} (-1)^{k-1} \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant 3} \mathbb{P}(A_1 \cap \cdots \cap A_k) \\
&= \sum_{1 \leqslant i \leqslant 3} \mathbb{P}(A_i) - \sum_{1 \leqslant i_1 < i_2 \leqslant 3} \mathbb{P}(A_{i_1} \cap A_{i_2})
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{1 \leqslant i_1 < i_2 < i_3 \leqslant 3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\
&= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) \\
&\quad - \mathbb{P}(A_2 \cap A_3) + \mathbb{P}(A_1 \cap A_2 \cap A_3)\,.
\end{aligned}
$$

So, the sum $\sum_{1 \leqslant i_1 < \cdots < i_k \leqslant 2}$ is to be interpreted as the sum going through all k-tiples $(i_1, \ldots, i_k)$ of numbers $\{1, \ldots, n\}$ with no repetition (inqualities are strict). As we have seen in section 2, there are $\binom{n}{k}$ such k-tiples, so the sum will have $\binom{n}{k}$ summands.

Proof. We will prove the result only for $n = 2$ (the proof of the induction step in the general case is similar, but messier!). We write

$$A_1 \cup A_2 = (A_1 - B) \cup B \cup (A_2 - B),$$

where $B = A_1 \cap A_2$. The sets $A_1 - B, B, A_2 - B$ are all disjoint, so we can write

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}\left((A_1 - B) \cup B \cup (A_2 - B)\right) = \mathbb{P}(A_1 - B) + \mathbb{P}(B) + \mathbb{P}(A_2 - B)$$

using finite additivity. We know by proposition 3.2 that $\mathbb{P}((A_1 - B) = \mathbb{P}(A_1) - \mathbb{P}(B)$ and similarly $\mathbb{P}(A_2 - B) = \mathbb{P}(A_2) - \mathbb{P}(B)$. Replacing these to the formula above, we get

$$\mathbb{P}(A_1 \cup A_2) = (\mathbb{P}(A_1) - \mathbb{P}(B)) + \mathbb{P}(B) + (\mathbb{P}(A_2) - \mathbb{P}(B)) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(B),$$

which proces the claim. □

Proposition 3.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $A, B \in \mathcal{F}$ and $A \subseteq B$, then*

$$\mathbb{P}(A) \leqslant \mathbb{P}(B)\,.$$

Proof. Since $A \subseteq B$, it follows that $\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$ or, equivalently, $\mathbb{P}(A) = \mathbb{P}(B) - \mathbb{P}(B - A) \leqslant \mathbb{P}(B)$, where the inequality follows from the fact that probabilities are always non-negative. □

Proposition 3.5 (Boole's inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If*

$A_1, \ldots, A_n \in \mathcal{F}$, *then*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} \mathbb{P}(A_i) \quad (*)$$

Proof. We proceed by induction. For $n = 2$, notice that

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \underbrace{\mathbb{P}(A_1 \cap A_2)}_{\geqslant 0} \leqslant \mathbb{P}(A_1) + \mathbb{P}(A_2)\,.$$

so that (*) holds for $n = 2$. Assume now (*) holds $\forall\, j \leqslant n$, then we need to prove it holds for $n + 1$ events. Let $A_1, \ldots, A_{n+1} \in \mathcal{F}$, then arguing as above we have

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{j=1}^{n+1} A_j\right) &= \mathbb{P}\left(\left(\bigcup_{j=1}^{n} A_j\right) \cup A_{n+1}\right) \\
&= \mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) + \mathbb{P}\left(A_{n+1}\right) - \mathbb{P}\left(\left(\bigcup_{j=1}^{n} A_j\right) \cap A_{n+1}\right) \\
&\leqslant \mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) + \mathbb{P}(A_{n+1}) \\
&\leqslant \sum_{j=1}^{n} \mathbb{P}(A_j) + \mathbb{P}(A_{n+1}) = \sum_{j=1}^{n+1} \mathbb{P}(A_j)\,.
\end{aligned}$$

□

4 Conditional probability and independence

4.1 Conditional probability

Example 4.1. Suppose that before rolling a fair die, you bet one pound that the outcome is 3. Your friend sees the result before you and tells you that the die shows an even number. Would you continue your bet or withdraw from it? What about if you are told that the outcome is odd? How does this partial information about the outcome changes the probability?

We model the probability space corresponding to rolling a fair die by taking $\Omega = \{1, 2, \ldots, 6\}$, $\mathcal{F}$ to be the power set and $\mathbb{P}$ as the uniform probability on it.

Then, the event our friend tells us happened is

$$B = \{2, 4, 6\}, \quad \text{and its probability is} \qquad \mathbb{P}(B) = \frac{3}{6} = \frac{1}{2} > 0$$

The favourable event for us is

$$A = \{3\} \quad \text{and its probability is} \qquad \mathbb{P}(A) = \frac{1}{6}\,.$$

Knowing that the outcome is even can be interpreted as changing the sample space from Ω to B. Intuitively, we would assume that the probability on the new sample space remains uniform, but the probability of each outcome changes from $\frac{1}{6}$ to $\frac{1}{3}$ since there are now only 3 outcomes. Given that our preferred outcome 3 is not in the new sample space, we would expect the probability of getting 3 to be 0 and thus it would make sense to withdraw from the bet. If, in the other hand, we were told that the outcome is odd, then we could reformalise the probability space as one with sample space $B^c = \{1, 3, 5\}$ and we would expect the probability of winning the bet to be $\frac{1}{3}$, as it is one of 3 possible outcomes.

What about if we bet on $\{2, 3\}$? Then we would be looking at the number of ways we can still win, divided by the number of outcomes still possible. So, according to our intuition, we would expect the updated probability given that event B happened to be

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

But is this a well-defined probability?

Proposition 4.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ s.t. $\mathbb{P}(B) > 0$. Let $\mathbf{P}_B : \mathcal{F} \to \mathbb{R}$ such that*

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

Then $\mathbf{P}_B$ is a probability measure.

Proof. We need to check that all properties of probability measures are satisfied.

(*i*) First, we need to show that the $\mathbf{P}_B$ is defined for every $A \in \mathcal{F}$ and it takes values on $[0, 1]$, i.e. $\mathbf{P}_B$ is a map from $\mathcal{F}$ to $[0, 1]$ as it should.

- Let $A \in \mathcal{F}$. We have assumed that $B \in \mathcal{F}$ and thus $A \cap B \in \mathcal{F}$ as the event space is closed under intersections. So $\mathbb{P}(A \cap B)$ is well defined and since $\mathbb{P}(B) > 0$, their ratio $\mathbf{P}_B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ is well defined.
- $A \cap B \subseteq B$ and thus $\mathbb{P}(A \cap B) \leqslant \mathbb{P}(B)$ (proposition 3.4). It follows that $\mathbf{P}_B(A) \leqslant 1$. Similarly, since $\mathbb{P}(A \cap B) > 0$ and $\mathbb{P}(B) > 0$, it follows that $\mathbf{P}_B(A) > 0$. So, $\mathbf{P}_B(A) \in [0, 1]$.

(*ii*)
$$\mathbf{P}_B(\Omega) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1\,,$$
as required.

(*iii*) (countable additivity) Let $A_n \in \mathcal{F}$ for every $n \geqslant 1$, such that $A_n \cap A_m = \emptyset$ for all $n \neq m$ (disjoint events). Then
$$\begin{aligned}\mathbf{P}_B\left(\bigcup_{n=1}^{\infty} A_n\right) &= \frac{1}{\mathbb{P}(B)}\mathbb{P}\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \cap B\right) \\ &= \frac{1}{\mathbb{P}(B)}\mathbb{P}\left(\bigcup_{n=1}^{\infty} (A_n \cap B)\right) \quad (**)\end{aligned}$$

Now, since $A_n \in \mathcal{F}$ and $B \in \mathcal{F}$ for all $n \geqslant 1$, it follows that $A_n \cap B \in \mathcal{F}$ for all $n \geqslant 1$. Moreover, the events $A_n \cap B$ are disjoint. Indeed, for $n \neq m$
$$(A_n \cap B) \cap (A_m \cap B) \subseteq A_n \cap A_m = \emptyset\,.$$

Since $\mathbb{P}$ is a probability measure, it is countably additive, which implies
$$(**) = \quad \frac{1}{\mathbb{P}(B)} \sum_{n=1}^{\infty} \mathbb{P}(A_n \cap B) = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \mathbb{P}(A_n|B) = \sum_{n=1}^{\infty} \mathbf{P}_B(A_n)\,.$$

□

Definition 4.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. For $A \in \mathcal{F}$, the *conditional probability of A given B* is denoted by $\mathbb{P}(A|B)$ and is defined as
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad (*)$$

Exercise 4.1. An experiment consists of tossing a fair coin 7 times.

(a) Describe the probability space associated to it.

(b) Let E be the event corresponding to getting a prime number of heads. What is $\mathbb{P}(E)$?

(c) Let B be the event "H occurs at least 6 times". What is $\mathbb{P}(E|B)$?

Solution.

(a) $\Omega = \{(a_1, \ldots, a_7) : a_i \in \{H, T\}\} = S_{2,7}(\{H, T\})$, $\mathcal{F}$ the power set of Ω and $\mathbb{P}$ the uniform probability, i.e. $\mathbb{P}$ is such that

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

Recall that $|\Omega| = |S_{2,7}(\{H, T\})| = 2^7$.

(b) For $i = 1, \ldots, 7$ let A_i be the event "we get exactly i heads". The elements of A_i can be uniquely characterised via the position H appearing in the sequence. Hence by the fundamental counting principle $|A_i| = \binom{7}{i}$. Thus

$$\mathbb{P}(A_i) = \frac{1}{2^7}\binom{7}{i}$$

Now, notice that $A_i \cap A_j = \emptyset$ for $i \neq j$ (no outcome has both i and j heads) and

$$E = A_2 \cup A_3 \cup A_5 \cup A_7\,.$$

Then, by finite additivity

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(A_2) + \mathbb{P}(A_3) + \mathbb{P}(A_5) + \mathbb{P}(A_7)\\ &= \binom{7}{2}\frac{1}{2^7} + \binom{7}{3}\frac{1}{2^7} + \binom{7}{5}\frac{1}{2^7} + \binom{7}{7}\frac{1}{2^7} = \frac{78}{128}\,.\end{aligned}$$

(c) B is the event "H appears at least 6 times", so $B = A_6 \cup A_7$. Notice that,

$$\begin{aligned}\mathbb{P}(B) = \mathbb{P}(A_6) + \mathbb{P}(A_7) &= \binom{7}{6}\frac{1}{2^7} + \binom{7}{7}\frac{1}{2^7}\\ = \frac{7!}{6!1!}\cdot\frac{1}{2^7} + \frac{7!}{6!0!}\cdot\frac{1}{2^7} &= \frac{7+1}{2^7} = \frac{7+1}{2^7} = \frac{8}{2^7} = \frac{1}{2^4} > 0\,.\end{aligned}$$

Now, we can compute $\mathbb{P}(E|B)$. By definition,

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

Since $E \cap B = (A_2 \cup A_3 \cup A_5 \cup A_7) \cap (A_6 \cup A_7) = A_7$, we have

$$\mathbb{P}(E|B) = \frac{\mathbb{P}(A_7)}{\mathbb{P}(B)} = \frac{\frac{1}{2^7}}{\frac{1}{2^4}} = \frac{2^4}{2^7} = \frac{1}{8}\,.$$

Example 4.2. A student buys 2 apples, 3 bananas and 5 coconuts. Every day the student chooses a fruit uniformly at random and eats it.

The sample space is the set of all triplets that can be constructed with the available fruits, with each outcome corresponding to the fruit eaten on each day. Since by the end of the three days we have full information, so the event space is the power set of the sample space. We define the events $A_i = \{$the student eats an apple on day i$\}$, $B_i = \{$the student eats a banana on day i$\}$ and $C_i = \{$the student eats a coconut on day i$\}$.

(a) What is the probability that the student eats a coconut in day 1 and a banana in day 2? The event 'the student eats a coconut in day 1 **and** a banana in day 2' corresponds to the event $C_1 \cap B_2$. Note that the way information about the probability is encoded is through conditional probabilities: the statement '**every day** the student chooses a fruit **uniformly at random** and eats it' can be interpreted as the conditional probability of choosing any of the remaining fruits uniformly at random, so we know that

$$\mathbb{P}(B_2|C_1) = \frac{3}{9}.$$

It follows from the definition of conditional probability that

$$\mathbb{P}(C_1 \cap B_2) = \mathbb{P}(B_2|C_1)\mathbb{P}(C_1) = \frac{3}{9}\frac{5}{10} = \frac{1}{6}.$$

Writing the probability of an intersection of two events as a product of a conditional probability and a probability is called the 'multiplication rule' and can be extended to intersections of more than two events. For example, let us consider the following question.

(b) What is the probability that on the third day the student will eat the last apple? Since there are exactly two apples, that means that the student will it the first apple on either day 1 or day 2. So, if A is the event 'student eats last apple on the third day', we can write

$$A = (A_1 \cap A_2^c \cap A_3) \cup (A_1^c \cap A_2 \cap A_3)\,.$$

Notice that the events $A_1 \cap A_2^c \cap A_3$ and $A_1^c \cap A_2 \cap A_3$ are disjoint, therefore

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_1 \cap A_2^c \cap A_3) + \mathbb{P}(A_1^c \cap A_2 \cap A_3) \\ &= \mathbb{P}(A_1)\mathbb{P}(A_2^c|A_1)\mathbb{P}(A_3|A_1 \cap A_2^c) + \mathbb{P}(A_1^c)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1^c \cap A_2) \\ &= \frac{2}{10} \cdot \frac{8}{9} \cdot \frac{1}{8} + \frac{8}{10} \cdot \frac{2}{9} \cdot \frac{1}{8} = \frac{1}{45} + \frac{1}{45} = \frac{2}{45}\,,\end{aligned}$$

by using the multiplication rule twice.

Proposition 4.2 (Multiplication Rule)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $A_1, \ldots, A_n \in \mathcal{F}$ *s.t.* $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$*. Then,*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \ldots \mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1})\,.$$

Proof. Notice that for $k = 1, \ldots, n-1$, $A_1 \cap \cdots \cap A_k \supset A_1 \cap \cdots \cap A_{n-1}$. Hence, by Proposition 3.4 and by assumption

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_k) \geqslant \mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0\,.$$

This ensures that all the conditional probabilities at the right hand side are well-defined. The result follows by a direct application of the definition of conditional probability on the right-hand-side:

$$\begin{aligned}&\mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \ldots \mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1}) \\ &= \mathbb{P}(A_1)\frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)}\frac{\mathbb{P}(A_3 \cap A_1 \cap A_2)}{\mathbb{P}(A_1 \cap A_2)} \cdots \frac{\mathbb{P}(A_1 \cap \cdots \cap A_n)}{\mathbb{P}(A_1 \cap \cdots \cap A_{n-1})} \\ &= \mathbb{P}(A_1 \cap \cdots \cap A_n)\,.\end{aligned}$$

□

4.2 Law of Total Probability

Example 4.3 (4.2 continued). Suppose that we are now asked to compute the probability that the student eats a coconut on day 2. To compute the probability, we need to condition on what happened in day 1, but going through all possible options. In this case, there are two options that affect the computation of the conditional probability: whether the student also had a coconut on day 1 (event

C_1) or not (event C_1^c). So

$$\mathbb{P}(C_2) = \mathbb{P}(C2|C_1) \cdot \mathbb{P}(C_1) + \mathbb{P}(C_2|C_1^c) \cdot \mathbb{P}(C_1^c) = \frac{4}{9} \cdot \frac{5}{10} + \frac{5}{9} \cdot \frac{5}{10} = \frac{1}{2}.$$

Where is this formula coming from? We write

$$C_2 = (C_2 \cap C_1) \cup (C_2 \cap C_1^c).$$

so, from finite additivity, it follows that

$$\mathbb{P}(C_2) = \mathbb{P}(C2 \cap C_1) + \mathbb{P}(C_2 \cap C_1^c).$$

By applying the multiplication rule to the conditional probabilities above, we get the formula which is a specific example of the *law of total probabilities*.

The law of total probabilities allows us to compute the probability of an event, by conditioning on all possible instances of a 'different event', or, more formally, on every set in a partition of the sample space.

Definition 4.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $B_n \in \mathcal{F}$ for all $n = 1, \ldots, N\}$ (where N is either finite or infinite). Then, the collection of all B_n, $\{B_n : n = 1, \ldots, N\}$, is called a *partition* of Ω if

- $B_n \neq \emptyset \quad \forall\ n = 1, \ldots, N.$
- $B_n \cap B_m = \emptyset \qquad \forall n \neq m$
- $\bigcup_{n=1}^{N} B_n = \Omega.$

So, a partition is a collection of non-empty, disjoint events that span the whole space.

Proposition 4.3 (Law of Total Probability)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{B_n : n = 1, \ldots, N\}$ with N is either finite or infinite, be a partition of Ω such that $\mathbb{P}(B_n) > 0, \quad \forall\ n = 1, \ldots, N$. Then, for all $A \in \mathcal{F}$*

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

Proof. Notice that since $\{B_n : n = 1, \ldots, N\}$ forms a partition of Ω, we have

$$A = A \cap \Omega = A \cap \bigcup_{n=1}^{N} B_n = \bigcup_{n=1}^{N} (A \cap B_n) .$$

Further, since B_n's are disjoint so are $\{A \cap B_n : n = 1 \ldots N\}$, therefore by finite/countable additivity, we have

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{n=1}^{N} A \cap B_n \right) = \sum_{n=1}^{N} \mathbb{P}(A \cap B_n) = \sum_{n=1}^{N} \mathbb{P}(A|B_n)\mathbb{P}(B_n) .$$

In the last equality we use the definition of conditional probability with the assumption that $\mathbb{P}(B_n) > 0 \quad \forall\, n = 1, \ldots, N$. □

Example 4.4. A student is faced with a multiple choice question, with 4 choices. The student either knows the answer or chooses one of the answers uniformly at random. The probability that the student knows the answer is $\frac{2}{3}$.

