Appendix: Background on Probability Theory

In this appendix, we review a number of basic probabilistic tools that can be needed for option pricing and hedging. We refer the reader to Pitman (1999) Jacod and Protter (2000), Devore (2003), for additional background on probability theory.

A.1 Probability Sample Space and Events

We will need the following notation coming from set theory. Given A and B to abstract sets, "$A \subset B$" means that A is contained in B, and in this case, $B \setminus A$ denotes the set of elements of B which do not belong to A. The property that the element ω belongs to the set A is denoted by "$\omega \in A$", and given two sets A and Ω such that $A \subset \Omega$, we let $A^c = \Omega \setminus A$ denote the *complement* of A in Ω. The finite set made of n elements $\omega_1, \ldots, \omega_n$ is denoted by $\{\omega_1, \ldots, \omega_n\}$, and we will usually distinguish between the element ω and its associated singleton set $\{\omega\}$.

A probability sample space is an abstract set Ω that contains the possible outcomes of a random experiment.

Examples

i) Coin tossing: $\Omega = \{H, T\}$.

$\circlearrowleft$

815

ii) Rolling one die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

iii) Picking one card at random in a pack of 52: $\Omega = \{1, 2, 3, \ldots, 52\}$.

iv) An integer-valued random outcome: $\Omega = \mathbb{N} = \{0, 1, 2, \ldots\}$.

In this case the outcome $\omega \in \mathbb{N}$ can be the random number of trials needed until some event occurs.

v) A nonnegative, real-valued outcome: $\Omega = \mathbb{R}_+$.

In this case the outcome $\omega \in \mathbb{R}_+$ may represent the (nonnegative) value of a continuous random time.

vi) A random continuous parameter (such as time, weather, price or wealth, temperature, ...): $\Omega = \mathbb{R}$.

vii) Random choice of a continuous path in the space $\Omega = \mathcal{C}(\mathbb{R}_+)$ of all continuous functions on $\mathbb{R}_+$.

In this case, $\omega \in \Omega$ is a function $\omega : \mathbb{R}_+ \longrightarrow \mathbb{R}$ and a typical example is the graph $t \longmapsto \omega(t)$ of a stock price over time.

Product spaces:

Probability sample spaces can be built as product spaces and used for the modeling of repeated random experiments.

i) Rolling two dice: $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$.

In this case a typical element of Ω is written as $\omega = (k, l)$ with $k, l \in \{1, 2, 3, 4, 5, 6\}$.

ii) A finite number n of real-valued samples: $\Omega = \mathbb{R}^n$.

In this case the outcome ω is a vector $\omega = (x_1, \ldots, x_n) \in \mathbb{R}^n$ with n components.

Note that to some extent, the more complex Ω is, the better it fits a practical and useful situation, *e.g.* $\Omega = \{H, T\}$ corresponds to a simple coin tossing experiment while $\Omega = \mathcal{C}(\mathbb{R}_+)$ the space of continuous functions on $\mathbb{R}_+$ can be applied to the modeling of stock markets. On the other hand, in many cases and especially in the most complex situations, we will *not* attempt to specify Ω explicitly.

Events

An event is a collection of outcomes, which is represented by a subset of Ω. In what follows we consider collections of events, called *σ-algebras* (or *σ-fields*), according to the following definition.

Definition A.1. *A collection $\mathcal{G}$ of events is a σ-algebra on Ω provided that it satisfies the following conditions:*

(i) $\emptyset \in \mathcal{G}$,

(ii) For all countable sequences $(A_n)_{n \geqslant 1}$ such that $A_n \in \mathcal{G}$, $n \geqslant 1$, we have $\bigcup_{n \geqslant 1} A_n \in \mathcal{G}$,

(iii) $A \in \mathcal{G} \implies (\Omega \setminus A) \in \mathcal{G}$,

where $\Omega \setminus A := \{\omega \in \Omega \ : \ \omega \notin A\}$.

Note that Properties (ii) and (iii) above also imply the stability of σ-algebras under intersections, as

$$\bigcap_{n \geqslant 1} A_n = \left(\bigcup_{n \geqslant 1} A_n^c \right)^c \in \mathcal{G}, \tag{A.1}$$

for all countable sequences $A_n \in \mathcal{G}$, $n \geqslant 1$.

The collection of all events in Ω will often be denoted by $\mathcal{F}$. The empty set $\emptyset$ and the full space Ω are considered as events but they are of less importance because Ω corresponds to "any outcome may occur" while $\emptyset$ corresponds to an absence of outcome, or no experiment.

In the context of stochastic processes, two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F} \subset \mathcal{G}$ will refer to two different amounts of information, the amount of information associated to $\mathcal{F}$ being here lower than the one associated to $\mathcal{G}$.

The formalism of σ-algebras helps in describing events in a short and precise way.

Examples

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

The event $A = \{2, 4, 6\}$ corresponds to

"the result of the experiment is an even number".

ii) Taking again $\Omega = \{1, 2, 3, 4, 5, 6\}$,

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

defines a σ-algebra on Ω which corresponds to the knowledge of parity of an integer picked at random from 1 to 6.

Note that in the set-theoretic notation, an event A is a subset of Ω, *i.e.* $A \subset \Omega$, while it is an element of $\mathcal{F}$, *i.e.* $A \in \mathcal{F}$. For example, we have $\Omega \supset \{2, 4, 6\} \in \mathcal{F}$, while $\{\{2, 4, 6\}, \{1, 3, 5\}\} \subset \mathcal{F}$.

iii) Taking

$$\mathcal{G} := \{\Omega, \emptyset, \{2,4,6\}, \{2,4\}, \{6\}, \{1,2,3,4,5\}, \{1,3,5,6\}, \{1,3,5\}\} \supset \mathcal{F},$$

defines a σ-algebra on Ω which is bigger than $\mathcal{F}$ and includes the parity information contained in $\mathcal{F}$, in addition to information on whether the outcome of the experiment is equal to 6 or not.

iv) Take

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

In this case, the collection $\mathcal{F}$ of all possible events is given by

$$\begin{aligned}
\mathcal{F} = \{&\emptyset, \{(H,H)\}, \{(T,T)\}, \{(H,T)\}, \{(T,H)\}, &\text{(A.2)}\\
&\{(T,T),(H,H)\}, \{(H,T),(T,H)\}, \{(H,T),(T,T)\},\\
&\{(T,H),(T,T)\}, \{(H,T),(H,H)\}, \{(T,H),(H,H)\},\\
&\{(H,H),(T,T),(T,H)\}, \{(H,H),(T,T),(H,T)\},\\
&\{(H,T),(T,H),(H,H)\}, \{(H,T),(T,H),(T,T)\}, \Omega\}.
\end{aligned}$$

Note that the set $\mathcal{F}$ of all events considered in (A.2) above has altogether

$$1 = \binom{n}{0} \text{ event of cardinality } 0,$$

$$4 = \binom{n}{1} \text{ events of cardinality } 1,$$

$$6 = \binom{n}{2} \text{ events of cardinality } 2,$$

$$4 = \binom{n}{3} \text{ events of cardinality } 3,$$

$$1 = \binom{n}{4} \text{ event of cardinality } 4,$$

with $n = 4$, for a total of

$$16 = 2^n = \sum_{k=0}^{4} \binom{4}{k} = 1 + 4 + 6 + 4 + 1$$

events. The collection of events

$$\mathcal{G} := \{\emptyset, \{(T,T),(H,H)\}, \{(H,T),(T,H)\}, \Omega\}$$

defines a sub σ-algebra of $\mathcal{F}$, which corresponds to the restricted information "the results of two coin tossings are different".

Exercise: Write down the set of all events on $\Omega = \{H,T\}$.

Note also that (H, T) is different from (T, H), whereas $\{(H, T), (T, H)\}$ is equal to $\{(T, H), (H, T)\}$.

In addition, we will distinguish between the *outcome* $\omega \in \Omega$ and its associated *event* $\{\omega\} \in \mathcal{F}$, which satisfies $\{\omega\} \subset \Omega$.

A.2 Probability Measures

Definition A.2. *A* probability measure *is a mapping* $\mathbb{P} : \mathcal{F} \longrightarrow [0, 1]$ *that assigns a probability* $\mathbb{P}(A) \in [0, 1]$ *to any event* $A \in \mathcal{F}$, *with the properties*

a) $\mathbb{P}(\Omega) = 1$, *and*

b) $\mathbb{P} \left(\bigcup_{n \geqslant 1} A_n \right) = \sum_{n \geqslant 1} \mathbb{P}(A_n)$, *whenever* $A_k \cap A_l = \emptyset$, $k \neq l$.

Property (b) above is named the *law of total probability*. It states in particular that we have

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$$

when the subsets $A_1, \ldots, A_n$ of Ω are disjoint, and

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \tag{A.3}$$

if $A \cap B = \emptyset$. We also have the *complement rule*

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

When A and B are not necessarily disjoint, we can write

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

which extends to arbitrary families of events $(A_i)_{i \in I}$ indexed by a finite set I as the inclusion-exclusion principle

$$\mathbb{P} \left(\bigcup_{i \in I} A_i \right) = \sum_{J \subset I} (-1)^{|J|+1} \mathbb{P} \left(\bigcap_{j \in J} A_j \right), \tag{A.4}$$

and

$$\mathbb{P} \left(\bigcap_{j \in J} A_j \right) = \sum_{I \subset J} (-1)^{|I|+1} \mathbb{P} \left(\bigcup_{i \in I} A_i \right). \tag{A.5}$$

The triple

$$(\Omega, \mathcal{F}, \mathbb{P}) \tag{A.6}$$

↺

is called a *probability space*, and was introduced by A.N. Kolmogorov (1903-1987). This setting is generally referred to as the *Kolmogorov framework*.

A property or event is said to hold $\mathbb{P}$-*almost surely* (also written $\mathbb{P}$-*a.s.*) if it holds with probability equal to one.

Example

Take

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

and

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

The *uniform* probability measure $\mathbb{P}$ on $(\Omega,\mathcal{F})$ is given by setting

$$\mathbb{P}(\{(T,T),(H,H)\}) := \frac{1}{2} \quad \text{and} \quad \mathbb{P}(\{(H,T),(T,H)\}) := \frac{1}{2}.$$

In addition, we have the following convergence properties.

