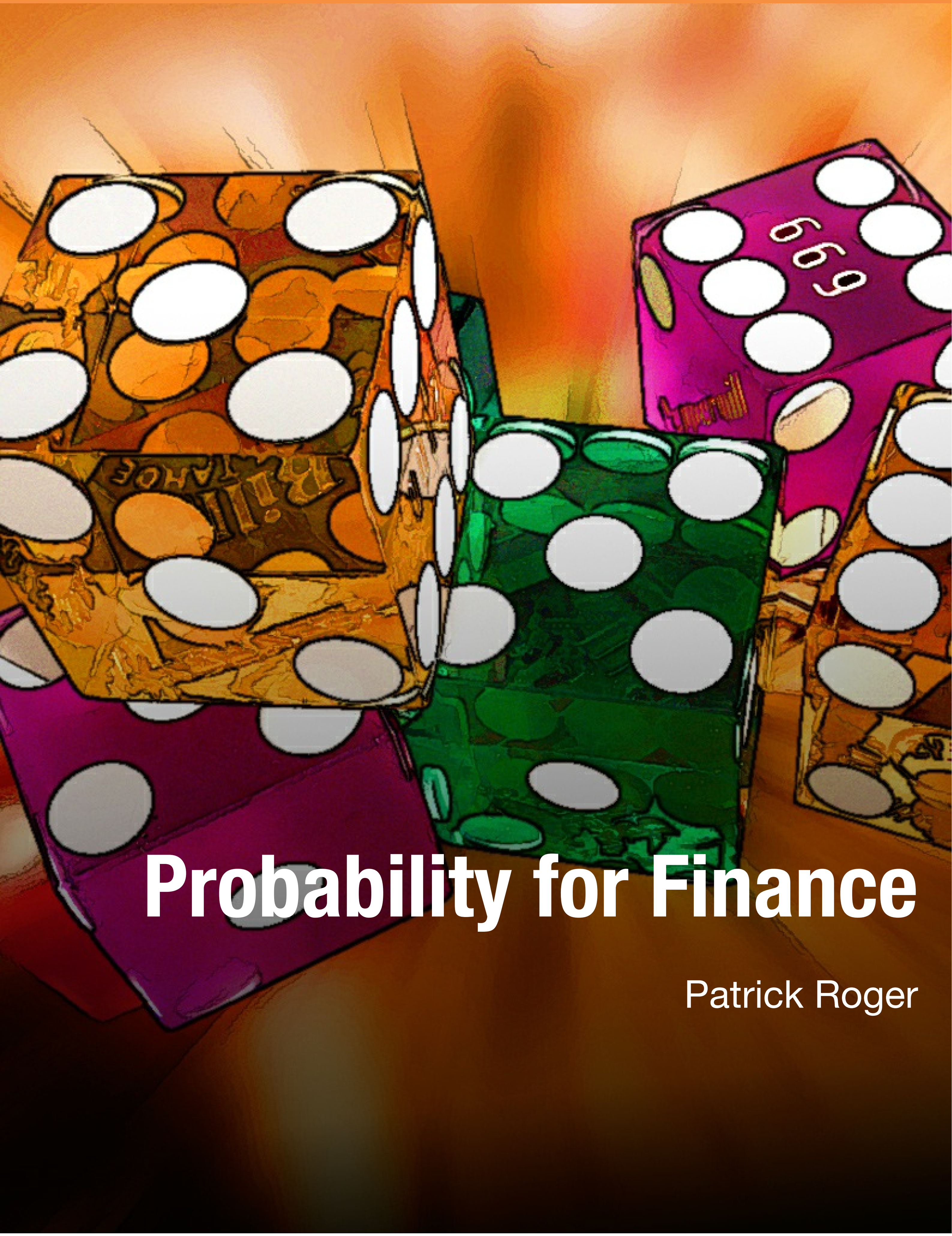

Probability for Finance

Patrick Roger

bookboon
The eBook company

Probability for Finance

Patrick Roger

Strasbourg University, EM Strasbourg Business School

May 2010

Probability for Finance

ISBN 978-87-7681-589-9

Contents

Introduction

This book is intented to be a technical support for students in finance. It is the reason why it is entitled "Probability for finance". Our purpose is to provide the essentials tools of probability theory useful to understand financial models. Consequently, almost all the examples illustrating probability results are taken from the fields of economics and finance. It means that we assume readers have elementary knowledge in finance and microeconomics, but also in elementary linear algebra and analysis.

Since the seminal work of Markowitz (1952) on portfolio diversification, mathematical models of financial markets have tremendously developed. The Capital Asset Pricing Model of Sharpe (1964), Lintner (1965) and Mossin (1966) was established in the sixties and continuous-time finance also started at the end of the same decade (Merton, 1969, 1971). Option pricing models, following the Black-Scholes-Merton model (Black and Scholes, 1973, Merton, 1973)) have given rise to a systematic mathematical approach of the pricing of derivative contracts. Sophisticated financial products have been created; they have generated a demand for valuation models. These models are essentially based on mathematics, and more precisely, on probability theory and stochastic processes.

Today, any finance student has to deal with a lot of mathematical concepts, some of them being very sophisticated and going beyond what is taught in undergraduate programs in economics and management. This book tries to fill the gap between what students actually know, and what they should know to enter the universe of financial models. One of our objectives is to present these tools in a pedagogical way, but it does not mean that the reading will be easy. Hard work is required to manage the tools in a performing way.

The book is divided in four chapters. Chapter 1 is devoted to probability spaces and random variables. Its purpose is to explain how to describe the uncertainty on financial markets and to specify how prices and returns can be written in a mathematical consistent way. Prices and returns can also be summarized by some numbers measuring their average value, the dispersion of possible future values or the relationship between the returns of different stocks. These quantities are called the moments of the probability distribution of prices or returns. Their presentation is developed in chapter 2.

Beyond the synthetic presentation of random variables through moments, a more detailed approach is necessary to specify how possible future values are disseminated along the real line. It means that it is useful to characterize probability distributions of economic variables like returns, interest rates or exchange rates. It is done in chapter 3 where we present the essential distributions appearing in the financial literature.

Finally, economic agents acquire (costly or freely) information over time. New information changes beliefs about the likelihood of future events or, in other words, changes the perceived probabilities of possible future events. The probability distributions of relevant economic variables are then modified. It is the reason why a part of chapter 4 is devoted to conditional distributions and conditional expectations.

The second part of this last chapter introduces limit theorems and convergence, in order to make a smooth transition between one-period models and multi-period models[1]. In fact, there are essentially two categories of financial models. They can be distinguished by the way time is measured. In discrete-time models, markets are open on a (finite or countable) set of dates when in continuous-time models, markets are always open. It is then important to check if a continuous-time model is the limit (in a sense to be defined) of a sequence of discrete-time models in which the duration between two transaction dates shrink to zero.

[1] These multiperiod models and the corresponding mathematical tools are described in Roger (2010).

Chapter 1

Probability spaces and random variables

1.1 Measurable spaces and probability measures

To start with a simple approach, we assume that economic agents live in a one-period economy with a starting date $t = 0$ and a end-date $T = 1$. Some financial securities (assets) are traded on the market at date 0 and generate payoffs at date 1. The description of these payoffs and the valuation of the corresponding securities at date 0 are the essential building blocks of a financial model.

Depending on the number of assets and on the complexity you desire for the model, you will authorize a number of possible terminal situations for the payoffs and, more generally, for the whole economy. This set of possibilities is called the **set of states of nature** and denoted Ω[1]. Ω may be finite or infinite depending on how you want to describe the market. The subsets of Ω are made of states of nature which describe information about the possible situation at date T. For example, if there is only one risky asset traded on the market, a range of terminal prices for this asset is associated to a subset of states of nature. For technical reasons we don't detail here, not all the subsets of Ω can be considered in a model when Ω is "too large"[2]. If $\mathcal{P}(\Omega)$

[1] In mathematical books devoted to probability theory, Ω is often called the sample space of a random experiment. In our context the random experiment in which we pick some situations is the economy or the financial market.

[2] When Ω is not countable, a probability measure cannot be defined on $\mathcal{P}(\Omega)$ in a consistent way. It will be justified in section1.1.3.

denotes the set of subsets of Ω, we restrict the model to subsets satisfying some reasonable properties allowing to define a probability measure. It is the reason why we need the (maybe) abstract concept of σ-algebra[3].

1.1.1 σ-algebra (or tribe) on a set Ω

Definition 1 *Let Ω denote a set of states of nature and $\mathcal{P}(\Omega)$ the set of subsets of Ω; a* ***σ-algebra*** *(also called a* ***tribe****) on Ω is a subset $\mathcal{A}$ of $\mathcal{P}(\Omega)$ satisfying:*

1) $\Omega \in \mathcal{A}$

2) $\forall\ B \in \mathcal{A}$, $B^c \in \mathcal{A}$ where B^c is the complement of B defined by $B^c = \{\omega \in \Omega / \omega \notin B\}$. $\mathcal{A}$ is then closed under complementation)

3) For any sequence $(B_n, n \in \mathbb{N})$ of elements of $\mathcal{A}$, $\bigcup_{n=1}^{+\infty} B_n \in \mathcal{A}$. In other words, $\mathcal{A}$ is closed under countable unions.

The pair $(\Omega, \mathcal{A})$ is called a ***measurable space*** *and the elements of $\mathcal{A}$ are called* ***events****. An event containing only one state of nature is an* ***elementary event****.*

At date $T = 1$, only one elementary event (a state of nature) ω occurs. An event A is said to be true if $\omega \in A$ and A is false if $\omega \notin A$.

Even when Ω is finite, it is possible to define several tribes on Ω. For example if Ω contains 4 states, that is $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, we could choose $\mathcal{A} = \{\emptyset, \Omega\}$ which is the most simple (and the smallest) tribe on Ω or $\mathcal{A}' = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\}$ or $\mathcal{A} = \mathcal{P}(\Omega)$, etc..

From definition 1, we easily get the following proposition.

Proposition 2 *Let $\mathcal{A}$ be a tribe on Ω;*

1) for any sequence $(B_n, n \in \mathbb{N})$ of elements in $\mathcal{A}$, the intersection $\cap_{n=1}^{+\infty} B_n \in \mathcal{A}$ ($\mathcal{A}$ is closed under countable intersections).

2) $\varnothing \in \mathcal{A}$.

[3] The symbol σ is often used in finance to denote the standard deviation of the return on a financial security (see chapter 2). Here, it has nothing to do with this usual interpretation but σ-algebra is the usual notation in probability theory for the notion defined below.

In a financial model, a tribe describes all the possible information conveyed about the state of nature which will eventually occur at the terminal date. To illustrate the point, we assume in a first step that Ω is finite.