(a) The student would like to compute the probability that they answer correctly. Let us start by defining the events of interest:

$$A = \{\text{student answers correctly}\}$$
$$B = \{\text{student knows the answer}\}$$

The information we are given about the probability are that 'the student either knows the answer (and thus answers correctly' or 'chooses one of the answers uniformly at random'. These can be expressed as $\mathbb{P}(A|B) = 1$ and $\mathbb{P}(A|B^c) = \frac{1}{4}$. We are also told that the probability that the student knows the answer is $\frac{2}{3}$. so $\mathbb{P}(B) = \frac{2}{3}$. Given this information, we are asked to find $\mathbb{P}(A)$.
Since B andB^c form a partition of the sample space, by applying the law of total probability we get

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

(b) The teacher would like to know the probability that the student knows the answer if they have answered correctly, so $\mathbb{P}(B|A)$. How can we use the

information we have to derive this? We write

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{2/3}{3/4} = \frac{8}{9}.$$

This is a particular example of what is known as Bayes' formula.

4.3 Bayes' Theorem

Theorem 4.1 (Bayes' Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{B_n : n = 1, \ldots, N\}$ with N is either finite or infinite, be a partition of Ω such that $\mathbb{P}(B_n) > 0 \quad \forall\, n = 1, \ldots, N$. Then for $A \in \mathcal{F}$ such that $\mathbb{P}(A) > 0$*

$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(A|B_n)\mathbb{P}(B_n)}{\sum_{j=1}^{N} \mathbb{P}(A|B_j)\mathbb{P}(B_j)} \quad \forall\, n = 1, \ldots, N\,.$$

Proof. By definition of conditional probability and since A is such that $\mathbb{P}(A) > 0$ then by the definition of conditional probability and the law of total probability:

$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(B_n \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_n)\mathbb{P}(B_n)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_n)\mathbb{P}(B_n)}{\sum_{j=1}^{N} \mathbb{P}(A|B_j)\mathbb{P}(B_j)}\,. \qquad \square$$

Example 4.5 (False Positives). A disease has incidence of 1 in 100 over the population. The available diagnostic test is such that

- if you have the disease, the test is positive with probability $\frac{72}{100}$
- if you don't have the disease, the test is positive with probability $\frac{5}{1000}$.

A person gets a positive result. What is the probability they actually have the disease?

The two events of interest are $D = \{$the person has the disease$\}$ and $P = \{$the person tests positive$\}$. We are interested in $\mathbb{P}(D|P)$. The information we are given is $\mathbb{P}(D) = \frac{1}{100}$, $\mathbb{P}(P|D) = \frac{72}{100}$ and $\mathbb{P}(P|D^c) = \frac{5}{1000}$. By Bayes' Theorem

$$\mathbb{P}(D|P) = \frac{\mathbb{P}(P|D)\mathbb{P}(D)}{\mathbb{P}(P|D)\mathbb{P}(D) + \mathbb{P}(P|D^c)\mathbb{P}(D^c)} \approx 0.59\,.$$

Bayes' theorem allows us to compute the conditional probability of one event, given another in terms of the reverse conditional probabilities. It is particularly useful in Statistics, leading to a whole area called Bayesian Statistics: while in

probability, we are interested in computing probabilities given a 'model' (i.e. sufficient information that determine the probabilities), in statistics, we are interested in choosing a model, given the events that we observe. Bayes' theorem allows us to connect the two.

4.4 Independence

Definition 4.3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that events A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

One way to think of independence is that knowledge about occurrence of one of the events will neither increase nor decrease the chance that the other occurs. Indeed, assuming that $\mathbb{P}(B) > 0$, you can check that A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$ (exercise!). In particular, if A and B are independent, then A^c and B are also independent.

Remark 3. The notions of "independent" and "disjoint" events are very different. In fact, these notions are normally incompatible: two disjoint events are independent if and only if the probability of one of them is 0 (exercise!).

Definition 4.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A_1, A_2, \ldots, A_n$ be events. We say that the events $A_1, \ldots, A_n$ are *pairwise independent* if A_j and A_k are independent for every choice of j and k distinct. We say that the events $A_1, \ldots, A_n$ are *mutually independent*, if

$$\mathbb{P}(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = \mathbb{P}(A_{j_1})\mathbb{P}(A_{j_2}) \cdots \mathbb{P}(A_{j_k})$$

for every $k = 2, \ldots, n$ and every choice of $1 \leqslant j_1 < j_2 < \cdots < j_k \leqslant n$.

In case $n = 2$, pairwise independence is obviously the same as mutual independence. In case $n = 3$, pairwise independence means

$$\begin{aligned}
\mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_1)\mathbb{P}(A_2) \\
\mathbb{P}(A_1 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_3) \\
\mathbb{P}(A_2 \cap A_3) &= \mathbb{P}(A_2)\mathbb{P}(A_3).
\end{aligned}$$

whereas mutual independence means

$$\begin{aligned}
\mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_1)\mathbb{P}(A_2) \\
\mathbb{P}(A_1 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_3) \\
\mathbb{P}(A_2 \cap A_3) &= \mathbb{P}(A_2)\mathbb{P}(A_3) \\
\mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3).
\end{aligned}$$

This illustrates that mutual independence is stronger than pairwise independence. It is hard to write down the list for larger values of n. For instance, if $n = 5$, pairwise independence involves $\binom{5}{2} = 10$ conditions to be checked, and mutual independence involves $2^5 - 5 - 1 = 26$ conditions to be checked.

Example 4.6. Two dice are rolled. Let

$$\begin{aligned}
A_1 &= \{\text{the first die is even}\} \\
A_2 &= \{\text{the second die is odd}\} \\
A_3 &= \{\text{sum of the dice is 7}\}.
\end{aligned}$$

These events are pairwise independent, since

$$\begin{aligned}
\mathbb{P}(A_1 \cap A_2) &= \frac{1}{4} = \mathbb{P}(A_1)\mathbb{P}(A_2) \\
\mathbb{P}(A_1 \cap A_3) &= \frac{1}{12} = \mathbb{P}(A_1)\mathbb{P}(A_3) \\
\mathbb{P}(A_2 \cap A_3) &= \frac{1}{12} = \mathbb{P}(A_2)\mathbb{P}(A_3).
\end{aligned}$$

That means, for each pair of events in this family, knowledge about occurrence of one of them will not affect the odds that any of the other two occurs. In particular, neither A_1 or A_2 alone will affect the odds of A_3. However, knowing that A_1 and A_2 both occur will in fact increase the chance that A_3 occurs, as

$$\mathbb{P}(A_3|A_1 \cap A_2) = \frac{1}{3} \neq \frac{1}{6} = \mathbb{P}(A_3).$$

More formally,

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{1}{12} \neq \frac{1}{24} = \mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3).$$

Example 4.7. Toss three fair coins. Consider the events:

$$A_1 = \text{First coin comes up Heads}$$
$$A_2 = \text{Second coin gives the same as the first coin}$$
$$A_3 = \text{Second coin gives the same as the third coin}$$
$$A_4 = \text{Third coms up Tails}$$

Then:
These events are pairwise independent.
A_1, A_2 and A_3 are mutually independent.
A_1, A_2, A_3 and A_4 are not mutually independent.

5 Random variables

5.1 Definition

Very often, we are interested in a quantity that is determined as the result of a given experiment.

For example, consider a game of chance where two dice are rolled and you get a monetary reward given by the maximum value obtained among the two dice. How do you model this situation? As usual, each outcome is a pair $\omega = (\omega_1, \omega_2)$ where both ω_1 and ω_2 are in $\{1, 2, 3, 4, 5, 6\}$. That is, $\Omega = \{1, 2, 3, 4, 5, 6\}^2 = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$. The reward is determined by the values of ω_1 and ω_2 through the following table:

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

The key word here is 'determined': even though the outcome is random, it is only random because the outcome ω is random.

Mathematically, this means that the prize X can be written as a *function* of the outcome ω. In general, a random variable is a function

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

from the sample space to the set of real numbers. More formally, we require that conditions specified in terms of X be random events, that is, events that the observer can determine whether they occur or not.

Definition 5.1 (Random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *random variable* is a function $X : \Omega \to \mathbb{R}$ such that $\{\omega \in \Omega : X(\omega) \leqslant a\} \in \mathcal{F}$ for every $a \in \mathbb{R}$.

In the above example, we can write X explicitly as the function that assigns to each pair (ω_1, ω_2) their maximum value, that is $X((x, y)) = \max\{x, y\}$.

Using this tool, we can make statements such as

$$\mathbb{P}(X = 1) = \frac{1}{36},\ \mathbb{P}(X = 5) = \frac{1}{4}, \ldots$$

Notation. For convenience, we will use the shorthand notation

$$\{X = 5\} = \{\omega \in \Omega : X(\omega) = 5\}, \quad \text{and } \mathbb{P}(X = 5) = \mathbb{P}(\{X = 5\}), \quad \text{etc.}$$

It is very useful to consider the probability measure on the set of real numbers induced by a random variable. If we are only interested in the value of X, we can leave $(\Omega, \mathcal{F}, \mathbb{P})$ behind, and take the sample space to be $\mathbb{R}$.

Definition 5.2 (Distribution)**.** The *distribution* of a random variable X is the probability measure on $\mathbb{R}$ denoted $\mathbb{P}_X$ and given by

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

for subsets B in some event space on the set of real numbers.

In the previous example, $\mathbb{P}_X(B)$ is determined by its values

$$\mathbb{P}_X(\{k\}) = \frac{2k - 1}{36},\ k = 1, 2, 3, 4, 5, 6, \quad \mathbb{P}_X(\mathbb{R} \setminus \{1, 2, 3, 4, 5, 6\}) = 0.$$

Notation. The event space on $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is an event space that contains sets such as $\{x\}$ and $(a, b]$. We will not worry about describing $\mathcal{B}(\mathbb{R})$ in details. You can think about this being the collection of all sets that you can obtain by applying a countable number of set operations (union, intersection, complement) to intervals.

Remark 4. The space Ω could be more complicated than $\mathbb{R}$. In general, $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ can be simpler than $(\Omega, \mathcal{F}, \mathbb{P})$. Note that $\mathbb{P}_X$ was built from $\mathbb{P}$ and X. However, it is not possible to reconstruct $(\Omega, \mathcal{F}, \mathbb{P})$ nor X from $\mathbb{P}_X$.

Proposition 5.1. *The function $\mathbb{P}_X$ is a probability measure on $\mathbb{R}$.*

Proof. Recall Definition 3.3. We check the three conditions:

(*i*) $\mathbb{P}_X(B) = \mathbb{P}(X \in B) \in [0, 1]$.
(*ii*) $\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in \mathbb{R}\}) = \mathbb{P}(\Omega) = 1$.
(*iii*) If $B_1, B_2, B_3, \cdots \in \mathcal{B}(\mathbb{R})$ are disjoint, then

$$\begin{aligned}\mathbb{P}_X(\cup_{n=1}^{\infty} B_n) &= \mathbb{P}(\{\omega : X(\omega) \in \cup_{n=1}^{\infty} B_n\}) \\ &= \mathbb{P}(\cup_{n=1}^{\infty} \{\omega : X(\omega) \in B_n\}) \\ &= \textstyle\sum_{n=1}^{\infty} \mathbb{P}(\{\omega : X(\omega) \in B_n\}) \\ &= \textstyle\sum_{n=1}^{\infty} \mathbb{P}_X(B_n).\end{aligned}$$

In the above inequalities we used: definition of $\mathbb{P}_X$; that pre-image of the union is the union of the pre-image; that pre-image of disjoint sets are disjoint, combined with countable additivity of $\mathbb{P}$; definition of $\mathbb{P}_X$.

This proves the proposition. □

5.2 Discrete random variables

Definition 5.3 (Discrete random variable)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable. We say that X and $\mathbb{P}_X$ are *discrete* if there is a finite or countably infinite set $S \subseteq \mathbb{R}$ such that $\mathbb{P}(X \in \mathbb{R} \setminus S) = 0$.

Definition 5.4 (Probability mass function)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a discrete random variable. We define the *probability mass function* of

X as the function $p_X : \mathbb{R} \to [0,1]$ given by

$$p_X(x) = \mathbb{P}_X(\{x\}).$$

Note that p_X is built from $\mathbb{P}_X$, and p_X is simpler than $\mathbb{P}_X$, because p_X takes as input a number and $\mathbb{P}_X$ takes as input a set of numbers. We will see below that it is possible to reconstruct $\mathbb{P}_X$ from p_X in case $\mathbb{P}_X$ is discrete.

Definition 5.5 (Discrete support)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a discrete random variable. We define the *discrete support* of X, or, more precisely, the discrete support of its distribution $\mathbb{P}_X$, as the set

$$\{x \in \mathbb{R} : p_X(x) > 0\}.$$

As claimed above, in order to study $\mathbb{P}_X$, it is enough to know p_X.

Proposition 5.2. *Let X be a discrete random variable. Then*

$$\mathbb{P}_X(B) = \sum_{x \in B \cap D_X} p_X(x)$$

for every $B \in \mathcal{B}(\mathbb{R})$, where D_X denotes the discrete support of X.

Notation. Before writing down the proof, we need to explain the meaning of $\sum_{x \in B \cap D}$. Since the set $B \cap D$ is countable, we can write $B \cap D = \{x_1, x_2, x_3, \dots\}$, and we can take $\sum_{x \in B \cap D} p_X(x)$ to be $\sum_{k=1}^{\infty} p_X(x_k)$. We need to be careful here, because we used an an arbitrary "enumeration" of $B \cap D$. However, since the terms in the sum are non-negative, another enumeration would mean reordering the terms, and it does not affect the value of the sum.

Proof. By definition of $\mathbb{P}_X$ being discrete, there is a countable set $S \subseteq \mathbb{R}$ such that $\mathbb{P}_X(S^c) = 0$. We can decompose

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

The second term is zero, because, since $B \cap S^c \subseteq S^c$,

$$0 \leqslant \mathbb{P}_X(B \cap S^c) \leqslant \mathbb{P}_X(S^c) = 0.$$

Hence,

$$\mathbb{P}_X(B) = \mathbb{P}_X(B \cap S) = \mathbb{P}_X(\cup_{x \in B \cap S}\{x\}) = \sum_{x \in B \cap S} \mathbb{P}_X(\{x\}).$$

If we substitute D^c instead of B in the above formula, we get $\mathbb{P}_X(D^c) = \sum_{x \in D^c \cap S} \mathbb{P}_X(\{x\}) = 0$, because $\mathbb{P}_X(\{x\}) = p_X(x) = 0$ for every $x \in D^c$.

By the same argument,

$$\mathbb{P}_X(B) = \mathbb{P}_X(B \cap D) + \mathbb{P}_X(B \cap D^c) = \mathbb{P}_X(B \cap D) = \sum_{x \in B \cap D} p_X(x),$$

which is what we wanted to prove. □

It is convenient to specify the distribution of a random variable by saying what p_X is. When we say "let X be a discrete random variable with probability mass function such and such," what do we mean? Does this really describe a random variable? The next definition and proposition answer this question.

Definition 5.6 (Probability mass function)**.** A function $f : \mathbb{R} \to [0, 1]$ is a *probability mass function* if the set D given by $D = \{x : f(x) > 0\}$ is countable and $\sum_{x \in D} f(x) = 1$.

Proposition 5.3. *Let $f : \mathbb{R} \to [0, 1]$ be a probability mass function. Then there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a discrete random variable X such that $p_X(x) = f(x)$ for every $x \in \mathbb{R}$.*

Proof. Take $D = \{x : g(x) > 0\}$. Take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ and

$$\mathbb{P}(B) = \sum_{x \in B \cap D} f(x).$$

Finally, take $X(x) = x$. Then X is a random variable. Moreover,

$$\mathbb{P}_X(D^c) = \sum_{x \in D \cap D^c} f(x) = 0,$$

because the sum over an empty set always equals zero. So X is a discrete random variable. Let us check that $p_X = p$.

For $x \in D$, we have

$$p_X(x) = \mathbb{P}_X(\{x\}) = \mathbb{P}(\{x\}) = \sum_{z \in \{x\} \cap D} f(z) = \sum_{z \in \{x\}} f(z) = f(x)$$

because the sum of a single factor is that factor.

On the other hand, for $x \in D^c$, we have $f(x) = 0$ and

$$p_X(x) = \sum_{z \in \{x\} \cap D} f(z) = \sum_{z \in \emptyset} f(z) = 0 = f(x).$$

Hence, $p_X = p$ as claimed, and this completes the proof of the proposition. □

5.3 The most common discrete distributions

Definition 5.7 (Bernoulli distribution)**.** We say that a discrete random variable X has *Bernoulli distribution* with parameter $p \in [0, 1]$, denoted $X \sim \text{Bernoulli}(p)$, if its probability mass function is

$$p_X(x) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Example 5.1. Let $\Omega = \{H, T\}$ (heads or tails for a coin toss) with $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = 1 - p$, and let $X(H) = 1$ and $X(T) = 0$. Then, $X \sim \text{Bernoulli}(p)$.