1. Let $(A_n)_{n\in\mathbb{N}}$ be a *non-decreasing* sequence of events, *i.e.* $A_n \subset A_{n+1}$, $n \geqslant 0$. Then we have

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n). \tag{A.7}$$

2. Let $(A_n)_{n\in\mathbb{N}}$ be a *non-increasing* sequence of events, *i.e.* $A_{n+1} \subset A_n$, $n \geqslant 0$. Then we have

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n). \tag{A.8}$$

Theorem A.3. *Borel-Cantelli Lemma. Let* $(A_n)_{n\geqslant 1}$ *denote a sequence of events on* $(\Omega,\mathcal{F},\mathbb{P})$*, such that*

$$\sum_{n\geqslant 1} \mathbb{P}(A_n) < \infty.$$

Then we have

$$\mathbb{P}\bigg(\bigcap_{n\geqslant 1}\bigcup_{k\geqslant n} A_k\bigg) = 0,$$

i.e. the probability that A_n *occurs infinitely many times occur is zero.*

A.3 Conditional Probabilities and Independence

We start with examples.

Consider a population $\Omega = M \cup W$ made of a set M of men and a set W of women. Here the σ-algebra $\mathcal{F} = \{\Omega, \emptyset, W, M\}$ corresponds to the information given by gender. After polling the population, *e.g.* for a market survey, it turns out that a proportion $p \in [0, 1]$ of the population declares to like apples, while a proportion $1 - p$ declares to dislike apples. Let $A \subset \Omega$ denote the subset of individuals who like apples, while $A^c \subset \Omega$ denotes the subset individuals who dislike apples, with

$$p = \mathbb{P}(A) \quad \text{and} \quad 1 - p = \mathbb{P}(A^c),$$

e.g. $p = 60\%$ of the population likes apples. It may be interesting to get a more precise information and to determine

- the relative proportion $\dfrac{\mathbb{P}(A \cap W)}{\mathbb{P}(W)}$ of women who like apples, and
- the relative proportion $\dfrac{\mathbb{P}(A \cap M)}{\mathbb{P}(M)}$ of men who like apples.

Here, $\mathbb{P}(A \cap W)/\mathbb{P}(W)$ represents the probability that a randomly chosen woman in W likes apples, and $\mathbb{P}(A \cap M)/\mathbb{P}(M)$ represents the probability that a randomly chosen man in M likes apples. Those two ratios are interpreted as *conditional probabilities*, for example $\mathbb{P}(A \cap M)/\mathbb{P}(M)$ denotes the probability that a given individual likes apples *given that* he is a man.

For another example, suppose that the population Ω is split as $\Omega = Y \cup O$ into a set Y of "young" people and another set O of "old" people, and denote by $A \subset \Omega$ the set of people who voted for candidate A in an election. Here it can be of interest to find out the relative proportion

$$\mathbb{P}(A \mid Y) = \frac{\mathbb{P}(Y \cap A)}{\mathbb{P}(Y)}$$

of young people who voted for candidate A.

Definition A.4. *Given any two events* $A, B \subset \Omega$ *with* $\mathbb{P}(B) \neq 0$, *we call*

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

the probability of A *given* B, *or* conditionally to B.

Remark A.5. *We note that if* $\mathbb{P}(B) = 1$ *we have* $\mathbb{P}(A \cap B^c) \leqslant \mathbb{P}(B^c) = 0$, *hence* $\mathbb{P}(A \cap B^c) = 0$, *which implies*

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

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

We also recall the following property:

821

$$\mathbb{P}\left(B \cap \bigcup_{n \geqslant 1} A_n\right) = \mathbb{P}\left(\bigcup_{n \geqslant 1}(B \cap A_n)\right)$$
$$= \sum_{n \geqslant 1} \mathbb{P}(B \cap A_n)$$
$$= \sum_{n \geqslant 1} \mathbb{P}(B \mid A_n)\mathbb{P}(A_n)$$
$$= \sum_{n \geqslant 1} \mathbb{P}(A_n \mid B)\mathbb{P}(B),$$

for any family of disjoint events $(A_n)_{n \geqslant 1}$ with $A_i \cap A_j = \emptyset$, $i \neq j$, and $\mathbb{P}(B) > 0$, $n \geqslant 1$. This also shows that conditional probability measures are probability measures, in the sense that whenever $\mathbb{P}(B) > 0$, we have

a) $\mathbb{P}(\Omega \mid B) = 1$, and

b) $\mathbb{P}\left(\bigcup_{n \geqslant 1} A_n \,\middle|\, B\right) = \sum_{n \geqslant 1} \mathbb{P}(A_n \mid B)$, whenever $A_k \cap A_l = \emptyset$, $k \neq l$.

In particular, if $\bigcup_{n \geqslant 1} A_n = \Omega$, $(A_n)_{n \geqslant 1}$ becomes a *partition* of Ω and we get the *law of total probability*

$$\mathbb{P}(B) = \sum_{n \geqslant 1} \mathbb{P}(B \cap A_n) = \sum_{n \geqslant 1} \mathbb{P}(A_n \mid B)\mathbb{P}(B) = \sum_{n \geqslant 1} \mathbb{P}(B \mid A_n)\mathbb{P}(A_n),$$
(A.9)

provided that $A_i \cap A_j = \emptyset$, $i \neq j$, and $\mathbb{P}(B) > 0$, $n \geqslant 1$.

Remark. In general we have

$$\mathbb{P}\left(A \,\middle|\, \bigcup_{n \geqslant 1} B_n\right) \neq \sum_{n \geqslant 1} \mathbb{P}(A \mid B_n),$$

even when $B_k \cap B_l = \emptyset$, $k \neq l$. Indeed, taking for example $A = \Omega = B_1 \cup B_2$ with $B_1 \cap B_2 = \emptyset$ and $\mathbb{P}(B_1) = \mathbb{P}(B_2) = 1/2$, we have

$$1 = \mathbb{P}(\Omega \mid B_1 \cup B_2) \neq \mathbb{P}(\Omega \mid B_1) + \mathbb{P}(\Omega \mid B_2) = 2.$$

Independent events

Definition A.6. *Two events A and B such that $\mathbb{P}(A), \mathbb{P}(B) > 0$ are said to be* independent *if*

$$\mathbb{P}(A \mid B) = \mathbb{P}(A). \tag{A.10}$$

We note that the independence condition (A.10) is equivalent to

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

A.4 Random Variables

A real-valued random variable is a mapping[*]

$$X : \Omega \longrightarrow \mathbb{R}$$
$$\omega \longmapsto X(\omega)$$

from a probability sample space Ω into the state space $\mathbb{R}$. Given

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

a random variable and a (measurable)[†] subset A of $\mathbb{R}$, we denote by $\{X \in A\}$ the event

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

Examples

i) Let $\Omega := \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$, and consider the mapping

$$X : \Omega \longrightarrow \mathbb{R}$$
$$(k,l) \longmapsto k+l.$$

Then X is a random variable giving the sum of the two numbers appearing on each die.

ii) the time needed everyday to travel from home to work or school is a random variable, as the precise value of this time may change from day to day under unexpected circumstances.

iii) the price of a risky asset can be modeled using a random variable.

In what follows, we will often use the notion of *indicator function* $\mathbb{1}_A$ of an event $A \subset \Omega$.

Definition A.7. *For any $A \subset \Omega$, the indicator function $\mathbb{1}_A$ is the random variable*

$$\mathbb{1}_A : \Omega \longrightarrow \{0,1\}$$
$$\omega \longmapsto \ \mathbb{1}_A(\omega)$$

---

[*] See (MOE and UCLES 2022a, page 14) lines 4-5 or (MOE and UCLES 2022b, page 19) lines 4-5.

[†] Measurability of subsets of $\mathbb{R}$ refers to *Borel measurability*, a concept which will not be defined in this text.

defined by

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

Indicator functions satisfy the property

$$\mathbb{1}_{A \cap B}(\omega) = \mathbb{1}_A(\omega)\mathbb{1}_B(\omega), \qquad (A.11)$$

since

$$\begin{aligned}
\mathbb{1}_{A \cap B}(\omega) = 1 &\iff \omega \in A \cap B \\
&\iff \omega \in A \text{ and } \omega \in B \\
&\iff \mathbb{1}_A(\omega) = 1 \text{ and } \mathbb{1}_B(\omega) = 1 \\
&\iff \mathbb{1}_A(\omega)\mathbb{1}_B(\omega) = 1.
\end{aligned}$$

We also have

$$\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A \cap B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A\mathbb{1}_B,$$

and

$$\mathbb{1}_{A \cup B} = \mathbb{1}_A + \mathbb{1}_B, \qquad (A.12)$$

if $A \cap B = \emptyset$.

For example, if $\Omega = \mathbb{N}$ and $A = \{k\}$, for all $l \geqslant 0$ we have

$$\mathbb{1}_{\{k\}}(l) = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

Given X a random variable, we also let

$$\mathbb{1}_{\{X=n\}} = \begin{cases} 1 & \text{if } X = n, \\ 0 & \text{if } X \neq n, \end{cases}$$

and

$$\mathbb{1}_{\{X<n\}} = \begin{cases} 1 & \text{if } X < n, \\ 0 & \text{if } X \geqslant n. \end{cases}$$

A.5 Probability Distributions

The *probability distribution* of a random variable $X : \Omega \longrightarrow \mathbb{R}$ is the collection

$$\{\mathbb{P}(X \in A) \; : \; A \text{ is a measurable subset of } \mathbb{R}\}.$$

As the collection of *measurable* subsets of $\mathbb{R}$ coincides with the σ-algebra generated by the intervals in $\mathbb{R}$, the distribution of X can be reduced to the knowledge of the probabilities

$$\{\mathbb{P}(a < X \leqslant b) = \mathbb{P}(X \leqslant b) - \mathbb{P}(X \leqslant a) \ : \ a < b \in \mathbb{R}\},$$

or of the cumulative distribution functions

$$\{\mathbb{P}(X \leqslant a) \ : \ a \in \mathbb{R}\}, \quad \text{or} \quad \{\mathbb{P}(X \geqslant a) \ : \ a \in \mathbb{R}\},$$

see *e.g.* Corollary 3.8 in Çınlar (2011).

Two random variables X and Y are said to be independent under the probability $\mathbb{P}$ if their probability distributions satisfy

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

for all (measurable) subsets A and B of $\mathbb{R}$.

