# Chapter 1

# Basics of Probability

## 1.1 Random variables and their distributions

Probability is, at its core, the mathematical theory of quantitative measure of uncertainty: the age at which an individual dies, the time until an event is reported, the number of events recorded in a year, the amount that must ultimately be paid out. Each of these quantities is unknown in advance but becomes a definite real number once the underlying random experiment (a lifetime, a year of observation, a population) has run its course. The mathematical device that formalises "a numerical quantity depending on chance" is the random variable. Before we can define it, however, we must first fix precisely what we mean by underlying probabilistic model, i.e., a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. When modelling the underlying experiment we usually have a set $\Omega$ of all possible outcomes, a collection $\mathcal{A}$ of all events (an event is a set of outcomes) related to the experiment and a function $\mathbb{P}\colon \mathcal{A} \to [0,1]$ that measures the likelihood or probability of each event taking place. In mathematical terms, we have the following definition.

**Definition 1.1.1.** Let $\Omega$ be any set, $\mathcal{A}$ a collection of subsets of $\Omega$ and $\mathbb{P}\colon \mathcal{A} \to [0,1]$ a function on $\mathcal{A}$. We say that the triple $(\Omega, \mathcal{A}, \mathbb{P})$ is a **probability space** on $\Omega$ if $\mathcal{A}$ is a **$\sigma$-algebra** of $\Omega$ and $\mathbb{P}$ is a **probability measure**. Recall that $\mathcal{A}$ is a $\sigma$-algebra if the following conditions are satisfied:

1. $\varnothing \in \mathcal{A}$,
2. $A \in \mathcal{A}$ iff $A^c \in \mathcal{A}$,
3. $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_n A_n \in \mathcal{A}$.

Also recall that $\mathbb{P}$ is a probability measure if it satisfies the conditions:

1. $0 = \mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$ for any $A \in \mathcal{A}$,
2. if $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint, then $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$.

The 2nd condition is called the **$\sigma$-additivity**.

**Example 1.1.1.** Consider the following example: $\Omega = \{\text{head}, \text{tail}\}$ for the set of outcomes of the random experiment of tossing a coin. The $\sigma$-algebra of this experiment is the collection of all possible events:

$$\mathcal{A} = 2^{\Omega} = \{\varnothing, \{\text{head}\}, \{\text{tail}\}, \{\text{head}, \text{tail}\}\}.$$

The notation $2^{\Omega}$ is used to denote the power set[1] of $\Omega$.

If the coin is perfect, we can define $\mathbb{P}$ on $\mathcal{A}$ as follows:

$$\mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(\{\text{head}\}) = \mathbb{P}(\{\text{tail}\}) = \frac{1}{2}, \quad \mathbb{P}(\Omega) = 1.$$

Of course, other choices for $\mathcal{A}$ and $\mathbb{P}$ would give a different models for the random experiment. For instance, one could define the probability $\mathbb{P}(\{\text{tail}\}) = \frac{2}{3}$ for a more tail bias coin, or simply define $\mathcal{A}$ to be the trivial $\sigma$-algebra $\mathcal{A} = \{\varnothing, \Omega\}$. For this particular case of $\Omega$, it is easy to see that these two $\sigma$-algebras (the trivial one and the power set) are the only possible $\sigma$-algebras. The trivial $\sigma$-algebra is not that interesting since we can only ask two questions: was the outcome a head or tail? Neither of these two?!

**Example 1.1.2.** To give a more complicated example, consider now the experiment of picking a real number at random from the interval $[0, 1]$. Clearly, the set of outcome of this experiment must be $\Omega = [0, 1]$. The collection of events of this experiment is more complicated because, in principle, we can ask any questions we like, for instance: was the number 0.5? was the number rational? irrational? algebraic? the inverse of a Pisot number?! Therefore, we are left to think that the $\sigma$-algebra $\mathcal{A}$ for this experiment must be the power set of $[0, 1]$, i.e., $\mathcal{A} = 2^{[0,1]}$. Now, we wish to define a probability measure for every subset of $\mathbb{R}$, that is, a function $\mathbb{P}\colon \mathcal{A} \to [0, 1]$ that satisfies the axioms 1 and 2 of a probability measure in Definition 1.1.1. Because our experiment is picking a point at random, it is natural to ask that $\mathbb{P}$ further satisfies the following two extra conditions:

1. *Compatible with length*, i.e., the probability that the point belongs to an interval $[a, b]$ is equal to its length: $\mathbb{P}([a, b]) = b - a$ for every $0 \leq a \leq b \leq 1$;

[1] The power set of a set is the collection of all subsets of that set. This collection is also known as the discrete $\sigma$-algebra.

2. *Invariant by translations*, i.e., the probability that a point belongs to a subset $A$ is the same if we translate the set $A$ by some number $b$, $\mathbb{P}(A+b) = \mathbb{P}(A)$ for every $b$ where $A+b = \{a+b : a \in A\}$.

It turns out that there is no such bonafide probability measure defined for every subset $A$ of the interval $[0,1]$. If you are curious about it, look for Vitali set. The problem is that we are too greedy, we are looking to define a probability for every subset of $[0,1]$ and that is simply not possible. In a first course of measure theory, one can show that there is a proper $\sigma$-algebra $\mathcal{M} \subsetneq 2^{[0,1]}$ (removing some pathological subsets of $[0,1]$ such as Vitali's sets) for which a probability measure $m$ on $[0,1]$ can be defined and satisfies the above extra conditions. The $\sigma$-algebra $\mathcal{M}$ is called the **Lebesgue $\sigma$-algebra** and $m$ is called the **Lebesgue measure** on $[0,1]$. The triple $([0,1], \mathcal{M}, m)$ is called the **Lebesgue space**. The $\sigma$-algebra $\mathcal{M}$ is rather big: contains all open subsets of $[0,1]$, all closed subsets, in particular all intervals (open or closed), unions/intersections of these, and so on... Quite often, we prefer to work with a smaller $\sigma$-algebra that already contains all the sets just mentioned, which is called the **Borel $\sigma$-algebra**, denoted by $\mathcal{B}$. To be precise, $\mathcal{B}$ is the smallest $\sigma$-algebra that contains all open subsets of $[0,1]$. An element $B \in \mathcal{B}$ is called a Borel set.

After this diversion, we step back and recall what we wanted to do: define a mathematical model for the experiment of picking a real number at random (with no preference at all) from the interval $[0,1]$. Therefore, we define the set of outcomes to be $\Omega = [0,1]$, the collection of events $\mathcal{A}$ to be the Borel $\sigma$-algebra $\mathcal{B}$ and the probability measure $\mathbb{P}$ to be the Lebesgue probability measure $m$ on $[0,1]$. The triple $([0,1], \mathcal{B}, m)$ is known as the **standard probability space of** $[0,1]$. Now we can ask questions about out random experiment. Pick $x \in [0,1]$ at random. What is the probability that $x \geq 0.5$? Well, we are asking about the probability of the event $[0.5,1]$, that is, what is the Lebesgue measure of $[0.5,1]$. Notice that the event $[0.5,1]$ is a Borel set (an interval). Because Lebesgue measure is compatible with length, $m([0.5,1]) = 1 - 0.5 = 0.5$. Thus, the probability is 0.5, which makes sense. And what is the probability of $x = 0.5$? As before, the event $\{0.5\}$ is a Borel set and $m(\{0.5\}) = m([0.5,0.5]) = 0.5 - 0.5 = 0$. So, the probability is zero. In fact, the probability that $x$ is equal to any given number is zero. We can further ask, what is the probability that $x > 0.5$? The answer is again 0.5, because $[0.5,1] = \{0.5\} \cup ]0.5,1]$, and we have, by $\sigma$-additivity, that $m(]0.5,1]) = m(\{0.5\}) + m(]0.5,1]) = m([0.5,1]) = 0.5$. Now, a more difficult question: what is the probability that $x$ is rational? That is, what is the Lebesgue measure of the set $R$ of rational numbers in $[0,1]$, i.e., $R = \{x \in [0,1] : x \text{ is rational}\}$. The set $R$ is also a Borel

set[2]. We can calculate the measure of $R$ by using the properties that define $m$. Indeed, enumerating $R$ as $a_1, a_2, \ldots$, by the $\sigma$-additivity of $m$, we have $m(R) = m(\bigcup_n \{a_n\}) = \sum_n m(\{a_n\}) = \sum_n 0 = 0$. Thus, the probability that $x$ is rational is zero! If you arrived at this point, then you should known by now that the probability that $x$ is irrational is...

With the underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ now in hand, we can give the central definition of this section.

**Definition 1.1.2** (Random variable)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A **random variable** (r.v.) is a function $X : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : X(\omega) \leqslant r\} \in \mathcal{A}, \qquad \text{for every } r \in \mathbb{R},$$

that is, $X$ is $\mathcal{A}$-measurable.

**Remark 1.1.3.** In the definition of a r.v., the probability $\mathbb{P}$ plays no role, only $(\Omega, \mathcal{A})$ is required. It is usual to say that $X$ is $\mathcal{A}$-measurable.

**Remark 1.1.4.** In order to ease the notation, it is customary in probability to write $\{X \leq r\}$ for the event in Definition 1.1.2. The measurability condition in Definition 1.1.2 is what allows us to speak of the probability of events such as $\{X \leqslant r\}$, $\{a < X \leqslant b\}$, or, more generally, $\{X \in B\}$ for any Borel set[3] $B \subset \mathbb{R}$: these are precisely the sets whose probabilities are guaranteed to be well defined once $\{X \leqslant r\} \in \mathcal{A}$ for all $r$. In fact, asking that $\{X \in B\} \in \mathcal{A}$ for every Borel set $B$ can be seen to be equivalent to the measurability condition in Definition 1.1.2. The name "random variable" is a misnomer, $X$ is neither random (it is a perfectly deterministic function of $\omega$) nor, in the usual sense, a variable. However, the terminology is so entrenched that renaming it would only cause confusion. Indeed, a constant function is a random variable!!

The distribution of $X$ is a probability measure on $\mathbb{R}$ that characterizes the "randomness" of $X$.

**Definition 1.1.3** (Distribution probability measure)**.** Let $X$ be a r.v. in $(\Omega, \mathcal{A}, \mathbb{P})$. The distribution probability measure (or simply distribution) of $X$ is the probability measure $P_X$ on $\mathbb{R}$ defined by the pushforward of $\mathbb{P}$ under $X$, i.e.,

$$P_X(B) = \mathbb{P}(X \in B), \quad B \in \mathcal{B}$$

[2]Every countable set is a Borel set. Indeed, each point set $\{a\}$ is Borel and the countable union of point sets is also Borel.

[3]The $\sigma$-algebra of Borel in $\mathbb{R}$ is the smallest $\sigma$-algebra of $\mathbb{R}$ that contains all open sets of $\mathbb{R}$. As before, we denote this $\sigma$-algebra by $\mathcal{B}$.

**Remark 1.1.5.** The distribution $P_X$ is also called the **law** of $X$.

An event $A \in \mathcal{A}$ is said have **full probability** if $\mathbb{P}(A) = 1$. Typically, a statement in probability can be expressed using an event, e.g. the r.v. $X$ is non-negative, corresponding to the event $\{X \geq 0\} = \{\omega \in \Omega \colon X(\omega) \geq 0\}$. Such a statement is said to hold **almost surely**, commonly abbreviated to a.s., if the corresponding event has full probability.

The **support** of a r.v. $X$, often denoted by $\operatorname{supp} X$, is the set of values it can take, i.e. the smallest closed set $S$ with full probability, i.e., $S \subset \mathbb{R}$ with $\mathbb{P}(X \in S) = 1$. Alternatively, it is equal to the intersection of all closed sets with full probability.

According to how many values they can take, random variables are classified as follows.

**Definition 1.1.4** (Discrete, continuous and mixed random variables)**.** A random variable $X$ is:

(a) **discrete** if its support is a countable set, i.e., finite or countably infinite set;

(b) **continuous** if $\mathbb{P}(X = x) = 0$ for every $x \in \mathbb{R}$;

(c) **mixed** if it is neither discrete nor continuous.

We will make this classification precise below, once we have introduced the distribution function. Typical random variables encountered in applications include: the age at death of a newborn; the time until an individual's status changes (e.g. recovery or attrition); the time to the first event of a given type affecting a randomly chosen subject; the severity (size) of a loss arising from an accident or an occurrence; the number of events recorded in a year for a randomly selected individual; the total amount paid out in a year on a randomly selected unit; and the value, at some future date, of an underlying financial index. All of these examples will recur throughout the chapter.

Everything we will ever want to know about a random variable's probabilistic behaviour — probabilities of intervals, moments, quantiles, tail behaviour — can, in principle, be recovered from a single function.

**Definition 1.1.5** (Distribution function)**.** The **(cumulative) distribution function**, or **cdf**, of a random variable $X$ is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = P_X(]-\infty, x]) = \mathbb{P}(X \leqslant x), \qquad x \in \mathbb{R}.$$

**Proposition 1.1.6** (Properties of a distribution function)**.** *Every distribution function $F_X$ satisfies:*

*(P1)* $0 \leqslant F_X(x) \leqslant 1$ *for all* $x \in \mathbb{R}$*;*

*(P2)* $F_X$ *is nondecreasing:* $a \leqslant b \implies F_X(a) \leqslant F_X(b)$*;*

*(P3)* $F_X$ *is right-continuous:* $\lim_{h \to 0^+} F_X(x+h) = F_X(x)$ *for all* $x$*;*

*(P4)* $\lim\limits_{x \to -\infty} F_X(x) = 0$ *and* $\lim\limits_{x \to +\infty} F_X(x) = 1$.

*Proof.* (P1) is immediate since $F_X(x) = \mathbb{P}(X \leqslant x)$ is a probability. For (P2), if $a \leqslant b$ then $\{X \leqslant a\} \subset \{X \leqslant b\}$, so monotonicity of $\mathbb{P}$ gives $F_X(a) \leqslant F_X(b)$. For (P3), let $h_n \downarrow 0$; then $\{X \leqslant x + h_n\}$ is a decreasing sequence of events with intersection $\{X \leqslant x\}$, so continuity of $\mathbb{P}$ from above[4] gives $F_X(x + h_n) \to F_X(x)$. For (P4), take $x_n \downarrow -\infty$: the events $\{X \leqslant x_n\}$ decrease to $\varnothing$, so $F_X(x_n) \to \mathbb{P}(\varnothing) = 0$; similarly, taking $x_n \uparrow +\infty$, the events $\{X \leqslant x_n\}$ increase to $\Omega$, so by continuity of $\mathbb{P}$ from below[5] we must have $F_X(x_n) \to \mathbb{P}(\Omega) = 1$. □

**Remark 1.1.7.** Conversely, any function $F : \mathbb{R} \to [0,1]$ satisfying (P1)–(P4) is the distribution function of some random variable (e.g. the generalised inverse[6] of $F$ defined on the standard probability space $([0,1], \mathcal{B}, m)$). Thus (P1)–(P4) *characterise* distribution functions completely, and in practice one often specifies a probabilistic model by writing down a function satisfying them directly, without reference to an underlying $\Omega$.

**Remark 1.1.8.** Using continuity of $\mathbb{P}$ from above, we have the following important identity:

$$\mathbb{P}(X = a) = F_X(a) - \lim_{x \to a^-} F_X(x).$$

This means that the probability $\mathbb{P}(X = a)$ is equal to the size of jump of $F_X$ at the point $a$. From this observation, we conclude that $X$ is continuous iff $F_X$ has no jumps, that is, $F_X$ is continuous.

**Example 1.1.9** (Four basic distributional shapes)**.** The following cdf's, all satisfying Proposition 1.1.6, illustrate the range of shapes a distribution function can take.

[4] Continuity of $\mathbb{P}$ from above: given any nested sequence of events $A_1 \supset A_2 \supset \cdots$ we have $\mathbb{P}(\bigcap_n A_n) = \lim_n \mathbb{P}(A_n)$.

[5] Continuity of $\mathbb{P}$ from below: given any nested sequence of events $A_1 \subset A_2 \subset \cdots$ we have $\mathbb{P}(\bigcup_n A_n) = \lim_n \mathbb{P}(A_n)$.

[6] $F^{-1}(p) = \inf\{x \in \mathbb{R} \colon F(x) \geq p\}$.