Definition 3 $\Gamma = \{B_1, ..., B_K\}$ *is a* ***partition*** *of* Ω *if:*
1) $B_i \cap B_j = \varnothing$ *when* $i \neq j$
2) $\cup_{i=1}^{K} B_i = \Omega$.

The information held at date 0 by an economic agent may be represented by a subset A included in Ω. It means the agent knows that the true state of nature is in A. However, this information does not completely remove uncertainty, because A may contain several states of nature.

Note that the set of all possible unions of elements of Γ, including Ω and $\varnothing$, is a tribe, called the tribe generated by Γ (the proof is left as an exercise). In fact, if a given set B_j of the partition is true (meaning that the state of nature which will occur is in B_j), all the unions of elements of Γ containing B_j are also true and all the unions of elements of Γ which do not contain B_j are known to be false. Obviously, $\varnothing$ is always false and Ω is always true.

Definition 4 *Let $\Gamma = \{B_1, ..., B_K\}$ be a finite partition of Ω; the tribe generated by Γ, denoted as $\mathcal{B}_\Gamma$, is the smallest tribe containing all the elements of Γ.*

The following proposition summarizes the properties of $\mathcal{B}_\Gamma$.

Proposition 5 *a) The elements of $\mathcal{B}_\Gamma$ are $\varnothing$, Ω, and all the unions of elements of Γ.*

b) Every tribe on a finite set Ω is generated by a partition.

c) $\mathcal{B}_\Gamma$ has 2^K elements[4].

1.1.2 Sub-tribes of $\mathcal{A}$

In multi-period financial models (that is, when $T > 1$) with a finite number of states of nature, the natural tribe to be chosen at the terminal date T is $\mathcal{P}(\Omega)$ since it contains all the elementary events. However, at date $t < T$, some uncertainty remains and it is relevant to consider sub-tribes of $\mathcal{P}(\Omega)$.

Definition 6 *A subset $\mathcal{A}'$ of $\mathcal{P}(\Omega)$ is a sub-tribe of $\mathcal{A}$ if $\mathcal{A}'$ is a subset of $\mathcal{A}$ containing Ω and satisfying points (1) to (3) of definition 1, where $\mathcal{A}$ is replaced by $\mathcal{A}'$.*

In other words, $(\Omega, \mathcal{A}')$ is itself a measurable space since $\mathcal{A}'$ satisfies all the properties of a tribe. For example, when $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, the tribe $\mathcal{A}' = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\}$ is a sub-tribe of $\mathcal{P}(\Omega)$.

It is easy to check that the three properties of definition 1 are satisfied. First, $\Omega \in \mathcal{A}'$; second, for any event B in $\mathcal{A}'$, B^c is in $\mathcal{A}'$ since $\{\omega_1, \omega_2\} = \{\omega_3, \omega_4\}^c$. Finally, any union of elements of $\mathcal{A}'$ is an element of $\mathcal{A}'$ because $\{\omega_1, \omega_2\} \cup \{\omega_3, \omega_4\} = \Omega$.

When Ω is a finite set, we saw in proposition 5 that any tribe is generated by a partition. Therefore, we can establish a link between two tribes $\mathcal{A}$ and $\mathcal{A}'$ such that $\mathcal{A}' \subset \mathcal{A}$ and the partitions Γ and Γ' generating these tribes.

[4] Generally speaking, even if $Card(\Omega)$ is infinite ($Card(\Omega)$ denotes the number of elements of Ω), $Card(\Omega) < Card(\mathcal{P}(\Omega))$. This result is due to Georg Cantor (1845-1918); it explains why in a preceding comment, we mentioned that if Ω is not countable, $\mathcal{P}(\Omega)$ is "too large".

Definition 7 *A partition* Γ *is finer than a partition* Γ' *if every element of* Γ' *is a union of elements of* Γ. *The partition* Γ *is then called a* ***refinement***[5] *of* Γ'.

Proposition 8 *Let* $\mathcal{A}'$ *be a sub-tribe of* $\mathcal{A}$*; the partition* Γ *generating* $\mathcal{A}$ *is a refinement of the partition* Γ' *generating* $\mathcal{A}'$.

This proposition is an obvious consequence of point (c) in proposition 5. As the number of elements in a tribe is always 2^K for some positive integer K, K is the number of elements of the partition generating the corresponding tribe. It follows that $\mathcal{A}'$ is a sub-tribe of $\mathcal{A}$; it contains less elements and it is also the case for the partition by which it is generated.

Example 9 *One of the most popular models to describe the time-evolution of stock prices is the so-called binomial model (Cox-Ross-Rubinstein, 1979). The price at a given date is obtained by multiplying the preceding price by* u (d), *meaning a price increase (decrease).*

Let $\Omega = \{uu; ud; du; dd\}$ *denote the set of possible trajectories of a stock price in the two-period binomial model.* $\mathcal{A}' = \{\varnothing; \{uu; ud\}; \{du; dd\}; \Omega\}$ *is a tribe on* Ω *and a sub-tribe of* $\mathcal{P}(\Omega)$. *In fact,* $\{du; dd\} = \{uu; ud\}^c$ *and* Ω *is the complement of the empty set; moreover,* $\{uu; ud\} \bigcup \{du; dd\} = \Omega \in \mathcal{A}$. *The price process of the stock is described on the figure below, where it is assumed that the initial price is equal to 1.*

$$\begin{array}{ccccc} & & & \nearrow & uu = u^2 \\ & & u & & \\ & \nearrow & & \searrow & ud \\ 1 & & & & \\ & \searrow & & \nearrow & du \\ & & d & & \\ & & & \searrow & dd = d^2 \end{array} \tag{1.1}$$

We observe that the subset $\{uu; ud\}$ *corresponds to the two price paths starting by an up-move. In the financial model, it simply means that after one period during which an up-move has been observed, investors know that the final state will be an element of* $\{uu; ud\}$. *The same remark is valid with the other subset when the stock price decreases at the end of the first period. Obviously, considered as numbers, the products* ud *and* du *are equal. However, when* ud *denotes a state of nature (corresponding to an up-move*

[5] Symetrically, Γ' is less fine than Γ.

followed by a down-move), it is not the same as du. We will come back on this point in the next section.

Example 10 *If* $\Omega = \mathbb{R}$, *the smallest tribe containing all open intervals is called the Borel tribe on* Ω *and denoted* $\mathcal{B}_{\mathbb{R}}$. *It is the commonly used tribe when one deals with* $\mathbb{R}$ *or an interval of* $\mathbb{R}$. *As it is a tribe,* $\mathcal{B}_{\mathbb{R}}$ *also contains all countable unions of open intervals, all closed intervals...and more generally all the subsets of the real line we need in a financial model.*

Even if the concept of tribe may seem abstract, the reader will understand in the next subsection why it is required to define correctly a probability measure.

1.1.3 Probability measures

During 2009, many economists were asked to provide predictions about the recovery of the economy in the months to come. They had comments like "a strong recovery is unlikely" or "we can expect a slow recovery in 2010". Likelihood of an event is usually measured by a number between 0 and 1. In the formalism of probability theory, the mapping linking events to such numbers is named a **probability measure** and is defined as follows.

Definition 11 *Let $(\Omega, \mathcal{A})$ be a measurable space; a probability measure on $\mathcal{A}$ is a mapping from $\mathcal{A}$ to $[0;1]$ satisfying:*

a) $P(\Omega) = 1$

b) For any sequence $(B_n, n \in \mathbb{N})$ of disjoint[6] *events in $\mathcal{A}$:*

$$P\left(\bigcup_{n=1}^{+\infty} B_n\right) = \sum_{n=1}^{+\infty} P(B_n)$$

c)The triple $(\Omega, \mathcal{A}, P)$ is called a ***probability space****. The event Ω is called the* ***sure event*** *and $\emptyset$ the* ***impossible event****.*

A probability measure being defined on events, it is necessary that a countable union of events is in the tribe for (b) not to be meaningless. Similarly, if we consider an event B and its complement B^c, point (b) implies

$$P(B) + P(B^c) = P(\Omega) = 1$$

from which we deduce $P(B^c) = 1 - P(B)$. The probability of the complement of a given event B is naturally defined, as soon as B^c is in the tribe. These remarks explain why defining tribes or σ-algebras was necessary.

The following proposition summarizes the properties induced by definition 11.

[6] Two events are disjoint if their intersection is empty.

Proposition 12 *Let $(\Omega, \mathcal{A}, P)$ denote a probability space:*

1) $P(\varnothing) = 0$

2) $\forall\ (B_1, B_2) \in \mathcal{A} \times \mathcal{A},\ B_1 \subseteq B_2 \Rightarrow P(B_1) \leq P(B_2)$

3) Let $(B_n, n \in \mathbb{N})$ be an increasing sequence ($B_n \subset B_{n+1}$) of elements in $\mathcal{A}$:

$$\lim_{n \to +\infty} P(B_n) = P\left(\bigcup_{n \in \mathbb{N}} B_n\right)$$

4) Let $(B_n, n \in \mathbb{N})$ be a decreasing sequence ($B_n \supset B_{n+1}$) of elements in $\mathcal{A}$:

$$\lim_{n \to +\infty} P(B_n) = P\left(\bigcap_{n \in \mathbb{N}} B_n\right)$$

5) $\forall\ B \in \mathcal{A},\ \ P(B^c) = 1 - P(B)$

Proof. 1) Ω and $\varnothing$ are disjoint, therefore $P(\Omega \bigcup \varnothing) = P(\Omega) + P(\varnothing) = P(\Omega) = 1$. It implies $P(\varnothing) = 0$

2) $B_1 \subseteq B_2 \Rightarrow P(B_2) = P\left(B_1 \bigcup (B_2 \bigcap B_1^c)\right) = P(B_1) + P(B_2 \bigcap B_1^c) \geq P(B_1)$

3) As $(B_n, n \in \mathbb{N})$ is increasing, the sequence $u_n = P\left(\bigcup_{p=1}^n B_p\right)$ is increasing and has an upper bound (lower or equal to $P(\Omega) = 1$), it then has a limit. But $(B_n, n \in \mathbb{N})$ being increasing for inclusion, the limit is nothing else than $P\left(\bigcup_{n \in \mathbb{N}} B_n\right)$.

4) As $(B_n, n \in \mathbb{N})$ is decreasing, the sequence $v_n = P\left(\bigcap_{p=1}^n B_p\right)$ is decreasing and has a lower bound (greater or equal to $P(\varnothing) = 0$), it then has a limit. But $(B_n, n \in \mathbb{N})$ being decreasing for inclusion, the limit is nothing else than $P\left(\bigcap_{n \in \mathbb{N}} B_n\right)$.