Distributions admitting a density

We say that the distribution of X admits a probability *density* distribution function $\varphi_X : \mathbb{R} \longrightarrow \mathbb{R}_+$ if, for all $a \leqslant b$, the probability $\mathbb{P}(a \leqslant X \leqslant b)$ can be written as

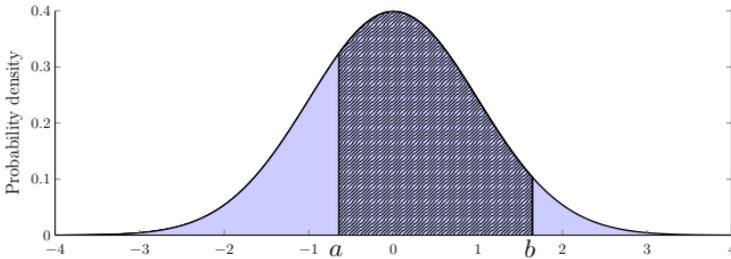$$\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b \varphi_X(x)dx.$$

Fig. A.1: Probability density function φ_X.

We also say that the distribution of X is absolutely continuous, or that X is an absolutely continuous random variable. This, however, does *not* imply that the density function $\varphi_X : \mathbb{R} \longrightarrow \mathbb{R}_+$ is continuous.

In particular, we always have

$$\int_{-\infty}^{\infty} \varphi_X(x)dx = \mathbb{P}(-\infty \leqslant X \leqslant \infty) = 1$$

for any probability density functions $\varphi_X : \mathbb{R} \longrightarrow \mathbb{R}_+$.

⟲   825

Remark A.8. *Note that if the distribution of X admits a probability density function φ_X, then for all $a \in \mathbb{R}$ we have*

$$\mathbb{P}(X = a) = \int_a^a \varphi_X(x)dx = 0, \tag{A.13}$$

and this is not a contradiction.

In particular, Remark A.8 shows that

$$\mathbb{P}(a \leqslant X \leqslant b) = \mathbb{P}(X = a) + \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a < X < b),$$

for $a \leqslant b$. Property (A.13) appears for example in the framework of lottery games with a large number of participants, in which a given number "a" selected in advance has a very low (almost zero) probability to be chosen.

The probability density function φ_X can be recovered from the Cumulative Distribution Functions (CDFs)

$$x \longmapsto F_X(x) := \mathbb{P}(X \leqslant x) = \int_{-\infty}^x \varphi_X(s)ds,$$

and

$$x \longmapsto 1 - F_X(x) = \mathbb{P}(X \geqslant x) = \int_x^\infty \varphi_X(s)ds,$$

as

$$\varphi_X(x) = \frac{\partial F_X}{\partial x}(x) = \frac{\partial}{\partial x}\int_{-\infty}^x \varphi_X(s)ds = -\frac{\partial}{\partial x}\int_x^\infty \varphi_X(s)ds, \qquad x \in \mathbb{R}.$$

Examples

i) The *uniform* distribution on an interval.

The probability density function of the uniform distribution on the interval $[a,b]$, $a < b$, is given by

$$\varphi(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x), \qquad x \in \mathbb{R}.$$

ii) The *Gaussian* distribution.

The probability density function of the standard normal distribution is given by

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}, \qquad x \in \mathbb{R}.$$

More generally, the probability density function of the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is given by

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

In this case, we write $X \simeq \mathcal{N}(\mu, \sigma^2)$.

iii) The *exponential* distribution.

The probability density function of the exponential distribution with parameter $\lambda > 0$ is given by

$$\varphi(x) := \lambda \mathbb{1}_{[0,\infty)}(x) \, e^{-\lambda x} = \begin{cases} \lambda \, e^{-\lambda x}, & x \geqslant 0 \\ 0, & x < 0. \end{cases} \qquad (A.14)$$

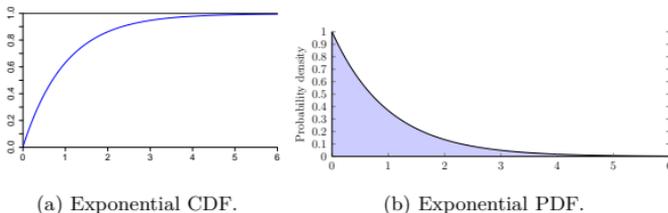
(a) Exponential CDF. (b) Exponential PDF.

Fig. A.2: Exponential CDF and PDF.

We also have

$$\mathbb{P}(X > t) = e^{-\lambda t}, \qquad t \geqslant 0. \qquad (A.15)$$

iv) The *gamma* distribution.

The probability density function of the gamma distribution is given by

$$\varphi(x) := \frac{a^\lambda}{\Gamma(\lambda)} \mathbb{1}_{[0,\infty)}(x) x^{\lambda-1} \, e^{-ax} = \begin{cases} \dfrac{a^\lambda}{\Gamma(\lambda)} x^{\lambda-1} \, e^{-ax}, & x \geqslant 0 \\ 0, & x < 0, \end{cases}$$

where $a > 0$ and $\lambda > 0$ are scale and shape parameters, and

$$\Gamma(\lambda) := \int_0^\infty x^{\lambda-1} \, e^{-x} dx, \qquad \lambda > 0,$$

is the gamma function.

v) The *Cauchy* distribution.

The probability density function of the Cauchy distribution is given by

$$\varphi(x) := \frac{1}{\pi(1+x^2)}, \qquad x \in \mathbb{R}.$$

vi) The *lognormal* distribution.

↺

827

The probability density function of the lognormal distribution is given by

$$\varphi(x) := \mathbb{1}_{[0,\infty)}(x)\frac{1}{x\sigma\sqrt{2\pi}}\,\mathrm{e}^{-(\mu-\log x)^2/(2\sigma^2)} = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}}\,\mathrm{e}^{-(\mu-\log x)^2/(2\sigma^2)}, & x \geqslant 0 \\ 0, & x < 0. \end{cases}$$

Exercise: For each of the above probability density functions φ, check that the condition

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1$$

is satisfied.

Joint densities

Given two absolutely continuous random variables $X : \Omega \longrightarrow \mathbb{R}$ and $Y : \Omega \longrightarrow \mathbb{R}$, we can form the $\mathbb{R}^2$-valued random variable (X,Y) defined by

$$(X,Y) : \Omega \longrightarrow \mathbb{R}^2$$
$$\omega \longmapsto (X(\omega), Y(\omega)).$$

We say that (X,Y) admits a joint probability density

$$\varphi_{(X,Y)} : \mathbb{R}^2 \longrightarrow \mathbb{R}_+$$

when

$$\mathbb{P}((X,Y) \in A \times B) = \mathbb{P}(X \in A \text{ and } Y \in B) = \int_B \int_A \varphi_{(X,Y)}(x,y)dxdy$$

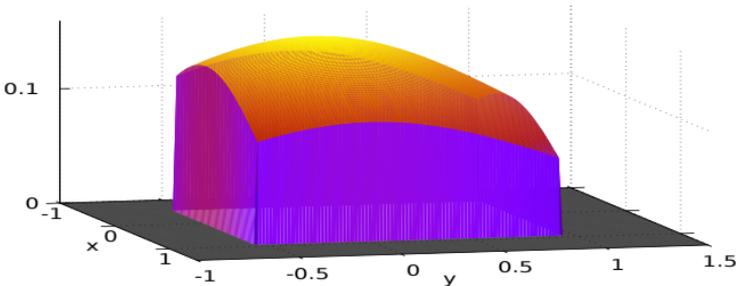for all *measurable* subsets A, B of $\mathbb{R}$, see Figure A.3.

Fig. A.3: Probability $\mathbb{P}((X,Y) \in [-0.5,1] \times [-0.5,1])$ computed as a volume integral.

The probability density function $\varphi_{(X,Y)}$ can be recovered from the joint cumulative distribution function

$$(x,y) \longmapsto F_{(X,Y)}(x,y) := \mathbb{P}(X \leqslant x \text{ and } Y \leqslant y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \varphi_{(X,Y)}(s,t)dsdt,$$

and

$$(x,y) \longmapsto \mathbb{P}(X \geqslant x \text{ and } Y \geqslant y) = \int_{x}^{\infty} \int_{y}^{\infty} \varphi_{(X,Y)}(s,t)dsdt,$$

as

$$\varphi_{(X,Y)}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x,y) \tag{A.16}$$

$$= \frac{\partial^2}{\partial x \partial y} \int_{-\infty}^{x} \int_{-\infty}^{y} \varphi_{(X,Y)}(s,t)dsdt \tag{A.17}$$

$$= \frac{\partial^2}{\partial x \partial y} \int_{x}^{\infty} \int_{y}^{\infty} \varphi_{(X,Y)}(s,t)dsdt,$$

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

The probability densities $\varphi_X : \mathbb{R} \longrightarrow \mathbb{R}_+$ and $\varphi_Y : \mathbb{R} \longrightarrow \mathbb{R}_+$ of $X : \Omega \longrightarrow \mathbb{R}$ and $Y : \Omega \longrightarrow \mathbb{R}$ are called the *marginal densities* of (X,Y), and are given by

$$\varphi_X(x) = \int_{-\infty}^{\infty} \varphi_{(X,Y)}(x,y)dy, \qquad x \in \mathbb{R}, \tag{A.18}$$

and

$$\varphi_Y(y) = \int_{-\infty}^{\infty} \varphi_{(X,Y)}(x,y)dx, \qquad y \in \mathbb{R}.$$

The conditional probability density $\varphi_{X|Y=y} : \mathbb{R} \longrightarrow \mathbb{R}_+$ of X given $Y = y$ is defined by

$$\varphi_{X|Y=y}(x) := \frac{\varphi_{(X,Y)}(x,y)}{\varphi_Y(y)}, \qquad x,y \in \mathbb{R}, \tag{A.19}$$

provided that $\varphi_Y(y) > 0$. In particular, X and Y are independent if and only if

$$\varphi_{X|Y=y}(x) = \varphi_X(x), \quad i.e., \quad \varphi_{(X,Y)}(x,y) = \varphi_X(x)\varphi_Y(y), \qquad x,y \in \mathbb{R}.$$

Example

If $X_1, \ldots, X_n$ are independent exponentially distributed random variables with parameters $\lambda_1, \ldots, \lambda_n$ we have

$$\begin{aligned} \mathbb{P}(\min(X_1, \ldots, X_n) > t) &= \mathbb{P}(X_1 > t, \ldots, X_n > t) \\ &= \mathbb{P}(X_1 > t) \cdots \mathbb{P}(X_n > t) \\ &= \mathrm{e}^{-(\lambda_1 + \cdots + \lambda_n)t}, \qquad t \geqslant 0, \tag{A.20} \end{aligned}$$

↻

hence $\min(X_1, \ldots, X_n)$ is an exponentially distributed random variable with parameter $\lambda_1 + \cdots + \lambda_n$.