Definition 5.8 (Geometric distribution)**.** We say that a discrete random variable X has *geometric distribution* with parameter $p \in (0, 1]$, denoted $X \sim \text{Geom}(p)$, if its probability mass function is

$$p_X(x) = \begin{cases} p \cdot (1 - p)^{x-1}, & x \in \mathbb{N}, \\ 0, & \text{otherwise.} \end{cases}$$

To see that p_X is indeed a probability mass function, note that

$$\sum_{k=0}^{\infty} p_X(k) = p \cdot \sum_{k=1}^{\infty} (1 - p)^{k-1} = p \cdot \sum_{\ell=0}^{\infty} (1 - p)^{\ell} = p \cdot \frac{1}{1 - (1 - p)} = p \cdot \frac{1}{p} = 1.$$

Random variables having the geometric distribution arise in the following situation. Suppose that we repeatedly perform trials, each of which can be a success or a failure. Assume that the trials are independent, and the probability of success is the same in each of them, equal to p. Then, the number of trials performed until the first success is obtained has a geometric distribution with parameter p.

Remark 5. Some rare references use a different definition for the geometric distribution with parameter p: they take the distribution on $\mathbb{N}_0$ (instead of $\mathbb{N}$) and probability mass function $\tilde{p}_X(k) = p \cdot (1-p)^k$, for $k \in \mathbb{N}_0$. A random variable with probability mass function $\tilde{p}_X$ counts the number of *failed* trials performed before the first success is obtained. So, in case a success is obtained already in the first trial, then the number of failed trials is zero.

Example 5.2. We roll a die repeatedly until we roll a 6 for the first time. Let X be the total number of times we roll the die. Then, $X \sim \text{Geom}(\frac{1}{6})$.

Definition 5.9 (Binomial distribution)**.** We say that a discrete random variable X has *binomial distribution* with parameters $n \in \mathbb{N}_0$ and $p \in [0,1]$, denoted $X \sim \text{Binom}(n,p)$, if its probability mass function is

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in \{0,1,\ldots,n\}, \\ 0, & \text{otherwise.} \end{cases}$$

Note that p_X is indeed a probability mass function since

$$1 = (p + (1-p))^n = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Example 5.3. Roll a die ten times, and let X the number of times a 5 or a 6 is rolled. Then, $X \sim \text{Binom}(10, \frac{1}{3})$.

Recall that $0^0 = 1$ and $0! = 1$.

Definition 5.10 (Poisson distribution)**.** We say that a discrete random variable X has *Poisson distribution* with parameter $\lambda \geqslant 0$, denoted by

$X \sim \text{Poisson}(n, p)$, if its probability mass function is

$$p_X(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, & x \in \mathbb{N}_0, \\ 0, & \text{otherwise.} \end{cases}$$

To show that p_X is indeed a probability mass function, we compute

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Random variables that count rare occurrences among many trials (such as: number of accidents in a road throughout a year, number of typos in a book page) typically follow the Poisson distribution. More precisely, a Poisson random variable can be used to approximate a Binomial(n, p) when n is large, p is small, and $np = \lambda$ is fixed. Indeed, if $X \sim \text{Binom}(n, p)$ for $p = \frac{\lambda}{n}$, then

$$\begin{aligned} \mathbb{P}(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\to \frac{\lambda^k}{k!} e^{-\lambda}, \quad \text{for } n \text{ large} \end{aligned}$$

where the approximation follows from the following approximations for large n:

$$n(n-1)...(n-k+1)/n^k \to 1, \quad \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}, \quad \frac{\lambda}{n} \to 0\,.$$

Brief review

Probability spaces

Ω – *sample space*: elements $\omega \in \Omega$ are outcomes, Ω is a non-empty set.

$\mathcal{F}$ – *event space*: elements $A \in \Omega$ are events ($A \subseteq \Omega$)

Must satisfy three conditions:

- $\mathcal{F} \neq \emptyset$
- $A^c \in \mathcal{F}$ for every $A \in \mathcal{F}$
- $(\cup_{n=1}^{\infty} A_n) \in \mathcal{F}$ for every sequence of events $A_1, A_2, A_3, \ldots$

Consequences of these conditions:

- $\Omega \in \mathcal{F}$ (indeed, take $A \in \mathcal{F}$, $A \cup A^c = \Omega$)
- $\emptyset \in \mathcal{F}$ (indeed, $\Omega^c = \emptyset$)
- $(\cap_{n=1}^{\infty} A_n) \in \mathcal{F}$ for every sequence of events $A_1, A_2, A_3, \ldots$
  (indeed, $\cap_{n=1}^{\infty} A_n = (\cup_{n=1}^{\infty} A_n^c)^c$)

$\mathbb{P}$ – *probability measure*: $\mathbb{P} : \mathcal{F} \to \mathbb{R}$.

Must satisfy three conditions:

- $\mathbb{P}(A) \geqslant 0$ for all $A \in \mathcal{F}$
- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}$ is countably additive: $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$
  for every sequence of *disjoint* events $A_1, A_2, A_3, \ldots$

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

Example 5.4 (Uniform probability spaces). Ω a finite set, $\mathcal{F} = \mathcal{P}(\Omega)$, $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$.

Example 5.5. There is no probability space to model the experiment "pick an integer at random". Much as we would like to say that an integer X chosen at random will be even with probability $\frac{1}{2}$ and the last digit in its representation in decimal will be 7 with probability $\frac{1}{10}$, there is no probability space that can model that. More precisely, for $\Omega = \mathbb{Z}$, $\mathcal{F} = \mathbb{P}(\mathbb{Z})$, there is no probability measure $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ such that $\mathbb{P}(\{j\}) = \mathbb{P}(\{k\})$ for every $j, k \in \mathbb{Z}$. Indeed, if $\mathbb{P}(\{j\}) > 0$ then $\mathbb{P}(\mathbb{Z}) = \sum x \in \mathbb{Z}\mathbb{P}(\{x\}) = +\infty$, and if $\mathbb{P}(\{j\}) = 0$ then $\mathbb{P}(\mathbb{Z}) = \sum x \in \mathbb{Z}\mathbb{P}(\{x\}) = 0$, and in either case $\mathbb{P}$ violates the requirements to be a probability measure, that $\mathbb{P}(\mathbb{Z}) = 1$.

Combinatorics

Finite set A, $n = |A|$.

Sequences of length k of elements of A:

$$|S_{n,k}(A)| = n^k,$$

because in order to specify an element of $S_{n,k}(A)$ we need to make k choices, in each choice we have n options.

Permutations (or reordering) of elements of A:

$$n!,$$

because we need to make n choices, and each time the number of options decreases. (read "n factorial")

Ordering of length k of elements of A:

$$|O_{n,k}(A)| = \frac{n!}{(n-m)!},$$

because we can obtain any element of $O_{n,k}(A)$ by permuting the elements of A, keeping the first m elements, and forgetting about the ordering of the $n-m$ remaining ones.

Subsets of A having cardinality k:

$$|C_{n,k}(A)| = \binom{n}{k} = \frac{n!}{k!(n-k)!},$$

because we can obtain any element of $C_{n,k}(A)$ by taking the first m elements in a permutation of A, then keeping the first m elements, and then forgetting about the ordering of these m elements as well as the $n-m$ remaining ones. (read "n choose k")

Decompose of A into *labelled* sets $A_1, \ldots, A_r$ such that $|A_j| = k_j$ for $j = 1, \ldots, r$, where $k_1 + \cdots + k_r = n$.

$$\frac{n!}{k_1!k_2!\cdots k_r!},$$

because we can obtain partition by first permuting all elements of A, taking A_1 to be the first k_1 elements of this permutation, A_2 to be the next k_2 elements and so on, and then forgetting about the ordering of each block. This is the number of ways in which n labelled balls can be distributed into r labelled buckets under the constraint that, for every $j = 1, \ldots, r$, bucket number j gets k_j balls. The combinations $C_{n,k}(A)$ are just a particular case of two buckets labelled "in" and "out".

To give a more rigorous justification for the terms in the denominator (which

we called "forget about the ordering"), we can obtain the number by reasoning backwards, as follows. We will produce a permutation of $A = \{1, \ldots, n\}$ in two ways. The first way has three steps:

(a) Choose k elements of A to go first, and let the remaining $n - k$ go after. There are x possibilities.
(b) Permute the first k elements. There are $k!$ possibilities.
(c) Permute the remaining $n - k$ elements. There are $(n - k)!$ possibilities.

In total, there are $x \cdot k! \cdot (n - k)!$ possibilities. The second way is straightforward: just permute the n elements already, there are $n!$ possibilities. Hence, $n! = x \cdot k! \cdot (n - k)!$, so we have just found what x is! Therefore, $|C_{n,k}(A)| = \frac{n!}{k!(n-k)!}$.

Sampling

Population with $n = n_1 + n_2$ individuals, where n_1 are individuals of Type 1 and n_2 are individuals of Type 2.

If we take a sample of k individuals, *without replacement*, then the chance of picking k_1 elements of Type 1 (and thus $k_2 = k - k_1$ individuals of Type 2) is

$$\frac{\binom{n_1}{k_1} \cdot \binom{n_2}{k_2}}{\binom{n}{k}}.$$

This can be obtained by assuming that the individuals were sampled simultaneously (so $|\Omega| = \binom{n}{k}$) or one after the other (so $|\Omega| = \frac{n!}{(n-k)!}$), and both approaches give the same probability (as expected!). Indeed, the second approach gives

$$\frac{\binom{k}{k_1} \frac{n_1!}{(n_1-k_1)!} \frac{n_2!}{(n_2-k_2)!}}{\frac{n!}{(n-k)!}}$$

which simplifies to the first formula if we expand the first term.

If we take a sample of k individuals, *with replacement*, then the chance of picking k_1 elements of Type 1 (and thus $k_2 = k - k_1$ individuals of Type 2) is

$$\binom{k}{k_1} \left(\frac{n_1}{n}\right)^{k_1} \left(1 - \frac{n_1}{n}\right)^{k_2}.$$

This can be only be achieved by assuming that the individuals were sampled one

after the other, so that we can put them back into the population (so $|\Omega| = n^k$). The above formula is obtained after rewriting

$$\frac{\binom{k}{k_1} n_1^{k_1} n_2^{k_2}}{n^k}$$

in a more convenient (or meaningful) form.

Observe that, if X denotes the number of individuals of Type 1 in the sample with replacement, then

$$X \sim \text{Binom}(k, \tfrac{n_1}{n}),$$

which can be checked by matching the above formula with the probability mass function of a binomial random variable.

Random variables

Random variable:

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

with the requirement that $\{\omega : X(\omega) \leqslant a\} \in \mathcal{F}$ for every $a \in \mathbb{R}$.

The *distribution* of X is $\mathbb{P}_X : \mathcal{B} \to \mathbb{R}$, given by

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega : X(\omega) \in B\})$$

is a probability measure on $\mathbb{R}$, where $\mathcal{B}$ is the event space on $\mathbb{R}$.

A random variable X is *discrete* if there is a countable set S such that $\mathbb{P}_X(S^c) = 0$.

For discrete X, we define the *probability mass function of* X as the function $p_X : \mathbb{R} \to \mathbb{R}$ given by

$$p_X(x) = \mathbb{P}_X(\{x\}).$$

and the *discrete support* of X as the set $\{x \in \mathbb{R} : p_X(x) > 0\}$.

If X is discrete, then

$$\mathbb{P}_X(B) = \sum_{x \in B \cap D_X} p_X(x)$$

for all $B \in \mathcal{B}$, where D_X denotes the discrete support of X.

We say that a function $f : \mathbb{R} \to \mathbb{R}$ is a *probability mass function* if $f(x) \geqslant 0$ for

all $x \in \mathbb{R}$, the set $\{x \in \mathbb{R} : f(x) > 0\}$ is countable and

$$\sum_{x:f(x)>0} f(x) = 1.$$

Given a probability mass function f, it is possible to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X such that f is the probability mass function of X.

6 Expectation

If we roll a fair die many times, we expect that each of the six possible outcomes will appear about one-sixth of the time, and thus the average of the numbers obtained would be approximately

$$\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{7}{2}.$$

We call this number $\frac{7}{2}$ the *expectation* of X.

6.1 Definition and examples

Definition 6.1 (Expectation)**.** Let X be a discrete random variable. We define the *expectation* of X, denoted $\mathbb{E}[X]$, as the real number given by

$$\mathbb{E}[X] = \sum_{x:\mathbb{P}(X=x)>0} x \cdot \mathbb{P}(X = x),$$

as long as this sum converges absolutely, otherwise $\mathbb{E}[X]$ is not defined.

Terminology. To say that $\sum_n a_n$ *converges absolutely* means that $\sum_n |a_n|$ converges.

Definition 6.2 (Integrable)**.** We say that a discrete random variable X is *integrable* if is expectation is defined, that is, if the sum

$$\sum_{x:\mathbb{P}(X=x)>0} |x| \cdot \mathbb{P}(X = x)$$

is convergent.

Notation. The fact that $\mathbb{E}[X]$ depends on X is made obvious by the fact that "X" appears in "$\mathbb{E}[X]$" and the fact that that it depends on $\mathbb{P}$ is made somewhat apparent by the fact that "$\mathbb{E}$" and "$\mathbb{P}$" use the same font.

Example 6.1. Toss a fair coin 4 times and count the number of Heads.

$$\begin{aligned}\mathbb{E}[X] &= 0\cdot\mathbb{P}(X=0)+1\cdot\mathbb{P}(X=1)+2\cdot\mathbb{P}(X=2)+\\ &\quad +3\cdot\mathbb{P}(X=3)+4\cdot\mathbb{P}(X=4)\\ &= 0\cdot\frac{1}{16}+1\cdot\frac{4}{16}+2\cdot\frac{6}{16}+3\cdot\frac{4}{16}+4\cdot\frac{1}{16}\\ &= 2.\end{aligned}$$

Example 6.2 (Indicator function). Let $A\in\mathcal{F}$ and take X given by

$$X(\omega)=\begin{cases}1, & \omega\in A,\\ 0, & \omega\in A^c.\end{cases}$$

Such a function is called *indicator function* of the set A and is denoted $\mathbb{1}_A$. In this case, $\mathbb{E}[X]=0\times\mathbb{P}(A^c)+1\times\mathbb{P}(A)=\mathbb{P}(A)$. That is,

$$\mathbb{E}[\mathbb{1}_A]=\mathbb{P}(A).$$

Example 6.3. Roll a fair die twice and add the observed values.

$$\begin{aligned}\mathbb{E}[X] = 2\times\frac{1}{36}+3\times\frac{2}{36}+4\times\frac{3}{36}+5\times\frac{4}{36}+6\times\frac{5}{36}+7\times\frac{6}{36}+\\ +8\times\frac{5}{36}+9\times\frac{4}{36}+10\times\frac{3}{36}+11\times\frac{2}{36}+12\times\frac{1}{36}=7.\end{aligned}$$

Example 6.4. Take 3 cards from a deck of 52, one after the other and without replacement, and count how many are queens.

$$\begin{aligned}\mathbb{E}[X] = 0\times\frac{48\cdot47\cdot46}{52\cdot51\cdot50}+1\times\frac{3\cdot48\cdot47\cdot4}{52\cdot51\cdot50}+\\ +2\times\frac{3\cdot48\cdot4\cdot3}{52\cdot51\cdot50}+3\times\frac{4\cdot3\cdot2}{52\cdot51\cdot50}=\frac{3}{13}.\end{aligned}$$

Example 6.5 (Poisson). If $X\sim\text{Poisson}(\lambda)$, then

$$\mathbb{E}[X]=\sum_{n=0}^{\infty}n\frac{\lambda^n e^{-\lambda}}{n!}=\sum_{n=1}^{\infty}\frac{\lambda^n e^{-\lambda}}{(n-1)!}=$$

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

Therefore, the expectation of a random variable distributed as Poisson(λ) is λ.

Example 6.6 (Binomial). If $X \sim \text{Binom}(n, p)$, then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{n} k\tbinom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} n\tbinom{n-1}{k-1} p^k (1-p)^{n-k} \\
&= n \sum_{j=0}^{n-1} \tbinom{n-1}{j} p^{j+1} (1-p)^{n-j-1} = np \sum_{j=0}^{n-1} \tbinom{n-1}{j} p^j (1-p)^{n-1-j} \\
&= np[p + (1-p)]^{n-1} = np.
\end{aligned}$$

Example 6.7 (Geometric). Suppose $X \sim \text{Geom}(p)$. We will compute

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} n(1-p)^{n-1} p$$

by differentiating a power series. Writing $x = 1 - p$, to develop as follows:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{n=1}^{\infty} n \cdot p \cdot (1-p)^{n-1} = p \sum_{n=0}^{\infty} n \cdot x^{n-1} \\
&= p \sum_{n=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x}\big[x^n\big] = p \cdot \frac{\mathrm{d}}{\mathrm{d}x}\Big[\sum_{n=0}^{\infty} x^n\Big] = p \cdot \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1}{1-x}\right] \\
&= p \cdot \Big(-(1-x)^{-2} \Big) \cdot (-1) = \frac{1}{p}.
\end{aligned}$$

Therefore, the expectation of a random variable distributed as Geom(p) is $\frac{1}{p}$. We know that the power series $\sum_n x^n$ converges if $|x| < 1$, and we are accepting a property which says that the power series can be differentiated term by term within that range.