1. *Age at death.* A Gompertz-–Makeham law of mortality for the age at death $X$ of a newborn, with constants $\alpha > 0$, $\beta > 0$ and $\lambda \geq 0$,

$$F_X(x) = 1 - \exp\left[-\lambda x - \frac{\alpha}{\beta}(e^{\beta x} - 1)\right], \qquad x \geqslant 0,$$

and $F_X(x) = 0$ for $x < 0$: a continuous cdf.

2. *Magnitude of a loss.* $X$ is the size of a loss arising from an accident or occurrence, with

$$F_X(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \qquad x \geqslant 0, \quad \mu \in \mathbb{R}, \quad \sigma > 0,$$

($F_X(x) = 0$ for $x < 0$), where $\Phi$ is the standard normal cdf: again continuous (this is the *lognormal* law, Section 1.7).

3. *Event count.* $X$ is the number of events (e.g. accidents, arrivals, or defects) recorded in a year for a randomly chosen individual, taking values $0, 1, 2, 3, 4, \ldots$ with $F_X$ a step function, e.g.

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 0.818731, & 0 \leqslant x < 1, \\ 0.982477, & 1 \leqslant x < 2, \\ 0.998852, & 2 \leqslant x < 3, \\ 0.999943, & 3 \leqslant x < 4, \\ 1, & x \geqslant 4 : \end{cases}$$

a discrete cdf.

4. *Total loss amount.* $X$ is the total amount paid out in a year for a randomly chosen individual, allowing for the possibility of no events at all,

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - 0.2\, e^{-0.001x}, & x \geqslant 0. \end{cases}$$

Here $F_X(0) = 0.8 > 0$: there is a jump of size 0.8 at the origin (the probability of paying nothing), superimposed on an otherwise continuously increasing curve. We revisit this last example in Section 1.2, once we have the vocabulary to describe it precisely.

## 1.2 Discrete, continuous and mixed random variables

### 1.2.1 Discrete random variables

**Definition 1.2.1** (Probability function)**.** Let $X$ be a discrete random variable with $\operatorname{supp} X = \{x_1, x_2, \ldots\}$. Its **probability function** (or probability mass function, pmf) is

$$f_X(x) = \begin{cases} \mathbb{P}(X = x_j), & x = x_j,\ j = 1, 2, \ldots \\ 0, & \text{otherwise.} \end{cases}$$

The distribution function is recovered from the probability function by summation,

$$F_X(x) = \sum_{y \leqslant x} f_X(y),$$

which is precisely why the cdf of a discrete random variable is a (right-continuous) step function, as in item 3 of Example 1.1.9: it jumps by $f_X(x_j)$ at each mass point $x_j$ and is constant in between. The law of $X$ is the convex combination of Dirac masses:

$$P_X = \sum_{x \in \operatorname{supp} X} f_X(x)\delta_x$$

where $\delta_x$ is the Dirac mass at $x$, i.e., the probability measure

$$\delta_x(B) = \begin{cases} 1, & x \in B \\ 0, & x \notin B \end{cases}.$$

### 1.2.2 Absolutely continuous random variables

**Definition 1.2.2** (Probability density function)**.** A continuous random variable $X$ is **absolutely continuous** if there exists a function $f_X : \mathbb{R} \to [0, \infty)$, called its **probability density function** (pdf), such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du, \qquad x \in \mathbb{R}.$$

Equivalently, $f_X$ is a pdf if and only if

(i) $f_X(x) \geqslant 0$ for all $x$, and

(ii) $\displaystyle\int_{-\infty}^{+\infty} f_X(x)\,dx = 1.$

Whenever $F_X$ is differentiable (which, for the densities considered in this course, is everywhere except possibly a finite set of points) we have $f_X(x) = F_X'(x)$; at any remaining points $f_X$ is left undefined (its value there does not affect any probability computed by integration). In any case, the law of $X$ is

$$p_X([a,b]) = \int_a^b f_X(u)\,du.$$

We are implicitly assuming that $f_X$ is a bounded function and Riemann integrable in every interval. We will extend this later on.

### 1.2.3 Decomposition of a distribution function, and mixed random variables

Not every random variable of practical interest is purely discrete or purely continuous: item 4 of Example 1.1.9 is a case in point, combining a point mass at 0 (no event occurs) with a continuously distributed positive loss amount. The general theory of distribution functions accommodates this through a decomposition theorem, which we state without proof (it is a special case of the Lebesgue decomposition theorem of measure theory).

**Theorem 1.2.1** (Decomposition of a distribution function)**.** *Every distribution function $F_X$ can be written as*

$$F_X(x) = p_1 F^{(d)}(x) + p_2 F^{(ac)}(x) + p_3 F^{(sc)}(x), \qquad x \in \mathbb{R},$$

*where $p_1, p_2, p_3 \geqslant 0$, $p_1+p_2+p_3 = 1$, $F^{(d)}$ is the distribution function of a discrete random variable, $F^{(ac)}$ is the distribution function of an absolutely continuous random variable (i.e. it has a density in the sense of Definition 1.2.2), and $F^{(sc)}$ is **singular continuous**: continuous, but with derivative equal to zero almost everywhere.*

Singular continuous distribution functions are pathological objects (the classical example is built from the Cantor set) that do not arise in the models used in this course, and we exclude them from now on by assuming $p_3 = 0$.

**Definition 1.2.3** (Mixed random variable)**.** A random variable $X$ whose distribution function has the form

$$F_X(x) = p\,F^{(d)}(x) + (1-p)\,F^{(ac)}(x), \qquad 0 < p < 1,$$

with $F^{(d)}$ discrete and $F^{(ac)}$ absolutely continuous, is called **mixed**.

**Example 1.2.2** (The total loss amount is a mixed random variable)**.** Return to item 4 of Example 1.1.9, $F_X(x) = 1 - 0.2e^{-0.001x}$ for $x \geqslant 0$. Writing

$$F_X^{(d)}(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geqslant 0 \end{cases} \qquad \text{(degenerate at 0)},$$

$$F_X^{(ac)}(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-0.001x}, & x \geqslant 0 \end{cases} \qquad \text{(Exponential}(\theta = 1000)\text{)},$$

one checks that $F_X(x) = 0.8\, F_X^{(d)}(x) + 0.2\, F_X^{(ac)}(x)$ for every $x$: with probability $p = 0.8$ no event occurs (mass at 0), and with probability $1 - p = 0.2$ a strictly positive, exponentially distributed amount is paid.

## 1.3 Expected value

We have so far described a random variable $X$ through its distribution function $F_X$, and distinguished the discrete, continuous and mixed cases. We now introduce the single most important summary of $X$: its **expected value**. The classical statistics textbook approach separates the expected value in two definitions — "$\sum_x x f_X(x)$ for discrete $X$, $\int x f_X(x)\, dx$ for continuous $X$" — treating the two cases separately and declaring $E(X)$ to be a sum or an integral, without saying why these are instances of the same underlying construction and without covering the mixed case at all (recall Example 1.2.2). More importantly, this naive approach does not tell us when the resulting expression is even meaningful. To fix both problems we define $E(X)$ properly, as an abstract **Lebesgue integral** of $X$ with respect to the probability measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$. The construction is carried out in stages of increasing generality; a full treatment belongs to a measure theory course, so here we recall the definitions and the statements of the results we need, referring to a standard text (e.g. Billingsley, *Probability and Measure*) for the proofs we omit.

### 1.3.1 Integrating non-negative random variables

**Definition 1.3.1** (Simple random variable)**.** A random variable $X$ is **simple** if it takes only finitely many values $a_1, \ldots, a_n \in \mathbb{R}$. Writing $A_i = \{X = a_i\} \in \mathcal{A}$ (so that $A_1, \ldots, A_n$ partition $\Omega$), $X$ admits the **canonical representation** $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$.

**Definition 1.3.2** (Integral of a nonnegative simple random variable)**.** If $X = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i} \geq 0$ is simple (canonical representation, so $a_i \geq 0$ for every $i$), its expected value is defined as

$$E(X) := \sum_{i=1}^{n} a_i \, \mathbb{P}(A_i) \in [0, \infty).$$

**Definition 1.3.3** (Integral of a nonnegative random variable)**.** For a (not necessarily simple) random variable $X \geq 0$, define

$$E(X) := \sup \big\{ E(Y) : Y \text{ simple}, \ 0 \leq Y \leq X \big\} \in [0, +\infty].$$

**Remark 1.3.1.** This supremum is exactly what is needed to make the definition consistent with the simple case and with the intuitive idea of "area under the graph of $X$".

**Remark 1.3.2.** Every $X \geq 0$ is the pointwise limit of a nondecreasing sequence of nonnegative simple random variables. For instance

$$X_n := \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \, \mathbf{1}_{\{k/2^n \leq X < (k+1)/2^n\}} + n \, \mathbf{1}_{\{X \geq n\}}, \qquad n \in \mathbb{N},$$

satisfies $0 \leq X_1 \leq X_2 \leq \cdots$ and $X_n \uparrow X$ pointwise on $\Omega$ as $n \to \infty$. As we recall next, $E(X) = \lim_n E(X_n)$ for *any* such approximating sequence, so Definition 1.3.3 does not depend on any particular choice of approximation.

### 1.3.2 Signed random variables, integrability, and the definition of $E(X)$

To extend the integral from nonnegative to general (signed) random variables, we split $X$ into its positive and negative parts.

**Definition 1.3.4** (Positive and negative parts)**.** For a random variable $X$, let $X^+ := \max(X, 0) \geq 0$ and $X^- := \max(-X, 0) \geq 0$. Then

$$X = X^+ - X^- \qquad \text{and} \qquad |X| = X^+ + X^-,$$

and $E(X^+)$, $E(X^-)$, $E(|X|)$ are all well defined in $[0, +\infty]$ by Definition 1.3.3.

**Definition 1.3.5** (Integrability and expected value)**.** A random variable $X$ is **integrable** if $E(X^+) < \infty$ and $E(X^-) < \infty$; equivalently (since the integral

of nonnegative random variables is additive, so $E(|X|) = E(X^+) + E(X^-)$), if $E(|X|) < \infty$. In that case, the **expected value**, or **mean**, of $X$ is

$$E(X) := E(X^+) - E(X^-) \ \in \ \mathbb{R}.$$

We also say, interchangeably, that "$X$ has finite mean" or that "$E(X)$ exists and is finite".

**Remark 1.3.3.** If exactly one of $E(X^+), E(X^-)$ is finite, $E(X) := E(X^+) - E(X^-)$ is still well defined as an extended real number in $\{-\infty\} \cup \mathbb{R} \cup \{+\infty\}$, but $X$ is not called integrable; if both are infinite, $E(X)$ is undefined (an "$\infty - \infty$" indeterminacy). This is exactly the "absolute convergence" proviso used informally throughout elementary treatments: a series $\sum_x x f_X(x)$ or integral $\int x f_X(x)\, dx$ that is convergent but not *absolutely* convergent does not define $E(X)$, precisely because it depends on the (irrelevant) order in which the positive and negative contributions are added up.

**Remark 1.3.4.** In a course of measure theory, the expected value as we've just defined is called the **Lebesgue integral** of $X$ over the measure $\mathbb{P}$. The following notation is often used

$$E(X) = \int X \, d\mathbb{P}.$$

**Proposition 1.3.5** (Basic properties of the expectation)**.** *Let $X, Y$ be integrable random variables and $a, b \in \mathbb{R}$. Then:*

*(a)* ***(Linearity)*** *$aX + bY$ is integrable and $E(aX + bY) = aE(X) + bE(Y)$.*

*(b)* ***(Monotonicity)*** *if $X \leq Y$ a.s., then $E(X) \leq E(Y)$; in particular $|E(X)| \leq E(|X|)$.*

*(c) if $\mathbb{P}(X = c) = 1$ for a constant $c \in \mathbb{R}$, then $E(X) = c$.*

These properties are inherited directly from the corresponding (easily checked) properties of the integral of nonnegative random variables in Definitions 1.3.2–1.3.3, via the decomposition $X = X^+ - X^-$; we omit the routine verification.

### 1.3.3 Convergence theorems

The question of when $\lim_n E(X_n) = E(\lim_n X_n)$ holds, that is, when limits and expectations may be interchanged, has no simple answer for the Riemann integral, and is the main technical reason probability theory is built on the

Lebesgue integral instead. We recall the three basic results governing this question; each is progressively used throughout the rest of these notes, usually without further comment.

But first, we need to define precisely the mode of convergence.

**Definition 1.3.6** (Pointwise convergence)**.** We say that a sequence $(X_n)_{n\geq 1}$ of r.v. converges (pointwise) almost surely (a.s.) to a random variable $X$ and write $X_n \to X$ a.s. if

$$\mathbb{P}(\{\omega \in \Omega \colon \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1.$$

**Theorem 1.3.6** (Monotone Convergence Theorem, MCT)**.** *If $0 \leq X_1 \leq X_2 \leq \cdots$ is a nondecreasing sequence of nonnegative random variables and $X_n \to X$ a.s., then*

$$E(X_n) \uparrow E(X) \qquad (\textit{possibly } +\infty).$$

*Proof.* By monotonicity of the integral (Proposition 1.3.5), $0 \leq E(X_1) \leq E(X_2) \leq \cdots$, so $(E(X_n))_{n\geq 1}$ is a monotone sequence and $E(X_n) \uparrow \sup_n E(X_n)$. We just need to show that $E(X) = \sup_n E(X_n)$. Because $X_n \leq X$ for every $n \geq 1$, we must have $E(X_n) \leq E(X)$. Thus, $\sup_n E(X_n) \leq E(X)$. To show the reverse inequality, we proceed as follows. Let $Y \leq X$ be a simple random variable. Given any $\varepsilon \in ]0,1[$, consider the events $E_n = \{X_n \geq \varepsilon Y\}$ for $n \geq 1$. Because $X_1 \leq X_2 \leq \cdots$, we have $E_1 \subset E_2 \subset \cdots$. Moreover, $\mathbb{P}(\bigcup_n E_n) = 1$, because $X_n \to X$. Thus, $E(X_n) \geq E(\mathbf{1}_{E_n} X_n) \geq \varepsilon E(\mathbf{1}_{E_n} Y)$. Taking $n \to \infty$ on both sides of the previous inequality we get $\sup_n E(X_n) \geq \varepsilon E(Y)$. Because $\varepsilon$ is arbitrary close to 1, $\sup_n E(X_n) \geq E(Y)$. Finally, taking the sup over all simple r.v. $Y \leq X$, we obtain $\sup_n E(X_n) \geq E(X)$. □

**Theorem 1.3.7** (Fatou's Lemma)**.** *For any sequence $(X_n)_{n\geq 1}$ of nonnegative random variables,*

$$E\Big(\liminf_{n\to\infty} X_n\Big) \;\leq\; \liminf_{n\to\infty} E(X_n).$$

*Proof.* It is a corollary of MCT. We leave it as an exercise. □

**Theorem 1.3.8** (Dominated Convergence Theorem, DCT)**.** *Let $(X_n)_{n\geq 1}$ be random variables with $X_n \to X$ a.s., and suppose there is an integrable random variable $Y \geq 0$ with $|X_n| \leq Y$ a.s. for every $n$. Then $X$ is integrable and*

$$E(X_n) \to E(X).$$

**Remark 1.3.9.** The MCT is where the real magic happens (it is, essentially, the statement that Definition 1.3.3 is well behaved under limits); Fatou's lemma and the DCT are both short consequences of it. As an illustration, and because the DCT is used repeatedly later in these notes, we derive the DCT from Fatou's lemma.