From the joint probability density function of (X_1, X_2) given by

$$\varphi_{(X_1, X_2)}(x, y) = \varphi_{X_1}(x)\varphi_{X_2}(y) = \lambda_1 \lambda_2 \, e^{-\lambda_1 x - \lambda_2 y}, \qquad x, y \geqslant 0,$$

we can write

$$\begin{aligned}
\mathbb{P}(X_1 < X_2) &= \mathbb{P}(X_1 \leqslant X_2) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{y} \varphi_{(X_1, X_2)}(x, y) dx dy \\
&= \lambda_1 \lambda_2 \int_{0}^{\infty} \int_{0}^{y} e^{-\lambda_1 x - \lambda_2 y} dx dy \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2},
\end{aligned} \tag{A.21}$$

and we note that

$$\mathbb{P}(X_1 = X_2) = \lambda_1 \lambda_2 \int_{\{(x,y) \in \mathbb{R}_+^2 \; : \; x = y\}} e^{-\lambda_1 x - \lambda_2 y} dx dy = 0.$$

Discrete distributions

We only consider integer-valued random variables, *i.e.* the distribution of X is given by the values of $\mathbb{P}(X = k)$, $k \geqslant 0$.

Examples

i) The *Bernoulli* distribution.

We have

$$\mathbb{P}(X = 1) = p \qquad \text{and} \qquad \mathbb{P}(X = 0) = 1 - p, \tag{A.22}$$

where $p \in [0, 1]$ is a parameter.

Note that any Bernoulli random variable $X : \Omega \longrightarrow \{0, 1\}$ can be written as the indicator function

$$X = \mathbb{1}_A$$

on Ω with $A = \{X = 1\} = \{\omega \in \Omega \; : \; X(\omega) = 1\}$.

ii) The *binomial* distribution.

We have

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

where $n \geqslant 1$ and $p \in [0, 1]$ are parameters and $\binom{n}{k} = n!/(k!(n-k)!)$, $0 \leqslant k \leqslant n$.

iii) The *geometric* distribution.

In this case, we have

$$\mathbb{P}(X = k) = (1-p)p^k, \qquad k \geqslant 0, \qquad (A.23)$$

where $p \in (0,1)$ is a parameter. For example, if $(X_k)_{k \in \mathbb{N}}$ is a sequence of independent Bernoulli random variables with distribution (A.22), then the random variable,[*]

$$T_0 := \inf\{k \geqslant 0 \ : \ X_k = 0\}$$

can denote the duration of a game until the time that the wealth X_k of a player reaches 0. The random variable T_0 has the geometric distribution (A.23) with parameter $p \in (0,1)$.

iv) The *negative binomial* (or *Pascal*) distribution.

We have

$$\mathbb{P}(X = k) = \binom{k + r - 1}{r - 1}(1-p)^r p^k, \qquad k \geqslant 0, \qquad (A.24)$$

where $p \in (0,1)$ and $r \geqslant 1$ are parameters. Note that the sum of $r \geqslant 1$ independent geometric random variables with parameter p has a negative binomial distribution with parameter (r, p). In particular, the negative binomial distribution recovers the geometric distribution when $r = 1$.

v) The *Poisson* distribution.

We have

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}, \qquad k \geqslant 0,$$

where $\lambda > 0$ is a parameter.

The probability that a discrete nonnegative random variable $X : \Omega \longrightarrow \mathbb{N} \cup \{+\infty\}$ is finite is given by

$$\mathbb{P}(X < \infty) = \sum_{k \geqslant 0} \mathbb{P}(X = k), \qquad (A.25)$$

and we have

$$1 = \mathbb{P}(X = \infty) + \mathbb{P}(X < \infty) = \mathbb{P}(X = \infty) + \sum_{k \geqslant 0} \mathbb{P}(X = k).$$

Remark A.9. *The distribution of a discrete random variable cannot admit a probability density. If this were the case, by Remark A.8 we would have*

---

[*] The notation "inf" stands for "infimum", meaning the smallest $n \geqslant 0$ such that $X_n = 0$, if such an n exists.

$\mathbb{P}(X = k) = 0$ *for all* $k \geqslant 0$ *and*

$$1 = \mathbb{P}(X \in \mathbb{R}) = \mathbb{P}(X \in \mathbb{N}) = \sum_{k \geqslant 0} \mathbb{P}(X = k) = 0,$$

which is a contradiction.

Given two discrete random variables X and Y, the conditional distribution of X given $Y = k$ is given by

$$\mathbb{P}(X = n \mid Y = k) = \frac{\mathbb{P}(X = n \text{ and } Y = k)}{\mathbb{P}(Y = k)}, \qquad n \geqslant 0,$$

provided that $\mathbb{P}(Y = k) > 0$, $k \geqslant 0$.

A.6 Expectation of Random Variables

The *expectation*, or *expected value*, of a random variable X is the mean, or average value, of X. In practice, expectations can be even more useful than probabilities. For example, knowing that a given equipment (such as a bridge) has a failure probability of 1.78493 out of a billion can be of less practical use than knowing the expected lifetime (*e.g.* 200000 years) of that equipment.

For example, the time $T(\omega)$ to travel from home to work/school can be a random variable with a new outcome and value every day, however we usually refer to its expectation $\mathbb{E}[T]$ rather than to its sample values that may change from day to day.

Expected value of a Bernoulli random variable

Any Bernoulli random variable $X : \Omega \longrightarrow \{0,1\}$ can be written as the indicator function $X := \mathbb{1}_A$ where A is the event $A = \{X = 1\}$, and the parameter $p \in [0,1]$ of X is given by

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

The expectation of a Bernoulli random variable with parameter p is defined as

$$\mathbb{E}[\mathbb{1}_A] := 1 \times \mathbb{P}(A) + 0 \times \mathbb{P}(A^c) = \mathbb{P}(A). \quad (A.26)$$

Expected value of a discrete random variable

Next, let $X : \Omega \longrightarrow \mathbb{N}$ be a discrete random variable. The expectation $\mathbb{E}[X]$ of X is defined as the sum

$$\mathbb{E}[X] = \sum_{k \geqslant 0} k \mathbb{P}(X = k), \tag{A.27}$$

in which the possible values $k \geqslant 0$ of X are weighted by their probabilities. More generally we have

$$\mathbb{E}[\phi(X)] = \sum_{k \geqslant 0} \phi(k) \mathbb{P}(X = k),$$

for all sufficiently summable functions $\phi : \mathbb{N} \longrightarrow \mathbb{R}$.[*]

The expectation of the indicator function $X = \mathbb{1}_A = \mathbb{1}_{\{X=1\}}$ can be recovered from (A.27) as

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{1}_A] = 0 \times \mathbb{P}(\Omega \setminus A) + 1 \times \mathbb{P}(A) = 0 + \mathbb{P}(A) = \mathbb{P}(A).$$

Note that the expectation is a *linear* operation, *i.e.* we have

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y], \qquad a, b \in \mathbb{R}, \tag{A.28}$$

provided that

$$\mathbb{E}[|X|] + \mathbb{E}[|Y|] < \infty.$$

Examples

i) Expected value of a Poisson random variable with parameter $\lambda > 0$:

$$\mathbb{E}[X] = \sum_{k \geqslant 0} k \mathbb{P}(X = k) = \mathrm{e}^{-\lambda} \sum_{k \geqslant 1} k \frac{\lambda^k}{k!} = \lambda \, \mathrm{e}^{-\lambda} \sum_{k \geqslant 0} \frac{\lambda^k}{k!} = \lambda. \tag{A.29}$$

ii) Estimating the expected value of a Poisson random variable using **R**:

Taking $\lambda := 2$, we can use the following **R** code:

```
1  poisson_samples <- rpois(100000, lambda = 2)
2  poisson_samples; mean(poisson_samples)
```
Listing A.1: **R** code - Mean of Poisson samples.

```
1  import numpy as np
2  poisson_samples = np.random.poisson(lam=2, size=100000)
   print(poisson_samples); print(poisson_samples.mean())
```
Listing A.2: Python code - Mean of Poisson samples.

_____
[*] This is the "law of the unconscious statistician" (LOTUS).

Given $X : \Omega \longrightarrow \mathbb{N} \cup \{+\infty\}$ a discrete nonnegative random variable X, we have

$$\mathbb{P}(X < \infty) = \sum_{k \geqslant 0} \mathbb{P}(X = k),$$

and

$$1 = \mathbb{P}(X = \infty) + \mathbb{P}(X < \infty) = \mathbb{P}(X = \infty) + \sum_{k \geqslant 0} \mathbb{P}(X = k),$$

and in general

$$\mathbb{E}[X] = +\infty \times \mathbb{P}(X = \infty) + \sum_{k \geqslant 0} k \mathbb{P}(X = k).$$

In particular, $\mathbb{P}(X = \infty) > 0$ implies $\mathbb{E}[X] = \infty$, and the finiteness condition $\mathbb{E}[X] < \infty$ implies $\mathbb{P}(X < \infty) = 1$, however the converse is *not true*. For example, assume that X has the geometric distribution

$$\mathbb{P}(X = k) := \frac{1}{2^{k+1}}, \qquad k \geqslant 0, \tag{A.30}$$

with parameter $p = 1/2$, and

$$\mathbb{E}[X] = \sum_{k \geqslant 0} \frac{k}{2^{k+1}} = \frac{1}{4} \sum_{k \geqslant 1} \frac{k}{2^{k-1}} = \frac{1}{4} \frac{1}{(1 - 1/2)^2} = 1 < \infty.$$

Letting $\phi(X) := 2^X$, we have

$$\mathbb{P}(\phi(X) < \infty) = \mathbb{P}(X < \infty) = \sum_{k \geqslant 0} \frac{1}{2^{k+1}} = 1,$$

and

$$\mathbb{E}[\phi(X)] = \sum_{k \geqslant 0} \phi(k) \mathbb{P}(X = k) = \sum_{k \geqslant 0} \frac{2^k}{2^{k+1}} = \sum_{k \geqslant 0} \frac{1}{2} = +\infty,$$

hence the expectation $\mathbb{E}[\phi(X)]$ is *infinite* although $\phi(X)$ is *finite* with probability one.[*]

Conditional expectation

The notion of expectation takes its full meaning under conditioning. For example, the expected return of a random asset usually depends on information such as economic data, location, etc. In this case, replacing the expectation

---

[*] This is the St. Petersburg paradox.

by a conditional expectation will provide a better estimate of the expected value.