6.2 Properties of the expectation

In the above examples, $2 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}$, $7 = \frac{7}{2} + \frac{7}{2}$, $\frac{3}{13} = \frac{1}{13} + \frac{1}{13} + \frac{1}{13}$, and $np = p + \cdots + p$. This is not a coincidence. It comes from the fact that

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

for integrable discrete random variables X and Y.

Theorem 6.1. *Let X and Y be integrable discrete random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then:*

(1) $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$ *for every* $A \in \mathcal{F}$,
(2) *If* $0 \leqslant Z \leqslant X$ *for all* $\omega \in \Omega$ *then* $0 \leqslant \mathbb{E}[Z] \leqslant \mathbb{E}[X]$,
(3) $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

We say that the expectation is *unitary, monotone*, and *linear*.

We will see a proof further down the road.

Example 6.8. In Example 6.1, we can define X_1, X_2, X_3 and X_4 as the indicator function of the events that the first, second, third and forth tosses of the coin came up Heads, respectively. Since $X = X_1 + X_2 + X_3 + X_4$ we can obtain the expectation using linearity, as in

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3] + \mathbb{E}[X_4] = \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} = 2,$$

instead of computing the probability mass function of X.

Example 6.9. In Example 6.3, observe that $X = Y + Z$, where Y and Z represent the result of the first and second dice. Thus

$$\mathbb{E}[X] = \mathbb{E}[Y] + \mathbb{E}[Z] = \frac{7}{2} + \frac{7}{2} = 7.$$

Example 6.10. In Example 6.4, observe that $X = X_1 + X_2 + X_3$, where X_k is the indicator of whether the k-th card is a queen. Unlike the previous examples, notice that here X_1, X_2 and X_3 are not "independent" (a precise notion of independence will be introduced further down the road). Nevertheless, each one of them individually satisfies $\mathbb{E}[X_k] = \frac{1}{13}$, and we can compute

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3] = \frac{3}{13}.$$

Example 6.11. In Example 6.6, observe that X has the same distribution as $X_1 + \cdots + X_n$, where each X_k is distributed as Bernoulli(p), and therefore

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = (p + \cdots + p) = np.$$

In the previous examples, it was easier to compute the expectation using linearity than using the distribution of X directly. In many other cases, describing the distribution of X in a way that allows us to compute the expectation can be very hard or even impractical, but it may still be possible to compute the expectation using linearity.

Example 6.12. A drawer contains 10 pairs of socks, all different from each other. Someone opens the drawer in the dark and takes 6 socks from it. What is the expectation of X, the number of pairs formed by the socks taken? It is more convenient to suppose that the socks are drawn in order, from 1-st to 6-th. We will count how many of them has a pair that was also drawn This will give a number N which is twice the number of pairs, because each pair will be counted twice, so $N = 2X$. Now observe that $N = X_1 + \cdots + X_6$, where $X_k = \mathbb{1}_{A_k}$ and A_k is the event that pair of the k-th sock taken has also been taken. Then $\mathbb{P}(A_k) = \frac{5}{19}$ (exercise!) and thus $\mathbb{E}[N] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_6] = 6 \cdot \frac{5}{19}$. Therefore, $\mathbb{E}[X] = \mathbb{E}[\frac{N}{2}] = \frac{15}{19}$. The combinatorics involved in showing that $\mathbb{P}(A_k) = \frac{5}{19}$ may not be very easy, but it is much easier than describing the distribution of N.

6.3 Function of a random variable

Proposition 6.1. *Let X be a discrete random variable, and let $g : \mathbb{R} \to \mathbb{R}$ be any function. Then*

$$\mathbb{E}[g(X)] = \sum_{x \in D_X} g(x) \cdot \mathbb{P}(X = x),$$

if this sum converges absolutely, and $\mathbb{E}[g(X)]$ is undefined if not. In the sum, D_X denotes the discrete support of X.

Example 6.13. Suppose $p_X(x) = \frac{1}{3}$ for $x = 1, 2, 3$. Let us compute $\mathbb{E}[(X-2)^2]$ in two ways. For the function $g(x) = (x-2)^2$, we want to compute $\mathbb{E}[g(X)]$.

The first way goes as follows. Define the random variable $Z = g(X) = (X-2)^2$, and compute $\mathbb{E}[Z]$ from the definition. We start by computing $p_Z(0) = \mathbb{P}(X \in$

$\{2\}) = \frac{1}{3}$ and $p_Z(1) = \mathbb{P}(X \in \{1,3\}) = \frac{2}{3}$, obtaining the table

z	$p_Z(z)$
0	$\frac{1}{3}$
1	$\frac{2}{3}$

and finally $\mathbb{E}[Z] = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}$. For the second way, just write down

x	$g(x)$	$p_X(x)$
1	1	$\frac{1}{3}$
2	0	$\frac{1}{3}$
3	1	$\frac{1}{3}$

and compute $\mathbb{E}[g(X)] = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3}$.

Example 6.14. If $X \sim \text{Poisson}(\lambda)$, then

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

Although a lot of algebraic computation was involved, the alternative would be worse: define $Z = X^2$, describe the discrete support of Z, find an expression for $p_Z(z)$, write $\mathbb{E}[Z] = \sum_z z \cdot p_Z(z)$ and then try to evaluate the sum.

Example 6.15. Suppose $X \sim \text{Geom}(p)$. As before, we will compute

$$\mathbb{E}[X^2] = \sum_{n=1}^{\infty} n^2 (1-p)^{n-1} p$$

by differentiating two power series. To do that, we write $x = 1 - p$ and develop

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{n=1}^{\infty} n^2 \cdot p \cdot x^{n-1} \\
&= p \sum_{n=1}^{\infty} n \cdot x^{n-1} + p \sum_{n=1}^{\infty} n \cdot (n-1) \cdot x^{n-1} \\
&= p \sum_{n=0}^{\infty} n \cdot x^{n-1} + px \sum_{n=0}^{\infty} n \cdot (n-1) \cdot x^{n-2} \\
&= p \sum_{n=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x}[x^n] + px \sum_{n=0}^{\infty} \frac{\mathrm{d}^2}{\mathrm{d}x^2}[x^n] \\
&= p \cdot \frac{\mathrm{d}}{\mathrm{d}x}\Big[\sum_{n=0}^{\infty} x^n\Big] + px \cdot \frac{\mathrm{d}^2}{\mathrm{d}x^2}\Big[\sum_{n=0}^{\infty} x^n\Big] \\
&= p \frac{1}{(1-x)^2} + px \frac{2}{(1-x)^3} \\
&= \frac{1}{p} + 2 \cdot \frac{1-p}{p^2} \\
&= \frac{2-p}{p^2}.
\end{aligned}$$

As before, we know that the power series $\sum_n x^n$ converges if $|x| < 1$, and we are accepting a property which says that the power series can be differentiated term by term within that range.

Example 6.16. Suppose $X \sim \mathrm{Geom}(p)$. For which values of t is e^{tX} integrable, and what is the value of $\mathbb{E}[e^{tX}]$? We can write

$$\mathbb{E}[e^{tX}] = \sum_{n=1}^{\infty} e^{tn} \cdot p \cdot (1-p)^{n-1} = pe^t \sum_{k=0}^{\infty} [e^t \cdot (1-p)]^{n-1} = \frac{pe^t}{1 - [e^t \cdot (1-p)]}.$$

This can be rewritten as

$$\mathbb{E}[e^{tX}] = \frac{p}{e^{-t} + p - 1},$$

and it is defined if $e^t \cdot (1-p) < 1$, or alternatively $t < \ln \frac{1}{1-p}$ and is undefined otherwise.

6.4 Variance

Here we introduce another fundamental quantity that describes the distribution of a random variable. While $\mathbb{E}[X]$ gives the mean number of X, we now define a quantity that quantifies the degree of dispersion of X from its mean value.

Definition 6.3 (square-integrable random variables)**.** We say that a discrete random variable X is *square-integrable* if X^2 is integrable, which means that the sum

$$\sum_{x:\mathbb{P}(X=x)>0} x^2 \cdot \mathbb{P}(X=x)$$

is convergent. Note that square-integrable random variables are automatically integrable, because $|x| \leqslant 1 + x^2$.

Definition 6.4 (Variance)**.** Let X be a square-integrable discrete random variable and denote $\mu = \mathbb{E}[X]$. We define the *variance* of X as

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

Even though this formula may be the best definition to understand the properties of variance, very often there is a more convenient way to compute it:

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

which we get by expanding $\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - \mu^2$.

Example 6.17 (Poisson). Suppose $X \sim \mathrm{Poisson}(\lambda)$. Then

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

so the variance of a Poisson random variable equals its expectation.

Example 6.18 (Geometric). Suppose $X \sim \mathrm{Geom}(p)$. Then

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

Example 6.19 (Bernoulli). Suppose $X \sim \mathrm{Bernoulli}(p)$. Then

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

Note that

$$\mathrm{Var}(aX) = a^2 \cdot \mathrm{Var}(X)$$

which means that $\mathrm{Var}(X)$ is not in the same unit of measure as X. For instance, if X is measured in meters, then $\mathbb{E}[X]$ is also measured in meters but $\mathrm{Var}(X)$ is measured in squared meters.

To quantify the dispersion of X in the same units as X, we need to take the square root.

Definition 6.5 (Standard deviation)**.** Let X be a square-integrable discrete random variable. We define the *standard deviation of* X as

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Unlike the variance, the standard deviation satisfies $\sigma(aX) = |a| \cdot \sigma(X)$.

The standard deviation of a Poisson random variable is $\sqrt{\lambda}$, of a Bernoulli random variable is $\sqrt{p(1-p)}$, and of a geometric random variable is $\sqrt{p^{-2} - p^{-1}}$.

7 Multivariate discrete distributions

7.1 Joint probability mass function of two variables

Definition 7.1 (Joint probability mass function)**.** Given two discrete random variables X and Y, we define the *joint probability mass function of* X *and* Y, denoted $p_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$, and given by

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

or, more formally, $\mathbb{P}(\{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\})$.

Example 7.1. Roll two dice, let X denote the large value and Y denote the smaller value. Then $p_{X,Y}(x, y)$ for $x = 1, \ldots, 6$ and $y = 1, \ldots, 6$ is given by the

table

	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$
2	$\frac{0}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$
3	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$
4	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$
5	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{1}{36}$	$\frac{2}{36}$
6	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{0}{36}$	$\frac{1}{36}$

,

and $p(x, y) = 0$ if x or y is not in $\{1, 2, 3, 4, 5, 6\}$.

Observe that

$$\begin{aligned} p_X(x) &= \mathbb{P}(X = x) \\ &= \sum_{y \in D_Y} \mathbb{P}(X = x, Y = y) + \mathbb{P}(X = x, Y \notin D_Y) \\ &= \sum_{y \in D_Y} \mathbb{P}(X = x, Y = y) \\ &= \sum_{y \in D_Y} p_{X,Y}(x, y) \end{aligned}$$

for every $x \in \mathbb{R}$, where D_Y denotes the discrete support of Y.

Terminology (Marginal probability mass function)**.** The above formula to compute the probability mas function of X from the joint probability mass function of X and Y is called *marginal probability mass function.*

Example 7.2. A bag contains 1 red, 2 greens and 2 blue balls. We pick 2 balls from the bag, without replacement. Let X be the number of green balls picked, and Y be the number of red balls picked. Then the joint probability mass function of X and Y is given by the central cell of the table below:

$y \backslash x$	0	1	2	total
0	0.1	0.4	0.1	0.6
1	0.2	0.2	0	0.4
total	0.3	0.6	0.1	1

.

By adding each column we find the marginal probability mass function of X, which is given by $p_X(0) = 0.3$, $p_X(1) = 0.6$, and $p_X(2) = 0.1$. By adding each

row we find the marginal probability mass function of Y, which is given by $p_X(0) = 0.6$ and $p_X(1) = 0.4$.

7.2 Expectation in the discrete bivariate case

Proposition 7.1 (Expectation in the bivariate case)**.** *Let X and Y be discrete random variables, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be any function. Then*

$$\mathbb{E}[g(X,Y)] = \sum_{x \in D_X} \sum_{y \in D_Y} g(x,y) \cdot \mathbb{P}(X = x, Y = y),$$

if this sum converges absolutely, and $\mathbb{E}[g(X,Y)]$ is undefined if not. The sets D_X and D_Y in the formula denote the discrete supports of X and Y.

Proof. Let $Z = g(X,Y)$. We first observe that, for each $z \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(Z = z) &= \mathbb{P}\big(g(X,Y) = z\big) \\
&= \mathbb{P}\big((X,Y) \in g^{-1}(z)\big) \\
&= \sum_{(x,y) \in g^{-1}(z) \cap D} \mathbb{P}\big(X = x, Y = y\big),
\end{aligned}$$

where $D = \{(x,y) : \mathbb{P}(X = x, Y = y) > 0\}$. The support of Z is given by the image $D_Z = g(D)$, which is countable. Finally,

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{z \in D_Z} z \cdot \mathbb{P}(Z = z) \\
&= \sum_{z \in D_Z} \sum_{(x,y) \in g^{-1}(z) \cap D} z \cdot \mathbb{P}\big(X = x, Y = y\big) \\
&= \sum_{z \in D_Z} \sum_{(x,y) \in g^{-1}(z) \cap D} g(x,y) \cdot \mathbb{P}\big(X = x, Y = y\big) \\
&= \sum_{(x,y) \in D} g(x,y) \cdot \mathbb{P}\big(X = x, Y = y\big) \\
&= \sum_{x \in D_X} \sum_{y \in D_Y} g(x,y) \cdot \mathbb{P}(X = x, Y = y),
\end{aligned}$$

and the sum converges absolutely if and only if Z is integrable. This concludes the proof. □

Proof of Proposition 6.1. If we take $Y = 0$ and apply Proposition 7.1 with $\tilde{g}(x, y) = g(x)$, we get

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\tilde{g}(X, Y)] \\ &= \sum_{x \in D_X} \sum_{y \in \{0\}} \tilde{g}(x, y) \cdot \mathbb{P}(X = x, Y = y) \\ &= \sum_{x \in D_X} g(x) \cdot \mathbb{P}(X = x),\end{aligned}$$

and the sums converge absolutely if and only if $\mathbb{E}[X]$ is defined. This concludes the proof of Proposition 6.1. □

Corollary 7.1. *The expectation is linear.*

Proof. Suppose X and Y are integrable, and let $a, b \in \mathbb{R}$. Using Proposition 7.1 with $g(x, y) = ax + by$, $g(x, y) = x$ and $g(x, y) = y$, we get

$$\begin{aligned}\mathbb{E}[aX + bY] &= \sum_{x \in D_X} \sum_{y \in D_Y} (ax + by) \cdot \mathbb{P}(X = x, Y = y) \\ &= a \sum_{x \in D_X} \sum_{y \in D_Y} x \cdot \mathbb{P}(X = x, Y = y) + b \sum_{x \in D_X} \sum_{y \in D_Y} y \cdot \mathbb{P}(X = x, Y = y) \\ &= a\,\mathbb{E}[X] + b\,\mathbb{E}[Y],\end{aligned}$$

which is what we wanted to prove. □

7.3 Independent discrete random variables

Definition 7.2 (Independence)**.** Two discrete random variables X and Y are *independent* if

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

for every $x, y \in \mathbb{R}$.

Example 7.3. Toss a fair coin 5 times. Let X be the number of Heads in the first three tosses, and Y be the number of Heads in the last two tosses. Then the

joint probability mass function is shown in the table

$y\backslash x$	0	1	2	3	total
0	1/32	3/32	3/32	1/32	1/4
1	2/32	6/32	6/32	2/32	1/2
2	1/32	3/32	3/32	1/32	1/4
total	1/8	3/8	3/8	1/8	1

Notice how each entry in the middle of the table is given by the product of its column total and row total, which means exactly $p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y)$.

Definition 7.3 (Pairwise independence)**.** We say that a collection of discrete random variables $X_1, X_2, X_3, \ldots$ is *pairwise independent* if X_j and X_k are independent for every $j \neq k$.

Definition 7.4 (Mutual independence)**.** We say that a collection of discrete random variables $X_1, X_2, X_3, \ldots$ is *mutually independent* if, for every k and every $x_1, x_2, \ldots, x_k$, we have

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_2 = x_2) \cdots \mathbb{P}(X_k = x_k).$$

Theorem 7.1 (Expectation of independent random variables)**.** *If X and Y are independent integrable discrete random variables, then XY is integrable and*

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Proof. Using Proposition 7.1,

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x \in D_X} \sum_{y \in D_Y} xy \cdot \mathbb{P}(X = x, Y = y) \\
&= \sum_{x \in D_X} x \cdot \Big(\sum_{y \in D_Y} y \cdot \mathbb{P}(X = x)\mathbb{P}(Y = y) \Big) \\
&= \sum_{x \in D_X} x \cdot \mathbb{P}(X = x) \cdot \Big(\sum_{y \in D_Y} y \cdot \mathbb{P}(Y = y) \Big) \\
&= \Big(\sum_{y \in D_Y} y \cdot \mathbb{P}(Y = y) \Big) \cdot \Big(\sum_{x \in D_X} x \cdot \mathbb{P}(X = x) \Big) \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y].
\end{aligned}$$

In order to use Proposition 7.1 we should have known that XY was integrable

in the first place. This can be checked using exactly the same development with $|x|$ instead of x and $|y|$ instead of y. This proves the theorem. □

Example 7.4. Roll a fair die twice and multiply the values observed.

$$\begin{aligned}\mathbb{E}[X] = \frac{1}{36}\big(&1\cdot 1 + 2\cdot 2 + 3\cdot 2 + 4\cdot 3 + 5\cdot 2 + 6\cdot 4 + \\ &+ 8\cdot 2 + 9\cdot 1 + 10\cdot 2 + 12\cdot 4 + 15\cdot 2 + 16\cdot 1 + \\ &+ 18\cdot 2 + 20\cdot 2 + 24\cdot 2 + 25\cdot 1 + 30\cdot 2 + 36\cdot 1\big) = \frac{49}{4}.\end{aligned}$$

A simpler solution is to observe that $X = YZ$, where Y and Z represent the first and second rolling of the die. Using the above theorem,

$$\mathbb{E}[X] = \mathbb{E}[Y]\cdot\mathbb{E}[Z] = \frac{7}{2}\cdot\frac{7}{2} = \frac{49}{4}.$$

Notice how the computation was simplified.