*Proof.* Since $|X_n| \le Y$ for every $n$ and $X_n \to X$ a.s., also $|X| \le Y$ a.s., so $E(|X|) \le E(Y) < \infty$ and $X$ is integrable. Both $Y + X_n \ge 0$ and $Y - X_n \ge 0$ a.s., so Fatou's lemma applies to each sequence:

$$E(Y)+E(X) = E\big(\liminf_n (Y+X_n)\big) \le \liminf_n E(Y+X_n) = E(Y)+\liminf_n E(X_n),$$

$$E(Y)-E(X) = E\big(\liminf_n (Y-X_n)\big) \le \liminf_n E(Y-X_n) = E(Y)-\limsup_n E(X_n).$$

Since $E(Y) < \infty$, it may be cancelled from both sides of each inequality, giving $E(X) \le \liminf_n E(X_n)$ from the first and $\limsup_n E(X_n) \le E(X)$ from the second. Combining,

$$\limsup_n E(X_n) \;\le\; E(X) \;\le\; \liminf_n E(X_n) \;\le\; \limsup_n E(X_n),$$

forcing equality throughout, so $E(X_n) \to E(X)$. □

### 1.3.4 Expectation via the distribution measure

Recall from Section 1.1 that the **law** (or **distribution**) of $X$ is the probability measure $P_X$ on $(\mathbb{R}, \mathcal{B})$ given by $P_X(B) = \mathbb{P}(X \in B)$, $B \in \mathcal{B}$, and that $P_X$ determines, and is determined by, $F_X$. Being itself a probability measure, $P_X$ has its own Lebesgue integral, built exactly as above, but now on $(\mathbb{R}, \mathcal{B}, P_X)$ rather than $(\Omega, \mathcal{A}, \mathbb{P})$. The next result shows that expectations of (functions of) $X$, computed as integrals over the abstract space $\Omega$, are always equal to concrete integrals over $\mathbb{R}$ against $P_X$; in particular, this finally makes precise, and unifies, the elementary discrete and continuous formulas for $E(X)$ and $E[g(X)]$. In the following, we will consider measurable[7] functions $g : \mathbb{R} \to \mathbb{R}$.

**Theorem 1.3.10** (Change of variables)**.** *Let $X$ be a random variable and $g : \mathbb{R} \to \mathbb{R}$ measurable. Then $g(X)$ is integrable if and only if $\int_{\mathbb{R}} |g(x)|\, dP_X(x) < \infty$, in which case*

$$E[g(X)] \;=\; \int_\Omega g(X)\, d\mathbb{P} \;=\; \int_{\mathbb{R}} g(x)\, dP_X(x). \tag{1.3.1}$$

[7] More precisely, $\mathcal{B}$-measurable. That is, $\{g \in B\} \in \mathcal{B}$ for every $B \in \mathcal{B}$.

*In particular, taking* $g = \mathrm{id}$,

$$E(X) = \int_{\mathbb{R}} x \, dP_X(x).$$

*Proof.* We argue by the standard procedure of measure theory: verify the identity for indicators, extend to simple functions, then to nonnegative functions, then to general integrable functions.

If $g = \mathbf{1}_B$ for a Borel set $B$, then $g(X) = \mathbf{1}_{\{X \in B\}}$ and, directly from the definition of $P_X$,

$$E[g(X)] = \mathbb{P}(X \in B) = P_X(B) = \int_{\mathbb{R}} \mathbf{1}_B \, dP_X,$$

which is (1.3.1) for indicators. By linearity (Proposition 1.3.5(a), applied on each side), (1.3.1) extends to every nonnegative simple $g = \sum_i a_i \mathbf{1}_{B_i}$. For general measurable $g \geq 0$, approximate $g$ pointwise from below by a nondecreasing sequence of nonnegative simple functions $g_n \uparrow g$ (exactly as in the remark after Definition 1.3.3, now on $\mathbb{R}$ instead of $\Omega$); then $g_n(X) \uparrow g(X)$ pointwise on $\Omega$, and the MCT (Theorem 1.3.6), applied once on $\Omega$ to $E[g_n(X)] \to E[g(X)]$ and once on $\mathbb{R}$ to $\int g_n \, dP_X \to \int g \, dP_X$, carries (1.3.1) from the $g_n$ to $g$. Finally, for general measurable $g$, apply the nonnegative case to $g^+$ and $g^-$ separately and subtract, which is valid exactly when $E[g(X)^+] = \int g^+ \, dP_X$ and $E[g(X)^-] = \int g^- \, dP_X$ are not both infinite, i.e. exactly when $\int |g| \, dP_X = E[g(X)^+] + E[g(X)^-] < \infty$, i.e. $g(X)$ is integrable. □

**Corollary 1.3.11** (Recovering the elementary formulas)**.** *Let* $g : \mathbb{R} \to \mathbb{R}$ *be measurable.*

*(a) If* $X$ *is discrete with probability function* $f_X$*, then* $P_X(\{x\}) = f_X(x)$ *for every* $x \in \operatorname{supp} X$*, and Theorem 1.3.10 gives*

$$E[g(X)] = \sum_{x \in \operatorname{supp} X} g(x) \, f_X(x),$$

*provided* $\sum_x |g(x)| f_X(x) < \infty$*.*

*(b) If* $X$ *is absolutely continuous with density* $f_X$*, then* $P_X(B) = \int_B f_X(x) \, dx$ *for every Borel set* $B$*, and Theorem 1.3.10 gives*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \, f_X(x) \, dx,$$

*provided* $\int_{-\infty}^{\infty} |g(x)| f_X(x) \, dx < \infty$*.*

*Proof.* (a) A discrete $P_X$ is precisely a countable sum of point masses, $P_X = \sum_{x \in \operatorname{supp} X} f_X(x)\,\delta_x$, so $\int_{\mathbb{R}} g\,dP_X = \sum_{x \in \operatorname{supp} X} g(x) f_X(x)$ directly from Definitions 1.3.2–1.3.3 applied on $\mathbb{R}$ (approximate the countable sum by finite partial sums, which are simple functions, and pass to the limit via the MCT for the nonnegative parts $g^+, g^-$). (b) That $\int_B f_X\,dx = P_X(B)$ for every Borel $B$ is precisely the statement that $X$ has density $f_X$; the stated formula for $\int g\,dP_X$ is then the standard fact, again proved by the standard procedure applied in the proof of Theorem 1.3.10 (indicators, simple functions, MCT, signed decomposition), that integration against a measure with a density reduces to ordinary (Riemann or Lebesgue) integration against that density. □

**Remark 1.3.12.** Corollary 1.3.11 recovers exactly the formulas used informally in elementary treatments and stated again, for convenience, in Section 1.6 below. Now, as a consequence of a single, general definition of $E(X)$ that also covers the mixed case of Example 1.2.2, writing $F_X = pF_X^{(d)} + (1-p)F_X^{(ac)}$ as in Definition 1.2.3, the law $P_X$ decomposes as $P_X = pP_X^{(d)} + (1-p)P_X^{(ac)}$, and linearity of the integral gives

$$E[g(X)] = p \sum_x g(x) f_X^{(d)}(x) + (1-p) \int_{-\infty}^{\infty} g(x) f_X^{(ac)}(x)\,dx,$$

a weighted combination of the discrete and continuous formulas, each applied to its own piece of the mixture.

## 1.4 Multivariate random variables and independence

Most quantities of interest in applications (e.g. the total loss of a population, the joint lifetimes of a pair of individuals, the correlation between the frequency and severity of events) depend on several random variables observed simultaneously. We now extend the previous definitions to this setting, restricting attention, for notational simplicity, to the bivariate case $(X, Y)$; the extension to $n$ variables is entirely analogous.

**Definition 1.4.1** (Joint distribution)**.** The **joint distribution law** of a bivariate random variable (random vector) $(X, Y)$ is

$$P_{X,Y}(E) = \mathbb{P}(\{(X, Y) \in E\}), \quad \text{Borel set } E \subset \mathbb{R}^2,$$

and the **joint distribution function** is

$$\begin{aligned} F_{X,Y}(x,y) &= P_{X,Y}(]-\infty,x]\times]-\infty,y]) \\ &= \mathbb{P}(X \leqslant x, Y \leqslant y), \qquad (x,y) \in \mathbb{R}^2. \end{aligned}$$

**Remark 1.4.1.** When $E = C \times D$ with $C$ and $D$ being Borel sets of $\mathbb{R}$, then

$$P_{X,Y}(C \times D) = \mathbb{P}(X \in C, Y \in D).$$

**Definition 1.4.2** (Joint probability and density functions)**.** If $(X,Y)$ is discrete, its **joint probability function** is $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$. If $(X,Y)$ is (jointly) absolutely continuous, its **joint density** is a function $f_{X,Y} \geqslant 0$ such that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du, \qquad (x,y) \in \mathbb{R}^2,$$

equivalently $f_{X,Y}(x,y) = \dfrac{\partial^2 F_{X,Y}}{\partial x\,\partial y}(x,y)$ whenever this exists. For Borel sets $C, D \subset \mathbb{R}$,

$$\mathbb{P}(X \in C, Y \in D) = \iint_{\{(x,y):x\in C, y\in D\}} f_{X,Y}(x,y)\,dx\,dy.$$

**Definition 1.4.3** (Marginal distributions)**.** The **marginal** distribution of $X$ (and, symmetrically, of $Y$) is obtained from the joint law by summing out, or integrating out, the other variable:

$$P_X(B) = P_{X,Y}(B \times \mathbb{R}) \quad \text{and} \quad F_X(x) = F_{X,Y}(x, +\infty).$$

In the special cases of discrete and absolutely continuity,

$$\begin{aligned} f_X(x) &= \sum_{\text{all } y_i} f_{X,Y}(x,y_i) \qquad \text{(discrete)}, \\ f_X(x) &= \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy \qquad \text{(abs continuous)}. \end{aligned}$$

**Definition 1.4.4** (Conditional probability and density functions)**.** For $\mathbb{P}(Y = a) > 0$ (discrete case) or $f_Y(a) > 0$ (abs continuous case),

$$\mathbb{P}(X = x \mid Y = a) = \frac{\mathbb{P}(X = x, Y = a)}{\mathbb{P}(Y = a)}, \qquad f_{X|Y=a}(x) = \frac{f_{X,Y}(x,a)}{f_Y(a)}.$$

**Proposition 1.4.2** (Law of total probability)**.**

$$f_X(x) = \sum_{all\ y_i} \mathbb{P}(X = x \mid Y = y_i)\, \mathbb{P}(Y = y_i) \qquad (\textit{discrete}),$$

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

*Proof.* In the discrete case, by Definitions 1.4.3 and 1.4.4,

$$f_X(x) = \sum_{y_i} f_{X,Y}(x, y_i) = \sum_{y_i} \mathbb{P}(X = x \mid Y = y_i)\mathbb{P}(Y = y_i),$$

since $\mathbb{P}(X = x, Y = y_i) = \mathbb{P}(X = x \mid Y = y_i)\mathbb{P}(Y = y_i)$ whenever $\mathbb{P}(Y = y_i) > 0$ (and both sides are 0 otherwise). The continuous case is identical, replacing sums by integrals and using $f_{X,Y}(x, y) = f_{X|Y=y}(x) f_Y(y)$. □

**Definition 1.4.5** (Independence)**.** Random variables $X$ and $Y$ are **independent** if

$$P_{X,Y}(C \times D) = P_X(C) P_Y(D)$$

for all Borel sets $C, D \subset \mathbb{R}$. Equivalently,

$$\mathbb{P}(X \in C, Y \in D) = \mathbb{P}(X \in C)\, \mathbb{P}(Y \in D).$$

The following proposition shows that independence can be expressed in terms of distribution and density functions.

**Proposition 1.4.3.** *If $X, Y$ are independent, then $F_{X,Y}(x, y) = F_X(x) F_Y(y)$ for all $(x, y)$; moreover $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all $(x, y)$ (equality of probability functions in the discrete case, of densities in the continuous case). Conversely, either factorisation implies independence. Furthermore, if $X, Y$ are independent and $G, H : \mathbb{R} \to \mathbb{R}$ are measurable, then $G(X)$ and $H(Y)$ are independent random variables.*

*Proof.* Taking $C = (-\infty, x]$, $D = (-\infty, y]$ in Definition 1.4.5 gives $F_{X,Y}(x, y) = F_X(x) F_Y(y)$ directly. In the discrete case, take $C = \{x\}$, $D = \{y\}$ to get $f_{X,Y}(x, y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = f_X(x) f_Y(y)$. In the continuous case, differentiate $F_{X,Y}(x, y) = F_X(x) F_Y(y)$ with respect to $x$ and $y$ to obtain $f_{X,Y}(x, y) = f_X(x) f_Y(y)$; the converse direction integrates the factorised density back up to $F_{X,Y}(x, y) = F_X(x) F_Y(y)$, from which Definition 1.4.5 for rectangles $C = (-\infty, x]$, $D = (-\infty, y]$ follows, and hence, by the standard extension from generating rectangles to all Borel sets, for general $C, D$. For the closure statement, note that $\{G(X) \in C'\} = \{X \in G^{-1}(C')\}$ and $\{H(Y) \in D'\} = \{Y \in H^{-1}(D')\}$ for Borel $C', D'$; since $G^{-1}(C')$ and $H^{-1}(D')$ are themselves Borel sets, independence of $G(X)$ and $H(Y)$ follows directly from Definition 1.4.5 applied to $C = G^{-1}(C')$, $D = H^{-1}(D')$. □

## 1.5 Conditional Expectation

Expectation, as constructed in Section 1.3, is the best constant approximation of $X$: among constants $c \in \mathbb{R}$, it is the value of $c$ minimising $E\big[(X-c)^2\big]$ (for $X$ with finite variance). **Conditional expectation** refines this question: given that some partial information about the outcome $\omega \in \Omega$ is available, for instance, the value taken by another random variable $Y$, what is the best approximation of $X$ available with that information? The answer, $E(X \mid Y)$, or, more generally, $E(X \mid \mathcal{G})$ for a sub-$\sigma$-algebra $\mathcal{G}$ representing an arbitrary piece of information, is again a random variable, since it depends on what, exactly, has been observed. We will define the conditional expectation in several steps of increasing generality. First, we define $E(X \mid Y)$ directly and elementarily when $Y$ is discrete, then identify the abstract structure this definition exhibits, and finally use that structure, together with the Radon–Nikodym theorem, to define $E(X \mid \mathcal{G})$ in full generality. Conditional expectation, together with families of sub-$\sigma$-algebras called **filtrations**, is one of the central tools of the theory of stochastic processes and Markov chains developed in Chapters 5 and 6.

### 1.5.1 Conditional expectation given a discrete random variable

Let $Y$ be a discrete random variable with support $\{y_1, y_2, \ldots\}$ and $\mathbb{P}(Y = y_n) > 0$ for every $n$, and let $X$ be an integrable random variable (Definition 1.3.5).

**Definition 1.5.1** (Conditional expectation given $Y = y_n$)**.** The **conditional expectation of $X$ given the event $\{Y = y_n\}$** is

$$E(X \mid Y = y_n) := \frac{1}{\mathbb{P}(Y = y_n)} \int_{\{Y = y_n\}} X \, d\mathbb{P}.$$

**Definition 1.5.2** (Conditional expectation given $Y$)**.** The **conditional expectation of $X$ given $Y$** is the random variable $E(X \mid Y) : \Omega \to \mathbb{R}$ defined by

$$E(X \mid Y)(\omega) := E(X \mid Y = y_n) \qquad \text{whenever } \omega \in \{Y = y_n\}.$$

Equivalently, $E(X \mid Y) = g(Y)$ where $g(y_n) := E(X \mid Y = y_n)$; in particular $E(X \mid Y)$ is constant on each event $\{Y = y_n\}$ and takes finitely or countably many values, one for each value of $Y$.

**Remark 1.5.1.** If $X$ is itself discrete, this reduces, via Definition 1.4.4 for the conditional probability function $\mathbb{P}(X = x \mid Y = y_n)$, to

$$E(X \mid Y = y_n) = \sum_{x \in \operatorname{supp} X} x\, \mathbb{P}(X = x \mid Y = y_n).$$

Later on, we will deduce a similar formula for the absolutely continuous case.

**Example 1.5.2.** Two coins, worth 20 and 50 units, are tossed independently and fairly. Let $X$ be the total value, among the two coins, of those that land heads up, and let $Y$ be the value of the first coin (20) if it lands heads, and 0 otherwise. Writing $H/T$ for heads/tails and listing outcomes as (first coin, second coin),

$$\{Y = 20\} = \{HT, HH\}, \qquad \{Y = 0\} = \{TT, TH\},$$

each with probability $\frac{1}{2}$. On $\{Y = 0\}$, $X$ equals 0 (outcome $TT$) or 50 (outcome $TH$), each with conditional probability $\frac{1}{2}$, so $E(X \mid Y = 0) = \frac{1}{2}(0) + \frac{1}{2}(50) = 25$; on $\{Y = 20\}$, $X$ equals 20 (outcome $HT$) or 70 (outcome $HH$), each with conditional probability $\frac{1}{2}$, so $E(X \mid Y = 20) = \frac{1}{2}(20) + \frac{1}{2}(70) = 45$. Hence $E(X \mid Y)$ is the random variable taking the value 25 on $\{Y = 0\}$ and 45 on $\{Y = 20\}$.