For instance, life expectancy is a natural example of a conditional expectation since it typically depends on location, gender, and other parameters.

The *conditional expectation* of a finite discrete random variable $X : \Omega \longrightarrow \mathbb{N}$ given an event A is defined by

$$\mathbb{E}[X \mid A] = \sum_{k \geqslant 0} k \mathbb{P}(X = k \mid A) = \sum_{k \geqslant 1} k \frac{\mathbb{P}(X = k \text{ and } A)}{\mathbb{P}(A)}.$$

Lemma A.10. *Given an event A such that $\mathbb{P}(A) > 0$, we have*

$$\mathbb{E}[X \mid A] = \frac{1}{\mathbb{P}(A)} \mathbb{E}[X \mathbb{1}_A]. \tag{A.31}$$

Proof. The proof is done only for $X : \Omega \longrightarrow \mathbb{N}$ a discrete random variable, however (A.31) is valid for general real-valued random variables. By Relation (A.11) we have

$$
\begin{aligned}
\mathbb{E}[X \mid A] &= \sum_{k \geqslant 0} k \mathbb{P}(X = k \mid A) \\
&= \frac{1}{\mathbb{P}(A)} \sum_{k \geqslant 0} k \mathbb{P}(\{X = k\} \cap A) = \frac{1}{\mathbb{P}(A)} \sum_{k \geqslant 0} k \mathbb{E}\big[\mathbb{1}_{\{X=k\} \cap A}\big] \\
&= \frac{1}{\mathbb{P}(A)} \sum_{k \geqslant 0} k \mathbb{E}\big[\mathbb{1}_{\{X=k\}} \mathbb{1}_A\big] = \frac{1}{\mathbb{P}(A)} \mathbb{E}\left[\mathbb{1}_A \sum_{k \geqslant 0} k \mathbb{1}_{\{X=k\}}\right] \\
&= \frac{1}{\mathbb{P}(A)} \mathbb{E}\big[\mathbb{1}_A X\big],
\end{aligned}
$$

where we used the relation

$$X = \sum_{k \geqslant 0} k \mathbb{1}_{\{X=k\}}$$

which holds since X takes only integer values. $\qquad \square$

Example

i) For example, consider $\Omega = \{1, 3, -1, -2, 5, 7\}$ with the non-uniform probability measure given by

$$\mathbb{P}(\{-1\}) = \mathbb{P}(\{-2\}) = \mathbb{P}(\{1\}) = \mathbb{P}(\{3\}) = \frac{1}{7}, \ \mathbb{P}(\{5\}) = \frac{2}{7}, \ \mathbb{P}(\{7\}) = \frac{1}{7},$$

and the random variable

$$X : \Omega \longrightarrow \mathbb{Z}$$

given by

$$X(k) = k, \qquad k = 1, 3, -1, -2, 5, 7.$$

Here, $\mathbb{E}[X \mid X > 1]$ denotes the expected value of X given

$$A = \{X > 1\} = \{3, 5, 7\} \subset \Omega,$$

i.e. the mean value of X given that X is strictly positive. This conditional expectation can be computed as

$$
\begin{aligned}
&\mathbb{E}[X \mid X > 1] \\
&= 3 \times \mathbb{P}(X = 3 \mid X > 1) + 5 \times \mathbb{P}(X = 5 \mid X > 1) + 7 \times \mathbb{P}(X = 7 \mid X > 1) \\
&= \frac{3 + 2 \times 5 + 7}{4} \\
&= \frac{3 + 5 + 5 + 7}{7 \times 4/7} \\
&= \frac{1}{\mathbb{P}(X > 1)} \mathbb{E}\big[X \mathbb{1}_{\{X>1\}}\big],
\end{aligned}
$$

where $\mathbb{P}(X > 1) = 4/7$ and the truncated expectation $\mathbb{E}\big[X \mathbb{1}_{\{X>1\}}\big]$ is given by $\mathbb{E}\big[X \mathbb{1}_{\{X>1\}}\big] = (3 + 2 \times 5 + 7)/7$.

ii) Estimating a conditional expectation using R:

```
1  geo_samples <- rgeom(100000, prob = 1/4)
   mean(geo_samples)
3  mean(geo_samples[geo_samples<10])
```

Listing A.3: R code - Conditional geometric distribution mean.

```
1  import numpy as np
   geo_samples = np.random.geometric(p=1/4, size=100000)
3  print(geo_samples.mean())
   print(geo_samples[geo_samples < 10].mean())
```

Listing A.4: Python code - Conditional geometric distribution mean.

Taking $p := 3/4$, we have

$$\mathbb{E}[X] = (1 - p) \sum_{k \geqslant 1} k p^k = \frac{p}{1 - p} = 3,$$

and

$$
\begin{aligned}
\mathbb{E}[X \mid X < 10] &= \frac{1}{\mathbb{P}(X < 10)} \mathbb{E}\big[X \mathbb{1}_{\{X<10\}}\big] \\
&= \frac{1}{\mathbb{P}(X < 10)} \sum_{k=0}^{9} k \mathbb{P}(X = k)
\end{aligned}
$$

$$= \frac{1}{\displaystyle\sum_{k=0}^{9} p^k} \sum_{k=1}^{9} kp^k$$

$$= \frac{p(1-p)}{1-p^{10}} \frac{\partial}{\partial p} \sum_{k=0}^{9} p^k$$

$$= \frac{p(1-p)}{1-p^{10}} \frac{\partial}{\partial p} \left(\frac{1-p^{10}}{1-p} \right)$$

$$= \frac{p(1-p^{10}-10(1-p)p^9)}{(1-p)(1-p^{10})}$$

$$\simeq 2.4032603455.$$

If the random variable $X : \Omega \longrightarrow \mathbb{N}$ is independent* of the event A, we have

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

and we naturally find

$$\mathbb{E}[X \mid A] = \mathbb{E}[X]. \tag{A.32}$$

Taking $X = \mathbb{1}_A$ with

$$\mathbb{1}_A : \Omega \longrightarrow \{0,1\}$$
$$\omega \longmapsto \mathbb{1}_A := \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A, \end{cases}$$

shows that, in particular,

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

One can also define the conditional expectation of X given $A = \{Y = k\}$, as

$$\mathbb{E}[X \mid Y = k] = \sum_{n \geqslant 0} n\mathbb{P}(X = n \mid Y = k),$$

where $Y : \Omega \longrightarrow \mathbb{N}$ is a discrete random variable.

Proposition A.11. *Given X a discrete random variable such that $\mathbb{E}[|X|] < \infty$, we have the relation*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]], \tag{A.33}$$

---

* *i.e.,* $\mathbb{P}(\{X = k\} \cap A) = \mathbb{P}(\{X = k\})\mathbb{P}(A)$ for all $k \geqslant 0$.

↺

which is sometimes referred to as the tower property.

Proof. We have

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \sum_{k \geqslant 0} \mathbb{E}[X \mid Y = k]\mathbb{P}(Y = k) \\
&= \sum_{k \geqslant 0} \sum_{n \geqslant 0} n\mathbb{P}(X = n \mid Y = k)\mathbb{P}(Y = k) \\
&= \sum_{n \geqslant 0} n \sum_{k \geqslant 0} \mathbb{P}(X = n \text{ and } Y = k) \\
&= \sum_{n \geqslant 0} n\mathbb{P}(X = n) = \mathbb{E}[X],
\end{aligned}
$$

where we used the marginal distribution

$$
\mathbb{P}(X = n) = \sum_{k \geqslant 0} \mathbb{P}(X = n \text{ and } Y = k), \qquad n \geqslant 0,
$$

that follows from the *law of total probability* (A.9) with $A_k = \{Y = k\}$, $k \geqslant 0$.
□

Taking

$$
Y = \sum_{k \geqslant 0} k \mathbb{1}_{A_k},
$$

with $A_k := \{Y = k\}$, $k \geqslant 0$, from (A.33) we also get the *law of total expectation*

$$
\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X \mid Y]] && (A.34) \\
&= \sum_{k \geqslant 0} \mathbb{E}[X \mid Y = k]\mathbb{P}(Y = k) \\
&= \sum_{k \geqslant 0} \mathbb{E}[X \mid A_k]\mathbb{P}(A_k).
\end{aligned}
$$

Example

Life expectancy in Singapore is $\mathbb{E}[T] = 80$ years overall, where T denotes the lifetime of a given individual chosen at random. Let $G \in \{m, w\}$ denote the gender of that individual. The statistics show that

$$
\mathbb{E}[T \mid G = m] = 78 \quad \text{and} \quad \mathbb{E}[T \mid G = w] = 81.9,
$$

and we have

$$
\begin{aligned}
80 &= \mathbb{E}[T] \\
&= \mathbb{E}[\mathbb{E}[T|G]]
\end{aligned}
$$

$$= \mathbb{P}(G = w)\mathbb{E}[T \mid G = w] + \mathbb{P}(G = m)\mathbb{E}[T \mid G = m]$$
$$= 81.9 \times \mathbb{P}(G = w) + 78 \times \mathbb{P}(G = m)$$
$$= 81.9 \times (1 - \mathbb{P}(G = m)) + 78 \times \mathbb{P}(G = m),$$

showing that

$$80 = 81.9 \times (1 - \mathbb{P}(G = m)) + 78 \times \mathbb{P}(G = m),$$

i.e.