Proposition 7.2. *Let $X_1, \ldots, X_n$ be pairwise independent square-integrable discrete random variables. Then*

$$\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathrm{Var}(X_i)\,.$$

Proof. Let $\mu_i := \mathbb{E}[X_i],\ i = 1, 2, \ldots, n$ and set $\mu := \sum_{i=1}^{n}\mu_i = \mathbb{E}[\sum_{i=1}^{n} X_i]$ by linearity of expectation. Then,

$$\begin{aligned}\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) &= \mathbb{E}\Big[\Big(\sum_{i=1}^{n} X_i - \mu\Big)^2\Big] \\ &= \mathbb{E}\Big[\Big(\sum_{i=1}^{n}(X_i - \mu_i)\Big)^2\Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu_i)(X_j - \mu_j)\Big] \\ &= \mathbb{E}\Big[\sum_{i=1}^{n}(X_i - \mu_i)^2\Big] + \mathbb{E}\Big[\sum_{i\neq j}(X_i - \mu)(X_j - \mu_j)\Big] \\ &= \sum_{i=1}^{n}\mathbb{E}[(X_i - \mu_i)^2] + \sum_{i\neq j}\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)],\end{aligned}$$

$$= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \Big(\mathbb{E}[X_i X_j] - \mu_i \mu_j \Big)$$
$$= \sum_{i=1}^{n} \mathrm{Var}(X_i). \qquad \square$$

8 The law of averages

One of the main topics of Probability is the "law of averages." It says that the sum of a large number of pairwise independent variables tends to be close to the expectation. More precisely, the observed average $\frac{X_1+\cdots+X_n}{n}$, which is random, will approximate the theoretical average $\mathbb{E}[\frac{X_1+\cdots+X_n}{n}]$, which is deterministic!

The simplest manifestation of this phenomenon concerns relative frequencies. Imagine we run an experiment many times, under the same condition, and count how many resulted in success and how many resulted in failure. Taking X_n to be the indicator that the n-th trial resulted in success, the relative frequency of success is given exactly by While writing this preamble, the lecturers simulated tossing a fair coin a million times, and obtained Heads 499,947 times. Repeating the same procedure, they obtained Heads 499.508 times, then 500.318 times, then 500.512 times. Obviously, something is happening here. As predicted by the law of averages, the relative frequency was always very close to the probability of obtaining Heads in each toss of the coin, which is exactly $\frac{1}{2}$. Intuitively, we tend to associate probability of success to relative frequency of successes, and this association is almost ingrained in our thinking.

More generally, the relevant result of each experiment does not need to be 0 or 1 to represent failure or success. It can be any random variable. Before writing this paragraph, the lecturers rolled a die 10 times, and the sum of the values obtained as 38. Repeating the same procedure, they obtained 33, 28, 37 and finally 43. It doesn't quite look like this sum is very well concentrated near any deterministic value. However, the lectures proceeded to simulating 10,000 rolls of the die, and the sum of the results was 35,082. Repeating this procedure, they obtained as sum 34.769, then 35.419, and finally 34.691. As predicted by the law of averages, when the number of rolls of the dice was large, the observed average was always close to the theoretical average, given by $\mathbb{E}[X_1] = \frac{7}{2}$.

In the theory of Probability, the law of averages is not just a tale or a mysterious

phenomenon, it is a theorem that can have many different formulations. Here we will consider the simplest possible version.

Theorem 8.1 (Law of averages)**.** *Let $X_1, X_2, X_3, \dots$ be pairwise independent square-integrable discrete random variables with the same mean μ and same variance σ^2. Then, for every $a > 0$, and $n \in \mathbb{N}$,*

$$\mathbb{P}\Big(\mu - a \leqslant \frac{X_1 + \cdots + X_n}{n} \leqslant \mu + a\Big) \geqslant 1 - \frac{\sigma^2}{a^2\, n}.$$

The law of averages is commonly known as the Law of Large Numbers.

We draw the attention to the fact that, no matter how small a is, this probability can be made as close to 1 as we wish by taking n large enough (of course, if a is too small, we will need n to be really very large).

Our next goal is now to understand how such an estimate for probabilities even came about. So far we used probabilities to compute expectations and standard deviations, and now suddenly we are using the standard deviation to make extremely interesting estimates about probabilities!

This endeavour will have two parts: understanding what is the variance of the sum of many random variables, and understanding how the variance of a random variable provides estimates on the probability that it deviates from its mean.

9 Covariance

9.1 Definition

What happens to the variance when we add variables?

Example 9.1. Toss five fair coins. Let X be the number of Heads among the first two coin tosses, and Y be the number of Heads among the last three coin tosses. Let $Z = X + Y$ the the total number of Heads among all five tosses. After computations (that we omit), we get $\mathrm{Var}(X) = \frac{1}{2}$, $\mathrm{Var}(Y) = \frac{3}{4}$ and $\mathrm{Var}(Z) = \frac{5}{4}$. In this case, the relation $\mathrm{Var}(X + Y) = \mathrm{Var}(X + Y)$ has been verified. Is this a coincidence? Under which conditions is this relation verified?

Example 9.2. Suppose $X \sim \mathrm{Bernoulli}(\frac{1}{2})$ and let $Y = X$. In this case, we have $\mathrm{Var}(X) = \frac{1}{4}$, $\mathrm{Var}(Y) = \frac{1}{4}$, $\mathrm{Var}(X + Y) = 1$, so $\mathrm{Var}(X + Y) \neq \mathrm{Var}(X + Y)$.

To better understand what happens to $\mathrm{Var}(X+Y)$, we expand:

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

Definition 9.1 (Covariance)**.** Suppose X and Y are square-integrable discrete random variables. We define the *covariance of X and Y* as

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

From the previous calculations, we see that $\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)$ if and only if $\mathrm{Cov}(X,Y)=0$. When this condition is satisfied, we say that X and Y are *uncorrelated*.

9.2 Properties

Let us see the main properties of covariance. By switching X and Y,

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

By substituting X in the place of Y, we get

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$$

Proposition 9.1. *Suppose X, Y and Z are square-integrable discrete random variables. Then*

$$\mathrm{Cov}(aX+bY,Z) = a\,\mathrm{Cov}(X,Z)+b\,\mathrm{Cov}(Y,Z)$$

for every $a,b\in\mathbb{R}$.

Proof. Just expand and group:

$$\begin{aligned}\mathrm{Cov}(aX+bY,Z) &= \mathbb{E}[((aX+bY)-\mathbb{E}[aX+bY])(Z-\mathbb{E}[Z])] \\ &= \mathbb{E}[(aX-\mathbb{E}[aX]+bY-\mathbb{E}[bY])(Z-\mathbb{E}[Z])] \\ &= \mathbb{E}[(aX-\mathbb{E}[aX])(Z-\mathbb{E}[Z])+(bY-\mathbb{E}[bY])(Z-\mathbb{E}[Z])]\end{aligned}$$

$$= a\,\mathbb{E}[(X - \mathbb{E}[X])(Z - \mathbb{E}[Z])] + b\,\mathbb{E}[(Y - \mathbb{E}[Y])(Z - \mathbb{E}[Z])]$$
$$= a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z),$$

proving the identity. □

Corollary 9.1. *Suppose $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are square-integrable discrete random variables. Then*

$$\mathrm{Cov}\Big(\sum_{j=1}^{n} a_j X_j\,,\,\sum_{k=1}^{m} b_k Y_k\Big) = \sum_{j=1}^{n}\sum_{k=1}^{m} a_j b_k \mathrm{Cov}(X_j, Y_k)$$

for every $a_1, \ldots, a_n, b_1, \ldots, b_m \in \mathbb{R}$.

Proof. Using the previous proposition repeatedly and symmetry,

$$\begin{aligned}\mathrm{Cov}\Big(\sum_{j=1}^{n} a_j X_j\,,\,\sum_{k=1}^{m} b_j Y_j\Big) &= \sum_{j=1}^{n} a_j \mathrm{Cov}\Big(X_j, \sum_{k=1}^{m} b_j Y_j\Big)\\ &= \sum_{j=1}^{n} a_j \mathrm{Cov}\Big(\sum_{k=1}^{m} b_j Y_j, X_j\Big)\\ &= \sum_{j=1}^{n}\sum_{k=1}^{m} a_j b_j \mathrm{Cov}(Y_j, X_j)\\ &= \sum_{j=1}^{n}\sum_{k=1}^{m} a_j b_j \mathrm{Cov}(X_j, Y_j),\end{aligned}$$

which proves the stated identity. □

Corollary 9.2. *Let $X_1, \ldots, X_n$ be square-integrable discrete random variables. Then*

$$\mathrm{Var}\Big(\sum_{k=1}^{n} X_k\Big) = \sum_{k=1}^{n} \mathrm{Var}(X_k) + 2 \sum_{1 \leqslant j < k \leqslant n} \mathrm{Cov}(X_j, X_k)$$

Proof. Using the previous properties,

$$\begin{aligned}\mathrm{Var}\Big(\sum_{k=1}^{n} X_k\Big) &= \mathrm{Cov}\Big(\sum_{j=1}^{n} X_j, \sum_{k=1}^{n} X_k\Big)\\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \mathrm{Cov}(X_j, X_k)\end{aligned}$$

$$= \sum_{k=1}^{n} \mathrm{Cov}(X_k, X_k) + \sum_{j \neq k} \mathrm{Cov}(X_j, X_k)$$
$$= \sum_{k=1}^{n} \mathrm{Var}(X_k) + 2 \sum_{1 \leqslant j < k \leqslant n} \mathrm{Cov}(X_j, X_k). \qquad \square$$

Definition 9.2. We say that a collection of square-integrable discrete random variables $X_1, X_2, X_3, \ldots$ is *uncorrelated* if $\mathrm{Cov}(X_j, X_k) = 0$ for every $j \neq k$.

Corollary 9.3. *If $X_1, \ldots, X_n$ are uncorrelated discrete random variables, then*

$$\mathrm{Var}\Big(\sum_{k=1}^{n} X_k\Big) = \sum_{k=1}^{n} \mathrm{Var}(X_k).$$

Proof. Apply the previous formula and note that the covariance of different terms is zero. $\square$

This gives us the "square root law": if $X_1, \ldots, X_n$ are uncorrelated discrete random variables with the same mean μ and variance σ^2, then

$$\mathbb{E}\big[\tfrac{X_1 + \cdots + X_n}{n}\big] = \mu \qquad \text{and} \qquad \sigma\big(\tfrac{X_1 + \cdots + X_n}{n}\big) = \tfrac{\sigma}{\sqrt{n}}.$$

This starts to explain why the law of averages emerges when we add many random variables.

9.3 Sums of pairwise independent variables

Perhaps a more convenient expression for covariance:

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

If X and Y are independent, then, by Theorem 7.1, $\mathrm{Cov}(X, Y) = 0$.

Corollary 9.4. *If $X_1, \ldots, X_n$ are pairwise independent square-integrable discrete random variables, then*

$$\mathrm{Var}\Big(\sum_{k=1}^{n} X_k\Big) = \sum_{k=1}^{n} \mathrm{Var}(X_k).$$

By the previous observation, a family of pairwise independent discrete random variables is also uncorrelated.

10 Chebyshev's inequality

We will now see how the mean and standard deviation of a random variable allow us to make some estimates on probabilities involving the random variable.

10.1 Markov's inequality

In order to get there, we start with something more modest: Markov's inequality. It allows us to say something about the distribution of a random variable using knowledge of its expectation.

Theorem 10.1 (Markov's inequality)**.** *Let X be an integrable non-negative discrete random variable. Then,*

$$\mathbb{P}(X > x) \leqslant \frac{\mathbb{E}[X]}{x}$$

for every $x > 0$.

Proof. Fix $x > 0$. Define the random variable

$$Y := \begin{cases} x & \text{if } X \geqslant x; \\ 0 & \text{otherwise.} \end{cases}$$

We have that $X \geqslant Y$, because:

- if $X \geqslant x$, then $Y = x$, so $X \geqslant Y$;
- if $X \in [0, x)$, then $Y = 0$, so $X \geqslant Y$.

This also gives $\mathbb{E}[X] \geqslant \mathbb{E}[Y]$. Next, note that Y is a discrete random variable (it only attains the values 0 and x) with

$$p_Y(x) = \mathbb{P}(X \geqslant x), \qquad p_Y(0) = \mathbb{P}(X < x).$$

Hence,

$$\mathbb{E}[X] \geqslant \mathbb{E}[Y] = 0 \cdot p_Y(0) + x \cdot p_Y(x) = x \cdot p_Y(x) = x \cdot \mathbb{P}(X \geqslant x).$$

Rearranging this, we obtain the desired inequality. □

Example 10.1. Suppose a company produces an average of 50 items per week. Can you estimate the probability that a particular week's production exceeds 75 items? Let X be the number of items the company produces each week. By the statement, we know that $X \geqslant 0$ and $\mathbb{E}[X] = 50$. Hence the assumptions of Markov's inequality are satisfied and we can deduce an estimate on the requested probability, which is $\mathbb{P}(X \geqslant 75)$. Hence,

$$\mathbb{P}(X \geqslant 75) \leqslant \frac{\mathbb{E}[X]}{75} = \frac{50}{75} = \frac{2}{3}\,.$$

It is interesting to note that Markov's inequality does not always give a useful bound. Indeed, if X is a non-negative random variable with expectation equal to μ, and $x \in (0, \mu]$, then in the inequality

$$\mathbb{P}(X \geqslant x) \leqslant \frac{\mu}{x},$$

the right-hand side is larger than 1, so the bound only tells us that the probability is smaller than or equal to 1, but we knew that already!

10.2 Chebyshev's inequality

While Markov's inequality gives a bound on the probability that a random variable is large, Chebyshev's inequality gives a bound on the probability that a random variable is far from its expectation.

Theorem 10.2 (Chebyshev's inequality)**.** *Let X be a square-integrable discrete random variable. Then,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant a) \leqslant \frac{\mathrm{Var}(X)}{a^2}$$

for every $a > 0$.

We emphasize that here no assumption is made concerning the sign of X.

Proof. Let $a > 0$. Define $Y := (X - \mathbb{E}[X])^2$. Then, Y is non-negative and

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

In particular, Y is integrable. Next, note that the following events are the same:

$$\{|X - \mathbb{E}[X]| \geqslant a\} = \{(X - \mathbb{E}[X])^2 \geqslant a^2\} = \{Y \geqslant a^2\}.$$

Hence, by Markov's inequality we have

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant a) = \mathbb{P}(Y \geqslant a^2) \leqslant \frac{\mathbb{E}[Y]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}.$$

This concludes the proof of the theorem. □

Example 10.2. Suppose $\mathbb{E}[X] = 10$ and $\sigma(X) = 2$. Let us find an estimate on the probability that $6 \leqslant X \leqslant 14$. Taking $x = 4$ in Chebyshev's inequality,

$$\mathbb{P}(6 \leqslant X \leqslant 14) \geqslant \mathbb{P}(6 < X < 14) = 1 - \mathbb{P}(|X - \mathbb{E}[X]| \geqslant 4) \geqslant 1 - \frac{2^2}{4^2} = \frac{3}{4}.$$

10.3 Proof of the law of averages

Recall that $X_1, X_2, X_3, \ldots$ are pairwise independent square-integrable discrete random variables with the same mean μ and same variance σ^2.

We want to show that, for every $a > 0$, and $n \in \mathbb{N}$,

$$\mathbb{P}\Big(\mu - a \leqslant \frac{X_1 + \cdots + X_n}{n} \leqslant \mu + a\Big) \geqslant 1 - \frac{\sigma^2}{a^2\, n}.$$

Using linearity of expectation and the fact that all X_k have the same mean μ,

$$\mathbb{E}\Big[\frac{X_1 + \cdots + X_n}{n}\Big] = \frac{1}{n}\mathbb{E}\big[X_1 + \cdots + X_n\big] = \frac{1}{n} \cdot \big(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\big) = \mu.$$

Using Corollary 9.4 and the fact that all X_k have the same variance σ^2,

$$\begin{aligned}\mathrm{Var}\Big(\frac{X_1 + \cdots + X_n}{n}\Big) &= \frac{1}{n^2}\mathrm{Var}\big(X_1 + \cdots + X_n\big) \\ &= \frac{1}{n^2}\big(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\big)\end{aligned}$$

$$= \frac{1}{n^2} \cdot n\sigma^2$$
$$= \frac{\sigma^2}{n}.$$

Using Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}\Big(\mu - a \leqslant \tfrac{X_1+\cdots+X_n}{n} \leqslant \mu + a\Big) &= 1 - \mathbb{P}\Big(\Big|\frac{X_1 + \cdots + X_n}{n} - \mu\Big| > a\Big) \\ &\geqslant 1 - \frac{\mathrm{Var}(\frac{X_1+\cdots+X_n}{n})}{a^2} \\ &= 1 - \frac{\sigma^2}{a^2 \cdot n},\end{aligned}$$

which concludes the proof.