**Proposition 1.5.3** (Law of total expectation, discrete case)**.**

$$E(X) = \sum_n E(X \mid Y = y_n)\, \mathbb{P}(Y = y_n) = E\big[E(X \mid Y)\big].$$

*Proof.* By Definition 1.5.1 and $\sigma$-additivity of the integral over the partition $\{\{Y = y_n\}\}_n$ of $\Omega$,

$$\sum_n E(X \mid Y = y_n)\mathbb{P}(Y = y_n) = \sum_n \int_{\{Y = y_n\}} X \, d\mathbb{P} = \int_\Omega X \, d\mathbb{P} = E(X),$$

which is the first equality; the second is the same sum rewritten as $\sum_n g(y_n)\mathbb{P}(Y = y_n) = E[g(Y)] = E[E(X \mid Y)]$, using Definition 1.5.2 and Corollary 1.3.11(a). □

## 1.5.2 The information contained in a random variable

Definition 1.5.2 produces a specific random variable, constant on each of the events $\{Y = y_n\}$. These events, together with their complements, unions, and so on, form exactly the collection of events about which knowing the value of $Y$ lets us decide whether they occurred or not. This collection is a $\sigma$-algebra in its own right, and is what we will mean, precisely, by the *information contained in* $Y$.

**Definition 1.5.3** ($\sigma$-algebra generated by a random variable)**.** The **$\sigma$-algebra generated by a random variable** $Y : \Omega \to \mathbb{R}$ is

$$\sigma(Y) := \big\{Y^{-1}(B) : B \subset \mathbb{R} \text{ Borel}\big\} = \big\{\{Y \in B\} : B \subset \mathbb{R} \text{ Borel}\big\}.$$

**Remark 1.5.4.** One checks directly from Definition 1.5.3 that $\sigma(Y)$ is indeed a $\sigma$-algebra on $\Omega$ (it contains $\varnothing = Y^{-1}(\varnothing)$, is closed under complements since $Y^{-1}(B^c) = (Y^{-1}(B))^c$, and closed under countable unions since $Y^{-1}(\bigcup_n B_n) = \bigcup_n Y^{-1}(B_n)$), and that it is the *smallest* sub-$\sigma$-algebra of $\mathcal{A}$ with respect to which $Y$ is measurable: any $\sigma$-algebra rendering $Y$ measurable must, by definition of measurability, contain every set $Y^{-1}(B)$. Consistently with Definition 1.1.2, $\sigma(Y)$ is generated by the sets $\{Y \leqslant r\}$, $r \in \mathbb{R}$, alone. Informally, $\sigma(Y)$ collects precisely the events that can be decided once the value of $Y(\omega)$ is known, formalizing the notion "the information contained in (or conveyed by) $Y$". When $Y$ is discrete with support $\{y_1, y_2, \ldots\}$, $\sigma(Y)$ is generated by the countable partition $\{\{Y = y_n\}\}_n$ of $\Omega$, and its elements are exactly the unions of blocks of this partition (together with $\varnothing$); this matches the description above of the events decidable from $Y$'s value.

**Remark 1.5.5.** More generally, $\sigma(Y_1, \ldots, Y_n) := \sigma\big(\bigcup_{i=1}^n \sigma(Y_i)\big)$, is the $\sigma$-algebra generated jointly by several random variables, representing the information conveyed by observing all of them.

Two properties of $E(X \mid Y)$, both immediate from Definitions 1.5.1–1.5.2 and 1.5.3, turn out to characterise it completely and will motivate the general definition below:

(i) $E(X \mid Y)$ is $\sigma(Y)$-measurable: it is constant on each block $\{Y = y_n\}$ of the partition generating $\sigma(Y)$, hence measurable with respect to $\sigma(Y)$ (though typically not with respect to any smaller $\sigma$-algebra).

(ii) For every $A \in \sigma(Y)$, $A$ is a (countable) union of blocks $\{Y = y_n\}$, and summing the defining identity of Definition 1.5.1 over those blocks gives

$$\int_A E(X \mid Y)\, d\mathbb{P} = \int_A X\, d\mathbb{P}.$$

That is, $E(X \mid Y)$ is the (essentially unique, as we verify below) $\sigma(Y)$-measurable random variable whose integral over every $\sigma(Y)$-event agrees with that of $X$ itself: the best $\sigma(Y)$-measurable approximation to $X$, in the sense that it reproduces exactly as much of $X$ as the information in $Y$ allows.

### 1.5.3 The general definition, via the Radon–Nikodym theorem

Properties (i)–(ii) above make no reference to $Y$ being discrete, or even to a generating random variable at all, only to a sub-$\sigma$-algebra of $\mathcal{A}$. This suggests the following definition, valid for an arbitrary sub-$\sigma$-algebra and arbitrary (not necessarily discrete) integrable $X$.

**Definition 1.5.4** (Conditional expectation given a $\sigma$-algebra)**.** Let $X$ be an integrable random variable and let $\mathcal{G} \subset \mathcal{A}$ be a sub-$\sigma$-algebra. A random variable $Z$ is called **a conditional expectation of $X$ given $\mathcal{G}$** if

(a) $Z$ is $\mathcal{G}$-measurable, and

(b) $\displaystyle\int_A Z\, d\mathbb{P} = \int_A X\, d\mathbb{P}$ for every $A \in \mathcal{G}$.

As the following theorem shows, there is one and only one conditional expectation given a $\sigma$-algebra $\mathcal{G}$. Thus, we shall denote such uniquely defined random variable $Z$ by $E(X \mid \mathcal{G})$.

**Theorem 1.5.6** (Existence and uniqueness)**.** *For every integrable $X$ and every sub-$\sigma$-algebra $\mathcal{G} \subset \mathcal{A}$, a random variable $Z$ satisfying Definition 1.5.4 exists, and is unique up to almost-sure equality: if $Z, Z'$ both satisfy (a)–(b), then $\mathbb{P}(Z = Z') = 1$. Moreover $Z$ is integrable, with $E(|Z|) \le E(|X|)$.*

**Remark 1.5.7.** We recall, without proof (a full treatment belongs to a measure theory course; see e.g. Billingsley, *Probability and Measure*, Section. 34), the idea behind Theorem 1.5.6. Suppose first $X \ge 0$. The set function $\nu(A) := \int_A X\, d\mathbb{P}$, $A \in \mathcal{G}$, defines a (finite) measure on $(\Omega, \mathcal{G})$ which is *absolutely continuous* with respect to $\mathbb{P}$ restricted to $\mathcal{G}$: $\mathbb{P}(A) = 0 \implies \nu(A) = 0$ for $A \in \mathcal{G}$. The **Radon–Nikodym theorem** then guarantees the existence of a $\mathcal{G}$-measurable density, i.e. a $\mathcal{G}$-measurable random variable $Z \ge 0$ with $\nu(A) = \int_A Z\, d\mathbb{P}$ for every $A \in \mathcal{G}$. Thus, conditions (a)–(b) are satisfied. For general integrable $X$, apply this to $X^+$ and $X^-$ separately and set $Z := Z^+ - Z^-$. Uniqueness up to a.s. equality follows the same test-set argument used repeatedly below (Proposition 1.5.9(b)): if $Z, Z'$ both satisfy (a)–(b), then $A := \{Z > Z'\} \in \mathcal{G}$ satisfies $\int_A Z\, d\mathbb{P} = \int_A X\, d\mathbb{P} = \int_A Z'\, d\mathbb{P}$, which forces $\mathbb{P}(A) = 0$ since $Z > Z'$ on $A$; symmetrically $\mathbb{P}(Z < Z') = 0$.

With existence and uniqueness (up to a.s. equality, which we no longer mention explicitly) secured, we may finally define $E(X \mid Y)$ for an arbitrary, not necessarily discrete, random variable $Y$, consistently with Definition 1.5.2.

**Definition 1.5.5** (Conditional expectation given a random variable, general case)**.** For any random variable $Y$, $E(X \mid Y) := E\big(X \mid \sigma(Y)\big)$, where $\sigma(Y)$ is as in Definition 1.5.3. As before, $E(X \mid Y)$ is $\sigma(Y)$-measurable, hence (Remark 1.5.4 and the fact, standard in measure theory, that every $\sigma(Y)$-measurable random variable is of the form $g(Y)$ for some measurable $g : \mathbb{R} \to \mathbb{R}$) there is a function $y \mapsto E(X \mid Y = y)$ with $E(X \mid Y)(\omega) = E(X \mid Y = y)$ whenever $Y(\omega) = y$.

**Proposition 1.5.8** (Consistency with the elementary formulas)**.** *Definition 1.5.5 recovers the familiar formulas:*

*(a) If $Y$ is discrete, $E(X \mid Y = y_n)$ as given by Definition 1.5.5 coincides with Definition 1.5.1.*

*(b) If $(X, Y)$ are jointly continuous with conditional density $f_{X|Y=y}$ as in Definition 1.4.4, then*

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

*Proof.*

(a) When $Y$ is discrete, $\sigma(Y)$ is generated by the countable partition $\{\{Y = y_n\}\}_n$ (Remark 1.5.4), and any $\sigma(Y)$-measurable random variable is constant on each block $\{Y = y_n\}$; conditions (a)–(b) of Definition 1.5.4, restricted to the generating blocks $A = \{Y = y_n\} \in \sigma(Y)$, read exactly $\mathbb{P}(Y = y_n)\, Z|_{\{Y=y_n\}} = \int_{\{Y=y_n\}} X\, d\mathbb{P}$, i.e. Definition 1.5.1; this determines $Z$ on every block and hence, since the blocks partition $\Omega$, on all of $\Omega$, so the two definitions agree.

(b) This is the continuous analogue of (a), with sums replaced by integrals: one verifies directly, using $f_{X,Y}(x, y) = f_{X|Y=y}(x) f_Y(y)$ (Definition 1.4.4) and Fubini's theorem, that $Z(\omega) := \int x\, f_{X|Y=Y(\omega)}(x)\, dx$ is $\sigma(Y)$-measurable and satisfies, for every Borel $B \subset \mathbb{R}$ and $A = \{Y \in B\} \in \sigma(Y)$,

$$\begin{aligned}\int_A Z\, d\mathbb{P} &= \int_B \Big( \int_{-\infty}^{+\infty} x f_{X|Y=y}(x)\, dx \Big) f_Y(y)\, dy \\ &= \int_B \int_{-\infty}^{+\infty} x\, f_{X,Y}(x, y)\, dx\, dy \\ &= \int_A X\, d\mathbb{P},\end{aligned}$$

which is Definition 1.5.4(b); Definition 1.5.4(a) holds by construction, so $Z = E(X \mid Y)$.

□

### 1.5.4 Basic properties

Throughout, $\mathcal{G}, \mathcal{H} \subset \mathcal{A}$ denote sub-$\sigma$-algebras and all random variables are assumed integrable (so that every conditional expectation below is well defined by Theorem 1.5.6); all identities involving conditional expectations hold almost surely.

**Proposition 1.5.9** (Basic properties of conditional expectation)**.** *Let $X, X_1, X_2$ be integrable random variables, $\alpha_1, \alpha_2 \in \mathbb{R}$, and $\mathcal{H} \subset \mathcal{G} \subset \mathcal{A}$ sub-$\sigma$-algebras. Then:*

*(a)* ***(Linearity)*** $E(\alpha_1 X_1 + \alpha_2 X_2 \mid \mathcal{G}) = \alpha_1 E(X_1 \mid \mathcal{G}) + \alpha_2 E(X_2 \mid \mathcal{G})$.

*(b)* ***(Positivity/monotonicity)*** *if $X \geq 0$ a.s. then $E(X \mid \mathcal{G}) \geq 0$ a.s.; more generally, if $X_1 \leq X_2$ a.s. then $E(X_1 \mid \mathcal{G}) \leq E(X_2 \mid \mathcal{G})$ a.s.*

*(c)* ***(Law of total expectation)*** $E\big[E(X \mid \mathcal{G})\big] = E(X)$ *(taking $\mathcal{G} = \{\varnothing, \Omega\}$ recovers $E(X \mid \{\varnothing, \Omega\}) = E(X)$ itself).*

*(d)* ***(Taking out what is known)*** *if $W$ is $\mathcal{G}$-measurable and $WX$ is integrable, then $E(WX \mid \mathcal{G}) = W\,E(X \mid \mathcal{G})$; in particular, for $Y$ a random variable and $g : \mathbb{R} \to \mathbb{R}$ measurable with $g(Y)X$ integrable, $E\big(g(Y)X \mid Y = y\big) = g(y)E(X \mid Y = y)$.*

*(e)* ***(Independence)*** *if $X$ is independent of $\mathcal{G}$ (i.e. $\mathbb{P}(\{X \in B\} \cap A) = \mathbb{P}(X \in B)\mathbb{P}(A)$ for all Borel $B$ and $A \in \mathcal{G}$), then $E(X \mid \mathcal{G}) = E(X)$; in particular, if $X$ and $Y$ are independent, $E(X \mid Y) = E(X)$, i.e. $E(X \mid Y = y) = E(X)$ for every $y$.*

*(f)* ***(Tower property)*** $E\big[E(X \mid \mathcal{G}) \mid \mathcal{H}\big] = E(X \mid \mathcal{H})$.

*Proof.* (a) Both sides are $\mathcal{G}$-measurable (a linear combination of $\mathcal{G}$-measurable random variables is $\mathcal{G}$-measurable), and for $A \in \mathcal{G}$, by linearity of the integral (Proposition 1.3.5(a)) and Definition 1.5.4(b) applied to $X_1$ and $X_2$ separately,

$$\int_A \big(\alpha_1 E(X_1 \mid \mathcal{G}) + \alpha_2 E(X_2 \mid \mathcal{G})\big)\, d\mathbb{P} = \alpha_1 \int_A X_1\, d\mathbb{P} + \alpha_2 \int_A X_2\, d\mathbb{P} = \int_A (\alpha_1 X_1 + \alpha_2 X_2)\, d\mathbb{P};$$

by uniqueness (Theorem 1.5.6), the left-hand side equals $E(\alpha_1 X_1 + \alpha_2 X_2 \mid \mathcal{G})$.

(b) Suppose $X_1 \le X_2$ a.s. and let $A := \{E(X_1 \mid \mathcal{G}) > E(X_2 \mid \mathcal{G})\}$. Since $E(X_1 \mid \mathcal{G})$ and $E(X_2 \mid \mathcal{G})$ are both $\mathcal{G}$-measurable, so is $A$, hence $A \in \mathcal{G}$ and Definition 1.5.4(b) applies to each:

$$\int_A E(X_1 \mid \mathcal{G})\, d\mathbb{P} = \int_A X_1\, d\mathbb{P} \;\le\; \int_A X_2\, d\mathbb{P} = \int_A E(X_2 \mid \mathcal{G})\, d\mathbb{P},$$

using monotonicity of the integral (Proposition 1.3.5(b)) in the middle step. But on $A$ itself, $E(X_1 \mid \mathcal{G}) > E(X_2 \mid \mathcal{G})$ by definition of $A$; this is compatible with the displayed inequality only if $\mathbb{P}(A) = 0$. Hence $E(X_1 \mid \mathcal{G}) \le E(X_2 \mid \mathcal{G})$ a.s. Taking $X_1 = 0, X_2 = X$ gives positivity.

(c) Take $A = \Omega \in \mathcal{G}$ in Definition 1.5.4(b): $E\big[E(X \mid \mathcal{G})\big] = \int_\Omega E(X \mid \mathcal{G})\, d\mathbb{P} = \int_\Omega X\, d\mathbb{P} = E(X)$. When $\mathcal{G} = \{\varnothing, \Omega\}$, the only $\mathcal{G}$-measurable random variables are the (a.s.) constants (Definition 1.5.4(a)), and the constant $c = E(X \mid \mathcal{G})$ must then equal $E(X)$ by this same identity, applied with $\mathcal{G} = \{\varnothing, \Omega\}$.