$$\mathbb{P}(G = m) = \frac{81.9 - 80}{81.9 - 78} = \frac{1.9}{3.9} = 0.487.$$

Variance

The *variance* of a random variable X is defined by

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

provided that $\mathbb{E}\big[|X|^2\big] < \infty$. If $(X_k)_{k=1,\ldots,n}$ is a sequence of independent random variables, we have

$$\mathrm{Var}\left[\sum_{k=1}^n X_k\right] = \mathbb{E}\left[\left(\sum_{k=1}^n X_k\right)^2\right] - \left(\mathbb{E}\left[\sum_{k=1}^n X_k\right]\right)^2$$
$$= \mathbb{E}\left[\sum_{k=1}^n X_k \sum_{l=1}^n X_l\right] - \mathbb{E}\left[\sum_{k=1}^n X_k\right]\mathbb{E}\left[\sum_{l=1}^n X_l\right]$$
$$= \mathbb{E}\left[\sum_{k=1}^n \sum_{l=1}^n X_k X_l\right] - \sum_{k=1}^n \sum_{l=1}^n \mathbb{E}[X_k]\mathbb{E}[X_l]$$
$$= \sum_{k=1}^n \mathbb{E}\big[X_k^2\big] + \sum_{1 \leqslant k \neq l \leqslant n} \mathbb{E}[X_k X_l] - \sum_{k=1}^n (\mathbb{E}[X_k])^2 - \sum_{1 \leqslant k \neq l \leqslant n} \mathbb{E}[X_k]\mathbb{E}[X_l]$$
$$= \sum_{k=1}^n \big(\mathbb{E}\big[X_k^2\big] - (\mathbb{E}[X_k])^2\big)$$
$$= \sum_{k=1}^n \mathrm{Var}[X_k]. \tag{A.35}$$

Random sums

In what follows, we consider $Y : \Omega \longrightarrow \mathbb{N}$ an *a.s.* finite, integer-valued random variable, *i.e.* we have $\mathbb{P}(Y < \infty) = 1$ and $\mathbb{P}(Y = \infty) = 0$. Based on the tower property of conditional expectations (A.33) or ordinary conditioning,

the expectation of a random sum $\sum_{k=1}^{Y} X_k$, where $(X_k)_{k\in\mathbb{N}}$ is a sequence of random variables, can be computed from the *tower property* (A.33) or from the *law of total expectation* (A.34) as

$$\mathbb{E}\left[\sum_{k=1}^{Y} X_k\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{Y} X_k \,\Big|\, Y\right]\right]$$

$$= \sum_{n\geqslant 0} \mathbb{E}\left[\sum_{k=1}^{Y} X_k \,\Big|\, Y=n\right]\mathbb{P}(Y=n)$$

$$= \sum_{n\geqslant 0} \mathbb{E}\left[\sum_{k=1}^{n} X_k \,\Big|\, Y=n\right]\mathbb{P}(Y=n),$$

and if the sequence $(X_k)_{k\in\mathbb{N}}$ is (mutually) independent of Y, this yields

$$\mathbb{E}\left[\sum_{k=1}^{Y} X_k\right] = \sum_{n\geqslant 0} \mathbb{E}\left[\sum_{k=1}^{n} X_k\right]\mathbb{P}(Y=n)$$

$$= \sum_{n\geqslant 0} \mathbb{P}(Y=n)\sum_{k=1}^{n}\mathbb{E}[X_k].$$

Random products

Similarly, for a random product we will have, using the independence of Y with $(X_k)_{k\in\mathbb{N}}$,

$$\mathbb{E}\left[\prod_{k=1}^{Y} X_k\right] = \sum_{n\geqslant 0} \mathbb{E}\left[\prod_{k=1}^{n} X_k\right]\mathbb{P}(Y=n) \tag{A.36}$$

$$= \sum_{n\geqslant 0} \mathbb{P}(Y=n)\prod_{k=1}^{n}\mathbb{E}[X_k],$$

where the last equality requires the (mutual) independence of the random variables in the sequence $(X_k)_{k\geqslant 1}$.

Distributions admitting a density

Given a random variable X whose distribution admits a probability density $\varphi_X : \mathbb{R} \longrightarrow \mathbb{R}_+$ we have

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

and more generally,

$$\mathbb{E}[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)\varphi_X(x)dx, \tag{A.37}$$

for all sufficiently integrable function ϕ on $\mathbb{R}$. For example, if X has a standard normal distribution we have

$$\mathbb{E}[\phi(X)] = \int_{-\infty}^{\infty} \phi(x)\,\mathrm{e}^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

Examples

a) In case X has a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, we have

$$\mathbb{E}[\phi(X)] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \phi(x)\,\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}dx. \tag{A.38}$$

b) The uniform random variable U on $[0,1]$ satisfies $\mathbb{E}[U] = 1/2 < \infty$ and

$$\mathbb{P}(1/U < \infty) = \mathbb{P}(U > 0) = \mathbb{P}(U \in (0,1]) = 1,$$

however we have

$$\mathbb{E}[1/U] = \int_0^1 \frac{dx}{x} = +\infty,$$

and $\mathbb{P}(1/U = +\infty) = \mathbb{P}(U = 0) = 0$.

c) If the random variable X has an exponential distribution with parameter $\mu > 0$ we have

$$\mathbb{E}\big[\mathrm{e}^{\lambda X}\big] = \mu \int_0^{\infty} \mathrm{e}^{\lambda x}\,\mathrm{e}^{-\mu x}dx = \begin{cases} \dfrac{\mu}{\mu - \lambda} < \infty & \text{if } \mu > \lambda, \\ +\infty, & \text{if } \mu \leqslant \lambda. \end{cases}$$

<u>Exercise</u>: In case $X \simeq \mathcal{N}(\mu, \sigma^2)$ has a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, check that

$$\mu = \mathbb{E}[X] \qquad \text{and} \qquad \sigma^2 = \mathbb{E}\big[X^2\big] - (\mathbb{E}[X])^2.$$

When $(X,Y) : \Omega \longrightarrow \mathbb{R}^2$ is a $\mathbb{R}^2$-valued couple of random variables whose distribution admits a probability density $\varphi_{X,Y} : \mathbb{R}^2 \longrightarrow \mathbb{R}_+$ we have

$$\mathbb{E}[\phi(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x,y)\varphi_{X,Y}(x,y)dxdy,$$

for all sufficiently integrable function ϕ on $\mathbb{R}^2$.

The expectation of an absolutely continuous random variable satisfies the same linearity property (A.28) as in the discrete case.

The conditional expectation of an absolutely continuous random variable can be defined as

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \varphi_{X|Y=y}(x) dx$$

where the conditional probability density $\varphi_{X|Y=y}(x)$ is defined in (A.19), with the relation

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] \tag{A.39}$$

which is called the *tower property* and holds as in the discrete case, since

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] \varphi_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \varphi_{X|Y=y}(x) \varphi_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} \varphi_{(X,Y)}(x, y) dy dx \\
&= \int_{-\infty}^{\infty} x \varphi_X(x) dx = \mathbb{E}[X],
\end{aligned}
$$

where we used Relation (A.18) between the probability density of (X, Y) and its marginal X.

For example, an exponentially distributed random variable X with probability density function (A.14) has the expected value

$$\mathbb{E}[X] = \lambda \int_0^{\infty} x \, e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Proposition A.12. *(Fatou's lemma). Let* $(F_n)_{n \in \mathbb{N}}$ *be a sequence of* non-negative *random variables. We have*

$$\mathbb{E}\Big[\liminf_{n \to \infty} F_n\Big] \leqslant \liminf_{n \to \infty} \mathbb{E}[F_n].$$

In particular, Fatou's lemma shows that if in addition the sequence $(F_n)_{n \in \mathbb{N}}$ converges with probability one and the sequence $(\mathbb{E}[F_n])_{n \in \mathbb{N}}$ converges in $\mathbb{R}$ then we have

$$\mathbb{E}\Big[\lim_{n \to \infty} F_n\Big] \leqslant \lim_{n \to \infty} \mathbb{E}[F_n].$$

Moment Generating Functions

Characteristic functions

The *characteristic function* of a random variable X is the function

$$\Psi_X : \mathbb{R} \longrightarrow \mathbb{C}$$

defined by

$$\Psi_X(t) = \mathbb{E}\big[e^{itX}\big], \qquad t \in \mathbb{R}.$$

The characteristic function Ψ_X of a random variable X with probability density function $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ satisfies

$$\Psi_X(t) = \int_{-\infty}^{\infty} e^{ixt} \varphi(x) dx, \qquad t \in \mathbb{R}.$$

On the other hand, if $X : \Omega \longrightarrow \mathbb{N}$ is a discrete random variable we have

$$\Psi_X(t) = \sum_{n \geqslant 0} e^{itn} \mathbb{P}(X = n), \qquad t \in \mathbb{R}.$$

One of the main applications of characteristic functions is to provide a characterization of probability distributions, as in the following theorem.

Theorem A.13. *Two random variables $X : \Omega \longrightarrow \mathbb{R}$ and $Y : \Omega \longrightarrow \mathbb{R}$ have same distribution if and only if*

$$\Psi_X(t) = \Psi_Y(t), \qquad t \in \mathbb{R}.$$

Theorem A.13 is used to identify or to determine the probability distribution of a random variable X, by comparison with the characteristic function Ψ_Y of a random variable Y whose distribution is known.

The characteristic function of a random vector (X, Y) is the function $\Psi_{X,Y} : \mathbb{R}^2 \longrightarrow \mathbb{C}$ defined by

$$\Psi_{X,Y}(s,t) = \mathbb{E}\big[e^{isX+itY}\big], \qquad s, t \in \mathbb{R}.$$

Theorem A.14. *The random variables $X : \Omega \longrightarrow \mathbb{R}$ and $Y : \Omega \longrightarrow \mathbb{R}$ are* independent *if and only if*

$$\Psi_{X,Y}(s,t) = \Psi_X(s)\Psi_Y(t), \qquad s, t \in \mathbb{R}.$$

A random variable X has a Gaussian distribution with mean μ and variance σ^2 if and only if its characteristic function satisfies

$$\mathbb{E}\big[e^{i\alpha X}\big] = e^{i\alpha\mu - \alpha^2\sigma^2/2}, \qquad \alpha \in \mathbb{R}. \tag{A.40}$$

From Theorems A.13 and A.14 we deduce the following proposition.