11 Correlation coefficient

The covariance is a useful quantity that describes how two random variables vary together. However, it has one disadvantage: it is not scale invariant. To explain what this means, suppose that X and Y are two random variables, both measuring lengths in meters. Assume that U and V give the same measurements as X and Y, respectively, but in centimetres, that is, $U = 100X$ and $V = 100Y$. Then,

$$\mathrm{Cov}(U, V) = \mathrm{Cov}(100X, 100Y) = 100 \cdot 100 \cdot \mathrm{Cov}(X, Y) = 10^4 \cdot \mathrm{Cov}(U, V).$$

This means that changing the scale also changes the covariance. To obtain a scale-invariant quantity, we make the following definition.

Definition (Correlation coefficient)**.** Let X and Y be square-integrable discrete random variables with positive variance. The *correlation coefficient* between X and Y is defined as

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)}.$$

As promised, the correlation coefficient does not change when we rescale or shift the random variables.

Proposition. *Let X and Y be square-integrable random variables with positive*

variance. For every $a, b, c, d \in \mathbb{R}$ with $a, c > 0$, we have

$$\rho(aX + b,\ cY + d) = \rho(X, Y).$$

Proof. We first note that the covariance between a constant and any other random variable is equal to zero. Indeed,

$$\mathrm{Cov}(b, Y) = \mathbb{E}[bY] - \mathbb{E}[b] \cdot \mathbb{E}[Y] = b \cdot \mathbb{E}[Y] - b \cdot \mathbb{E}[Y] = 0.$$

Therefore,

$$\mathrm{Cov}(aX + b,\ cY + d) = ac \cdot \mathrm{Cov}(X, Y).$$

Substituting this into the formula for the correlation coefficient,

$$\begin{aligned} \rho(aX + b,\ cY + d) &= \frac{\mathrm{Cov}(aX + b,\ cY + d)}{\sigma(aX + b) \cdot \sigma(cY + d)} \\ &= \frac{ac \cdot \mathrm{Cov}(X, Y)}{a \cdot \sigma(X) \cdot c \cdot \sigma(Y)} \\ &= \frac{\mathrm{Cov}(X, Y)}{\sigma(X) \cdot \sigma(Y)} = \rho(X, Y). \end{aligned}$$ □

Also, since $\sigma(-Y) = \sigma(Y)$ and $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$, the correlation coefficient also satisfies.

$$\rho(X, Y) = \rho(Y, X) \text{ and } \rho(X, -Y) = -\rho(X, Y).$$

The next proposition further describes in what sense the correlation coefficient $\rho(X, Y)$ is a dimensionless index that quantifies how well X and Y are aligned, see Figure 11.1 for a very visual description.

Proposition. *Let X and Y be square-integrable random variables with positive variance. Then*

$$-1 \leqslant \rho(X, Y) \leqslant 1.$$

In the extreme cases, $\rho(X, X) = 1$ and $\rho(X, -X) = -1$.

Proof. Observe that

$$\left(\frac{X - \mathbb{E}[X]}{\sigma(X)} - \frac{Y - \mathbb{E}[Y]}{\sigma(Y)} \right)^2 \geqslant 0$$

Taking expectation and expanding,

$$\frac{\mathbb{E}[(X-\mathbb{E}[X])^2]}{\sigma^2(X)}+\frac{\mathbb{E}[(Y-\mathbb{E}[Y])^2]}{\sigma^2(Y)}-2\frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sigma(X)\sigma(Y)}\geqslant 0,$$

which means

$$2\rho(X,Y)\leqslant 2,$$

so $\rho(X,Y)\leqslant 1$. The same argument with $-Y$ instead of Y gives $\rho(X,Y)=-\rho(X,-Y)\leqslant 1$, whence $\rho(X,Y)\geqslant -1$. Finally,

$$\rho(X,X)=\frac{\operatorname{Cov}(X,Y)}{\sigma(X)\cdot\sigma(X)}=1$$

and $\rho(X,-X)=-\rho(X,X)=-1$. □

12 Central Limit Theorem

Theorem 12.1 (Central Limit Theorem)**.** *Let $X_1, X_2, X_3, \ldots$ be mutually independent square-integrable discrete random variables with the same distribution. Denote their mean by μ and variance by $\sigma^2>0$. Then, for every $a<b$*

$$\mathbb{P}\Big(a\leqslant\frac{X_1+\cdots+X_n-n\cdot\mu}{\sigma\cdot\sqrt{n}}\leqslant b\Big)\approx\int_a^b\tfrac{1}{\sqrt{2\pi}}e^{-x^2/2}\,\mathrm{d}x.$$

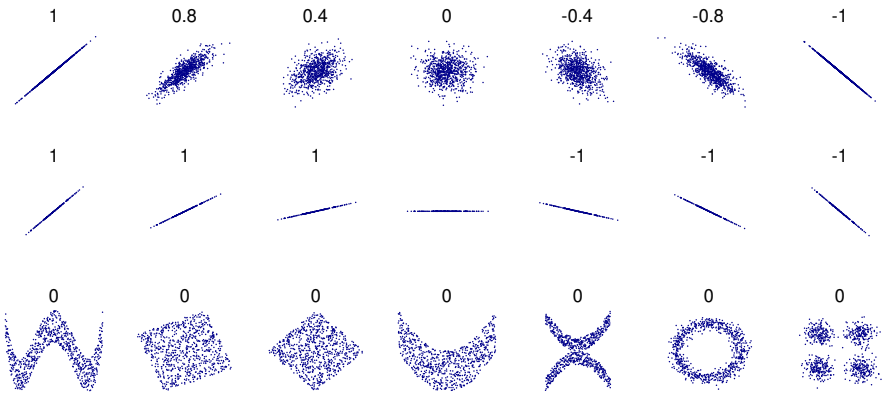


Figure 11.1: Illustration of $\rho(X,Y)$ assuming the pair (X,Y) has the same probability to be each point in the depicted cloud. (taken form Wikipedia)

The approximation "$\approx$" means that the probability gets as close to the integral (shown in Figure 12.1) as we wish if we pick n large enough.

This remarkable phenomenon is in the heart of statistics and most natural sciences. It says that, *regardless of the distribution of* X, if we add enough many samples of X together, we only see its mean μ and variance σ^2.

We will not prove the Central Limit Theorem in this module as some more advanced tools are needed.

In case $X \sim \text{Bernoulli}(\frac{1}{2})$, which corresponds to tossing a fair coin, we can see visually how the distribution of $X_1 + \cdots + X_n$ approximates this function $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, it is illustrated in Figure 12.2.

Example 12.1. When counting the votes in a very close election, 25,301 votes have been counted: 12,636 for Candidate A and 12,665 for Candidate B. There are still 400 votes to be counted. What is the probability that Candidate B wins the election? Assuming each vote is a fair coin toss, the question are are asking is about

$$\mathbb{P}(X_1 + \cdots + X_{400} \geqslant 215),$$

where $X_1, \ldots, X_n$ are independent with distribution $\text{Bernoulli}(\frac{1}{2})$. By symmetry, this is the same as $\mathbb{P}(X_1 + \cdots + X_{400} \leqslant 185)$, and therefore it equals

$$\frac{1}{2} \cdot \Big[1 - \mathbb{P}(185 < X_1 + \cdots + X_{400} < 215)\Big].$$

Since $\mu = \frac{1}{2}$ and $\sigma = \frac{1}{2}$, we conveniently rewrite the event to get

$$\frac{1}{2} - \frac{1}{2} \cdot \mathbb{P}\Big(-1.5 < \frac{X_1 + \cdots + X_{400} - 400 \cdot \mu}{\sigma\sqrt{400}} < 1.5\Big)$$

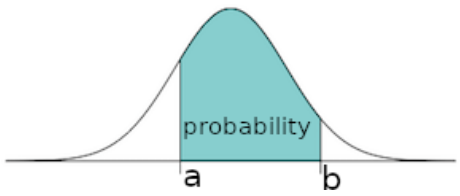


Figure 12.1: Graph of $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and the probability that $Z \in [a, b]$ represented by the greenish area between points a and b.

and, using the Central Limit Theorem, we approximate it by

$$\frac{1}{2} - \frac{1}{2}\int_{-1.5}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x \approx 0.07.$$

You should not try to compute this integral at home, the only way to get this value is by looking at a table, we will see more about this later. So the answer is 0.07, or 7%. Note that we only gave a rough answer with one significant figure. In order to get more precision than that, some more careful considerations would be necessary and would be the topic of more advanced modules.

We can also use "one-sided versions" of the Central Limit Theorem. This way, the previous example is simplified:

$$\begin{aligned}\mathbb{P}(X_1 + \cdots + X_{400} \geqslant 215) &= \mathbb{P}\Big(\frac{X_1 + \cdots + X_{400} - 400 \cdot \mu}{\sigma\sqrt{400}} \geqslant 1.5\Big) \\ &\approx \int_{1.5}^{+\infty} \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x \\ &\approx 0.07.\end{aligned}$$

If we rewrite the Central Limit Theorem as

$$\mathbb{P}\Big(\mu + \frac{\sigma}{\sqrt{n}}a \leqslant \frac{X_1 + \cdots + X_n}{n} \leqslant \mu + \frac{\sigma}{\sqrt{n}}b\Big) \approx \int_a^b \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x,$$

we get a good description of the statistical behaviour of the observed average

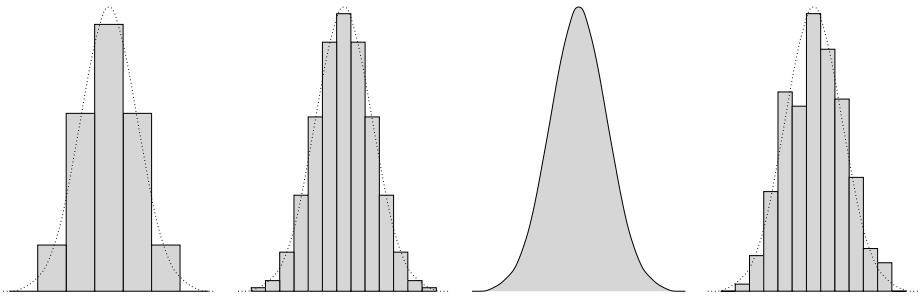


Figure 12.2: The first two graphs are the probability mass functions of $\mathrm{Binom}(n, \frac{1}{2})$ rescaled so as to show $\pm 3\sigma$. The third graph is the graph of $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The fourth graph displays the relative frequencies in a random sample of 200 independent random variables with distribution $\mathrm{Binom}(16, \frac{1}{2})$.

$\frac{X_1+\cdots+X_n}{n}$. As predicted by the law of averages, the observed average is concentrated around μ, but now we can say something more precise. The observed average fluctuates as $\mu + \sigma\frac{1}{\sqrt{n}}Z$, where Z is this "thing" described by

$$\mathbb{P}(a \leqslant Z \leqslant b) = \int_a^b \tfrac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x.$$

Random variables described in terms of integrals are called *continuous random variables*, the topic of the next section.

13 Continuous random variables

13.1 Probability density function

Consider the random variable X informally defined by: "Let X be a number chosen in the interval $(0,1)$ uniformly at random." Upon careful thought, we see that this description is a bit puzzling. The word "uniformly" should mean that any number $x \in (0,1)$ is equally likely to be picked, that is, $\mathbb{P}(X = x)$ should be the same for each x. However, there are infinitely many $x \in (0,1)$, so this would force $\mathbb{P}(X = x)$ to be zero. This will indeed be the case for all random variables with continuous distributions, which we now define.

Definition 13.1. We say that a random variable X is *continuous with probability density function* $f_X : \mathbb{R} \to \mathbb{R}$ if

$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(x)\,\mathrm{d}x$$

for every $a < b \in \mathbb{R}$.

A density f_X specifies the "probability per unit length." It is somewhat analogous to the probability mass function, but not exactly. The probability that X is in a small interval of length Δx is given by $f_X(x)\Delta x$ rather than $f_X(x)$ and f_X itself may take very large values on small intervals. So it is "$f_X(x)\Delta x$" which is the analogous to $p_X(x)$.

A density function necessarily satisfies

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1.$$

13.2 Uniform variables

Let $a, b \in \mathbb{R}$ with $a < b$. A random variable X has the (continuous) uniform distribution on (a, b) if it has

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

We write $X \sim \mathcal{U}(a, b)$.

Observe that, for any interval $[c, d] \subseteq [a, b]$, we have

$$\mathbb{P}(X \in [c, d]) = \int_c^d \frac{1}{b-a}\mathrm{d}x = \frac{d-c}{b-a},$$

that is, the interval $[c, d]$ is assigned a probability given by the *proportion* of length that it has inside the interval $[a, b]$. On the other hand, if $c < a$ and $d \in [a, b]$, then

$$\mathbb{P}(X \in [c, d]) = \frac{d-a}{b-a}$$

because the part of the interval $[c, d]$ that is not overlapping with $[a, b]$ does not count.

13.3 Normal distribution

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. A random variable X has normal (or Gaussian) distribution with parameters μ and σ^2 if it has probability density function given by

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

for every $x \in \mathbb{R}$. We write $X \sim \mathcal{N}(\mu, \sigma^2)$.

The parameter μ gives the centre of the density function, and the parameter

σ^2 specifies the scale of how this density is being stretched. Through a change of variables in the integral, we can see that $X \sim \mathcal{N}(\mu, \sigma^2)$ is equivalent to $X = \mu + \sigma \cdot Z$ with $Z \sim \mathcal{N}(0, 1)$. Indeed, defining $Z = \frac{X-\mu}{\sigma}$,

$$\begin{aligned}\mathbb{P}(a \leqslant Z \leqslant b) &= \mathbb{P}(\mu + a\sigma \leqslant X \leqslant \mu + b\sigma) \\ &= \int_{\mu+a\sigma}^{\mu+a\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x \\ &= \int_a^b \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} \,\mathrm{d}z.\end{aligned}$$

Remark 6. It is not easy to see that $\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1$. By substituting,

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \,\mathrm{d}x = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{\pi}} \cdot e^{-u^2} \,\mathrm{d}u.$$

So it is enough to check that $\int_{-\infty}^{+\infty} e^{-x^2} = \sqrt{\pi}$. In most textbooks, this is done using polar coordinates for integrals on $\mathbb{R}^2$, but we don't want to use that. Instead, we use a different trick: switch the iterated integrals in

$$\int_0^{+\infty} \left(\int_0^{+\infty} y e^{-(1+x^2)y^2} \mathrm{d}y \right) \mathrm{d}x = \int_0^{+\infty} \left(\int_0^{+\infty} y e^{-x^2y^2} e^{-y^2} \mathrm{d}x \right) \mathrm{d}y,$$

which we can do because the integrand is non-negative. The first integral can be computed by

$$\lim_{z\to+\infty} \int_0^z y e^{-(1+x^2)y^2} \mathrm{d}y = \lim_{z\to+\infty} \frac{-1}{2(1+x^2)} \left[e^{-(1+x)^2y^2} \right]_0^z = \frac{1}{2(1+x^2)}$$

and

$$\int_0^{+\infty} \left(\int_0^{+\infty} y e^{-(1+x^2)y^2} \mathrm{d}y \right) \mathrm{d}x = \lim_{z\to+\infty} \int_0^z \frac{1}{2(1+x^2)} \mathrm{d}x = \lim_{z\to+\infty} \frac{\arctan z}{2} = \frac{\pi}{4}.$$

The second integral can be rewritten as

$$\int_0^{+\infty} \left(\int_0^{+\infty} e^{-u^2} \mathrm{d}u \right) e^{-y^2} \mathrm{d}y.$$

This way we conclude that $\int_0^{+\infty} e^{-x^2} = \frac{\sqrt{\pi}}{2}$. By symmetry, $\int_{-\infty}^{+\infty} e^{-x^2} = \sqrt{\pi}$, which is what we were after.

13.4 Exponential lifetimes

Let $\lambda > 0$. A random variable X has the exponential distribution with parameter λ if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0, \\ 0 & \text{if } x \leqslant 0. \end{cases}$$

We write $X \sim \mathrm{Exp}(\lambda)$.

Note that

$$\mathbb{P}(X > t) = e^{-\lambda t}.$$

The exponential distribution is commonly used to model the lifetime of entities that have a *lack of memory* property, normally inanimate objects that experience no ageing effect. To explain what this means, let us think of light bulbs. Suppose that the lifetime of light bulbs of a particular brand has an Exponential(λ) distribution (assume that we turn on the light and don't turn it off until the bulb burns out). Then, the memoryless property means that: regardless of whether the bulb has just been activated, or it has been active for a certain amount of time, the distribution of the remaining lifetime is the same. Mathematically, this is expressed by the following identity, which holds for all $s, t \geqslant 0$:

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

13.5 Expectation

Definition 13.2 (Expectation)**.** Let X be a continuous random variable with density f_X. We define the *expectation* of X, denoted $\mathbb{E}[X]$, as the real number given by

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, \mathrm{d}x,$$

as long as this integral converges absolutely. If the integral is absolutely convergent, we say that X is *integrable*, otherwise $\mathbb{E}[X]$ is not defined.