(d) We use the standard argument, as in the proof of Theorem 1.3.10. Suppose first $W = \mathbf{1}_B$ for some $B \in \mathcal{G}$. Then $W\,E(X \mid \mathcal{G})$ is $\mathcal{G}$-measurable, and for $A \in \mathcal{G}$, using $A \cap B \in \mathcal{G}$ and Definition 1.5.4(b),

$$\int_A W\,E(X \mid \mathcal{G})\, d\mathbb{P} = \int_{A\cap B} E(X \mid \mathcal{G})\, d\mathbb{P} = \int_{A\cap B} X\, d\mathbb{P} = \int_A WX\, d\mathbb{P},$$

so $W\,E(X \mid \mathcal{G})$ satisfies Definition 1.5.4 for $WX$, hence equals $E(WX \mid \mathcal{G})$ by uniqueness. By linearity (part (a)), the identity extends to $\mathcal{G}$-measurable simple $W$; for general $\mathcal{G}$-measurable $W$, approximate $W^+, W^-$ by nondecreasing sequences of nonnegative $\mathcal{G}$-measurable simple random variables and pass to the limit via the Monotone/Dominated Convergence Theorems, exactly as in the proof of Theorem 1.3.10. The particular case follows by taking $\mathcal{G} = \sigma(Y)$ and $W = g(Y)$, which is $\sigma(Y)$-measurable.

(e) Let $Z := E(X)$, a constant r.v., hence trivially $\mathcal{G}$-measurable. For $A \in \mathcal{G}$, independence of $X$ and $\mathcal{G}$ gives, i.e. applying the definition of independence to $\mathbf{1}_A$ (which is $\sigma(A) \subset \mathcal{G}$-measurable) and $X$,

$$\int_A X\, d\mathbb{P} = E(X\mathbf{1}_A) = E(X)\mathbb{P}(A) = \int_A Z\, d\mathbb{P},$$

using independence in the second equality. Thus $Z$ satisfies Definition 1.5.4(a)–(b) for $\mathcal{G}$, so $Z = E(X \mid \mathcal{G})$ by uniqueness. Taking $\mathcal{G} = \sigma(Y)$ gives the case of a random variable $Y$: $X$ independent of $Y$ means, by definition (Definition 1.4.5), $X$ independent of every event $\{Y \in B\} \in \sigma(Y)$, hence independent of $\sigma(Y)$ in the sense required.

(f) Both $E\big[E(X \mid \mathcal{G}) \mid \mathcal{H}\big]$ and $E(X \mid \mathcal{H})$ are $\mathcal{H}$-measurable by construction. For $A \in \mathcal{H}$, since $\mathcal{H} \subset \mathcal{G}$ we also have $A \in \mathcal{G}$, so Definition 1.5.4(b) applies twice, once to $E(X \mid \mathcal{G})$ against $\mathcal{H}$ and once to $X$ itself against $\mathcal{G}$:

$$\int_A E\big[E(X \mid \mathcal{G}) \mid \mathcal{H}\big]\, d\mathbb{P} = \int_A E(X \mid \mathcal{G})\, d\mathbb{P} = \int_A X\, d\mathbb{P},$$

which is exactly Definition 1.5.4(b) for $E(X \mid \mathcal{H})$. By uniqueness, $E\big[E(X \mid \mathcal{G}) \mid \mathcal{H}\big] = E(X \mid \mathcal{H})$. □

**Corollary 1.5.10** (Law of total expectation, general case)**.** *For any discrete random variable $Y$,*

$$E(X) = \sum_n E(X \mid Y = y_n)\mathbb{P}(Y = y_n),$$

*and for absolutely continuous $Y$ with density $f_Y$,*

$$E(X) = \int_{-\infty}^{+\infty} E(X \mid Y = y) f_Y(y)\, dy.$$

*Proof.* Rewrite $E[E(X \mid Y)] = E[g(Y)]$, with $g(y) = E(X \mid Y = y)$, via Corollary 1.3.11. □

### 1.5.5 Conditional probability

Let $\mathcal{G} \subset \mathcal{A}$ denote a sub-$\sigma$-algebra.

**Definition 1.5.6.** The conditional probability of $X$ given $\mathcal{G}$ is

$$\mathbb{P}(X \in B \mid \mathcal{G}) := E(\mathbf{1}_{\{X \in B\}} \mid \mathcal{G}).$$

This definition generalizes the basic definition of condition probability $\mathbb{P}(C \mid D)$ for two events $C$ and $D$. By setting $\mathcal{G} = \sigma(Y)$, the information of $Y$, we denote the conditional probability of $X$ on $Y$ by $\mathbb{P}(X \in B \mid Y)$. Using the law of total expectation (Proposition 1.5.9(c)), we have the following well-known formula:

$$\mathbb{P}(X \in B) = E(\mathbf{1}_{\{X \in B\}}) = E(E(\mathbf{1}_{\{X \in B\}} \mid Y)) = \int \mathbb{P}(X \in B \mid Y = y) f_Y(y)\, dy.$$

A similar formula exists for discrete $Y$.

**Remark 1.5.11** (Looking ahead: filtrations and Markov chains)**.** The tower property, Proposition 1.5.9(f), is the algebraic heart of the Markov property. When we study stochastic processes $(X_t)_{t\in T}$ from Chapter 5 onward, the role played here by a single sub-$\sigma$-algebra $\mathcal{G}$ will be played by an entire increasing family $(\mathcal{F}_t)_{t\in T}$ of sub-$\sigma$-algebras: a **filtration** with $\mathcal{F}_t$ representing all the information available to an observer of the process up to time $t$ (formally, $\mathcal{F}_t = \sigma(X_s : s \in T, s \le t)$, the natural filtration of the process). The Markov property of Chapter 6 is, at heart, the statement that conditioning on the whole past $\mathcal{F}_t$ is, for a Markov process, the same as conditioning on the present state $X_t$ alone; and the general (measure-theoretic) formulation of the Markov property given in Chapter 5 is written exactly as $\mathbb{P}(X_{t+h} \in B \mid \mathcal{F}_t) = \mathbb{P}(X_{t+h} \in B \mid X_t)$, an equality of conditional expectations (of $\mathbf{1}_{\{X_{t+h}\in B\}}$) in the sense of Definition 1.5.4. The tower property is precisely the tool used there to reconcile this general formulation with the more elementary, discrete-state definition of a Markov chain.

## 1.6 Moments, quantiles, hazard rate and mode

### 1.6.1 Mean and moments of a single random variable

**Remark 1.6.1** (Mean, recalled)**.** The **mean**, or **expected value**, $\mu_X = E(X)$ of $X$ was defined rigorously, as a Lebesgue integral, in Section 1.3 (Definition 1.3.5), where $X$ is called integrable, or is said to have finite mean, exactly when $E(|X|) < \infty$. By Corollary 1.3.11, this reduces to the familiar formulas

$$\mu_X = \sum_x x\, f_X(x) \quad \text{(discrete)},$$
$$\mu_X = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx \quad \text{(abs continuous)},$$

valid precisely when the sum or integral is absolutely convergent (i.e. $\sum |x| f_X(x) < \infty$, resp. $\int |x| f_X(x)\, dx < \infty$). These are the two special cases of (1.3.1) with $g = \mathrm{id}$.

**Definition 1.6.1** (Raw and central moments)**.** For $k \in \mathbb{N}$, the **$k$th raw moment** of $X$ is $\mu'_k = E(X^k)$, and the **$k$th central moment** is $\mu_k = E\big[(X - \mu_X)^k\big]$, whenever these exist; by Corollary 1.3.11 (with $g(x) = x^k$,

resp. $g(x) = (x - \mu_X)^k$), both are computed, as with the mean, via

$$E[g(X)] = \sum_x g(x) f_X(x) \qquad \text{(discrete)},$$
$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx \qquad \text{(abs continuous)}.$$

**Definition 1.6.2** (Variance and related quantities)**.** The **variance** of $X$ is its second central moment,

$$\mathrm{Var}(X) = \sigma_X^2 = \mu_2 = E\big[(X - \mu_X)^2\big] = E(X^2) - \mu_X^2,$$

and its **standard deviation** is $\sigma_X = \sqrt{\sigma_X^2}$. Further standard summary quantities are the **coefficient of variation** $CV_X = \sigma_X/\mu_X$, the **skewness coefficient** $\gamma_X = \mu_3/\sigma_X^3$, and the **coefficient of kurtosis** $\mu_4/\sigma_X^4$.

### 1.6.2 Bivariate moments

**Definition 1.6.3** (Expectation of a function of two random variables)**.**

$$E[g(X,Y)] = \sum_x \sum_y g(x,y) f_{X,Y}(x,y) \qquad \text{(discrete)},$$
$$E[g(X,Y)] = \iint g(x,y) f_{X,Y}(x,y)\, dx\, dy \qquad \text{(abs continuous)}.$$

**Definition 1.6.4** (Covariance and correlation)**.** The **covariance** of $X$ and $Y$ is

$$\mathrm{Cov}(X,Y) = E\big[(X - \mu_X)(Y - \mu_Y)\big] = E(XY) - E(X)E(Y),$$

and their **correlation coefficient** is $\rho(X,Y) = \mathrm{Cov}(X,Y)/(\sigma_X \sigma_Y)$, defined whenever $\sigma_X, \sigma_Y \in (0, \infty)$.

**Proposition 1.6.2** (Variance of a sum)**.** *For any $X, Y$ with finite variance,*

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

*Proof.* Let $\mu_X = E(X)$, $\mu_Y = E(Y)$. Then

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

□

**Corollary 1.6.3.** *If $X, Y$ are independent, then* $\mathrm{Cov}(X, Y) = 0$ *and* $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

*Proof.* By properties (d) and (e) in Proposition 1.5.9,

$$E(XY \mid Y) = YE(X \mid Y) = YE(X).$$

Taking the expect value in both sides and applying property (c) of Proposition 1.5.9 to the right-hand-side, $E(XY) = E(YE(X)) = E(X)E(Y)$. So $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0$; the variance identity is then immediate from Proposition 1.6.2. □

**Remark 1.6.4.** The converse of Corollary 1.6.3 is *false*: $\mathrm{Cov}(X, Y) = 0$ does not imply independence. For instance, let $X$ take the values $-1, 0, 1$ each with probability $1/3$, and let $Y = X^2$. Then $X$ and $Y$ are clearly dependent (knowing $Y = 0$ tells us $X = 0$ with certainty), yet

$$\mathrm{Cov}(X, Y) = E(X^3) - E(X)E(X^2) = 0 - 0 \cdot E(X^2) = 0,$$

since $E(X) = E(X^3) = 0$ by symmetry.

### 1.6.3 Quantiles

Quantiles generalise the everyday notion of a "percentile" and are the basic building block of risk measures used throughout risk management and statistics (most notably Value-at-Risk, which is nothing but a quantile of a loss distribution).

**Definition 1.6.5** (Quantile)**.** Let $p \in (0, 1)$. A $p$**th quantile** of $X$ is any value $\pi_p \in \mathbb{R}$ satisfying

$$F_X(\pi_p^-) \leqslant p \leqslant F_X(\pi_p),$$

where $F_X(\pi_p^-) = \lim_{x \uparrow \pi_p} F_X(x)$ denotes the left limit of $F_X$ at $\pi_p$. The 0.5th quantile $\pi_{0.5}$ is called the **median** of $X$.

**Remark 1.6.5.** The interval $[F_X(\pi_p^-), F_X(\pi_p)]$ collapses to the single value $F_X(\pi_p)$ wherever $F_X$ is continuous at $\pi_p$, but the definition must allow for a nondegenerate interval because $F_X$ may jump over the level $p$ (if $X$ is discrete or mixed) or may be flat at height $p$ over an interval (if $X$ has a density vanishing on that interval), and in either case there is genuine non-uniqueness or ambiguity to record. If $F_X$ is continuous and strictly increasing, the definition collapses to the familiar $\pi_p = F_X^{-1}(p)$, the unique solution of $F_X(\pi_p) = p$.

### 1.6.4 Hazard Rate and Mode

In survival analysis and in reliability engineering one is often less interested in the density itself than in the *instantaneous risk* of the event of interest (death, failure, default) occurring at time $x$, given that it has not yet occurred. This is the content of the hazard rate.

**Definition 1.6.6** (Survival function)**.** The **survival function** of a random variable $X$ is

$$S_X(x) := \overline{F}_X(x) = \mathbb{P}(X > x) = 1 - F_X(x), \qquad x \in \mathbb{R}.$$

In the case of a random vector $(X, Y)$, the **joint survival function** is

$$S_{X,Y}(x, y) = \mathbb{P}(X > x, Y > y), \qquad (x, y) \in \mathbb{R}^2.$$

**Remark 1.6.6.** Unlike the univariate case, $S_{X,Y}(x, y) \neq 1 - F_{X,Y}(x, y)$ in general: by inclusion–exclusion,

$$1 - F_{X,Y}(x, y) = \mathbb{P}(X > x) + \mathbb{P}(Y > y) - \mathbb{P}(X > x, Y > y) \neq S_{X,Y}(x, y)$$

except in degenerate cases. This is a common and consequential source of error, worth remembering.

**Theorem 1.6.7** (Mean via the survival function)**.** *For any abs continuous random variable $X$ with finite mean,*

$$E(X) = -\int_{-\infty}^{0} F_X(x)\,dx + \int_{0}^{+\infty} S_X(x)\,dx.$$

*In particular, if $X \geqslant 0$,*

$$E(X) = \int_{0}^{+\infty} S_X(x)\,dx. \tag{1.6.2}$$

*Proof.* We prove (1.6.2) for a nonnegative continuous $X$; the general (signed) statement follow by the same argument applied separately to the positive and negative parts of $X$, using Fubini's theorem to justify the interchange of the order of integration/summation. Write, for $x \geqslant 0$, $S_X(x) = \int_x^\infty f_X(t)\,dt$. Then

$$\int_0^\infty S_X(x)\,dx = \int_0^\infty \int_x^\infty f_X(t)\,dt\,dx = \int_0^\infty \int_0^t f_X(t)\,dx\,dt = \int_0^\infty t\,f_X(t)\,dt = E(X),$$

where the interchange of the order of integration (Fubini/Tonelli, valid since the integrand is nonnegative) replaces the region $\{0 \leq x \leq t < \infty\}$ described first by "$x$ outer, $t \geq x$" and then by "$t$ outer, $0 \leq x \leq t$". □

**Remark 1.6.8.** A similar result can be obtained for discrete random variables. For instance, suppose that $\operatorname{supp} X = \{0, 1, 2, 3, \ldots\}$. Then

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

**Definition 1.6.7** (Hazard rate)**.** Let $X$ be a nonnegative absolutely continuous random variable with density $f_X$. The **hazard rate** of $X$ (also called the **intensity of mortality**, in survival analysis, or the **failure rate**, in reliability engineering) is

$$h_X(x) = \frac{f_X(x)}{S_X(x)}, \qquad x \text{ such that } S_X(x) > 0.$$

**Proposition 1.6.9.** *For $X$ as in Definition 1.6.7,*

$$h_X(x) = -\frac{S_X'(x)}{S_X(x)} = -\frac{d}{dx} \ln S_X(x).$$

*Proof.* Since $S_X = 1 - F_X$, we have $S_X'(x) = -F_X'(x) = -f_X(x)$, so $-S_X'(x)/S_X(x) = f_X(x)/S_X(x) = h_X(x)$. The equality with $-\frac{d}{dx} \ln S_X(x)$ is the chain rule, $\frac{d}{dx} \ln S_X(x) = S_X'(x)/S_X(x)$. □

Heuristically, $h_X(x)\,dx \approx \mathbb{P}(x < X \leqslant x + dx \mid X > x)$: the hazard rate is the conditional density of $X$ at $x$, given survival up to $x$. Proposition 1.6.9 shows that $h_X$ determines $\ln S_X$ up to a constant, and since $S_X(0) = 1$ for a nonnegative random variable, it in fact determines $S_X$ completely.