Proposition A.15. *Let $X \simeq \mathcal{N}(\mu, \sigma_X^2)$ and $Y \simeq \mathcal{N}(\nu, \sigma_Y^2)$ be independent Gaussian random variables. Then $X + Y$ also has a Gaussian distribution*

$$X + Y \simeq \mathcal{N}(\mu + \nu, \sigma_X^2 + \sigma_Y^2).$$

Proof. Since X and Y are independent, by Theorem A.14 the characteristic function Ψ_{X+Y} of $X + Y$ is given by

$$\begin{aligned}
\Psi_{X+Y}(t) &= \Psi_X(t)\Psi_Y(t) \\
&= e^{it\mu - t^2\sigma_X^2/2}\, e^{it\nu - t^2\sigma_Y^2/2} \\
&= e^{it(\mu+\nu) - t^2(\sigma_X^2 + \sigma_Y^2)/2}, \qquad t \in \mathbb{R},
\end{aligned}$$

where we used (A.40). Consequently, the characteristic function of $X + Y$ is that of a Gaussian random variable with mean $\mu + \nu$ and variance $\sigma_X^2 + \sigma_Y^2$ and we conclude by Theorem A.13. $\qquad\square$

Moment generating functions

The *moment generating function* of a random variable X is the function $\Phi_X : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$\Phi_X(t) := \mathbb{E}\big[e^{tX}\big],$$

for t in a neighborhood of 0. In particular, we have

$$\mathbb{E}[X^n] = \frac{\partial^n}{\partial t^n}\Phi_X(0), \qquad n \geqslant 1,$$

provided that $\mathbb{E}[|X|^n] < \infty$, and

$$\Phi_X(t) = \mathbb{E}\big[e^{tX}\big] = \sum_{n \geqslant 0} \frac{t^n}{n!}\mathbb{E}[X^n],$$

provided that $\mathbb{E}\big[e^{t|X|}\big] < \infty$, $t \in \mathbb{R}$, and for this reason the moment generating function G_X characterizes the *moments* $\mathbb{E}[X^n]$ of $X : \Omega \longrightarrow \mathbb{N}$, $n \geqslant 0$.

The moment generating function Φ_X of a random variable X with probability density function $f : \mathbb{R} \longrightarrow \mathbb{R}_+$ satisfies

$$\Phi_X(t) = \int_{-\infty}^{\infty} e^{xt}\varphi(x)dx, \qquad t \in \mathbb{R}.$$

For example, the moment generating functions (MGF) of a Gaussian random variable X with mean μ and variance σ^2 is given by

$$\mathbb{E}\big[e^{\alpha X}\big] = e^{\alpha\mu + \alpha^2\sigma^2/2}, \qquad \alpha \in \mathbb{R}. \tag{A.41}$$

Note that in probability, the moment generating function is written as a *bilateral* transform defined using an integral from $-\infty$ to $+\infty$.

A.7 Conditional Expectation

The construction of conditional expectations of the form $\mathbb{E}[X \mid Y]$ given above for discrete and absolutely continuous random variables can be generalized to σ-algebras.

Definition A.16. *Given $\mathcal{F}$ a σ-algebra on Ω, a random variable $X : \Omega \longrightarrow \mathbb{R}$ is said to be $\mathcal{F}$-measurable if*

$$\{X \leqslant x\} := \{\omega \in \Omega \; : \; X(\omega) \leqslant x\} \in \mathcal{F},$$

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

Intuitively, when X is $\mathcal{F}$-measurable, the knowledge of the values of X depends only on the information contained in $\mathcal{F}$. For example, when $\mathcal{F} = \sigma(A_1, \ldots, A_n)$ where $(A_n)_{n \geqslant 1}$ is a *partition* of Ω with $\bigcup_{n \geqslant 1} A_n = \Omega$, any $\mathcal{F}$-measurable random variable X can be written as

$$X(\omega) = \sum_{k=1}^{n} c_k \mathbb{1}_{A_k}(\omega), \qquad \omega \in \Omega,$$

for some $c_1, \ldots, c_n \in \mathbb{R}$.

Definition A.17. *Given $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space we let $L^2(\Omega, \mathcal{F})$ denote the space of $\mathcal{F}$-measurable and square-integrable random variables, i.e.*

$$L^2(\Omega, \mathcal{F}) := \big\{ X : \Omega \longrightarrow \mathbb{R} \; : \; \mathbb{E}\big[|X|^2\big] < \infty \big\}.$$

More generally, for $p \geqslant 1$ one can define the space $L^p(\Omega, \mathcal{F})$ of $\mathcal{F}$-measurable and p-integrable random variables as

$$L^p(\Omega, \mathcal{F}) := \big\{ X : \Omega \longrightarrow \mathbb{R} \; : \; \mathbb{E}[|X|^p] < \infty \big\}.$$

We define a *inner product* $\langle \cdot, \cdot \rangle_{L^2(\Omega, \mathcal{F})}$ between elements of $L^2(\Omega, \mathcal{F})$, as

$$\langle X, Y \rangle_{L^2(\Omega, \mathcal{F})} := \mathbb{E}[XY], \quad X, Y \in L^2(\Omega, \mathcal{F}). \qquad (A.42)$$

This inner product is associated to the norm $\| \cdot \|_{L^2(\Omega)}$ by the relation

$$\|X\|_{L^2(\Omega)} = \sqrt{\mathbb{E}\big[X^2\big]} = \sqrt{\langle X, X \rangle_{L^2(\Omega, \mathcal{F})}}, \qquad X \in L^2(\Omega, \mathcal{F}).$$

The norm $\| \cdot \|_{L^2(\Omega)}$ also defines the *mean-square* distance

$$\|X - Y\|_{L^2(\Omega)} = \sqrt{\mathbb{E}\big[(X - Y)^2\big]}$$

N. Privault

between random variables $X, Y \in L^2(\Omega, \mathcal{F})$, and it induces a notion of *orthogonality*, namely X is *orthogonal* to Y in $L^2(\Omega, \mathcal{F})$ if and only if

$$\langle X, Y \rangle_{L^2(\Omega, \mathcal{F})} = 0.$$

Proposition A.18. *The ordinary expectation* $\mathbb{E}[X]$ *achieves the minimum distance*

$$\left\| X - \mathbb{E}[X] \right\|_{L^2(\Omega)}^2 = \min_{c \in \mathbb{R}} \| X - c \|_{L^2(\Omega)}^2. \tag{A.43}$$

Proof. It suffices to differentiate

$$\frac{\partial}{\partial c} \mathbb{E}\big[(X - c)^2\big] = -2\mathbb{E}[X - c] = 0,$$

showing that the minimum in (A.43) is reached when $\mathbb{E}[X - c] = 0$, *i.e.* $c = \mathbb{E}[X]$. $\qquad\square$

Similarly to Proposition A.18, the conditional expectation will be defined by a distance minimizing procedure.

Definition A.19. *Given* $\mathcal{G} \subset \mathcal{F}$ *a sub σ-algebra of $\mathcal{F}$ and $X \in L^2(\Omega, \mathcal{F})$, the conditional expectation of X given $\mathcal{G}$, and denoted*

$$\mathbb{E}[X \mid \mathcal{G}],$$

is defined as the orthogonal projection *of X onto $L^2(\Omega, \mathcal{G})$.*
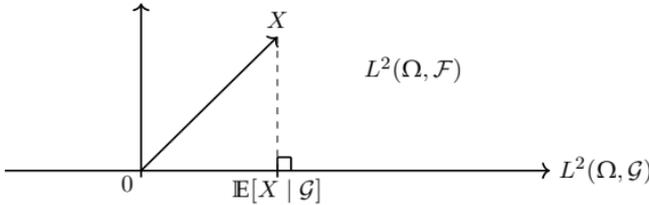
As a consequence of the uniqueness of the orthogonal projection onto the subspace $L^2(\Omega, \mathcal{G})$ of $L^2(\Omega, \mathcal{F})$, the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$ is characterized by the relation

$$\langle Y, X - \mathbb{E}[X \mid \mathcal{G}] \rangle_{L^2(\Omega, \mathcal{F})} = 0,$$

which rewrites as

$$\mathbb{E}[Y(X - \mathbb{E}[X \mid \mathcal{G}])] = 0,$$

i.e.

$$\mathbb{E}[YX] = \mathbb{E}[Y\mathbb{E}[X \mid \mathcal{G}]],$$

for all bounded and $\mathcal{H}$-measurable random variables Y, where $\langle \cdot, \cdot \rangle_{L^2(\Omega, \mathcal{F})}$ denotes the inner product (A.42) in $L^2(\Omega, \mathcal{F})$. Proposition A.20 extends Propo-

846

sition A.18 as a consequence of Definition A.19. See Theorem 5.1.4 page 197 of Stroock (2011) for an extension of the construction of conditional expectation to the space $L^1(\Omega, \mathcal{F})$ of integrable random variable.

Proposition A.20. *The conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$ realizes the minimum in mean-square distance between $X \in L^2(\Omega, \mathcal{F})$ and $L^2(\Omega, \mathcal{G})$, i.e. we have*

$$\|X - \mathbb{E}[X \mid \mathcal{G}]\|_{L^2(\Omega)} = \min_{Y \in L^2(\Omega, \mathcal{G})} \|X - Y\|_{L^2(\Omega)}. \qquad (A.44)$$

Proof. This is a consequence of the Pythagorean theorem written as

$$\|X - Y\|_{L^2(\Omega)} = \|X - \mathbb{E}[X \mid \mathcal{G}]\|_{L^2(\Omega)} + \|\mathbb{E}[X \mid \mathcal{G}] - Y\|_{L^2(\Omega)},$$

for any $Y \in L^2(\Omega, \mathcal{G})$. $\qquad \square$

The following proposition will often be used as a characterization of $\mathbb{E}[X \mid \mathcal{G}]$.

Proposition A.21. *Given $X \in L^2(\Omega, \mathcal{F})$, $Z := \mathbb{E}[X \mid \mathcal{G}]$ is the unique random variable Z in $L^2(\Omega, \mathcal{G})$ that satisfies the relation*

$$\mathbb{E}[YX] = \mathbb{E}[YZ] \qquad (A.45)$$

for all bounded and $\mathcal{G}$-measurable random variables Y.

We note that taking $Y = \mathbf{1}$ in (A.45) yields

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]. \qquad (A.46)$$

In particular, when $\mathcal{G} = \{\emptyset, \Omega\}$ we have $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X \mid \{\emptyset, \Omega\}]$ and

$$\mathbb{E}[X \mid \{\emptyset, \Omega\}] = \mathbb{E}[\mathbb{E}[X \mid \{\emptyset, \Omega\}]] = \mathbb{E}[X], \qquad (A.47)$$

because $\mathbb{E}[X \mid \{\emptyset, \Omega\}]$ is in $L^2(\Omega, \{\emptyset, \Omega\})$ and is *a.s.* constant. In addition, the conditional expectation operator has the following properties.

  i) $\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}]$ if Y depends only on the information contained in $\mathcal{G}$.