Example 13.1 (Uniform)*.* If $X \sim \mathcal{U}[a, b]$, then

$$\mathbb{E}[X] = \int_a^b x\, \frac{1}{b-a}\, \mathrm{d}x = \frac{a+b}{2}.$$

That is, the expectation of a random variable with uniform distribution on $[a, b]$ is the middle point of the interval.

Example 13.2 (Exponential). If $X \sim \mathrm{Exp}(\lambda)$, then, integrating by parts,

$$\mathbb{E}[X] = \int_0^{+\infty} x\lambda e^{-\lambda x}\,\mathrm{d}x = \lim_{u\to+\infty}\left[-xe^{-\lambda x} - \tfrac{1}{\lambda}e^{-\lambda x}\right]_0^u = \frac{1}{\lambda}.$$

Example 13.3 (Normal). Suppose $X \sim \mathcal{N}(0,1)$, Then, by substitution $u = x^2/2$,

$$\int_0^{+\infty} x\frac{e^{-x^2/2}}{\sqrt{2\pi}}\mathrm{d}x = \lim_{z\to+\infty}\left[\frac{-e^{-x^2/2}}{\sqrt{2\pi}}\right]_0^z = \frac{1}{\sqrt{2\pi}}.$$

By symmetry,

$$\int_{-\infty}^0 x\frac{e^{-x^2/2}}{\sqrt{2\pi}}\mathrm{d}x = -\frac{1}{\sqrt{2\pi}}$$

and, therefore $\mathbb{E}[X] = 0$.

Example 13.4 (Cauchy). Suppose X is a random variable with density

$$f_X(x) = \frac{1}{\pi\cdot(1+x^2)}.$$

Then

$$\int_0^{+\infty} xf_X(x)\,\mathrm{d}x = \int_0^{+\infty}\frac{x}{\pi\cdot(1+x^2)}\mathrm{d}x \geqslant \int_1^{+\infty}\frac{x}{\pi\cdot(1+x^2)}\mathrm{d}x,$$

and thus

$$\int_0^{+\infty} xf_X(x)\,\mathrm{d}x \geqslant \int_1^{+\infty}\frac{1}{2\pi x}\mathrm{d}x = \frac{1}{2\pi}\lim_{z\to\infty}\ln z = +\infty.$$

In this case $\mathbb{E}[X]$ is *not defined*, despite the symmetry. We finally have an example of a random variable that is *not integrable*!

13.6 Variance

Proposition 13.1. *Let X be a continuous random variable with density f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous function. Then*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)\,f_X(x)\,\mathrm{d}x$$

if the integral is absolutely convergent, and $\mathbb{E}[g(X)]$ *is undefined if not.*

The analogy with probability mass function is summarised in Table 1.

Example 13.5 (Uniform). If $X \sim \mathcal{U}[a, b]$, then

$$\mathbb{E}[X^2] = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{a^2+ab+b^2}{3}.$$

Example 13.6 (Exponential). If $X \sim \mathrm{Exp}(\lambda)$, then, integrating by parts twice,

$$\mathbb{E}[X^2] = \int_0^{+\infty} x^2\lambda e^{-\lambda x}\,\mathrm{d}x = \lim_{z\to+\infty}\left[-x^2e^{-\lambda x} - \tfrac{2x}{\lambda}e^{-\lambda x} - \tfrac{2}{\lambda^2}e^{-\lambda x}\right]_0^z = \frac{2}{\lambda^2}.$$

Example 13.7 (Normal). If $X \sim \mathcal{N}(0,1)$, then, integrating by parts,

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_{-\infty}^{+\infty} x^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}}\mathrm{d}x \\
&= 2\cdot\frac{1}{\sqrt{2\pi}}\int_0^{+\infty} x\cdot(xe^{-x^2/2})\,\mathrm{d}x \\
&= \frac{2}{\sqrt{2\pi}}\lim_{u\to+\infty}\left[-xe^{-x^2/2} + \int_0^u e^{-x^2/2}\,\mathrm{d}x\right]_0^u \\
&= \frac{2}{\sqrt{2\pi}}\int_0^{+\infty} e^{-x^2/2}\,\mathrm{d}x = 1.
\end{aligned}$$

p_X	f_X
$p_X : \mathbb{R} \to \mathbb{R}$	$f_X : \mathbb{R} \to \mathbb{R}$
$p_X(x) \geqslant 0\ \forall\, x \in \mathbb{R}$	$f_X(x) \geqslant 0\ \forall\, x \in \mathbb{R}$
$\mathbb{P}(X = x) = p_X(x)$	$\mathbb{P}(X = x) = 0$
$\mathbb{P}(a \leqslant X \leqslant b) = \sum_{a\leqslant x\leqslant b} p_X(x)$	$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f_X(x)\mathrm{d}x$
$p_X(x) = \mathbb{P}(X = x)$	defined implicitly by above
$\sum_x p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
$p_X(x) \leqslant 1\ \forall\, x$	$f_X(x)$ may be > 1
$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f_X(x)\mathrm{d}x$

Table 1: Probability mass function and probability density function.

Definition 13.3 (Square-integrable)**.** Like for discrete random variables, we say that a continuous random variable is *square-integrable* if X^2 is integrable.

As in the discrete case, if a random variable is square-integrable, then it is automatically integrable (because $|x| \leqslant 1 + x^2$).

Definition 13.4 (Variance)**.** Let X be a square-integrable continuous random variable with density f_X and mean $\mu = \mathbb{E}[X]$. We define the *variance* of X as

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

As we did for discrete random variables, we can expand the definition of variance to get an alternative formula:

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

which we use in the following examples.

Example 13.8 (Uniform). If $X \sim \mathcal{U}[a, b]$, then

$$\mathrm{Var}(X) = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{(b-a)^2}{12}.$$

Example 13.9 (Exponential). If $X \sim \mathrm{Exp}(\lambda)$, then

$$\mathrm{Var}(X) = \frac{2}{\lambda^2} - \Big(\frac{1}{\lambda}\Big)^2 = \frac{1}{\lambda^2}.$$

Example 13.10 (Normal). If $X \sim \mathcal{N}(0, 1)$, then

$$\mathrm{Var}(X) = 1 - 0^2 = 1.$$

14 A single theory for discrete and continuous

14.1 Cumulative distribution function

One way to specify a probability distribution on $\mathbb{R}$ is to say how much probability is to the left of each point x. In terms of a random variable X with the given distribution, this probability is a function of x.

Definition 14.1. Let X be a random variable. The *cumulative distribution function of* X is the function $F_X : \mathbb{R} \to [0, \infty)$ defined by

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

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

Other probabilities from F_X

We now observe that, although $F_X(x)$ is defined as $\mathbb{P}(X \leqslant x)$, it is possible to use F_X to obtain other probabilities involving X. Important formulas are

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leqslant x) = 1 - F_X(x)$$

and, for $x < y$,

$$\mathbb{P}(x < X \leqslant y) = \mathbb{P}(X \leqslant y) - \mathbb{P}(X \leqslant x) = F_X(y) - F_X(x).$$

Observe also that

$$\mathbb{P}(X = x) > 0 \quad \text{if and only if} \quad F_X \text{ has a jump at } x$$

and in that case, the size of the jump is the probability that $X = x$.

F_X determines $\mathbb{P}_X$

Proposition 14.1. *If X and Y are two random variables with $F_X = F_Y$, then X and Y have the same distribution.*

This proposition tells us that the cumulative distribution function indeed encodes the distribution of a random variable (in the sense that given the cumulative distribution function, there is only one distribution corresponding to it).

We have already seen that F_X determines $\mathbb{P}_X(\{x\})$ for each $x \in \mathbb{R}$ and $\mathbb{P}_X((a, b])$ for every $a < b \in \mathbb{R}$.

We lack the tools needed to prove that it determines $\mathbb{P}_X(B)$ for every $B \in \mathcal{B}$.

14.2 Discrete and continuous cases

To get a first idea of what a cumulative distribution function looks like, let us consider the case where X is discrete and has discrete support contained in $\mathbb{N}_0$, so that

$$\sum_{k=0}^{\infty} \mathbb{P}(X = k) = 1.$$

Then, first observe that $F_X(x) = \mathbb{P}(X \leqslant x) = 0$ for all $x < 0$. Next, and for every $x \in [0, 1)$,

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X < 0) + \mathbb{P}(X = 0) + \mathbb{P}(0 < X \leqslant x) = 0 + p_X(0) + 0 = p_X(0).$$

By arguing similarly, we conclude that

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ p_X(0) & \text{if } x \in [0, 1), \\ p_X(0) + p_X(1) & \text{if } x \in [1, 2), \\ p_X(0) + p_X(1) + p_X(2) & \text{if } x \in [2, 3) \\ \cdots \end{cases}$$

The graph of F_X looks like the one in Figure 14.1.

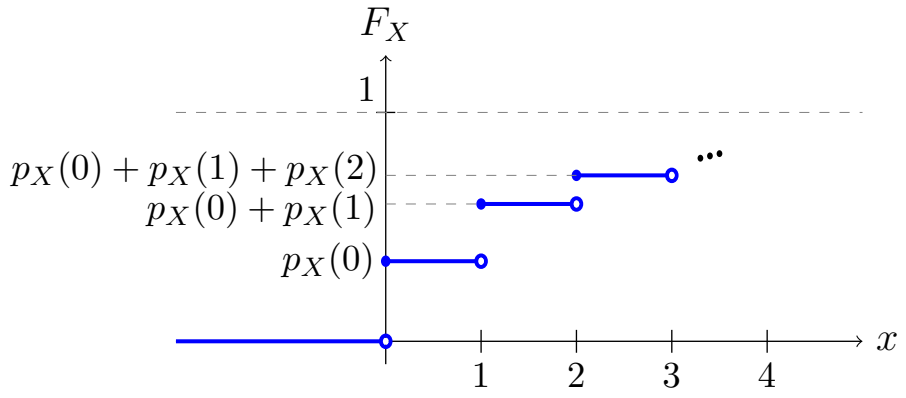


Figure 14.1: Cumulative distribution function of a discrete random variable.

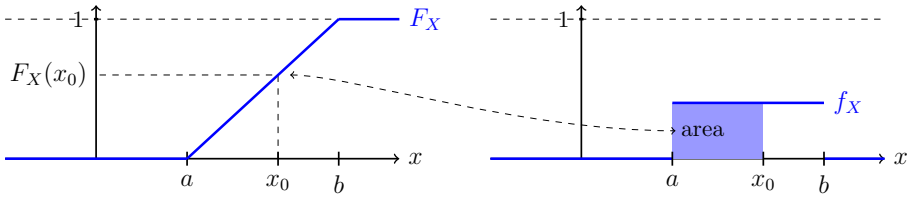


Figure 14.2: Cumulative distribution function of a uniform random variable.

If X is continuous, then

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(X \in (-\infty, x]) = \int_{-\infty}^{x} f_X(y)\mathrm{d}y.$$

The fundamental theorem of Calculus then implies that, at points where f_X is continuous,

$$f_X(x) = \tfrac{\mathrm{d}}{\mathrm{d}x} F_X'(x),$$

that is, the cumulative distribution function is differentiable and its derivative is the probability density function. The graph of F_X looks like the one in Figures 14.2 and 14.3.

14.3 Expectation and variance

It is possible to give a unified definition of expectation of a random variable, without assuming that it is discrete or that it has a density. There a magic formula using F_X that works simultaneously for any type of random variable. We will not bother giving such a formula, but it is important to keep in mind that

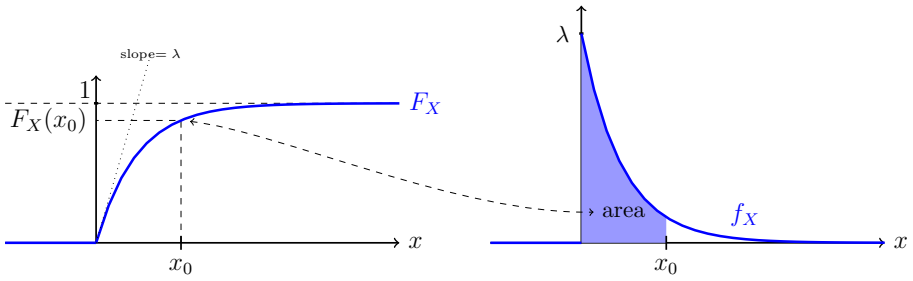


Figure 14.3: Cumulative distribution function of an exponential random variable.

expectation is something that can be defined for any bounded random variable (and, provided some sums or integrals are convergent, can also be defined for unbounded random variables). Again we say that X is *integrable* if $\mathbb{E}[X]$ is defined and finite.

This general definition of expectation still satisfies the three properties:

- Unitary: $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$,
- Monotone: If $0 \leqslant Z \leqslant X$ for all $\omega \in \Omega$ then $0 \leqslant \mathbb{E}[Z] \leqslant \mathbb{E}[X]$,
- Linear: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$

as long as X and Y are integrable.

We will not prove these properties. Of course, we could not possibly prove them since we have not even given the general definition of expectation. But even if we had written down the formula, with the current tools we would not be able to prove that the expectation is linear in general. The idea of the proof is the following: any random variables X and Y can be approximated by discrete random variables X' and Y' and, since $\mathbb{E}[X' + Y'] = \mathbb{E}[X'] + \mathbb{E}[Y']$, we conclude that $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Again, and X is *square-integrable* if $\mathbb{E}[X^2]$ is finite, and note that if X is square-integrable then it is automatically integrable (because $|x| \leqslant 1 + x^2$).

Definition 14.2 (Variance)**.** The *variance* of a square-integrable random variable X is defined as

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

We remark that Chebyshev's inequality holds for any square-integrable random variable. Indeed, in the proof given in Section 10.2, we only used the above properties of expectation and nothing else.

Definition 14.3 (Covariance)**.** We also define the *covariance* of two square-integrable random variables as

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

and say that they are *uncorrelated* if their covariance is zero.

Note that the covariance has all the properties stated in Section 9.2. Indeed, the proof of those properties only used the above three properties of expectation and nothing else. In particular Corollary 9.3 holds in general, that is,

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$$

provided that $X_1, \ldots, X_n$ are uncorrelated.

Independence is discussed in the next section.

15 Joint distributions and independence

15.1 Joint density

We now explain what it means for two random variables to follow a jointly continuous distribution.

Definition 15.1 (Joint density)**.** Two random variables X and Y defined in the same probability space are called *jointly continuous* with *joint probability density function* $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ if

$$\mathbb{P}(a_1 \leqslant X \leqslant a_2,\ b_1 \leqslant Y \leqslant b_2) = \int_{a_1}^{a_2} \Big(\int_{b_1}^{b_2} f_{X,Y}(x,y)\ \mathrm{d}y \Big) \mathrm{d}x$$

for every $a_1 < a_2$ and $b_1 < b_2$.

The *marginal density functions* are given by:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x.$$

Shortly put, marginal densities are obtained from the joint density by "integrating out" the other variable.

Example 15.1. Consider the square $Q = \{(x,y) \in \mathbb{R}^2 : |x| + |y| < 1\}$, and suppose a joint probability density function of X and Y is $f_{X,Y}(x,y) = \frac{1}{2}\mathbb{1}_Q(x,y)$. We can determine the marginal density function of X by integrating:

$$f_X(x) = \int_{-\infty}^{+\infty} \tfrac{1}{2}\mathbb{1}_Q(x,y)\,\mathrm{d}y = (1 - |x|)\mathbb{1}_{[-1,1]}(x).$$

If we integrate with respect to x, we obtain $f_Y(y) = (1 - |y|)\mathbb{1}_{[-1,1]}(y)$.

15.2 Joint cumulative distribution function

Definition 15.2. Let X, Y be random variables. The *joint cumulative distribution function of* (X, Y) is the function $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by

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

for every $x, y \in \mathbb{R}$.

For discrete, random variables, this definition reduces to

$$F_{X,Y}(x, y) = \sum_{s \leqslant x} \sum_{t \leqslant y} p_{X,Y}(s, t)$$

for every $x, y \in \mathbb{R}$. For jointly continuous random variables, the joint cumulative distribution function is given by

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t)\, \mathrm{d}t\, \mathrm{d}s.$$

In either case, it is possible to recover the joint probability mass function and the joint probability density function from the joint cumulative distribution function, but we will not dig into that.

15.3 Independence

Definition 15.3 (Independence of two random variables)**.** Given any two random variables X and Y we say that X *and* Y *are independent* if

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

for all $A, B \in \mathcal{B}$.

For discrete random variables, this definition is equivalent to

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

for every $x, y \in \mathbb{R}$, that we gave in Section 7.3.

Amazingly enough, the joint cumulative distribution function is able to capture whether two random variables are independent or not.

Proposition 15.1. *Two random variables X and Y are independent if and only if*

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

for every $x, y \in \mathbb{R}$.