5) Point (b) of definition 11 implies $P\left(B \bigcup B^c\right) = P(B) + P(B^c)$ since B and B^c are disjoint; but as $B \bigcup B^c = \Omega$, we deduce $P\left(B \bigcup B^c\right) = P(\Omega) = 1$ leads to $P(B^c) = 1 - P(B)$. ■

Example 13 *Let $Card(\Omega) = N$ and $\mathcal{A} = \mathcal{P}(\Omega)$; the uniform probability measure on $\mathcal{A}$ is the one which gives to every elementary event the same weight, that is*[7]*:*

$$\forall \omega \in \Omega, P(\omega) = \frac{1}{N}$$

This probability measure appears in simple experiments like coin tosses or simulation problems.

Example 14 *Let now Ω be the unit square $[0;1] \times [0;1]$; it is an uncountable subset of $\mathbb{R}^2$; it has to be equipped with the Borel σ-algebra, that is the tribe containing all open sets. In this framework the uniform probability distribution is characterized in the following way. If A is an event included in the unit square Ω, $P(A)$ is equal to the area of A. P obviously satisfies $P(\Omega) = 1$; P is usually called the Lebesgue probability measure on the unit square. The area of the union of two disjoint subsets of $[0;1] \times [0;1]$ is obviously equal to the sum of the areas of these subsets.*

It is important to mention that any finite or countable set of points in the square has a Lebesgue measure equal to 0. Generally speaking, the probability of any rectangle $B = [a;b] \times [c;d]$ is equal to $(d-c)(b-a) \leq 1$. This remark points out the intuition about the equivalence between the Lebesgue measure on the square and the uniform probability measure on a finite set. Imagine that a dart is thrown at random on the square[8]*, the probability that it falls in B is equal to $(d-c)(b-a)$.*

1.2 Conditional probability and Bayes theorem

In the following, the probability space we refer to is $(\Omega, \mathcal{A}, P)$ even if it is not explicitly recalled. When investors get some information, it means that they learn something about the state which will eventually occur. For example, they may learn that an event $B \subset \Omega$ is true. Consequently, they change the initial probability measure defined on $\mathcal{A}$ to take into account

[7] $P(\omega)$ is a commonly used simplified notation. In fact, to be rigorous, we should note $P(\{\omega\})$ because a probability measure is defined on events, not on states.

[8] In a comparison which has become famous, Burton Malkiel (*A Random Walk down Wall Street,* 1973) wrote: "A blindfolded monkey throwing darts at a newspaper's financial pages could select a portfolio that would do just as well as one carefully selected by the experts."

this new information. To formalize this process, we introduce conditional probabilities.

1.2.1 Independent events and independent tribes

Definition 15 *1) Two events B_1, B_2 in $\mathcal{A}$ are independent if* $P(B_1 \bigcap B_2) = P(B_1) \times P(B_2)$.

2) Let $B_2 \in \mathcal{A}$ such that $P(B_2) \neq 0$; the conditional probability of B_1 knowing B_2, denoted as $P(B_1 | B_2)$, is defined by:

$$P(B_1 | B_2) = \frac{P(B_1 \bigcap B_2)}{P(B_2)}$$

As said before, conditional probability has a natural interpretation. If you learn that the event B_2 occurs, you also know that the true state of nature is in B_2. Consequently, your evaluation of the probability that B_1 occurs is changed; the uncertainty is reduced to B_2, not to the whole set Ω. In particular, if $B_1 \bigcap B_2 = \varnothing$, you can be sure that B_1 will not occur. Therefore, the conditional probability of B_1 will be 0.

Analogously, if the two events B_1 and B_2 are independent, the occurrence of B_2 brings no information about the occurrence of B_1. In this case, the conditional probability of B_1 knowing B_2 should be equal to the unconditional probability...and it is obviously the case since:

$$P(B_1 | B_2) = \frac{P(B_1 \bigcap B_2)}{P(B_2)} = \frac{P(B_1) \times P(B_2)}{P(B_2)} = P(B_1)$$

Example 16 *To get an easily understandable illustration of independence, consider one more time $\Omega = [0;1] \times [0;1]$ equipped with its Borel tribe and the Lebesgue probability measure. Denote (x,y) a point in Ω; let $B_1 = \left[0; \frac{1}{2}\right] \times \left[\frac{1}{3}; 1\right]$ and $B_2 = \left[0; \frac{1}{3}\right] \times \left[0; \frac{1}{2}\right]$; we then have:*

$$\begin{aligned} P(B_1) &= \frac{1}{2} \times \frac{2}{3} = \frac{1}{3} \\ P(B_2) &= \frac{1}{3} \times \frac{1}{2} = \frac{1}{6} \end{aligned}$$

If B_2 *is known to occur and if* $(x, y) \in B_1$*, we know that* $x \in \left[0; \frac{1}{3}\right]$ *and* y *has a probability* $1/3$ *to be in the range* $\left[\frac{1}{3}; \frac{1}{2}\right]$. *In fact,* $(x, y) \in B_2$ *means that* $y \leq \frac{1}{2}$. *For* (x, y) *to belong to* B_1 *it is also necessary that* $y \geq \frac{1}{3}$, *consequently, knowing that* $y \in \left[0; \frac{1}{2}\right]$ *implies a* $1/3$ *probability that* $y \in \left[\frac{1}{3}; \frac{1}{2}\right]$.*We then deduce that* $P(B_1 | B_2) = \frac{1}{3}$. *As* $B_1 \bigcap B_2 = \left[0; \frac{1}{3}\right] \times \left[\frac{1}{3}; \frac{1}{2}\right]$, *we write:*

$$P(B_1 \bigcap B_2) = \left(\frac{1}{3} - 0\right) \times \left(\frac{1}{2} - \frac{1}{3}\right) = \frac{1}{18}$$

It leads to:

$$P(B_1 | B_2) = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{1}{3} = P(B_1)$$

B_1 *is then independent of* B_2.

Lemma 17 *Two disjoint events with non zero probability and different from Ω are not independent*[9].

This result is obvious because if B_1 and B_2 are disjoint with non zero probability, their intersection is empty and has zero probability. Therefore, the conditional probability is 0, different from the one of B_1, and the two events cannot be independent. Knowing that B_2 is true implies that B_1 is surely false.

Independence of events is then generalized to the independence of σ-algebras in the following way.

Definition 18 *Two sub-tribes $\mathcal{G}$ and $\mathcal{G}'$ of $\mathcal{A}$ are independent if :*

$$\forall B \in \mathcal{G}, \forall B' \in \mathcal{G}', P(B \cap B') = P(B) \times P(B')$$

Two tribes are independent if any pair of events in $\mathcal{G} \times \mathcal{G}'$ is independent. We advice readers to look for two independent sub-tribes of $\mathcal{P}(\Omega)$ on a set Ω with 4 equally likely states. Lemma 17 may be useful to build such an example.

1.2.2 Conditional probability measures

We mentioned before that, when an investor gets a piece of information, he changes his beliefs. In other words, he defines a new probability measure on $(\Omega, \mathcal{A})$. No information means that the only event you know to be true is Ω and the only event you know to be false is $\varnothing$. Consequently, getting a piece of information means that you know some proper subsets of Ω are true while some non empty others are false. Conditional probability measures formalize this process.

Proposition 19 *Let $B \in \mathcal{A}$ and define $\mathcal{A}_B$ by:*

$$\mathcal{A}_B = \left\{ A \bigcap B \text{ with } A \in \mathcal{A} \right\}; \tag{1.2}$$

$\mathcal{A}_B$ is a tribe on B. In other words, $(B, \mathcal{A}_B)$ is a measurable space.

[9] We mention this obvious result because in many occasions we noticed that some students had a tendency to mix the two notions, independence and void intersection.

Proof. $B \in \mathcal{A}_B$ because $\Omega \bigcap B = B$; let $(C_n, n \in \mathbb{N})$ be a sequence of sets written $C_n = A_n \bigcap B$; we can write:

$$\bigcup_{n\in\mathbb{N}} C_n = \bigcup_{n\in\mathbb{N}} \left(A_n \bigcap B\right) = \left(\bigcup_{n\in\mathbb{N}} A_n\right) \bigcap B$$

As $\mathcal{A}$ is a tribe, $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$ and then $\left(\bigcup_{n\in\mathbb{N}} A_n\right) \bigcap B \in \mathcal{A}_B$.

Denote now $C = A \bigcap B \in \mathcal{A}_B$ and C_B^c the complement of C in B. We have:

$$\begin{aligned} C_B^c &= \left(A \bigcap B\right)^c \bigcap B = \left(A^c \bigcap B\right) \bigcup \left(B^c \bigcap B\right) \\ &= A^c \bigcap B \in \mathcal{A}_B \end{aligned}$$

■

Proposition 20 *Let* $B \in \Omega$ *such that* $P(B) \neq 0$*; the mapping denoted* $P(.\,|B)$, *which associates* $P(B_1\,|B)$ *to any event* B_1, *is a probability measure on* $(B, \mathcal{A}_B)$.

Proof. First, remark that $P(B\,|B) = 1$. Let $(C_n, n \in \mathbb{N})$ be a sequence of disjoint sets in $\mathcal{A}_B$; we get:

$$P\left(\bigcup_{n\in\mathbb{N}} C_n\,|B\right) = \frac{P\left(\left(\bigcup_{n\in\mathbb{N}} C_n\right) \bigcap B\right)}{P(B)} = \frac{P\left(\bigcup_{n\in\mathbb{N}} \left(C_n \bigcap B\right)\right)}{P(B)} \qquad (1.3)$$

For every n, $C_n \subset B$, then the last term in the right-hand side of equation 1.3 is written as:

$$\frac{P\left(\bigcup_{n\in\mathbb{N}} C_n\right)}{P(B)} = \frac{\sum_{n\in\mathbb{N}} P\left(C_n\right)}{P(B)} = \frac{\sum_{n\in\mathbb{N}} P\left(C_n \bigcap B\right)}{P(B)} = \sum_{n\in\mathbb{N}} P\left(C_n\,|B\right)$$

■

Thinking to the time-dimension of financial models provides a natural interpretation of conditional probability measures. Assume that an information is disclosed at date t which reveals the occurrence of an event B. Economic agents take this information as granted and reallocate probabilities to events conditioned on this information arrival.

For example, statistics about the U.S economy (production, unemployment, etc.) are often disclosed on fridays at 8:30 AM, before the opening of

U.S markets. Due to the time-lag, this information comes on European markets at 1:30 PM or 2:30 PM, that is when the corresponding financial markets are open. Investors change their beliefs according to these new pieces of information and it may have important consequences on market prices, especially when the disclosed information is perceived as a surprise. In other words, after the disclosure agents work on the probability space $(B, \mathcal{A}_B, P(.\,|B))$.

1.2.3 Bayes theorem

To introduce Bayes theorem, consider the following example. On a large population of individuals, 1 over 10 000 suffers from a rare disease which can be diagnosed by a simple test. In 1% of cases, the test result is wrong, providing a positive result when the individual is healthy or a negative result when he is ill. What is the probability of being ill if you receive a positive test result?