   Proof. By the characterization (A.45) it suffices to show that

$$\mathbb{E}[H(XY)] = \mathbb{E}[H(Y\mathbb{E}[X|\mathcal{G}])], \qquad (A.48)$$

   for all bounded and $\mathcal{G}$-measurable random variables H, which implies $\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}]$.

   Relation (A.48) holds from (A.45) because the product HY is $\mathcal{G}$-measurable hence Y in (A.45) can be replaced with HY.

  ii) $\mathbb{E}[Y|\mathcal{G}] = Y$ when Y depends only on the information contained in $\mathcal{G}$.

   Proof. This is a consequence of point (i) above by taking $X := \mathbf{1}$.

iii) $\mathbb{E}[\mathbb{E}[X|\mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X|\mathcal{H}]$ if $\mathcal{H} \subset \mathcal{G}$, called the *tower property*.

Proof. First, we note that by (A.46), (iii) holds when $\mathcal{H} = \{\emptyset, \Omega\}$. Next, by the characterization (A.45) it suffices to show that

$$\mathbb{E}[H\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[H\mathbb{E}[X|\mathcal{H}]], \qquad (A.49)$$

for all bounded and $\mathcal{H}$-measurable random variables H, which will imply (iii) from (A.45).

In order to prove (A.49) we check that by point (i) above and (A.46) we have

$$\begin{aligned}\mathbb{E}[H\mathbb{E}[X|\mathcal{G}]] &= \mathbb{E}[\mathbb{E}[HX|\mathcal{G}]] = \mathbb{E}[HX] \\ &= \mathbb{E}[\mathbb{E}[HX|\mathcal{H}]] = \mathbb{E}[H\mathbb{E}[X|\mathcal{H}]],\end{aligned}$$

and we conclude by the characterization (A.45).

iv) $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$ when X "does not depend" on the information contained in $\mathcal{G}$ or, more precisely stated, when the random variable X is *independent* of the σ-algebra $\mathcal{G}$.

Proof. It suffices to note that for all bounded $\mathcal{G}$-measurable Y we have

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

and we conclude again by (A.45).

v) If Y depends only on $\mathcal{G}$ and X is independent of $\mathcal{G}$, then

$$\mathbb{E}[h(X,Y)|\mathcal{G}] = \mathbb{E}[h(X,x)]_{x=Y}. \qquad (A.50)$$

Proof. This relation can be proved using the tower property, by noting that for any bounded $K \in L^2(\Omega, \mathcal{G})$ we have

$$\begin{aligned}\mathbb{E}[K\mathbb{E}[h(x,X)]_{x=Y}] &= \mathbb{E}[K\mathbb{E}[h(x,X) \mid \mathcal{G}]_{x=Y}] \\ &= \mathbb{E}[K\mathbb{E}[h(Y,X) \mid \mathcal{G}]] \\ &= \mathbb{E}[\mathbb{E}[Kh(Y,X) \mid \mathcal{G}]] \\ &= \mathbb{E}[Kh(Y,X)],\end{aligned}$$

which yields (A.50) by the characterization (A.45).

The notion of conditional expectation can be extended from square-integrable random variables in $L^2(\Omega, \mathcal{F})$ to integrable random variables in $L^1(\Omega, \mathcal{F})$, cf. *e.g.* Theorem 5.1 in Kallenberg (2002).

Proposition A.22. *When the σ-algebra $\mathcal{G} := \sigma(A_1, A_2, \ldots, A_n)$ is generated by n disjoint events $A_1, A_2, \ldots, A_n \in \mathcal{F}$, we have*

$$\mathbb{E}[X \mid \mathcal{G}] = \sum_{k=1}^{n} \mathbb{1}_{A_k} \mathbb{E}[X \mid A_k] = \sum_{k=1}^{n} \mathbb{1}_{A_k} \frac{\mathbb{E}[X \mathbb{1}_{A_k}]}{\mathbb{P}(A_k)}.$$

Proof. It suffices to note that the $\mathcal{G}$-measurable random variables can be generated by indicator functions of the form $\mathbb{1}_{A_l}$, and that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{A_l} \sum_{k=1}^{n} \mathbb{1}_{A_k} \frac{\mathbb{E}[X \mathbb{1}_{A_k}]}{\mathbb{P}(A_k)} \right] &= \mathbb{E}\left[\mathbb{1}_{A_l} \frac{\mathbb{E}[X \mathbb{1}_{A_l}]}{\mathbb{P}(A_l)} \right] \\
&= \frac{\mathbb{E}[X \mathbb{1}_{A_l}]}{\mathbb{P}(A_l)} \mathbb{E}\left[\mathbb{1}_{A_l} \right] \\
&= \mathbb{E}\left[X \mathbb{1}_{A_l} \right], \qquad l = 1, 2, \ldots, n,
\end{aligned}$$

showing (A.45). The relation

$$\mathbb{E}[X \mid A_k] = \frac{\mathbb{E}[X \mathbb{1}_{A_k}]}{\mathbb{P}(A_k)}, \qquad k = 1, 2, \ldots, n,$$

follows from Lemma A.10. $\qquad\square$

For example, in case $\Omega = \{a, b, c, d\}$ and $\mathcal{G} = \{\emptyset, \Omega, \{a, b\}, \{c\}, \{d\}\}$, we have

$$\begin{aligned}
\mathbb{E}[X \mid \mathcal{G}] &= \mathbb{1}_{\{a,b\}} \mathbb{E}[X \mid \{a, b\}] + \mathbb{1}_{\{c\}} \mathbb{E}[X \mid \{c\}] + \mathbb{1}_{\{d\}} \mathbb{E}[X \mid \{d\}] \\
&= \mathbb{1}_{\{a,b\}} \frac{\mathbb{E}[X \mathbb{1}_{\{a,b\}}]}{\mathbb{P}(\{a, b\})} + \mathbb{1}_{\{c\}} \frac{\mathbb{E}[X \mathbb{1}_{\{c\}}]}{\mathbb{P}(\{c\})} + \mathbb{1}_{\{d\}} \frac{\mathbb{E}[X \mathbb{1}_{\{d\}}]}{\mathbb{P}(\{d\})}.
\end{aligned}$$

Regarding conditional probabilities we have similarly, for $A \subset \Omega = \{a, b, c, d\}$,

$$\begin{aligned}
\mathbb{P}(A \mid \mathcal{G}) &= \mathbb{1}_{\{a,b\}} \frac{\mathbb{P}(A \cap \{a, b\})}{\mathbb{P}(\{a, b\})} + \mathbb{1}_{\{c\}} \frac{\mathbb{P}(A \cap \{c\})}{\mathbb{P}(\{c\})} + \mathbb{1}_{\{d\}} \frac{\mathbb{P}(A \cap \{d\})}{\mathbb{P}(\{d\})} \\
&= \mathbb{1}_{\{a,b\}} \mathbb{P}(A \mid \{a, b\}) + \mathbb{1}_{\{c\}} \mathbb{P}(A \mid \{c\}) + \mathbb{1}_{\{d\}} \mathbb{P}(A \mid \{d\}).
\end{aligned}$$

In particular, if $A = \{a\} \subset \Omega = \{a, b, c, d\}$ we find

$$\begin{aligned}
\mathbb{P}(\{a\} \mid \mathcal{G}) &= \mathbb{1}_{\{a,b\}} \mathbb{P}(\{a\} \mid \{a, b\}) \\
&= \mathbb{1}_{\{a,b\}} \frac{\mathbb{P}(\{a\} \cap \{a, b\})}{\mathbb{P}(\{a, b\})} \\
&= \mathbb{1}_{\{a,b\}} \frac{\mathbb{P}(\{a\})}{\mathbb{P}(\{a, b\})}.
\end{aligned}$$

In other words, the probability of getting the outcome a is $\mathbb{P}(\{a\})/\mathbb{P}(\{a, b\})$ knowing that the outcome is either a or b, otherwise it is zero.

Notes on Stochastic Finance

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

a) Consider the function $x \mapsto x^+$ from $\mathbb{R}$ to $\mathbb{R}_+$, defined as

$$x^+ = \begin{cases} x & \text{if } x \geqslant 0, \\ 0 & \text{if } x \leqslant 0. \end{cases}$$

Compute $\mathbb{E}[X^+]$ as an integral.

b) Consider the function $x \mapsto (x-K)^+$ from $\mathbb{R}$ to $\mathbb{R}_+$, defined as

$$(x-K)^+ = \begin{cases} x-K & \text{if } x \geqslant K, \\ 0 & \text{if } x \leqslant K, \end{cases}$$

where $K \in \mathbb{R}$. Compute $\mathbb{E}[(X-K)^+]$ as an integral using the cumulative distribution function of the standard normal distribution

$$\Phi(x) := \int_{-\infty}^x \mathrm{e}^{-y^2/2}\frac{dy}{\sqrt{2\pi}}, \qquad x \in \mathbb{R}.$$

c) Consider the function $x \mapsto (K-x)^+$ from $\mathbb{R}$ to $\mathbb{R}_+$, defined as

$$(K-x)^+ = \begin{cases} K-x & \text{if } x \leqslant K, \\ 0 & \text{if } x \geqslant K, \end{cases}$$

where $K \in \mathbb{R}$. Compute $\mathbb{E}[(K-X)^+]$ using the cumulative distribution function Φ.

Exercise A.5   Let $X \simeq \mathcal{N}(0, v^2)$ be a centered Gaussian random variable with variance $v^2 > 0$.

a) Compute

$$\mathbb{E}\big[\mathrm{e}^{\sigma X}\mathbb{1}_{[K,\infty)}(x\,\mathrm{e}^{\sigma X})\big] = \frac{1}{\sqrt{2\pi v^2}}\int_{\sigma^{-1}\log(K/x)}^{\infty}\mathrm{e}^{\sigma y - y^2/(2v^2)}dy.$$

Hint. Use the completion of square identity

$$\sigma y - \frac{y^2}{v^2} = \frac{v^2\sigma^2}{4} - \left(\frac{y}{v} - \frac{v\sigma}{2}\right)^2.$$

b) Compute

$$\mathbb{E}[(e^{m+X} - K)^+] = \frac{1}{\sqrt{2\pi v^2}}\int_{-\infty}^{\infty}(e^{m+x} - K)^+\,\mathrm{e}^{-x^2/(2v^2)}dx.$$

851