**Theorem 1.6.10** (Recovery of the survival function from the hazard rate)**.** *Let $X$ be a nonnegative continuous random variable. Then*

$$S_X(x) = \exp\left(-\int_0^x h_X(t)\,dt\right), \qquad x \geqslant 0.$$

*Proof.* By Proposition 1.6.9, $\frac{d}{dt} \ln S_X(t) = -h_X(t)$. Integrating from 0 to $x$,

$$\ln S_X(x) - \ln S_X(0) = -\int_0^x h_X(t)\,dt.$$

Since $X \geqslant 0$, $S_X(0) = 1 - \mathbb{P}(X = 0)$ and, assuming (as is standard) $\mathbb{P}(X = 0) = 0$ so that $S_X(0) = 1$, we get $\ln S_X(0) = 0$ and the result follows by exponentiating. □

**Remark 1.6.11.** Theorem 1.6.10 is only valid, as stated, for nonnegative *continuous* random variables. For a mixed random variable the hazard rate is well defined only on the continuous part of the support (away from any mass points), and the identity of Theorem 1.6.10 must be corrected to account for the jumps of $F_X$.

**Example 1.6.12.** If $X \sim \mathrm{Exp}(\theta)$, then $S_X(x) = e^{-x/\theta}$ and $f_X(x) = \theta^{-1}e^{-x/\theta}$, so

$$h_X(x) = \frac{\theta^{-1}e^{-x/\theta}}{e^{-x/\theta}} = \frac{1}{\theta}, \qquad x \geqslant 0 :$$

the hazard rate of the exponential distribution is *constant*. This is the analytic expression of the memoryless property, to which we return in Example 1.9.2.

**Definition 1.6.8** (Mode)**.** A **mode** of $X$ is a value at which the probability (or density) function $f_X$ attains a (local) maximum. A distribution may have more than one mode, in which case it is called **multimodal**.

## 1.7 A catalogue of standard distributions

We record here, for later reference, the distributions used repeatedly throughout the course. Parametrisations are chosen to match the standard references on loss distributions (in particular Klugman–Panjer–Willmot, *Loss Models*) and are not always the "textbook default" ones (e.g. the Gamma and Beta distributions below are parametrised through an explicit scale parameter $\theta$, and the Pareto distribution is given in its shifted, Lomax-type form with support $(0, \infty)$); the reader is warned to check parametrisations carefully whenever consulting another source. Means and variances are standard computations, most of which are left as an exercise; two are worked out explicitly in Section 1.10 because of their relevance to tail behaviour.

### 1.7.1 Discrete distributions

**Bernoulli.** $X \sim B(1, q)$, $0 < q < 1$: $\mathbb{P}(X = 1) = q$, $\mathbb{P}(X = 0) = 1 - q$. $E(X) = q$, $\mathrm{Var}(X) = q(1 - q)$.

**Binomial.** $X \sim B(m, q)$, $m \in \mathbb{N}$, $0 < q < 1$:

$$f_X(k) = \binom{m}{k} q^k (1 - q)^{m-k}, \qquad k = 0, 1, \ldots, m.$$

$E(X) = mq$, $\mathrm{Var}(X) = mq(1 - q)$.

**Poisson.** $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$:

$$f_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

$E(X) = \text{Var}(X) = \lambda$.

**Negative Binomial.** $X \sim NB(\beta, r)$, $\beta, r > 0$:

$$f_X(k) = \binom{r+k-1}{k} \left(\frac{\beta}{1+\beta}\right)^k \left(\frac{1}{1+\beta}\right)^r, \qquad k = 0, 1, 2, \ldots,$$

where the generalised binomial coefficient is

$$\binom{r+k-1}{k} = \frac{r(r+1)\cdots(r+k-1)}{k!} = \frac{\Gamma(r+k)}{\Gamma(r)\,k!}, \qquad \Gamma(r) = \int_0^\infty t^{r-1}e^{-t}\,dt.$$

$E(X) = r\beta$, $\text{Var}(X) = r\beta(1+\beta)$.

**Geometric.** $X \sim NB(\beta, 1)$, $\beta > 0$:

$$f_X(k) = \left(\frac{\beta}{1+\beta}\right)^k \frac{1}{1+\beta}, \qquad k = 0, 1, 2, \ldots$$

$E(X) = \beta$, $\text{Var}(X) = \beta(1+\beta)$.

### 1.7.2 Continuous distributions

**Normal.** $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}, \qquad x \in \mathbb{R};$$

$E(X) = \mu$, $\text{Var}(X) = \sigma^2$. When $\mu = 0, \sigma = 1$ we call $X$ **standard normal**, write $X \sim N(0,1)$, and denote its cdf by $\Phi(x) = F_X(x)$ and its density by $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

**Lognormal.** $X \sim \text{Lognormal}(\mu, \sigma^2)$ if $Z = \ln X \sim N(\mu, \sigma^2)$:

$$F_X(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \qquad f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-(\ln x-\mu)^2/2\sigma^2}, \qquad x > 0.$$

$E(X) = e^{\mu+\sigma^2/2}$, $\text{Var}(X) = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}$.

**Gamma.** $X \sim \text{Gamma}(\alpha, \theta)$, shape $\alpha > 0$, scale $\theta > 0$:

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \frac{x^{\alpha-1}}{\theta^\alpha} e^{-x/\theta}, \qquad x > 0.$$

$E(X) = \alpha\theta$, $\text{Var}(X) = \alpha\theta^2$ (see Example 1.10.1).

**Exponential.** $X \sim \text{Exp}(\theta) := \text{Gamma}(1, \theta)$, $\theta > 0$: $f_X(x) = \theta^{-1} e^{-x/\theta}$, $x > 0$; $E(X) = \theta$, $\text{Var}(X) = \theta^2$.

**Pareto.** $X \sim \text{Pareto}(\alpha, \theta)$ (Lomax form), $\alpha, \theta > 0$:

$$f_X(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}, \qquad F_X(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha, \qquad x > 0.$$

$E(X) = \dfrac{\theta}{\alpha - 1}$ for $\alpha > 1$ (undefined otherwise), $\text{Var}(X) = \dfrac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$ (see Example 1.10.2).

**Uniform.** $X \sim \text{Uniform}(a, b)$, $a < b$: $f_X(x) = \dfrac{1}{b-a}$, $F_X(x) = \dfrac{x-a}{b-a}$, $a < x < b$; $E(X) = \dfrac{a+b}{2}$, $\text{Var}(X) = \dfrac{(b-a)^2}{12}$.

**Beta.** $X \sim \text{Beta}(a, b, \theta)$, $a, b, \theta > 0$:

$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\frac{x}{\theta}\right)^a \left(1 - \frac{x}{\theta}\right)^{b-1} \frac{1}{x},$$

$$F_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^{x/\theta} t^{a-1}(1-t)^{b-1}\, dt,$$

for $0 < x < \theta$. (The exponent in the density is normalised so that $F_X(x) \to 1$ as $x \to \theta^-$; some sources write the density with $a - 1$ in place of $a$ in the power of $x/\theta$ together with a compensating factor, which is algebraically equivalent.) $E(X) = \theta \dfrac{a}{a+b}$, $\text{Var}(X) = \theta^2 \dfrac{ab}{(a+b)^2(a+b+1)}$. Taking $a = b = 1$ recovers $X \sim \text{Uniform}(0, \theta)$.

**Chi-square.** $X \sim \chi^2_{(n)} := \text{Gamma}(n/2, 2)$, $n > 0$ (typically a positive integer, the "degrees of freedom"):

$$f_X(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \qquad x > 0.$$

$E(X) = n$, $\text{Var}(X) = 2n$. If $X_1, \ldots, X_n$ are i.i.d. $N(0, 1)$, then $\sum_{i=1}^n X_i^2 \sim \chi^2_{(n)}$.

**Student's $t$.** $X \sim t_{(n)}$, $n > 0$, defined by $X = U/\sqrt{V/n}$ with $U \sim N(0,1)$, $V \sim \chi^2_{(n)}$ independent. $E(X) = 0$ for $n > 1$, $\mathrm{Var}(X) = n/(n-2)$ for $n > 2$. As $n \to \infty$, $F_X(x \mid n) \to \Phi(x)$ for every $x$: the $t$-distribution converges to the standard normal.

**Snedecor's $F$.** $X \sim F_{(m,n)}$, $m, n > 0$, defined by $X = (U/m)/(V/n)$ with $U \sim \chi^2_{(m)}$, $V \sim \chi^2_{(n)}$ independent. If $X \sim F_{(m,n)}$ then $1/X \sim F_{(n,m)}$; if $T \sim t_{(n)}$ then $T^2 \sim F_{(1,n)}$.

**Remark 1.7.1.** The Gamma and Pareto distributions above will play a leading role in Section 1.10: they are used there as the canonical examples of, respectively, a light-tailed and a heavy-tailed loss distribution.

## 1.8 Generating Functions and Sums of Independent Random Variables

Generating functions package the entire sequence of moments of a random variable into a single function, and turn the problem of finding the distribution of a sum of independent random variables into simple multiplication. Both properties make them indispensable tools in risk theory, where aggregate losses are sums of individual loss amounts.

### 1.8.1 The moment generating function

**Definition 1.8.1** (Moment generating function)**.** The **moment generating function** (mgf) of $X$ is

$$M_X(r) = E\big(e^{rX}\big), \qquad r \in \mathbb{R},$$

for every $r$ for which the expectation is finite. (This set of $r$'s always contains $r = 0$, where $M_X(0) = 1$; for some distributions, e.g. the Pareto, it contains only $r = 0$, see Section 1.10.)

**Proposition 1.8.1** (Moment-generating property)**.** *If $M_X(r)$ is finite on an open interval containing 0, then $X$ has finite moments of all orders, $M_X$ is infinitely differentiable there, and*

$$E(X^k) = M_X^{(k)}(0) = \frac{d^k M_X}{dr^k}(r)\bigg|_{r=0}, \qquad k = 0, 1, 2, \ldots$$

*Proof.* Under the stated hypothesis one may differentiate under the expectation sign (this is the technical point; it is justified by dominated convergence, using the finiteness of $M_X$ on a neighbourhood of 0 to dominate the relevant difference quotients), giving $M_X^{(k)}(r) = E\big(X^k e^{rX}\big)$; evaluating at $r = 0$ yields $M_X^{(k)}(0) = E(X^k)$. Equivalently, expanding $e^{rX} = \sum_{k=0}^{\infty}(rX)^k/k!$ and exchanging sum and expectation (again justified by the finiteness hypothesis),

$$M_X(r) = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!}\, r^k,$$

and reading off $E(X^k)$ as $k!$ times the coefficient of $r^k$ in this Taylor series recovers $M_X^{(k)}(0) = E(X^k)$. □

The moment generating function of $X$ can be seen as the Laplace transform of the distribution of $X$. Therefore, the mgf determines uniquely the distribution under mild assumptions. The following result is stated without proof.

**Theorem 1.8.2.** *Suppose that $M_X(r)$ and $M_Y(r)$ are finite and agree on an open interval containing the origin. Then, $X$ and $Y$ have the same distribution.*

*Proof.* See e.g. Billingsley, *Probability and Measure*, Theorem. 22.2. □

### 1.8.2 The cumulant generating function

**Definition 1.8.2** (Cumulant generating function)**.** The **cumulant generating function** of $X$ is $\kappa_X(t) = \ln M_X(t)$, wherever $M_X(t)$ is finite and positive.

**Proposition 1.8.3.** *Wherever the relevant derivatives exist,*

$$\kappa_X'(0) = E(X), \qquad \kappa_X''(0) = \mathrm{Var}(X), \qquad \kappa_X'''(0) = \mu_3.$$

*Proof.* Write $m_k = M_X^{(k)}(0) = E(X^k)$ (Proposition 1.8.1), so $M_X(0) = m_0 = 1$. By the chain rule, $\kappa_X'(t) = M_X'(t)/M_X(t)$, so $\kappa_X'(0) = m_1/m_0 = m_1 = E(X)$. Differentiating again, $\kappa_X''(t) = \dfrac{M_X''(t)M_X(t) - M_X'(t)^2}{M_X(t)^2}$, so $\kappa_X''(0) = m_2 - m_1^2 = E(X^2) - E(X)^2 = \mathrm{Var}(X)$. Differentiating once more and evaluating at 0 (a direct, if slightly tedious, computation using $M_X(0) = 1$) gives $\kappa_X'''(0) = m_3 - 3m_1m_2 + 2m_1^3 = E\big[(X - \mu_X)^3\big] = \mu_3$. □

**Remark 1.8.4.** Proposition 1.8.3 is the reason $\kappa_X$ is useful in practice: its low-order Taylor coefficients at 0 are precisely the mean, variance, and (up to normalisation) skewness of $X$, and $\kappa_X$ is often easier to compute. Moreover, it is additive over independent summands (Theorem 1.8.6 below).

### 1.8.3 The probability generating function

For random variables supported on the nonnegative integers, it is often more convenient to work with the following variant of the mgf.

**Definition 1.8.3** (Probability generating function)**.** The **probability generating function** (pgf) of a discrete random variable $X$ is

$$P_X(z) = E\big(z^X\big), \qquad z \in \mathbb{R},$$

for every $z$ for which the expectation is finite. If the support of $X$ is $\{0, 1, 2, \ldots\}$, then

$$P_X(z) = \sum_{k=0}^{\infty} z^k \, \mathbb{P}(X = k),$$

which converges (absolutely) at least for $|z| \leqslant 1$.

The pgf is only defined for discrete $X$, since only in this case it is meaningful. Can be seen as the $z$-transform of the distribution of $X$. In this case, it is easy to see that determines in a unique way the distribution.

**Proposition 1.8.5.** *For $X$ supported on $\{0, 1, 2, \ldots\}$:*

*(a) $M_X(r) = P_X(e^r)$ and $P_X(z) = M_X(\ln z)$ (for $z > 0$ in the domain of convergence);*

*(b) $\mathbb{P}(X = k) = \dfrac{P_X^{(k)}(0)}{k!}$, $k = 0, 1, 2, \ldots$;*

*(c) the $k$th derivative of $P_X$ at $z = 1$ gives the $k$th **factorial moment**,*

$$P_X^{(k)}(1) = E\big[X(X-1)\cdots(X-k+1)\big].$$

*Proof.* (a) By definition, $P_X(e^r) = E\big((e^r)^X\big) = E(e^{rX}) = M_X(r)$; setting $z = e^r$, i.e. $r = \ln z$, gives $P_X(z) = M_X(\ln z)$. (b) Differentiating $P_X(z) = \sum_{j=0}^{\infty} z^j \mathbb{P}(X = j)$ term by term $k$ times and setting $z = 0$ kills every term except $j = k$ (all lower powers of $z$ vanish upon differentiating fewer than $j$ times and then setting $z = 0$ if $j > k$, and every term with $j < k$ is annihilated by the $k$th derivative), leaving $P_X^{(k)}(0) = k!\,\mathbb{P}(X = k)$. (c) Differentiating

$P_X(z) = \sum_j z^j \mathbb{P}(X = j)$ term by term $k$ times gives $P_X^{(k)}(z) = \sum_{j \geqslant k} j(j-1)\cdots(j-k+1) z^{j-k} \mathbb{P}(X = j)$; evaluating at $z = 1$ gives $\sum_j j(j-1)\cdots(j-k+1)\mathbb{P}(X = j) = E[X(X-1)\cdots(X-k+1)]$. □

### 1.8.4 Sums of independent random variables

**Definition 1.8.4** (Convolution)**.** Given random variables $X_1, \ldots, X_n$, the random variable $S_n = X_1 + \cdots + X_n$ is called their **sum**, or (when $X_1, \ldots, X_n$ are independent) their **convolution**. In many applications, if $X_i$ is the loss associated with unit $i$ of a population of $n$ units, $S_n$ is the population's aggregate loss.