If you think about the question too quickly (as many people do), you could answer that you have 99% chance of being ill. This answer is wrong; in fact the probability is only 1%. The question asks for the probability of being ill, conditional on a positive test, but most people provide the probability of getting a positive test, knowing that they are ill. To understand what is going on, consider a set of 10 000 people, one of them being ill (the mean proportion in the population). The test result being wrong 1% of the time, 100 people will get a positive result but only one person suffers from the disease. Consequently, a positive test result means that you have 1% chance of being ill.

Bayes theorem is the tool allowing to deal correctly with this kind of question.

Proposition 21 *Let $(B_1, B_2, ..., B_n)$ be a partition of Ω and $C \in \mathcal{A}$, all being non zero probability events; we then get:*

$$P(B_j \,|C\,) = \frac{P(C\,|B_j\,)P(B_j)}{\sum_{i=1}^{n} P(C\,|B_i\,)P(B_i)}$$

Proof. The subsets B_j define a partition, therefore:

$$C = \bigcup_{i=1}^{n} \left(C \bigcap B_i \right)$$

and we deduce:

$$P(C) = \sum_{i=1}^{n} P\left(C \bigcap B_i \right) = \sum_{i=1}^{n} P(C\,|B_i\,)P(B_i) \tag{1.4}$$

Moreover

$$P\left(C \bigcap B_j \right) = P(C\,|B_j\,)P(B_j) = P(B_j \,|C\,)P(C) \tag{1.5}$$

Replacing in (1.4) the value of $P(C)$ coming from equation (1.5) leads to the desired result. ■

Equation (1.4) shows that the probability of an event may be written as a weighted average of conditional probabilities, the loadings being the probabilities of the events in the partition.

Coming back to the above example, denote C the set of people getting a positive test, B_1 the set of ill people and $B_2 = B_1^c$ the set of healthy people.

$$P(B_1 \,|C\,) = \frac{P(C\,|B_1\,)P(B_1)}{P(C\,|B_1\,)P(B_1) + P(C\,|B_2\,)P(B_2)}$$

We know that

$$\begin{aligned} P(B_1) &= 10^{-4} \\ P(C\,|B_1\,) &= 0.99 \\ P(C\,|B_2\,) &= 0.01 \end{aligned}$$

Bayes theorem leads to:

$$P(B_1 \,|C\,) = \frac{0.99 \times 10^{-4}}{0.99 \times 10^{-4} + 0.01 \times (1 - 10^{-4})} \simeq 0.01$$

1.3 Random variables and probability distributions

1.3.1 Random variables and generated tribes

Intuitively, a random variable (a future stock price or a return) is a quantity not known in advance; in other words, its value depends on the elementary event which will eventually occur at the terminal date $T = 1$. It is then natural to describe this mathematical object as a mapping from Ω onto a set of numbers depending on the phenomenon we want to model. If, for example, the variable is a price, the relevant set of possible values is the set of positive real numbers[10] denoted $\mathbb{R}^+$. When the variable is a return (possibly taking negative values) the entire real line $\mathbb{R}$ is the set of possible values[11].

[10] A stock price cannot be negative because of the limited liability of shareholders.

[11] If a linear return is calculated as $(S_1 - S_0)/S_0$, the minimum possible value is -1. However, if a logarithmic return is used, defined by $\ln(S_1/S_0)$, the minimum value is $-\infty$.

However, one of the objectives of such a financial model is to assess probabilities to subsets of values taken by economic variables. For example, we could want to define the probability that the next day return on the S&P500 index will be in the range $[-2\%; 2\%]$. In the preceding section, we showed that a probability measure must be defined on a set of events (a tribe). The following definition translates these ideas in a consistent mathematical language .

Definition 22 *Let* $(\Omega, \mathcal{A})$ *and* $(E, \mathcal{B})$ *two measurable spaces; a* ***random variable*** *is a function defined on* Ω *taking values in* E $(X : \Omega \to E)$ *which satisfies:*

$$\forall B \in \mathcal{B},\ X^{-1}(B) \in \mathcal{A}$$

where the set $X^{-1}(B)$ *is defined by* $X^{-1}(B) = \{\omega \in \Omega \ / \ X(\omega) \in B\}$. X *is also called a* $\mathcal{A}$-***measurable function***.

Suppose that X is a stock return; in this case $E = \mathbb{R}$. The definition means that the reciprocal image of an interval of possible stock returns is in the tribe $\mathcal{A}$. Based on this assumption, it is possible (starting from a probability measure on $\mathcal{A}$) to define a probability measure P_X on $\mathcal{B}_{\mathbb{R}}$ by $P_X(B) = P\left(X^{-1}(B)\right)$. This induced probability measure will be defined more formally in the next section.

The notion of random variable allows to answer the abovementioned questions. For example, if X denotes the tomorrow closing value of the S&P500 index and if B is a range of possible index values, $P_X(B)$ is the probability that the index ends tomorrow in this range. The space E in which the random variables take their values is usually $\mathbb{R}$ or $\mathbb{R}^n$ or one of their subsets, like $\mathbb{R}^+$ or the set $\mathbb{N}$ of positive integers. If $E \subseteq \mathbb{R}$, we deal with real random variables, and if $E = \mathbb{R}^n$ we refer to random vectors.

When some information is obtained about the values of a random variable, we deduce that some events in $\mathcal{A}$ occur. For example, if X denotes a stock return and if we know that X is in $B = [-2\%; +2\%]$, we infer that the event $X^{-1}(B)$ occurs. More generally, observing the value of a random variable defines a list of events in $\mathcal{A}$ known to be true or false. This intuition is formally described in the following definition.

Definition 23 *Let* X *denote a random variable defined on* $(\Omega, \mathcal{A})$ *and taking values in* $(E, \mathcal{B})$. *The tribe generated by* X *(denoted* $\mathcal{B}_X$*) is the subset of* $\mathcal{A}$ *defined by:*

$$\mathcal{B}_X = \left\{A \in \mathcal{A} \quad / \quad \exists B \in \mathcal{B} \ , \ A = X^{-1}(B)\right\}$$

We let the reader check that $\mathcal{B}_X$ is a sub-tribe of $\mathcal{A}$, that is a subset of $\mathcal{A}$ satisfying the properties of definition 1.

Calling a random variable a $\mathcal{A}$-measurable function puts to light another important point; a function X may be a random variable when Ω is equipped with a given tribe and may not be a random variable with respect to another tribe. This point is fundamental when modelling the evolution of financial or economic variables over time. In fact, the information known at a given date t defines a sub-tribe (usually denoted $\mathcal{F}_t$) of the tribe generated by information known at date $s > t$. It means that agents do not forget what they knew in the past, or in technical notation, $\mathcal{F}_t \subset \mathcal{F}_s$.

For example, when we introduced the binomial model in the preceding section, we wrote states of nature as uu or ud meaning that, at date 2, investors remember that the stock price process started by an up-move. Finally, it is worth noting that if Ω is finite and $\mathcal{A} = \mathcal{P}(\Omega)$, any function from Ω to $\mathbb{R}$ is a random variable. The reason is that any subset of Ω and, *a fortiori,* the reciprocal image of any interval, is in $\mathcal{A}$.

Let us now illustrate these points with a simple example. Let $Card(\Omega) = 4$ and X, Y two random variables defined in table 1.1. The main difference between the two variables is related to the information they convey when we observe their values. All possible values of X are different; consequently, the state of nature is identified as soon as a value $X(\omega)$ is observed. On the contrary, suppose you observe $Y(\omega) = 2$. You cannot infer if the true state is ω_2 or ω_4. In more technical words, the tribe generated by Y is included in the one generated by X.

State	X	Y
ω_1	1	4
ω_2	6	2
ω_3	2	1
ω_4	3	2

Table 1.1: Definition of X and Y

More precisely, we can write :

$$\begin{aligned}\mathcal{B}_X &= \mathcal{P}(\Omega) \\ \mathcal{B}_Y &= \{\emptyset, \Omega, \{\omega_1\}, \{\omega_2, \omega_4\}, \{\omega_3\}, \{\omega_1, \omega_3\}, \{\omega_1, \omega_2, \omega_4\}, \{\omega_3, \omega_2, \omega_4\}\}\end{aligned}$$

We observe that $\mathcal{B}_Y$ does not separate states 2 and 4, simply because Y takes the same values on the two states[12].

Example 24 *Consider one more time the binomial evolution of a stock price; let S_t the date-t price of the stock:*

$$S_t = \begin{cases} uS_{t-1} \text{ with probability } p \\ dS_{t-1} \text{ with probability } 1-p \end{cases}$$

A two-period (three dates) model is represented by $\Omega = \{uu; ud; du; dd\}$ corresponding to the possible paths of the stock price. At date 0, agents know nothing about future prices, so the tribe generated by the initial price S_0 is $\mathcal{B}_0 = \{\varnothing, \Omega\}$. The only two events agents know to be true or false are $\varnothing$ and Ω.

At date 1, the price S_1 is observed and everybody knows which one of the two events $\{du; dd\}$ or $\{uu; ud\}$ occurs. The first (second) event means that a down(up)-state has been observed. The tribe $\mathcal{B}_1$ is then defined by :

$$\mathcal{B}_1 = \{\varnothing, \Omega, \{du; dd\}, \{uu; ud\}\}$$

[12] This kind of remark led Ross (1976) to show that index options are more "efficient" than options on individual stocks to complete a financial market.

It is a refinement of $\mathcal{B}_0$ since $\mathcal{B}_0 \subset \mathcal{B}_1$. It is nevertheless worth to note that,

even if $\mathcal{B}_1$ is a list of events relevant at date 1, it is built at date 0.

Finally, at date 2, an interesting phenomenon appears if the final price is $S_2 = udS_0$. It is not possible to know exactly the trajectory if S_1 has been "forgotten". In fact, an up-state followed by a down-state leads to the same final price as the reverse sequence (down-state followed by an up-state). Consequently, the relevant tribe $\mathcal{B}_2$ is the one generated by the pair of variables (S_1, S_2) and not by S_2 only. We let the reader determine $\mathcal{B}_{S_2}$ and show that this tribe is strictly included in the one generated by S_1 ***and*** *S_2.*

1.3.2 Independent random variables

Recall that two events A and B are independent if $P(A \cap B) = P(A) \times P(B)$ (definition 15).