Fragment of proof. For $A = (a, b]$ and $B = (c, d]$, we expand to get

$$\begin{aligned}
\mathbb{P}(a < X \leqslant b, c < Y \leqslant d) &= \mathbb{P}(a < X \leqslant b, Y \leqslant d) - \mathbb{P}(a < X \leqslant b, Y \leqslant c) \\
&= \mathbb{P}(X \leqslant b, Y \leqslant d) - \mathbb{P}(X \leqslant a, Y \leqslant d) - \mathbb{P}(a < X \leqslant b, Y \leqslant c) \\
&= F_{X,Y}(b, d) - F_{X,Y}(a, d) - [F_{X,Y}(b, c) - F_{X,Y}(a, c)] \\
&= F_X(b)F_Y(d) - F_X(a)F_Y(d) - [F_X(b)F_Y(c) - F_X(a)F_Y(c)] \\
&= [F_X(b) - F_X(a)] \cdot [F_Y(d) - F_Y(c)] \\
&= \mathbb{P}(a < X \leqslant b)\mathbb{P}(c < Y \leqslant d).
\end{aligned}$$

Therefore, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ when both A and B are given by a union of finitely many intervals of this form (open on the left and closed on the right). The case of more general sets A and B requires tools that we currently lack. □

For jointly continuous random variables, the situation is a bit trickier. Suppose X and Y have f_X and f_Y as density functions. Then they are independent if and only if they jointly continuous with a joint density function given by

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

for every $x, y \in \mathbb{R}$. However, if we are given a joint density function $f_{X,Y}$ and want to show that X and Y are *not* independent, it is not enough to find a single point (x, y) such that $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$. We need to check that $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$ for all $x \in [a, b]$ and all $y \in [c, d]$ for some non-degenerate intervals $[a, b]$ and $[c, d]$. This is because, unlike the probability mass

function, probability density functions are not unique, and we can modify them on a single point causing the identity displayed above to break down.

Theorem 15.1. *If X and Y are independent integrable random variables, then XY is integrable and*

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

In particular, independent square-integrable random variables are uncorrelated.

We have given a proof assuming X and Y are discrete. A proof that works in the general case requires tools that we currently lack. As for linearity of expectation, it is based on the fact that any random variable can be approximated by discrete random variables, reducing the problem to the case that we already know.

Definition 15.4 (Pairwise independence)**.** We say that a collection of random variables $X_1, X_2, X_3, \ldots$ is *pairwise independent* if X_j and X_k are independent for every $j \neq k$.

Definition 15.5 (Mutual independence)**.** We say that a collection of discrete random variables $X_1, X_2, X_3, \ldots$ is *mutually independent* if, for every k and every $A_1, \ldots, A_k \in \mathcal{B}$, we have

$$\mathbb{P}(X_1 \in A_1, \ldots, X_k \in A_k) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_k \in A_k).$$

There are analogous conditions for mutual independence when the random variables are continuous or discrete, and also an equivalent condition in terms of joint cumulative distribution functions. But we will not dig into this.

15.4 Covariance and the law of averages

As mentioned in the previous section, if $X_1, \ldots, X_n$ are uncorrelated random variables, then

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

Using this and Chebyshev's inequality, we can again prove the law of averages for any sequence of uncorrelated random variables with same mean μ and same

variance σ^2, without assuming that they are discrete. The proof is identical to the one seen in Section 10.3.

The Central Limit Theorem also holds for any sequence of mutually independent square-integrable random variables with the same distribution, without assuming that they are discrete.

16 Sums of independent random variables

Sums of independent random variables show up in many different contexts. Given two independent random variables X and Y, what is the distribution of $X + Y$?

If X and Y are both discrete, $X+Y$ is discrete and its probability mass function can be computed using the Law of Total Probability:

$$\begin{aligned} p_{X+Y}(z) &= \sum_x \mathbb{P}(X = x, Y = z - x) = \sum_x \mathbb{P}(X = x)\mathbb{P}(Y = z - x) \\ &= \sum_x p_X(x) p_Y(z - x). \end{aligned}$$

Example 16.1. Suppose $X \sim \text{Binom}(n, p)$ and $Y \sim \text{Binom}(m, p)$ The probability mass function of $X + Y$ can be obtained as

$$\begin{aligned} p_{X+Y}(k) &= \sum_{j=0}^{\infty} \mathbb{P}(X = j)\mathbb{P}(Y = k - j) \\ &= \sum_{j=0}^{k} \tbinom{n}{j} p^j (1-p)^{n-j} \tbinom{m}{k-j} p^{k-j} (1-p)^{m-k+j} \\ &= p^k (1-p)^{m+n-k} \sum_{j=0}^{k} \tbinom{n}{j} \tbinom{m}{k-j} \\ &= \tbinom{n+m}{k} p^k (1-p)^{m+n-k}. \end{aligned}$$

When the variables X and Y are independent and have densities f_X and f_Y, we have the analogous relation

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z - x)\, \mathrm{d}x.$$

Example 16.2 (Exponentials and Gamma). Let X and Y be independent, both with the Exponential distribution with parameter $\lambda > 0$, that is,

$$f_X(x) = f_Y(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Let $Z := X + Y$. We want to compute

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(z - x)\, \mathrm{d}x.$$

Now, note that the product inside the integral is equal to zero when $x < 0$ (since $f_X(x) = 0$ then) and when $x > z$ (since $f_Y(z - x) = 0$ then). The integral is then equal to

$$\int_0^z \lambda e^{-\lambda x} \cdot \lambda e^{-\lambda(z-x)}\, \mathrm{d}x = \lambda^2 \cdot \int_0^z e^{-\lambda z}\, \mathrm{d}x = \lambda^2 \cdot z \cdot e^{-\lambda z}.$$

The above distribution corresponds do a Gamma distribution with parameters 2 and λ. In general, Z has Gamma distribution with parameters n and λ if it has density given by $f_Z(z) = \frac{\lambda^n}{(n-1)!} \cdot z^{n-1} \cdot e^{-\lambda x}$ for $z \geqslant 0$.

The case when X and Y are normal is so important that we state it as a proposition.

Proposition 16.1. *Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. Then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Proof. Since $X_1 - \mu_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $X_2 - \mu_2 \sim \mathcal{N}(0, \sigma_2^2)$, we can suppose that $\mu_1 = \mu_2 = 0$. After long and laborious algebraic manipulations, it is possible to obtain

$$f_{X+Y}(z) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} e^{-\frac{(z-x)^2}{2\sigma_2^2}} e^{-\frac{x^2}{2\sigma_1^2}}\, \mathrm{d}x = \cdots = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \cdot e^{-\frac{z^2}{2(\sigma_1^2+\sigma_2^2)}}.$$

Therefore, f_{X+Y} is the density corresponding to the distribution $\mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$, which is what we wanted to show. □

17 Moments and moment generating functions

Definition 17.1. Given a random variable X, we define the *moment generating function of* X as the function M_X given by

$$M_X(t) = \mathbb{E}[e^{tX}]$$

for the values of t for which e^{tX} is integrable.

Example 17.1 (Geometric). If $X \sim \text{Geom}(p)$, then

$$M_X(t) = \sum_{n=1}^{\infty} e^{tn} p(1-p)^{n-1} = \begin{cases} \frac{p}{e^{-t}+p-1}, & t < \ln \frac{1}{1-p}, \\ +\infty, & t \geqslant \ln \frac{1}{1-p}. \end{cases}$$

Example 17.2 (Poisson). If $X \sim \text{Poisson}(\lambda)$, then

$$M_X(t) = \sum_{n=0}^{\infty} e^{tn} \frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

Example 17.3 (Normal). Let X be a normal random variable with parameters μ and σ^2, that is,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

The moment-generating function of X can be computed as follows:

$$\begin{aligned} M_X(t) = &= \int_{-\infty}^{\infty} e^{tx} \cdot f_X(x)\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}[(x-\mu)^2 - 2\sigma^2 tx]\right\}\, \mathrm{d}x \end{aligned} \tag{17.1}$$

We now complete the square:

$$\begin{aligned} (x-\mu)^2 - 2\sigma^2 tx &= x^2 - 2(\mu + \sigma^2 t) \cdot x + \mu^2 \\ &= x^2 - 2(\mu + \sigma^2 t) \cdot x + \mu^2 \ \pm (2\mu\sigma^2 t + \sigma^4 t^2) \\ &= (x - \mu - \sigma^2 t)^2 - 2\mu\sigma^2 t - \sigma^4 t^2. \end{aligned}$$

This gives

$$-\frac{1}{2\sigma^2}[(x-\mu)^2 - 2\sigma^2 tx] = -\frac{(x-\mu-\sigma^2 t)^2}{2\sigma^2} + t\mu + \frac{\sigma^2 t^2}{2}.$$

The integral in (17.1) then becomes

$$\exp\left\{t\mu + \frac{\sigma^2 t^2}{2}\right\} \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x-\mu-\sigma^2 t^2)^2}{2\sigma^2}\right\} \, \mathrm{d}x.$$

Now note that the function being integrated is the density function of $\mathcal{N}(\mu + \sigma^2 t,\ \sigma^2)$, so the integral equals 1. In conclusion,

$$M_X(t) = \exp\left\{t\mu + \frac{\sigma^2 t^2}{2}\right\}, \quad t \in \mathbb{R}.$$

Definition 17.2 (Moments)**.** We define the k-th moment of a random variable X as $\mathbb{E}[X^k]$ if X^k is integrable.

The name "moment generating function" comes from the following fact.

Proposition 17.1. *If $M_X(t)$ is defined on $(-a, a)$ for some $a > 0$ then X has all moments and they are given by*

$$\mathbb{E}[X^k] = M_X^{(k)}(0),$$

where $M_X^{(k)}$ denotes the k-th derivative of the function M_X.

We lack the tools to prove this proposition, but if we are willing to be cheeky, we can do:

$$\tfrac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t) = \tfrac{\mathrm{d}^k}{\mathrm{d}t^k} \mathbb{E}[e^{tX}] = \mathbb{E}\left[\tfrac{\mathrm{d}^k}{\mathrm{d}t^k} e^{tX}\right] = \mathbb{E}[X^k e^{tX}]$$

and, evaluating at $t = 0$ gives the proposition.

Example 17.4 (Geometric). If $X \sim \mathrm{Geom}(p)$, then

$$\mathbb{E}[X] = M_X'(0) = \tfrac{1}{p},\ \mathbb{E}[X^2] = M_X''(0) = \tfrac{2}{p^2} - \tfrac{1}{p},\ \mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \tfrac{1-p}{p^2}.$$

Example 17.5 (Poisson). If $X \sim \mathrm{Poisson}(\lambda)$, then

$$\mathbb{E}[X] = M_X'(0) = \lambda,\ \mathbb{E}[X^2] = M_X''(0) = \lambda^2 + \lambda,\ \mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \lambda.$$

Example 17.6 (Normal). Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Let us compute its mean (i.e. its expectation) and variance. In the previous example, we proved that $M_X(t)$ is defined for all $t \in \mathbb{R}$, with

$$M_X(t) = \exp\left\{\frac{t^2\sigma^2}{2} + \mu t\right\}.$$

We then use the above theorem to compute

$$\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\bigg|_{t=0} = \left[(t\sigma^2 + \mu) \cdot \exp\left\{\frac{t^2\sigma^2}{2} + \mu t\right\}\right]_{t=0} = \mu$$

and

$$\mathbb{E}[X^2] = \frac{\mathrm{d}^2}{\mathrm{d}^2 t} M_X(t)\bigg|_{t=0} = \cdots = \sigma^2 + \mu^2.$$

Hence,

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

This is the reason why X is said to be a normal random variable with mean μ and variance σ^2.

Proposition 17.2. *For every $a, b \in \mathbb{R}$,*

$$M_{aX+b}(t) = e^{tb} \cdot M_X(at)$$

for every t such that $M_X(t)$ is defined.

Proof. We compute

$$M_{aX+b}(t) = \mathbb{E}[e^{t(aX+b)}] = e^{tb} \cdot \mathbb{E}[e^{(at)X}] = e^{tb} \cdot M_X(at). \qquad \square$$

Proposition 17.3. *When X and Y are independent,*

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$$

for every t such that the moment generating functions are defined.

Proof. If X and Y are independent, then so are e^{tX} and e^{tY}. Therefore,

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} \cdot e^{tY}] = \mathbb{E}[e^{tX}] \cdot \mathbb{E}[e^{tY}] = M_X(t) \cdot M_Y(t). \quad \square$$

Example 17.7 (Sum of independent Poisson variables). Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent. Then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{\lambda(e^t-1)} e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)} = M_Z(t),$$

where $Z \sim \text{Poisson}(\lambda + \mu)$. Does it imply that $X + Y \sim \text{Poisson}(\lambda + \mu)$?

Th example above makes us wonder whether knowing the moment generating function of a random variable tells us what the distribution of the random variable is. This is indeed the case.

Theorem 17.1 (Moment generating function determines the distribution)**.** *Given two random variables X and Y, if there exists $a > 0$ such that $M_X(t)$ and $M_Y(t)$ are finite and coincide for every $x \in [-a, a]$, then X and Y have the same distribution.*

We also omit the proof of this theorem, and note that it requires tools that are even harder to build up than other proofs omitted in these notes.

Example 17.8 (Sum of independent Poisson variables). If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

Example 17.9 (Sum of independent normal variables). Let X and Y be two independent normal variables, with means μ_X, μ_Y and variances σ_X^2, σ_Y^2, respectively. Also let $a, b \in \mathbb{R}$ with $a \neq 0$. Let us determine the distributions of $aX + b$ and of $X + Y$. In the last lecture we showed that

$$M_X(t) = \exp\left\{\frac{t^2\sigma_X^2}{2} + t\mu_X\right\}, \qquad M_Y(t) = \exp\left\{\frac{t^2\sigma_Y^2}{2} + t\mu_Y\right\}.$$

We have that

$$M_{aX+b}(t) = e^{tb} \cdot M_X(at) = e^{tb} \cdot \exp\left\{a\mu_X t + \frac{a^2\sigma_X^2 t^2}{2}\right\} = \exp\left\{(a\mu_X + b)t + \frac{a^2\sigma_X^2 t^2}{2}\right\}.$$

Hence, $aX + b$ has the same moment-generating function as a $\mathcal{N}(a\mu + b, a^2\sigma^2)$ random variable. Since this moment-generating function is defined in a

neighbourhood of the origin (in fact, in the whole real line), we conclude that $aX + b \sim \mathcal{N}(a\mu_X + b, a^2\sigma_X^2)$.

Next, since X and Y are independent, we have

$$\begin{aligned} M_{X+Y}(t) = M_X(t) \cdot M_Y(t) &= \exp\left\{\mu_X t + \frac{\sigma_X^2 t^2}{2}\right\} \cdot \exp\left\{\mu_Y t + \frac{\sigma_Y^2 t^2}{2}\right\} \\ &= \exp\left\{(\mu_X + \mu_Y)t + \frac{(\sigma_X^2 + \sigma_Y^2)t^2}{2}\right\}. \end{aligned}$$

This shows that $X + Y$ has the same moment-generating function as an $\mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ random variable. Since this moment-generating function is defined in an open interval containing zero, we conclude that $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

A Useful sums

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j} \qquad a, b \in \mathbb{R}, n \in \mathbb{N}_0$$

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \qquad 0 < x < 1$$

$$\sum_{n=0}^{\infty} n x^{n-1} = \frac{1}{(1-x)^2} \qquad 0 < x < 1$$

$$\sum_{n=0}^{\infty} n(n-1) x^{n-2} = \frac{2}{(1-x)^3} \qquad 0 < x < 1$$

$$\sum_{n=0}^{\infty} n(n-1)(n-2) x^{n-3} = \frac{3!}{(1-x)^4} \qquad 0 < x < 1$$

$$\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x \qquad x \in \mathbb{R}$$

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \qquad n \in \mathbb{N}$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \qquad n \in \mathbb{N}$$

The first five are: binomial theorem, geometric series, derivative of geometric series, second derivative of geometric series, third derivative of geometric series. It turns out, it is legitimate to differentiate a series of the form $\sum_n a_n x^n$ term by term, but we are not concerned about the details of why this is true.

The fifth is the so-called "Taylor series" of the exponential function. To check the formula makes sense, observe that both sides give 1 for $x = 0$ and each side is equal to its own derivative. These two facts imply that both sides are equal for every x, but we are not concerned about the details of this either.

The last two formulas, once written down, can be proved by induction (suppose they are correct for a certain n, show that they are correct for $n+1$). If you are curious about how such formulas came about, they can be derived first making the educated guess that they should be given by polynomials one degree higher than the summand, and then using the first two or three terms to write down a system of equations for the coefficients.

B Exponentials beat polynomials

For all $x \geqslant 0$ and $n \in \mathbb{N}$,

$$e^x \geqslant 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$$

Proof. For $n = 0$, we already know that $e^x \geqslant 1$. For $n = 1$,

$$e^x = 1 + \int_0^x e^x \,\mathrm{d}x \geqslant 1 + \int_0^x 1 \,\mathrm{d}x = 1 + x$$

For $n = 2$,

$$e^x = 1 + \int_0^x e^x \,\mathrm{d}x \geqslant 1 + \int_0^x (1 + x)\mathrm{d}x = 1 + x + \tfrac{x^2}{2}$$

For $n = 3$,

$$e^x = 1 + \int_0^x e^x \,\mathrm{d}x \geqslant 1 + \int_0^x (1 + x + \tfrac{x^2}{2})\mathrm{d}x = 1 + x + \tfrac{x^2}{2} + \tfrac{x^3}{3!}.$$

The pattern is clear. □

This implies that

$$\frac{a_0 + a_1 x + \cdots + a_n x^n}{e^{ax}}$$

tends to 0 as $x \to +\infty$, for every $a > 0$.

Proof. Indeed, since $e^{ax} \geqslant \frac{a^{n+1}}{(n+1)!} x^{n+1}$, each term in

$$\frac{a_0}{e^{ax}} + \frac{a_1 x}{e^{ax}} + \cdots + \frac{a_n x^n}{e^{ax}}$$

is approaching zero as x increases. □

This is useful when computing improper integrals that include polynomials and exponential functions.