**Theorem 1.8.6.** *If $X_1, \ldots, X_n$ are independent, then*

$$M_{S_n}(r) = \prod_{i=1}^{n} M_{X_i}(r), \qquad P_{S_n}(z) = \prod_{i=1}^{n} P_{X_i}(z)$$

*(the second identity when each $X_i$ is supported on the nonnegative integers).*

*Proof.* $M_{S_n}(r) = E\big(e^{r(X_1+\cdots+X_n)}\big) = E\big(e^{rX_1}\cdots e^{rX_n}\big) = \prod_{i=1}^n E\big(e^{rX_i}\big) = \prod_{i=1}^n M_{X_i}(r)$, using independence in the third equality (Proposition 1.4.3 extended to $n$ variables and to the multiplicativity of expectation of independent random variables). The pgf identity is proved identically, replacing $e^{rX_i}$ by $z^{X_i}$. □

**Proposition 1.8.7** (Convolution formula)**.** *If $X, Y$ are independent continuous random variables with densities $f_X, f_Y$, then $Z = X + Y$ has density*

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y)\, dy.$$

*If $X, Y$ are independent, nonnegative, integer-valued, then $Z = X + Y$ has probability function*

$$f_Z(k) = \sum_{j=0}^{k} f_X(k-j) f_Y(j), \qquad k = 0, 1, 2, \ldots$$

*Proof.* In the continuous case, condition on $Y$: by Proposition 1.4.2 and independence,

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{Z|Y=y}(z) f_Y(y)\, dy = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y)\, dy,$$

since, conditionally on $Y = y$, $Z = X + y$ has the density of $X$ shifted by $y$, and this conditional density equals the unconditional one by independence of $X$ and $Y$. The discrete case is the same computation with sums in place of integrals: $\mathbb{P}(Z = k) = \sum_j \mathbb{P}(X = k - j \mid Y = j)\mathbb{P}(Y = j) = \sum_j f_X(k - j) f_Y(j)$, and only $j = 0, \dots, k$ contribute since $f_X(k - j) = 0$ for $j > k$. □

Theorem 1.8.6 is, in practice, the more useful of the two results: it is usually far easier to multiply generating functions and recognise the product than to evaluate the convolution integral or sum of Proposition 1.8.7 directly.

**Example 1.8.8.** If $X_1, \dots, X_n$ are independent with $X_i \sim \text{Poisson}(\lambda_i)$, then $P_{X_i}(z) = E(z^{X_i}) = \sum_k z^k e^{-\lambda_i} \lambda_i^k / k! = e^{\lambda_i (z-1)}$, so by Theorem 1.8.6,

$$P_{S_n}(z) = \prod_{i=1}^{n} e^{\lambda_i (z-1)} = e^{(\lambda_1 + \cdots + \lambda_n)(z-1)},$$

which is recognisable as the pgf of $\text{Poisson}(\lambda_1 + \cdots + \lambda_n)$: the sum of independent Poisson random variables is again Poisson, with the parameters adding.

### 1.8.5 The Central Limit Theorem

We close this section with a classical asymptotic result underlying the widespread use of the normal distribution as an approximation for aggregate quantities (such as the total loss of a large population).

**Theorem 1.8.9** (Central Limit Theorem)**.** *Let $X_1, X_2, \dots$ be independent and identically distributed random variables with finite mean $\mu$ and finite variance $\sigma^2 > 0$, and let $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for every $x \in \mathbb{R}$,*

$$\lim_{n \to \infty} \mathbb{P}\left( \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \leqslant x \right) = \Phi(x),$$

*where $\Phi$ is the standard normal cdf. In other words, regardless of the shape of the common distribution of the $X_i$, the (standardised) sample mean is approximately normally distributed for large $n$.*

*Proof.* First, we standardize each $X_i$ by setting $Y_i = \frac{X_i - \mu}{\sigma}$, so that $E(X_i) = 0$ and $E(Y_i^2) = 1$. Let $M(t) = E(e^{tY_i})$ denote the mgf of $Y_i$. Then, $M(0) = 1$, $M'(0) = 0$ and $M''(0) = 1$. Hence, as $t \to 0$,

$$M(t) = 1 + \frac{t^2}{2} + o(t^2).$$

Now, letting $Z_n = \frac{Y_1+\cdots+Y_n}{\sqrt{n}}$, using independence we get

$$M_{Z_n}(t) = M(t/\sqrt{n})^n.$$

Using the above expansion,

$$M(t/\sqrt{n}) = 1 + \frac{t^2}{2n} + o(1/n)$$

and taking the limit $n \to \infty$,

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + o(1/n)\right)^n \Longrightarrow e^{t^2/2},$$

which is precisely the mgf of $N(0,1)$. Because mgf determines the distribution, the CLT is proved. □

**Remark 1.8.10.** Equivalently, $S_n = X_1+\cdots+X_n$ is approximately $N(n\mu, n\sigma^2)$ for large $n$. Theorem 1.8.9 is the theoretical justification for the normal approximation routinely used to model large, reasonably homogeneous, populations of independent risks; Theorem 1.8.6 above gives an alternative route to the same conclusion when moment generating functions are available, since $\kappa_{S_n}(t) = n\kappa_{X_1}(t)$ (additivity of cumulant generating functions for i.i.d. sums) already shows how the shape of $S_n$'s distribution is governed, for large $n$, by the first two cumulants alone.

## 1.9 Residual Life, Censored and Limited Variables

We now turn to three constructions that are central to loss modelling: what remains of a loss (or a lifetime) beyond a threshold, what is actually paid once a threshold is applied, and what is paid once a limit caps the payment. Throughout this section $X \geqslant 0$ represents a lifetime or a loss amount.

### 1.9.1 Residual life (excess loss)

**Definition 1.9.1** (Residual life)**.** Let $X \geqslant 0$ and $d \geqslant 0$ with $S_X(d) > 0$. The **residual life** (or **future lifetime**) of $X$ at $d$ is the random variable $Y^P$, with distribution given by the conditional law of $X - d$ given $X > d$, i.e. with survival function

$$S_d(x) := \mathbb{P}(Y^P > x) = \mathbb{P}(X - d > x \mid X > d) = \frac{S_X(x+d)}{S_X(d)}, \qquad x \geqslant 0.$$

When $X$ represents a loss (rather than a lifetime), $Y^P$ is called the **excess loss** variable at $d$ (the amount by which the loss exceeds the threshold $d$, given that it does exceed it).

**Definition 1.9.2** (Mean residual life / mean excess loss)**.** The **mean residual life function** (or **mean excess loss function**) of $X$ is

$$e_X(d) = E\big[X - d \mid X > d\big] = E(Y^P), \qquad d \geqslant 0 \text{ with } S_X(d) > 0.$$

**Proposition 1.9.1.**

$$e_X(d) = \frac{\displaystyle\int_0^{+\infty} S_X(x+d)\,dx}{S_X(d)} = \frac{\displaystyle\int_d^{+\infty} S_X(x)\,dx}{S_X(d)},$$

*and, split by type,*

$$e_X(d) = \begin{cases} \dfrac{\displaystyle\int_d^{+\infty} (x-d) f_X(x)\,dx}{S_X(d)}, & X \textit{ abs continuous}, \\[2ex] \dfrac{\displaystyle\sum_{x>d} (x-d) f_X(x)}{S_X(d)}, & X \textit{ discrete}. \end{cases}$$

*Proof.* $Y^P$ is a nonnegative random variable with survival function $S_d(x) = S_X(x+d)/S_X(d)$ (Definition 1.9.1). Applying (1.6.2) of Theorem 1.6.7 to $Y^P$,

$$e_X(d) = E(Y^P) = \int_0^{+\infty} S_d(x)\,dx = \int_0^{+\infty} \frac{S_X(x+d)}{S_X(d)}\,dx = \frac{1}{S_X(d)} \int_0^{\infty} S_X(x+d)\,dx,$$

and the substitution $u = x + d$ turns the last integral into $\int_d^{+\infty} S_X(u)\,du$, giving the first display. For the split-by-type formula in the continuous case, apply Definition 1.9.2 directly: by definition of conditional density, $X - d \mid X > d$ has density $f_X(x+d)/S_X(d)$ for $x \geqslant 0$, so

$$e_X(d) = \int_0^{\infty} x\,\frac{f_X(x+d)}{S_X(d)}\,dx = \frac{1}{S_X(d)} \int_d^{\infty} (u-d) f_X(u)\,du$$

by the same substitution $u = x + d$; the discrete case is identical with sums in place of integrals. □

**Definition 1.9.3** (Higher moments of the residual life)**.** More generally, $e_X^k(d) := E\big[(X-d)^k \mid X > d\big]$, $k = 1, 2, \ldots$ (so $e_X^1(d) = e_X(d)$).

**Example 1.9.2** (The exponential distribution is memoryless)**.** If $X \sim \text{Exp}(\theta)$, then for $d \geqslant 0$, $x \geqslant 0$,

$$S_d(x) = \frac{S_X(x+d)}{S_X(d)} = \frac{e^{-(x+d)/\theta}}{e^{-d/\theta}} = e^{-x/\theta} = S_X(x) :$$

the residual life at any age $d$ has *exactly the same distribution* as $X$ itself. This is the **memoryless property** of the exponential distribution: knowing that an individual has survived to age $d$ tells us nothing about their remaining lifetime. Consequently $e_X(d) = \theta$ for every $d$: the mean excess loss function of the exponential is constant, matching the constant hazard rate found in Example 1.6.12.

**Example 1.9.3** (Residual life of the Pareto distribution)**.** If $X \sim \text{Pareto}(\alpha, \theta)$, then $S_X(x) = \left(\frac{\theta}{\theta+x}\right)^\alpha$, so

$$S_d(x) = \frac{S_X(x+d)}{S_X(d)} = \frac{(\theta+d)^\alpha/(\theta+d+x)^\alpha}{1} = \left(\frac{\theta+d}{\theta+d+x}\right)^\alpha, \qquad x \geqslant 0,$$

which is again the survival function of a $\text{Pareto}(\alpha, \theta+d)$ distribution: the Pareto family is closed under taking residual lives, the shape parameter $\alpha$ staying fixed while the scale shifts from $\theta$ to $\theta+d$. Consequently, using $E(X) = \theta/(\alpha-1)$ for $\alpha > 1$,

$$e_X(d) = \frac{\theta+d}{\alpha-1},$$

an *increasing, unbounded* function of $d$: unlike the exponential, the older a Pareto-distributed risk gets, the larger its expected remaining loss/lifetime. We will see in Section 1.10 that this increasing mean excess loss is a signature of a heavy-tailed distribution.

### 1.9.2 Left censored and shifted variables

**Definition 1.9.4** (Left censored and shifted variable)**.** Let $X \geqslant 0$ and $d \geqslant 0$. The **left censored and shifted** variable at $d$ is

$$Y^L = (X-d)_+ = (X-d) \vee 0 := \max(X-d, 0).$$

**Remark 1.9.4.** The variable $Y^L$ differs fundamentally from the residual life $Y^P$ of Definition 1.9.1: while $Y^P$ is only defined *given* $X > d$ and is (for continuous $X$) itself continuous with $\mathbb{P}(Y^P = 0) = 0$, the variable $Y^L$ is

defined unconditionally on the whole sample space and carries a point mass at 0,

$$\mathbb{P}(Y^L = 0) = \mathbb{P}(X \leqslant d) = 1 - S_X(d),$$

representing all outcomes in which $X$ does not exceed the threshold $d$. In risk-modelling terms, $Y^L$ is the natural model for a payment subject to a threshold $d$: if the loss $X$ does not exceed $d$, nothing is paid ($Y^L = 0$); if it does, the excess $X - d$ is paid.

**Theorem 1.9.5.** *For $k = 1, 2, \ldots$,*

$$E\big[(Y^L)^k\big] = E\big[(X-d)_+^k\big] = e_X^k(d)\, S_X(d),$$

*and, in particular,*

$$E(Y^L) = E\big[(X-d)_+\big] = e_X(d)\, S_X(d) = \int_d^{+\infty} S_X(x)\, dx. \tag{1.9.3}$$

*Proof.* Splitting the expectation according to whether $X > d$ or not, and using that $(X-d)_+^k = 0$ on $\{X \leqslant d\}$,

$$E\big[(X-d)_+^k\big] = E\big[(X-d)^k \mathbf{1}_{\{X>d\}}\big] = E\big[(X-d)^k \mid X > d\big]\, \mathbb{P}(X > d) = e_X^k(d)\, S_X(d),$$

using the definition of conditional expectation, $E[Z\mathbf{1}_A] = E[Z \mid A]\mathbb{P}(A)$. Setting $k = 1$ and invoking Proposition 1.9.1 for $e_X(d)S_X(d) = \int_d^\infty S_X(x)\, dx$ gives (1.9.3). □

### 1.9.3 Limited loss (limited expected value)

**Definition 1.9.5** (Limited loss variable)**.** Let $X \geqslant 0$ and $u > 0$. The **limited loss** (or **limited**) variable at $u$ is

$$Y = X \wedge u := \min(X, u).$$

Its expectation $E(X \wedge u)$ is called the **limited expected value** of $X$ at $u$.

**Proposition 1.9.6.**

$$E(X \wedge u) = \int_0^u S_X(x)\, dx.$$

*Proof.* $X \wedge u$ is a nonnegative random variable with survival function

$$\mathbb{P}(X \wedge u > x) = \begin{cases} S_X(x), & 0 \leqslant x < u, \\ 0, & x \geqslant u, \end{cases}$$

since $X \wedge u > x$ (for $x < u$) holds exactly when $X > x$, while $X \wedge u \leqslant u$ always. Applying (1.6.2) of Theorem 1.6.7 to $X \wedge u$ gives $E(X \wedge u) = \int_0^\infty \mathbb{P}(X \wedge u > x)\,dx = \int_0^u S_X(x)\,dx$. □

**Proposition 1.9.7.** *For every $u > 0$,*

$$X = (X \wedge u) + (X - u)_+ \qquad \textit{(pointwise, hence also in expectation)},$$

*so that*

$$E(X) = E(X \wedge u) + E\big[(X-u)_+\big] = \int_0^u S_X(x)\,dx + \int_u^{+\infty} S_X(x)\,dx.$$

*Proof.* If $X \leqslant u$ then $X \wedge u = X$ and $(X-u)_+ = 0$, so the sum is $X$; if $X > u$ then $X \wedge u = u$ and $(X-u)_+ = X - u$, so the sum is again $X$. Taking expectations and using Proposition 1.9.6 and Theorem 1.9.5 for the two terms recovers (1.6.2) split at $u$. □

**Remark 1.9.8.** Proposition 1.9.7 has an immediate risk-modelling reading: if a threshold $u$ acts as a *limit on the amount retained within a given layer*, then the total loss $X$ splits exactly into the part retained/limited at $u$ (the amount paid never exceeds $u$ under a limit, or the amount retained is at most $u$ under a threshold, depending on which side of the layer one considers) and the excess beyond $u$. Limited expected values of this kind are the basic building block used to model layered risk transfer and limits/thresholds.

## 1.10 Tails of Distributions

In risk modelling, what happens "in the tail", the relative likelihood of very large losses, matters far more than the bulk of the distribution: stability and risk-transfer decisions are driven by extreme, low-probability, high-severity events. Distributions with a larger probability of extreme values, i.e. a larger survival function $S_X(x) = \mathbb{P}(X > x)$ (also called the **tail probability**) for large $x$, are said to be **heavier-tailed**. This section develops four complementary ways of making this comparison precise: via moments, via limiting ratios, via the hazard rate, and via the mean excess loss function. As a rule of thumb, to be justified rigorously below, the Gamma/exponential family is **light-tailed**, the lognormal is **medium-tailed**, and the Pareto is **heavy-tailed**.

### 1.10.1 Comparison based on moments

**Example 1.10.1** (Moments of the Gamma distribution)**.** Let $X \sim \text{Gamma}(\alpha, \theta)$. For every $k > -\alpha$,

$$E(X^k) = \int_0^\infty x^k \, \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\theta} \, dx = \frac{1}{\theta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha+k-1} e^{-x/\theta} \, dx = \frac{\theta^k \Gamma(\alpha+k)}{\Gamma(\alpha)},$$

using the substitution $t = x/\theta$ and the definition of $\Gamma(\alpha+k) = \int_0^\infty t^{\alpha+k-1} e^{-t} \, dt$ (finite whenever $\alpha + k > 0$). In particular, taking $k = 1, 2$ and using $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(\alpha+2) = \alpha(\alpha+1)\Gamma(\alpha)$, one recovers $E(X) = \alpha\theta$ and $\text{Var}(X) = E(X^2) - E(X)^2 = \alpha(\alpha+1)\theta^2 - \alpha^2\theta^2 = \alpha\theta^2$, as claimed in Section 1.7. Every moment $E(X^k)$ of the Gamma distribution is finite, for every $k > 0$.