Definition 25 *Two random variables X and Y defined on $(\Omega, \mathcal{A}, P)$ and taking values in $(E, \mathcal{B})$ are independent if for any pair $(A, B) \in \mathcal{B}^2$, the events $X^{-1}(A)$ and $Y^{-1}(B)$ are independent.*

Example 26 *Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ and assume that the four states are equally likely*[13]*. Let X and Y be two random variables defined by:*

$$(X, Y) = \begin{bmatrix} 1 & 1 \\ -1 & 2 \\ 1 & 2 \\ -1 & 1 \end{bmatrix}$$

The question of independence can be first examined in an intuitive manner by asking if knowing the value taken by one of the two variables brings information on the value taken by the other. When $Y = 1$, X is equal to 1 or -1 with equal probabilities (states ω_1 and ω_4). But when $Y = 2$, we get exactly the same possible results for X (states ω_2 and ω_3). Consequently, knowing Y doesn't change the probabilities of the events in $\mathcal{B}_X$. The same arguments could be applied by exchanging the roles of the two variables.

The independence of X and Y can be easily proved by using definition 25. The important point is that the values of X and Y are not fundamental. What is crucial is the information revealed by the variables. In other words, if X and Y are independent, it is also the case for aX and bY where a and

[13] To be precise we should say "elementary events" instead of "states".

b are non-zero real numbers. The following proposition links independence of variables and independence of the tribes generated by these variables.

Proposition 27 *Two random variables X and Y are independent if and only if is the generated tribes $\mathcal{B}_X$ and $\mathcal{B}_Y$ are independent.*

1.3.3 Probability distributions and cumulative distributions

A real random variable X is, as defined before, a function defined on a measurable space $(\Omega, \mathcal{A})$ taking values in $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Any random variable X then allows to define a probability measure on $\mathcal{B}_{\mathbb{R}}$ starting from the probability measure P on $\mathcal{A}$.

Definition 28 *Let X denote a random variable defined on $(\Omega, \mathcal{A})$ with range in $(E, \mathcal{B})$;*

a) The probability distribution of X is a probability measure P_X on $\mathcal{B}$, defined by:

$$\forall B \in \mathcal{B},\ P_X(B) = P\left(X^{-1}(B)\right) = P\left(\{\omega \in \Omega \ / \ \ X(\omega) \in B\}\right)$$

b) When $(E, \mathcal{B}) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, the cumulative distribution function of X, denoted F_X, is a function from $\mathbb{R}$ to $[0;1]$ defined by:

$$F_X(x) = P\left(\{\omega \in \Omega \ / \ X(\omega) \leq x\}\right) = P_X\left((-\infty; x]\right)$$

$X(\omega)$	0	1	2	3
Proba	$\frac{\binom{7}{3}}{\binom{10}{3}} = \frac{35}{120}$	$\frac{3\times\binom{7}{2}}{\binom{10}{3}} = \frac{63}{120}$	$\frac{3\times\binom{7}{1}}{\binom{10}{3}} = \frac{21}{120}$	$\frac{1}{120}$

Table 1.2: Probability measure induced by X

This definition shows why it is interesting to define probabilities of events linked to economic variables. The probability measure defined on $\mathcal{A}$ is an abstraction because Ω is a general space describing all possible economic situations. In real-life problems, probabilities of the P_X-type, and the corresponding cumulative distribution functions, are often used in place of P.

Example 29 *Consider a simplified lotto game with 10 numbers, players chosing 3 numbers among the 10. Let X denote the random variable counting the number of correct numbers on a given ticket, the official draw being given. The relevant set Ω is the set of triples of different numbers between 1 and 10.*

Denote $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ the number of combinations of k numbers taken from a set of n numbers. The set Ω has then $\binom{10}{3} = \frac{10!}{3!(10-3)!} = 120$ elements. As Ω is finite, we can choose $\mathcal{A} = \mathcal{P}(\Omega)$ and $P(\omega) = \frac{1}{120}$.

The probability measure P_X is built as follows. First remark that X can take only 4 values from 0 to 3. Table 1.2 gives the probabilities of the events[14] $\{X = k\}$ for $k = 0, ..., 3$.

There is only one combination with three correct numbers, so $P(X = 3) = \frac{1}{120}$. Concerning $\{X = 2\}$, there are 3 possible pairs of winning numbers and

we then have to draw one number in the 7 losing numbers. Consequently $P(X = 2) = \frac{3\times 7}{120}$. Using the same argument for $\{X = 1\}$, there are three possible choices for 1 winning number among the 3 and $\binom{7}{2} = 21$ couples of losing numbers; we then deduce $P(X = 1) = \frac{63}{120}$. Finally, $P(X = 0) = \binom{7}{3}/\binom{10}{3} = 35/120$. We can check immediately:

$$63 + 35 + 21 + 1 = 120 \tag{1.6}$$

[14] Remember that $\{X = k\}$ is a shortcut for $\{\omega \in \Omega \text{ such that } X(\omega) = k\}$.

This simple example first shows how to build the space $(\Omega, \mathcal{A}, P)$ and, second, how to use binomial coefficients to define P_X. Moreover, $\mathcal{B}_X$ is a proper sub-tribe of $\mathcal{A}$, meaning that $\mathcal{B}_X \subsetneqq \mathcal{A}$. Even if a player is only interested in the number of winning numbers on his ticket, telling him this number does not completely reveal the state of nature (the official draw), except if he notched the three correct numbers.

Proposition 30 *The* ***cumulative distribution function*** *(CDF) F_X of a random variable X is an increasing, right-continuous function, satisfying:*

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

Proof. F_X is increasing due to point (2) of proposition 12 ($B_1 \subset B_2 \Rightarrow P(B_1) \leq P(B_2)$). If $x \leq y$, we have $(-\infty; x] \subseteq (-\infty; y]$ and then $P_X((-\infty; x]) \leq P_X((-\infty; y])$.

Right-continuity comes directly from the definition of $F_X(x)$ as the probability of the interval $(-\infty; x]$, closed on the right. In fact, let $(x_n, n \in \mathbb{N})$ denote a decreasing sequence converging to x. The sequence $B_n = (-\infty; x_n]$ is decreasing and converges to $B = (-\infty; x]$. Point (4) of proposition 12 allows to conclude.

The results about the two limits can be proved with the same approach, that is by using sequences going to $-\infty$ or $+\infty$ when n tends to infinity. ■

Example 31 *In finance, the CDF is used to define a popular risk measure, namely* ***Value at Risk*** *or VaR* [15]*. Banks must hold sufficient capital to face potential portfolio losses. To value the amount of required capital, the VaR(99%) is commonly use. It is defined as the number x such that:*

$$P(X \geq x) = 1 - F_X(x) = 0.99$$

where X denotes the return of the portfolio over a given period of time.

When X follows a continuous distribution, we get:

$$VaR(99\%) = F_X^{-1}(0.01)$$

In other words, the potential loss which may be borne with a probability of 99% is lower than $|x|$.

[15] See Jorion (2000) for developments on VaR.

Example 32 ***Stochastic dominance*** *is a concept which allows to rank financial assets*[16]*. It is defined as follows: a financial asset which pays X dominates an other financial asset paying Y, for stochastic dominance of degree 1 if:*

$$\forall x \in \mathbb{R}, \quad F_X(x) \leq F_Y(x)$$

where F_X and F_Y are the CDF of X and Y.

It can be proved that when X dominates Y, all agents characterized by a strictly increasing utility function prefer X to Y, independently of their risk attitude. This result is intuitive because the above inequality is equivalent to:

$$P(\{X \geq x\}) \geq P(\{Y \geq x\})$$

Consequently, for any value x, the probability of getting a payoff greater than x is greater for asset X than for asset Y. It explains the expression "stochastic dominance".

[16] See Huang-Litzenberger (1988), chapter 2.

1.3.4 Discrete and continuous random variables

Definition 33 *a) A random variable X is* ***discrete*** *if there exists a sequence $(x_n, n \in \mathbb{N})$ such that:*

$$\sum_{n\in\mathbb{N}} P\left(\{\omega \in \Omega \ \ / \ X(\omega) = x_n\}\right) = \sum_{n\in\mathbb{N}} P\left(X = x_n\right) = 1$$

$(x_n, n \in \mathbb{N})$ is called the ***support*** *of X.*

b) A random variable Y is ***continuous*** *if there exists a positive function f_Y, continuous (except at most on a finite or countable number of points) such that:*

$$F_Y(x) = \int_{-\infty}^{x} f_Y(y)dy$$

where F_Y is the CDF of Y. f_Y is called the probability density of Y (or, in short, the density).

Remark 34 *1) Proposition 30 shows that:*

$$\int_{-\infty}^{+\infty} f_Y(y)dy = 1$$

2) When a variable X is discrete, as in example 29, the CDF is a step function which exhibits jumps at the values in the support of X. A discrete variable can then be expressed as a combination of indicator functions defined as follows.

Definition 35 *Let $B \in \mathcal{A}$; the* ***indicator function*** *of the event B, denoted $\mathbf{1}_B$ is defined by:*

$$\begin{aligned} \mathbf{1}_B(\omega) &= 1 \quad \text{if } \omega \in B \\ &= 0 \ \text{otherwise} \end{aligned}$$

The name of these variables is natural because their value is 1 on a state ω to indicate that ω is an element of B.

Example 36 *Some financial assets can be modelized by means of indicator functions. For example, the Chicago Board Options Exchange trades* ***binary options*** *on the S&P500 index[17]. These contracts pay $100 to their holders when the index is above a given value (the strike price) at the maturity date. Suppose a strike price $K = 1000$ and denote X_T the value of the S&P500 at the maturity date T. The payoff of the binary contract is equal to $100 \times \mathbf{1}_{\{X_T \geq 1000\}}$ where $\{X_T \geq 1000\} = \{\omega \in \Omega \ / \ X_T(\omega) \geq 1000\}$.*

[17] see http://www.cboe.com/products/indexopts/bsz_spec.aspx

More generally, indicator functions allow to present discrete variables in an alternative way, that is as a linear combination of indicator functions.

Proposition 37 *Let X denote a discrete variable with finite support $\{x_1, ..., x_n\}$, with $x_i \neq x_j$ for $i \neq j$; there exists a partition $\Gamma = \{B_1, ..., B_n\}$ of Ω such that:*

$$X = \sum_{i=1}^{n} x_i \mathbf{1}_{B_i}$$

This result is obvious when defining $B_i = \{\omega \in \Omega \;/\; X(\omega) = x_i\}, i = 1, 2, ..., n$.

If $Card(\Omega) = N$ and $X(\omega_i) = x_i$, we get:

$$X = \sum_{i=1}^{N} x_i \mathbf{1}_{\{\omega_i\}} \tag{1.7}$$

In financial models with $Card(\Omega) < +\infty$, the financial assets whose payoffs may be described by $\mathbf{1}_{\{\omega_i\}}$ are called **Arrow-Debreu securities**. When all these securities are traded, equation (1.7) shows that every financial security is a portfolio of Arrow-Debreu securities. The market is said "complete" in this case.

1.3.5 Transformations of random variables

The question addressed in this section is the following: knowing the probability distribution of a given random variable X, can we deduce the probability distribution of the random variable $Y = g(X)$ where g is a sufficiently regular function? There are many economic and financial examples showing that this question is relevant.

- A derivative security is a contract whose payoff is a function g of the price of an underlying asset, like a stock or an index. For example, a call option on a stock, with exercise price K and maturity T, is a contract which pays $Y_T = g(X_T) = \max(X_T - K; 0)$ where X_T is the date-T price of the stock.

- An other more simple transformation lies in the link between prices and logarithmic returns. Assume to simplify that a stock price is equal to 1 at date 0 and equal to X_t at date-t. The logarithmic return on the interval $[0; t]$ is defined by:

$$Y_t = \ln\left(\frac{X_t}{1}\right) = \ln(X_t) \tag{1.8}$$

  The question is then to determine the probability distribution of the return, starting from a given distribution of the price.

- In microeconomic models, the future random wealth of an agent is transformed by a utility function to measure satisfaction. Moreover, to compare individual risk aversions, it is necessary to study concave transformations of utility functions.

The following proposition formally establishes the link between f_X and f_Y. We will illustrate this result in chapter 3 by determinig the density of a price starting from the density of a return and vice versa.

Proposition 38 *Let X denote a variable with density f_X and g a strictly monotone, continuously differentiable function from $\mathbb{R}$ to $\mathbb{R}$. The density f_Y of $Y = g(X)$ is defined by:*

$$\begin{aligned} f_Y(x) &= \frac{f_X(g^{-1}(x))}{|g'(g^{-1}(x))|} \quad \textit{if} \quad x \in Y(\Omega) \\ &= 0 \quad \textit{otherwise} \end{aligned}$$

where $Y(\Omega) = \{y \in \mathbb{R} \ / \ y = Y(\omega) \text{ for } \omega \in \Omega\}$.

Chapter 2

Moments of a random variable

The standard theory of portfolio choice, developed in the fifties by Harry Markowitz (1952) says that investors realize a tradeoff between return and risk when they build a portfolio. As return on a future period of time is a random variable, investors use a measure of return called "expected return", which is a weighted average of possible future returns. They measure risk by a simple function of the deviations of possible returns with respect to the expected return. It is called the variance of returns.

Mathematically speaking, investors realize a tradeoff between expectation and variance. The notions of expectation and variance of a random variable are then fundamental in most financial models. In this chapter, we first distinguish discrete and continuous variables to define expectation and variance, before presenting the genaral definition. We then move on to skewness and kurtosis which allow to analyze moresophisticated properties of stock returns like assymetry or "peakedness" of the distribution of returns.

We introduce covariances and correlations, which measure the intensity of the linear relationship between random variables. These tools are also important in portfolio management because of the principle of diversification. Diversifying risk allows investors to get a better expected return without taking more risk. It is then intuitive that including several stocks in a portfolio achieves better diversification when the returns on these stocks are not strongly linked, that is when the covariance between them is low.

2.1 Mathematical expectation

The mathematical expectation of a random variable is the technical translation of the intuitive concept of average of a sequence of numbers.

Consider one more time example 29 of the lotto ticket (chapter 1) and assume that the organizer of the game decides to attribute an equal share of prizes to the three categories of winners. Assume that 120 players have each bought one ticket (which costs \$1) and played different combinations. The amount of prizes to be shared is \$120. Each category of winners has to share \$40. The unique winner with three correct numbers receives \$40. The 21 winners with two correct numbers receive $\$\frac{40}{21}$ each and the 63 winners with only one correct number obtain $\$\frac{40}{63}$ each. Before the official draw, a player who wants to value his average gain weighs the amounts he can win by the corresponding probability of winning. Before the official draw, the player can expect the following average gain:

$$\$40 \times \frac{1}{120} + \$\frac{40}{21} \times \frac{21}{120} + \$\frac{40}{63} \times \frac{63}{120} = \$1$$

Obviously, in this overly simplified example, we find that the expected gain is equal to the initial price of the ticket because we assume that all players' stakes are redistributed (and they chose different combinations). In this case, we can say the game is fair. Real lottos are not that fair because the organizer keeps around half of the stakes.

2.1.1 Expectations of discrete and continuous random variables

Definition 39 *1) Let X denote a discrete random variable with support $\{x_1, ..., x_n\}$, $x_i \in \mathbb{R}$ for any i. Let $p_i = P(X = x_i)$ for $i = 1, ..., n$; the* ***expectation*** *of X under probability P (also called "first-order moment") is the quantity (if it exists), denoted $E(X)$, and defined by:*

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

2) Let X be a continuous variable with density f_X and CDF F_X; the ***expectation*** *of X under probability P is the quantity (if it exists), denoted $E(X)$, defined by:*

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x) dt = \int_{-\infty}^{+\infty} x dF_X(x)$$

If n and the x_i are finite in point (1) of definition 39, the expectation exists.

Moreover, if $X = \mathbf{1}_B$ then $E(X) = E(\mathbf{1}_B) = P(B)$.
Therefore, if $X=\sum_{i=1}^{n} x_i \mathbf{1}_{B_i}$ with $B_i = \{X = x_i\}$, we get :

$$E(X) = E\left(\sum_{i=1}^{n} x_i \mathbf{1}_{B_i}\right) = \sum_{i=1}^{n} x_i E\left(\mathbf{1}_{B_i}\right) = \sum_{i=1}^{n} x_i p_i$$

To be completely rigorous, we should note E_P instead of E since the expectation depends on the probability measure P. However, E is sufficient when there is no ambiguity about the probability measure under which expectation is defined.

Nevertheless, it is worth to notice that arbitrage pricing models are based on a probability change. So, financial models often distinguish the "real" probability measure, denoted P, and the "risk-neutral" probability measure, denoted Q. It explains why sometimes it is necessary to denote expectations as E_P or E_Q (section 2.4 gives more details about probability changes).

Discrete and continuous variables are important cases but a general definition is needed because many random variables are neither discrete nor continuous. For example, if X is discrete and Y continuous, $X + Y$ is neither continuous nor discrete.

2.1.2 Expectation: the general case

The expectation of a general random variable is built through a convergence process starting with variables with finite support (as in definition 39). It is then generalized to positive random variables and, finally, to general variables.

Definition 40 *Let $\mathcal{V}$ denote the set of random variables with finite support*[1] *defined on a probability space $(\Omega, \mathcal{A}, P)$ and X a positive (general) random variable. The* ***expectation*** *of X, denoted $E(X)$ or $\int_\Omega XdP$, is the quantity (if it exists) defined by*[2]*:*

$$E(X) = \int_\Omega XdP = \sup_{Y \in \mathcal{V}} \{E(Y),\ Y \leq X\}$$

We are able to calculate $E(Y)$ for any $Y \in \mathcal{V}$ because Y is discrete (definition 39). Therefore, the above definition means that the expectation of a positive random variable X may be calculated as the upper bound of the expectations of all discrete variables lower than X. The definition may also be interpreted by saying that a positive random variable can be written as the limit of a sequence of discrete random variables. In fact, the idea behing this definition is the same as the one used to define the Riemann integral of a function as the limit of the integrals of a sequence of step functions.

To understand the notation $\int_\Omega XdP$, remember how we defined $E(X)$ for a continuous variable. The expectation was the integral of x with respect to the CDF F_X. In the general case, the expectation is the integral of X with respect to P. It explains the notation as an integral on Ω.

To address the general case, that is when the sign of X is unknown, it is enough to remark that a random variable X may be decomposed as follows:

$$X = X^+ - X^-$$

with $X^+ = \max(X; 0)$ et $X^- = \max(-X; 0)$. The definition of $E(X)$ is then deduced from the preceding definition.

[1] Variables in $\mathcal{V}$ are also called simple variables.

[2] Assume that f is a function; the upper bound $\sup_{x \in A} f(x)$ is the lowest number greater than all the numbers $f(x)$ for $x \in A$.

Definition 41 *Let X denote a real random variable; the **expectation** of X, denoted $E(X)$, is the quantity (if it exists) defined by:*

$$E(X) = E(X^+) - E(X^-)$$

This definition is consistent with definition 40 since X^+ and X^- are positive.

In each of the preceding definitions we mentioned that $E(X)$ may not exist; when it exists, X is said integrable with respect to P, or simply integrable when there is no confusion about P. Finally, remark that $E(X)$ exists as soon as $E(|X|)$ exists, simply because $|X| = X^+ + X^-$.

The essential properties of expectations are summarized in the following proposition.

Proposition 42 *Let X, Z denote two integrable random variables and A, B two events in $\mathcal{A}$;*

1) $X = \mathbf{1}_A \Rightarrow E(X) = P(A)$

2) $0 \leq X \leq Z \Rightarrow 0 \leq E(X) \leq E(Z)$

3) $\{X \geq 0 \text{ and } A \subset B\} \Rightarrow E(X\mathbf{1}_A) \leq E(X\mathbf{1}_B)$

4)$\forall c \in \mathbb{R}, E(cX) = cE(X)$

5) $E(X + Z) = E(X) + E(Z)$

6) $|E(X)| \leq E(|X|)$

Proof. 1) $X = \mathbf{1}_A$ is equal to 1 with probability $P(A)$ and equal to 0 with probability $1 - P(A)$. A direct application of definition 39 gives the result $E(X) = P(A)$.

2) In definition 40 choose $Y = 0$ as the variable in $\mathcal{V}$. It follows that $E(X) \geq E(Y) = 0$.

In the same way, $Z \geq X$, implies that:

$$\sup_{Y \in \mathcal{V}} \{E(Y),\ Y \leq X\} \leq \sup_{Y \in \mathcal{V}} \{E(Y),\ Y \leq Z\} \tag{2.1}$$

As a consequence, $E(Z) \geq E(X)$.

3) We know that $A \subset B$ and $X \geq 0$, consequently $X\mathbf{1}_A \leq X\mathbf{1}_B$ and the result is a direct consequence of (2).

4) If $X \in \mathcal{V}$, the result is obvious and it is also the case if X is positive and $c > 0$. More generally, decompose X and cX in $X^+ - X^-$ and $(cX)^+ - (cX)^-$. It is worth to note that if c is negative, $(cX)^- = -cX^+$ and $(cX)^+ = -cX^-$, therefore:

$$\begin{aligned} E(cX) &= E\left((cX)^+\right) - E\left((cX)^-\right) \\ &= -cE\left(X^-\right) + cE\left(X^+\right) \\ &= -c\left(-E(X)\right) = cE(X) \end{aligned}$$

5) The proof is the same as in point (4) when considering, first, positive variables, and then decomposing X in $X^+ - X^-$.