**Example 1.10.2** (Moments of the Pareto distribution)**.** Let $X \sim \text{Pareto}(\alpha, \theta)$. For $k \geqslant 0$,

$$E(X^k) = \int_0^\infty x^k \, \frac{\alpha\theta^\alpha}{(\theta+x)^{\alpha+1}} \, dx = \int_0^\infty \frac{\alpha x^k}{\theta} \left(\frac{\theta}{\theta+x}\right)^{\alpha+1} dx.$$

Substituting $z = \theta/(\theta+x)$, so $x = \theta(1-z)/z$ and $dx = -\theta z^{-2} \, dz$, transforms this into

$$E(X^k) = \alpha\theta^k \int_0^1 z^{\alpha-k-1}(1-z)^k \, dz = \alpha\theta^k \, B(\alpha-k, k+1),$$

where $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} \, dt = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function, and the integral converges if and only if $\alpha - k > 0$. Using $B(\alpha-k, k+1) = \Gamma(\alpha-k)\Gamma(k+1)/\Gamma(\alpha+1) = k!\,\Gamma(\alpha-k)/(\alpha\Gamma(\alpha))$, we obtain

$$E(X^k) = \theta^k \, k! \, \frac{\Gamma(\alpha-k)}{\Gamma(\alpha)}, \qquad 0 \leqslant k < \alpha.$$

Thus the $k$th moment of a $\text{Pareto}(\alpha, \theta)$ random variable exists *if and only if* $k < \alpha$: only finitely many moments are finite (specifically, $\lfloor \alpha \rfloor$ of the positive-integer ones), and in particular the mgf $M_X(r) = E(e^{rX})$, which would require moments of *every* order to be finite (Proposition 1.8.1), does not exist for any $r > 0$.

The contrast between Examples 1.10.1 and 1.10.2 yields the first classification criterion.

**Definition 1.10.1** (Light- and heavy-tailed distributions, by moments)**.** A distribution is **light-tailed** if all its (positive-integer) moments are finite; it is **heavy-tailed** if only finitely many (or none) of its moments are finite.

### 1.10.2 Comparison based on limiting tail behaviour

**Definition 1.10.2** (Relative tail heaviness)**.** Distribution 1 has a **heavier tail** than distribution 2 if

$$\lim_{x\to+\infty} \frac{S_1(x)}{S_2(x)} = +\infty.$$

When $S_1(x), S_2(x) \to 0$ as $x \to \infty$ and are differentiable, L'Hôpital's rule shows this is equivalent to

$$\lim_{x\to+\infty} \frac{S_1'(x)}{S_2'(x)} = \lim_{x\to+\infty} \frac{f_1(x)}{f_2(x)} = +\infty,$$

so the comparison may equally be carried out on densities.

Definition 1.10.2 suggests using the exponential distribution, the simplest distribution with an exponentially decaying tail, as a tail benchmark.

**Definition 1.10.3** (Heavy-tailed distribution, via exponential comparison)**.** A nonnegative random variable $X$ (or its distribution) is **heavy-tailed** if its tail is not exponentially bounded, that is, if for every $\theta > 0$,

$$\lim_{x\to+\infty} e^{x/\theta}\, S_X(x) = +\infty.$$

**Proposition 1.10.3.** *$X$ is heavy-tailed in the sense of Definition 1.10.3 if and only if $M_X(r) = +\infty$ for every $r > 0$.*

*Proof.* Since $X \geq 0$,

$$\begin{aligned} E(e^{rX}) &= \int_0^\infty \mathbb{P}(e^{rX} > y)\, dy \\ &= \int_0^1 \mathbb{P}(e^{rX} > y)\, dy + \int_1^\infty \mathbb{P}(e^{rX} > y)\, dy \\ &= 1 + \int_1^\infty S_X(\frac{\log y}{r})\, dy \\ &= 1 + \int_0^\infty S_X(x) r e^{rx}\, dx. \end{aligned}$$

This gives the following identity, valid for $r > 0$ whenever the right-hand side is finite,

$$M_X(r) = 1 + r \int_0^{+\infty} e^{rx} S_X(x)\, dx.$$

If $M_X(r) < \infty$ for some $r > 0$, the integral on the right is finite, which forces $e^{rx}S_X(x) \to 0$; taking $\theta = 1/r$ then shows $\lim_x e^{x/\theta}S_X(x) = 0 \neq +\infty$, so $X$ is *not* heavy-tailed by Definition 1.10.3. Conversely, $M_X(r) = \infty$ for every $r > 0$ forces, via the same integral identity, $\int_0^\infty e^{rx}S_X(x)\,dx = \infty$ for every $r > 0$, which (given that $S_X$ is nonincreasing) is only possible if $e^{rx}S_X(x) \not\to 0$, hence (again using monotonicity considerations) $e^{rx}S_X(x) \to \infty$, for every $r = 1/\theta > 0$: precisely Definition 1.10.3. □

### 1.10.3 Comparison based on the hazard rate

**Proposition 1.10.4.** *Let $X \geqslant 0$ be continuous with hazard rate $h_X$. If $h_X(x) \to \infty$ as $x \to \infty$, then $X$ is light-tailed in the sense of Definition 1.10.3; if $h_X(x) \to 0$ as $x \to \infty$ (in particular, if $h_X$ is nonincreasing and not eventually constant and positive), $X$ is heavy-tailed. More generally, if $h_X$ is **increasing**, the tail of $X$ is regarded as **light**; if $h_X$ is **decreasing**, the tail of $X$ is regarded as **heavy** — by comparison with the constant hazard rate of the exponential (Example 1.6.12).*

*Proof.* By Theorem 1.6.10, $S_X(x) = \exp\big(-\int_0^x h_X(t)\,dt\big)$. Fix $\theta > 0$ and consider $g(x) := e^{x/\theta}S_X(x) = \exp\big(\int_0^x [\theta^{-1} - h_X(t)]\,dt\big)$. If $h_X(x) \to \infty$, then for any $\theta$ there is $x_0$ with $h_X(t) > 1/\theta$ for $t > x_0$, so $\int_0^x [\theta^{-1} - h_X(t)]dt \leq C - (h_X\text{-excess integral}) \to -\infty$ as $x \to \infty$, giving $g(x) \to 0$; since this holds for every $\theta > 0$, $X$ is not heavy-tailed by Proposition 1.10.3's criterion, i.e. it is light-tailed. Symmetrically, if $h_X(x) \to 0$, then for every $\theta > 0$ eventually $h_X(t) < 1/\theta$, so the exponent $\int_0^x [\theta^{-1} - h_X(t)]\,dt \to +\infty$, giving $g(x) \to \infty$ for every $\theta$: $X$ is heavy-tailed by Definition 1.10.3. □

### 1.10.4 Comparison based on the mean excess loss function

**Definition 1.10.4.** $X$ is regarded as **heavy-tailed** if its mean excess loss function $e_X(d)$ is (eventually) increasing in $d$, and **light-tailed** if $e_X(d)$ is (eventually) decreasing in $d$.

**Proposition 1.10.5.** *If the hazard rate $h_X$ is decreasing, then the mean excess loss function $e_X$ is increasing.*

*Proof.* By Proposition 1.9.1 and Theorem 1.6.10,

$$e_X(d) = \int_0^{+\infty} \frac{S_X(x+d)}{S_X(d)}\,dx,$$

where, for fixed $x \geqslant 0$,

$$\frac{S_X(x+d)}{S_X(d)} = \frac{\exp\big(-\int_0^{x+d} h_X(t)\,dt\big)}{\exp\big(-\int_0^{d} h_X(t)\,dt\big)} = \exp\left(-\int_d^{x+d} h_X(t)\,dt\right)$$
$$= \exp\left(-\int_0^{x} h_X(t+d)\,dt\right),$$

the last equality by the substitution $t \mapsto t+d$ inside the integral. Fix $x \geqslant 0$ and consider $\varphi(d) := \int_0^x h_X(t+d)\,dt$ as a function of $d$; differentiating under the integral sign, $\varphi'(d) = \int_0^x h_X'(t+d)\,dt \leqslant 0$ since $h_X$ is decreasing (so $h_X' \leqslant 0$), so $\varphi$ is nonincreasing in $d$. Hence $-\varphi(d)$ is nondecreasing in $d$, so $S_X(x+d)/S_X(d) = e^{-\varphi(d)}$ is nondecreasing in $d$ for each fixed $x$, and therefore so is its integral over $x \in [0,\infty)$: $e_X(d)$ is nondecreasing (increasing, in the generic case of strict monotonicity of $h_X$) in $d$. □

**Remark 1.10.6.** The converse of Proposition 1.10.5 is *false*: an increasing mean excess loss function does not force a decreasing hazard rate. (Klugman, Panjer and Willmot's *Loss Models* has a counterexample as an exercise.)

**Proposition 1.10.7** (Limit of the mean excess loss)**.**

$$\lim_{d\to+\infty} e_X(d) = \lim_{d\to+\infty} \frac{1}{h_X(d)},$$

*whenever the right-hand limit exists.*

*Proof.* By Proposition 1.9.1, $e_X(d) = \dfrac{\int_d^\infty S_X(x)\,dx}{S_X(d)}$, an indeterminate form $0/0$ as $d \to \infty$ (both numerator and denominator $\to 0$). Differentiating numerator and denominator with respect to $d$ and applying L'Hôpital's rule,

$$\lim_{d\to\infty} e_X(d) = \lim_{d\to\infty} \frac{-S_X(d)}{-f_X(d)} = \lim_{d\to\infty} \frac{S_X(d)}{f_X(d)} = \lim_{d\to\infty} \frac{1}{h_X(d)}.$$

□

### 1.10.5 The equilibrium distribution

**Definition 1.10.5** (Equilibrium distribution)**.** Let $X \geqslant 0$ with finite mean $E(X)$ and $S_X(0) = 1$. The **equilibrium distribution** of $X$ is the continuous distribution with density

$$f_e(x) = \frac{S_X(x)}{E(X)}, \qquad x > 0,$$

survival function $S_e(x) = \dfrac{\int_x^{+\infty} S_X(t)\,dt}{E(X)}$, and hazard rate $h_e(x) = f_e(x)/S_e(x)$.

**Proposition 1.10.8.**

$$h_e(x) = \frac{1}{e_X(x)}, \qquad x > 0.$$

*Proof.* By Definition 1.10.5 and Proposition 1.9.1,

$$h_e(x) = \frac{f_e(x)}{S_e(x)} = \frac{S_X(x)/E(X)}{\int_x^{\infty} S_X(t)\,dt/E(X)} = \frac{S_X(x)}{\int_x^{\infty} S_X(t)\,dt} = \frac{1}{e_X(x)}.$$

□

**Remark 1.10.9.** Proposition 1.10.8 says that the reciprocal of the mean excess loss function is itself a (bona fide) hazard rate: that of the equilibrium distribution. This has an interesting consequence: since a hazard rate determines its survival function uniquely (Theorem 1.6.10), the mean excess loss function $e_X$ determines the original distribution $S_X$ uniquely as well.

**Theorem 1.10.10** (Reconstruction of $S_X$ from $e_X$)**.** *Let $X \geqslant 0$ with $S_X(0) = 1$ and finite mean $E(X) = e_X(0)$. Then*

$$S_X(x) = \frac{e_X(0)}{e_X(x)} \exp\left( - \int_0^x \frac{dt}{e_X(t)} \right), \qquad x \geqslant 0.$$

*Proof.* By Theorem 1.6.10 applied to the equilibrium distribution, $S_e(x) = \exp\big( - \int_0^x h_e(t)\,dt\big) = \exp\big( - \int_0^x e_X(t)^{-1}\,dt\big)$, using Proposition 1.10.8. On the other hand, by Definition 1.10.5 and Proposition 1.9.1,

$$S_e(x) = \frac{\int_x^{\infty} S_X(t)\,dt}{E(X)} = \frac{e_X(x) S_X(x)}{e_X(0)}.$$

Equating the two expressions for $S_e(x)$ and solving for $S_X(x)$ gives the claimed identity. □

### 1.10.6 The equilibrium distribution and the coefficient of variation

We conclude with a result linking the monotonicity of the mean excess loss function, our tail-heaviness criterion of Definition 1.10.4, to the coefficient of variation $CV_X$ of Definition 1.6.2, thereby tying together the moment-based and mean-excess-based notions of tail weight introduced earlier in this section.

**Proposition 1.10.11.** *Assume $S_X(0) = 1$ and $E(X^2) < \infty$. If $e_X$ is (nondecreasing) increasing, then $CV_X \geqslant 1$; if $e_X$ is (nonincreasing) decreasing, then $CV_X \leqslant 1$.*

*Proof.* Suppose $e_X$ is nondecreasing, so $e_X(x) \geqslant e_X(0)$ for all $x \geqslant 0$. By Proposition 1.10.8–1.10.10's identity $e_X(x)/e_X(0) = S_e(x)/S_X(x)$ (obtained by dividing $S_e(x) = e_X(x)S_X(x)/e_X(0)$, established in the proof of Theorem 1.10.10, through by $S_X(x)$), this is equivalent to

$$S_e(x) \geqslant S_X(x), \qquad \text{for all } x \geqslant 0.$$

Integrating both sides over $x \in [0, \infty)$,

$$\int_0^\infty S_e(x)\,dx \geqslant \int_0^\infty S_X(x)\,dx = E(X),$$

using (1.6.2). But, since $S_e$ is itself the survival function of the equilibrium distribution with density $f_e(x) = S_X(x)/E(X)$, applying (1.6.2) again (this time to the equilibrium distribution) gives

$$\int_0^\infty S_e(x)\,dx = E_e(X) = \int_0^\infty x f_e(x)\,dx = \frac{1}{E(X)}\int_0^\infty x S_X(x)\,dx = \frac{1}{E(X)}\cdot\frac{E(X^2)}{2},$$

where the last equality is the standard identity $\int_0^\infty x S_X(x)\,dx = \frac{1}{2}E(X^2)$ (itself provable by the Fubini argument of Theorem 1.6.7, applied to $xS_X(x) = x\int_x^\infty f_X(t)\,dt$). Combining the two displays,

$$\frac{E(X^2)}{2E(X)} \geqslant E(X) \iff E(X^2) \geqslant 2E(X)^2 \iff \mathrm{Var}(X) = E(X^2) - E(X)^2 \geqslant E(X)^2$$
$$\iff CV_X = \frac{\sigma_X}{E(X)} \geqslant 1.$$

The decreasing case is symmetric, with all inequalities reversed. □

## 1.10.7 Summary

The four criteria developed in this section are compatible, and in fact form a chain of implications summarised in Table 1.1.

**Corollary 1.10.12.** *If $h_X$ is decreasing, then $e_X$ is increasing (Proposition 1.10.5) and hence $CV_X \geqslant 1$ (Proposition 1.10.11). Symmetrically, if $h_X$ is increasing, then $e_X$ is decreasing and $CV_X \leqslant 1$. None of the reverse implications hold in general.*

| **Criterion** | **Light tail** | **Heavy tail** |
|---|---|---|
| Moments (Definition 1.10.1) | all moments finite | only finitely many finite |
| Limiting ratio (Definition 1.10.2) | comparison $S_1(x)/S_2(x) \to \infty$ or not | |
| Hazard rate $h_X(x)$ | increasing | decreasing |
| Mean excess loss $e_X(x)$ | decreasing | increasing |
| Coefficient of variation | $CV_X \leqslant 1$ (typically) | $CV_X \geqslant 1$ (typically) |

Table 1.1: Tail-weight criteria and their typical verdicts.

**Example 1.10.13.** The Gamma distribution with $\alpha < 1$ has decreasing hazard rate (Exercise **??**), hence increasing mean excess loss and $CV_X \geqslant 1$; with $\alpha > 1$ it has increasing hazard rate, hence decreasing mean excess loss and $CV_X \leqslant 1$; at $\alpha = 1$ (the exponential) all quantities are constant and $CV_X = 1$ exactly, consistent with $\mathrm{Var}(X) = \theta^2 = E(X)^2$. The Pareto distribution, by contrast, always has decreasing hazard rate (Exercise **??**) and increasing mean excess loss (Example 1.9.3) for every choice of $\alpha, \theta$: it is unambiguously heavy-tailed by every criterion in Table 1.1, consistent with its failing even to possess a moment generating function (Example 1.10.2).