6) $|X| = X^+ + X^-$ implies $E\left(|X|\right) = E(X^+) + E(X^-) \geq |E(X^+) - E(X^-)|$; in fact, for two positive numbers x and y, we know that $x + y > x - y$ and $x + y > y - x$. ■

Remark 43 *When a sample* $(x_1, x_2, ..., x_n)$ *of a random variable* X *is observed, an estimate of* $E(X)$ *is the average* $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2.1.3 Illustration: Jensen's inequality and Saint-Petersburg paradox

The theory of decision making under uncertainty is based on a number of assumptions about agents' preferences. In the usual framework of microeconomic textbooks, agents maximize the expectation of a utility function by choosing amounts of consumption and investment. In a one-period model, agents consume at dates 0 and 1, date-1 consumption being financed by the payoffs of investments chosen at date 0. It follows that date-1 consumption is a random variable X and the agent maximizes $E[u(X)]$ under a budget constraint, where u stands for his utility function.

One of the usual assumptions is that agents are risk averse. It means that an agent being offered a lottery paying 0 or 100 with equal probabilities would prefer, instead of playing the lottery, to get $50 = \frac{1}{2}(0 + 100)$ for sure. In other words, such an agent prefers to get the expectation $E(X)$ for sure, instead of getting the random consumption X.

Risk aversion can then be characterized by:

$$E[u(X)] \leq u[E(X)]$$

where X and $u(X)$ are assumed integrable.

The Jensen's inequality in the following proposition allows to characterize the utility functions of risk averse agents.

Proposition 44 *Let X denote an integrable random variable and u a concave*[3] *function from $\mathbb{R}$ to $\mathbb{R}$ such that $u(X)$ is integrable; we then have:*

$$E[u(X)] \leq u[E(X)] \tag{2.2}$$

To illustrate this important result, consider a random variable taking two values x_1 and x_2 with probabilities p and $1 - p$. Jensen's inequality writes:

$$pu(x_1) + (1 - p)u(x_2) \leq u(px_1 + (1 - p)x_2)$$

The curve representing a concave function $u(x)$ is then always above the line joining $(x_1, u(x_1))$ and $(x_2, u(x_2))$.

[3] A function f is concave if for any (x, y) and any $\lambda \in [0; 1]$, $f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$

Remark 45 *a) If u is convex, inequality 2.2 is reversed.*

b) If u is strictly concave, the inequality is strict as soon as X is not always equal to its expectation (X is not a constant).

Strict concavity of utility functions has two different meanings. The first one is that marginal utility of consumption is decreasing because $u' > 0$ and $u'' < 0$. The more an agent consumes, the less a supplementary consumption unit brings satisfaction. This is one of the first assumptions we can find in undergraduate economics textbooks. But this decreasing marginal utility property has nothing to do with randomness and probability.

The second interpretation, appearing in Jensen's inequality, is risk aversion which is obviously linked to randomness and probability. It is then important to understand that the same mathematical property (concavity of the utility function) has two completely different interpretations. It is sometimes considered as a weakness of the expected utility theory.

Risk aversion and concavity of utility functions can be illustrated by the Saint-Petersburg paradox[4].

Example 46 ***The St-Petersburg paradox***

In a fair coin tossing, a player wins 2^n monetary units if Tails appears for the first time on the n-th toss and the game stops. When Heads occurs, the coin is flipped one more time.

Let N denote the random variable counting the number of tosses before the game stops. Each coin toss being fair, the probability of getting Heads (Tails) is 1/2. Consequently, $P(N = n) = \frac{1}{2^n}$ since we need a sequence of $n - 1$ Heads followed by a Tails (successive tosses are independent). But the gain is then equal to 2^n. Therefore, the random gain of the player, denoted X, has an expected value defined by:

$$E(X) = \sum_{n=1}^{+\infty} 2^n \times P(N = n) = \sum_{n=1}^{+\infty} 2^n \times \frac{1}{2^n} = +\infty$$

If economic agents were to maximize their expected wealth, they would be ready to pay ***any*** *price to play this game since the expected gain is infinite.*

[4] The initial contribution of Nicholas Bernoulli (1695-1726) concerning this paradox was published in 1738 and then published in English in 1954 in Econometrica.

All experiments show that people are much more risk averse. They accept to bet only a small amount of money to participate.

Bernoulli proposed an alternative to wealth maximization as the maximization of the expectation of a concave transformation of wealth, namely the logarithm of wealth. We get in this case:

$$E(\ln(X)) = \sum_{n=1}^{+\infty} \ln(2^n) \times \frac{1}{2^n} = \ln(2) \sum_{n=1}^{+\infty} \frac{n}{2^n}$$

and we know that:

$$\sum_{n=1}^{+\infty} \frac{n}{2^n} = \sum_{n=1}^{+\infty} \sum_{k=n}^{+\infty} \frac{1}{2^k} = 2$$

It follows that $E(\ln(X)) = 2\ln(2) = \ln(4)$*; the player is then indifferent between playing the game and getting 4 monetary units for sure. In other words, with a logarithmic utility function, he is ready to pay 4 units to play, not more, even if the expected value of the game is infinite. This famous example shows that other elements than the expected final wealth play a role in decision making.*

2.2 Variance and higher moments

As shown by the Saint Petersburg paradox, agents take into account, not only the expected final wealth, but also the risk associated with X. Many risk measures have been proposed in the financial literature but the most popular, especially in portfolio choice theory, is the variance of returns. This popularity is partly due to the statistical properties it carries and to the simplicity of the portfolio choice models it allows to build. The variance of returns is also a fundamental variable in option pricing models.

2.2.1 Second-order moments

Definition 47 *The* ***second-order moment*** *of a random variable X, denoted as $\mu_2(X)$ is the quantity (if it exists) defined by:*

$$\mu_2(X) = E(X^2)$$

When $\mu_2(X)$ exists, X is said square-integrable.

Definition 48 *Let X be a square-integrable random variable; the* ***variance*** *of X, denoted $V(X)$ or $\sigma^2(X)$ is the quantity*[5]*:*

$$V(X) = \sigma^2(X) = E\left[(X - E(X))^2\right]$$

$V(X)$ is also called the central second-order moment. In fact, if $Y = X - E(X)$, we get $V(X) = \mu_2(Y)$. Y is a zero-mean random variable, that is, $E(Y) = 0$.

Proposition 49 *Let X be a square-integrable random variable;*

$$V(X) = E\left[(X - E(X))^2\right] = E(X^2) - E(X)^2 \tag{2.3}$$

Proof.

$$\begin{aligned} E\left[(X - E(X))^2\right] &= E\left[X^2 - 2XE(X) + E(X)^2\right] && (2.4)\\ &= E\left(X^2\right) - 2E\left[XE(X)\right] + E(X)^2 && (2.5)\\ &= E(X^2) - 2E(X)^2 + E(X)^2 && (2.6)\\ &= E(X^2) - E(X)^2 && (2.7) \end{aligned}$$

■

[5]Now σ has nothing to do with σ-algebras presented in chapter 1. We hope that this common notation will not be confusing. In general it is not.

Example 50 *Suppose* $card(\Omega) = 4$, $P(\omega) = 0.25$ *for every* ω, *and* X *defined by:*

$$X = \begin{bmatrix} 2 \\ 3 \\ -1 \\ 0 \end{bmatrix}$$

$E(X) = 1$ *and the corresponding zero-mean variable* $Y = X - E(X)$ *is equal to:*

$$Y = \begin{bmatrix} 1 \\ 2 \\ -2 \\ -1 \end{bmatrix}$$

The variance of X *and* Y *are equal and given by:*

$$V(X) = V(Y) = 0,25 \times \left(1^2 + 2^2 + (-2)^2 + (-1)^2\right) = 2.5$$

This example also shows that variance is invariant by translation. In other words:

$$V(X) = V(X + c) \tag{2.8}$$

for any real number c.

Remark 51 *When a sample* $(x_1, x_2, ..., x_n)$ *of a random variable* X *is used in empirical studies, an unbiased estimate of the variance is given by:*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.9}$$

The coefficient $n-1$ *instead of* n *comes from the fact that the expectation of* X *is not known and replaced by its estimate* $\overline{x}$.

Definition 52 *Let* X *be a square-integrable variable; the* ***standard deviation*** *of* X, *denoted* $\sigma(X)$, *is the square root of* $V(X)$*:*

$$\sigma(X) = \sqrt{V(X)}$$

In Markowitz portfolio theory, agents are assumed to make their choices in the expectation-variance space or in the expectation-standard deviation space. For a given expected return, they minimize the variance (standard deviation) of their portfolio return. However, a portfolio contains several assets and the relationships between the individual returns must be taken into account to evaluate the variance of the portfolio[6]. These relationships are quantitatively measured by covariances and correlations. They are developed in section 2.3.

[6] To give an idea, the standard deviation of yearly U.S stock returns was around 20% over the 20th century.

2.2.2 Skewness and kurtosis

Variance is commonly used as a measure of risk in portfolio management. However, variance weighs in the same way returns above the mean and returns below the mean. Therefore, measuring risk by the variance of returns implicitly assumes that the distribution of returns is symmetrical with respect to the mean. In fact, it is intuitive that agents link risk more to potential losses than to potential gains. Consequently, when the distribution is not symmetrical, variance may not be a good measure of risk. Skewness is a way to measure the assymetry of a probability distribution and many empirical studies show that stock returns are negatively skewed, meaning that that they exhibit more large negative returns than large positive returns and more small positive returns than small negative returns.

In the next chapter, we will describe the main probability distributions appearing in financial models. The most commonly used to describe returns is the Gaussian distribution which is symmetrical with respect to its mean. However, when looking at stock returns, we also observe that extreme returns are more frequent than what is expected under the Gaussian distribution. Kurtosis is the standard measure to take into account these extreme returns.

Definition 53 *The **moment of order** n of a random variable X, denoted as $\mu_n(X)$ is the quantity (if it exists) defined by:*

$$\mu_n(X) = E(X^n)$$

Definition 54 *Let X denote a random variable with a finite moment of order 3. The **skewness** of X, denoted $Sk(X)$ is defined by:*

$$Sk(X) = \frac{\mu_3(X - E(X))}{\sigma(X)^3} \tag{2.10}$$

In empirical studies using a sample $(x_1, x_2, ..., x_n)$ of a random variable X, $Sk(X)$ is estimated by:

$$\widehat{Sk} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \overline{x}}{s} \right)^3 \tag{2.11}$$

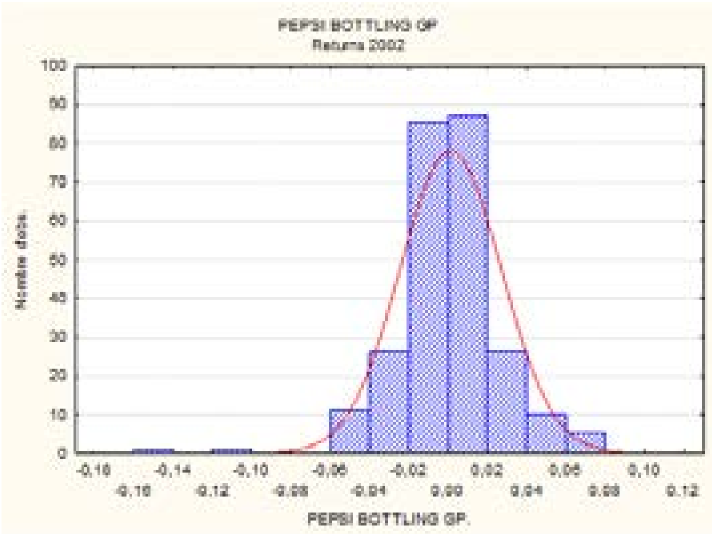

Figure 2.1: Histogram of 2002 PEPSI daily returns

Figure 2.1 shows the histogram of daily returns on the PEPSI stock in 2002. We observe some largely negative returns and $\widehat{Sk} = -0.73$ this year. Skewness is used in some portfolio choice models to take into account the assymetric perception of risk by investors. The thin line on figure 2.1 shows what we could expect if the distribution of returns were Gaussian (see chapter 3), that is symmetrical.

Definition 55 *The* ***kurtosis*** *of a random variable X is defined by:*

$$\kappa(X) = \frac{\mu_4(X - E(X))}{\sigma^4} \tag{2.12}$$

The ***excess kurtosis*** *of a random variable X is defined by:*

$$e\kappa(X) = \kappa(X) - 3 \tag{2.13}$$

In case of Gaussian returns, $\kappa = 3$. It explains the way excess kurtosis is defined. As mentioned before, $\kappa(X)$ measures the importance of extreme returns in a probability distribution. In the case of PEPSI returns, $\kappa(X) = 8.93$ which is very high with respect to what is expected for Gaussian returns. It is common for single stocks to observe a large kurtosis. When portfolios are considered, the diversification effect often generates returns with a lower kurtosis and a skewness closer to 0.

Before studying relationships between random variables by means of covariances and correlations, we give several important properties of the vector space of integrable/square-integrable random variables. Even if they seem abstract, they are very useful in general arbitrage pricing models.

2.3 The vector space of random variables

Let $\mathcal{L}^0(\Omega, \mathcal{A})$ denote the set of real random variables defined on $(\Omega, \mathcal{A})$. Addition of variables and multiplication by a real number may be intuitively defined by:

$$\begin{aligned} \forall\omega &\in \Omega,\ (X+Y)(\omega) = X(\omega) + Y(\omega) \\ \forall\omega &\in \Omega, \forall c \in \mathbb{R}, (cX)(\omega) = cX(\omega) \end{aligned}$$

It is quite obvious that $\mathcal{L}^0(\Omega, \mathcal{A})$ is a vector space, the null random variable being the neutral element for addition. This space is very general and it is impossible to enrich its structure without constraints. But, if we restrict the analysis to integrable variables, a norm can be defined on this subspace. Moreover, considering only square-integrable random variables allows to define an inner product, inducing nice geometrical properties.

It is worth to notice that integrable or square-integrable variables can be considered only if a probability measure P has been specified. Some technical precautions are needed to introduce P and to consider random variables as elements of vector spaces. They are presented in the next sub-section.

2.3.1 Almost surely equal random variables

When we say that two n-dimensional vectors in $\mathbb{R}^n$, say x and y, are equal, there is no ambiguity. It means that the distance $d(x, y)$ between x and y is zero, as soon as a metric[7] d has been defined on $\mathbb{R}^n$. Using the usual metric on $\mathbb{R}^n$, it means:

$$x = y \Leftrightarrow \sum_{i=1}^{n}(x_i - y_i)^2 = 0 \tag{2.14}$$

or, equivalently, $x_i = y_i$ for $i = 1, ..., n$.

Consider now the set of Riemann-integrable functions defined on an interval $[a; b]$. The usual metric on this space is defined by:

$$d(f, g) = \int_a^b |f(x) - g(x)|\, dx \tag{2.15}$$

In fact, it is not really a metric because we can have $d(f, g) = 0$ with $f \neq g$! Assume $f(x) = 0$ on $[a; b]$ and $g(x) = 0$ on $[a; b[$ but $g(b) = 1$. It

turns out that $d(f, g) = 0$ because f and g only differ on a set containing one point.

It would be the same if f and g were different on a finite or countable set of points. To deal with a "real" metric, it is necessary to avoid these cases. The problem is solved by defining a binary relation $\mathcal{R}$ as follows:

$f\mathcal{R}g$ if f and g are equal, except on a finite or countable set of points.

$\mathcal{R}$ is an equivalence relation. It is reflexive ($f\mathcal{R}f$), symmetric ($f\mathcal{R}g \Leftrightarrow g\mathcal{R}f$) and transitive ($f\mathcal{R}g$ and $g\mathcal{R}h \Rightarrow f\mathcal{R}h$).

Consequently, we do not define the metric d on the set of integrable functions, but on the $\mathcal{R}$-equivalence classes of integrable functions defined on $[a; b]$. If $\hat{f}$ and $\hat{g}$ stand for two equivalence classes containing respectively f and g, the metric $d(\hat{f}, \hat{g})$ can be defined by using formula **??**, for any pair (f, g) of functions belonging to $\hat{f} \times \hat{g}$.

The same "trick" is used on the space of integrable random variables, using the equivalence relation "almost-surely equal".

[7] A metric d on a space $\mathcal{S}$ is a mapping from $\mathcal{S} \times \mathcal{S}$ to $\mathbb{R}^+$ such that 1) $d(x, y) = 0$ iff $x = y$ 2) $d(x, y) = d(y, x)$ and 3) $d(x, z) \leq d(x, y) + d(y, z)$ (called the triangular inequality).

Definition 56 *Two random variables X and Y defined on $(\Omega, \mathcal{A}, P)$ are equal P-almost-surely(or P-almost everywhere)*[8] *if:*

$$P\left(\omega \in \Omega \ / \ X(\omega) = Y(\omega)\right) = 1$$

We could also write in short: $X = Y \ a.s \Leftrightarrow P(X = Y) = 1$.

An alternative presentation can be used, based on negligible events.

Definition 57 *Let $(\Omega, \mathcal{A}, P)$ a probability space; $A \in \mathcal{A}$ is P-negligible if $P(A) = 0$.*

Definition 56 is then equivalent to say that two variables are a.s equal if they differ on a negligible event.

$\mathcal{L}^1(\Omega, \mathcal{A}, P)$ will denote the set of P-integrable random variables defined on $(\Omega, \mathcal{A}, P)$.

Proposition 58 *The binary relation $\mathcal{R}$ defined on $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ by:*

$$X \mathcal{R} Y \Leftrightarrow X = Y \ P\text{-}a.s$$

is an equivalence relation.

[8] We note briefly P-a.s or P-a.e or simply a.e or a.s if no confusion can arise about the probability used in the model.

2.3.2 The space $L^1(\Omega, \mathcal{A}, P)$

Let now $L^1(\Omega, \mathcal{A}, P)$[9] be the set of $\mathcal{R}$-equivalence classes of random variables in $\mathcal{L}^1(\Omega, \mathcal{A}, P)$. In the same way, $L^0(\Omega, \mathcal{A}, P)$ denotes the set of $\mathcal{R}$-equivalence classes of random variables in $\mathcal{L}^0(\Omega, \mathcal{A})$.

Proposition 59 *1) $L^1(\Omega, \mathcal{A}, P)$ is a vector subspace of $L^0(\Omega, \mathcal{A}, P)$.*

2) The mapping from $L^1(\Omega, \mathcal{A}, P)$ to $\mathbb{R}_+$, denoted $X \to \|X\|_1$ and defined by:

$$X \to \|X\|_1 = E(|X|)$$

is a norm[10].

3) The mapping from L^1 to $\mathbb{R}$ which links X to $E(X)$, denoted $X \to E(X)$, is a positive linear mapping.

Proof. 1) The fact that $L^1(\Omega, \mathcal{A}, P)$ is a vector subspace of $L^0(\Omega, \mathcal{A}, P)$ is a direct consequence of points (4) and (5) in proposition 42.

2) We now prove that $X \to \|X\|_1$ is a norm. $\|X\|_1 = 0 \Leftrightarrow X = 0$ P-a.s. It is then sufficient to show that:

$$\|X + Y\|_1 \leq \|X\|_1 + \|Y\|_1$$

But for any $\omega \in \Omega$, we have $|X(\omega) + Y(\omega)| \leq |X(\omega)| + |Y(\omega)|$, therefore:

$$E(|X + Y|) \leq E(|X|) + E(|Y|)$$

It is also obvious to see that $\|\alpha X\|_1 = |\alpha| \|X\|_1$using the properties of the absolute value function.

3) Linearity and positivity of the expectation come directly from proposition 42. ■

Linearity of expectations is essential for economic interpretations. In fact, a large part of the financial literature on arbitrage uses, as we will see later on, a probability change to express prices as expectations of discounted

[9] or simply L^1 if no confusion can arise about the probability measure.
[10] A norm defined on a vector space $\mathcal{S}$ is a mapping from $\mathcal{S}$ to $\mathbb{R}^+$, denoted $\|.\|$, satisfying:
a) $\|x\| = 0$ if and only if $x = 0$
b) $\forall x \in \mathcal{S}, \forall c \in \mathbb{R}, \|cx\| = |c| \|x\|$
c) $\forall (x, y) \in \mathcal{S} \times \mathcal{S}$, $\|x + y\| \leq \|x\| + \|y\|$

future cash-flows. In this framework, the linearity property simply says that the value of a portfolio is equal to the sum of the values of the securities contained in the portfolio. It is also the intuitive idea behind the no-arbitrage assumption.You cannot buy two stocks for \$50 and immediately resell them \$30 each. The price of one stock must be \$25 if there are no arbitrage opportunities.

Proposition 59 allows to consider an integrable random variable as a vector (element of the vector space $L^1(\Omega, \mathcal{A}, P)$). As a norm induces a metric by $d_1(X,Y) = \|X - Y\|_1$, $L^1(\Omega, \mathcal{A}, P)$ is also a metric space as any normed vector space. Convergence associated to the metric d_1 is called L_1 convergence or convergence in mean.

Definition 60 *A sequence of random variables $(X_n, n \in \mathbb{N}^*)$ converges in L^1 to a limit $X \in L^1$ if and only if:*

$$\lim_{n \to +\infty} E\left(|X_n - X|\right) = 0$$

We then write $X_n \xrightarrow{L_1} X$.

Unfortunately, the L_1-norm is not deduced from an inner product. Some intuitive geometrical results, valid in finite-dimensional spaces like $\mathbb{R}^n$, are not true in L_1. It is especially the case for the projection theorem and the Riesz representation theorem. To keep these convenient properties true, it is neccessary to restrain the space to square-integrable variables.

2.3.3 The space $L^2(\Omega, \mathcal{A}, P)$

Let now $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ denote the vector space of square-integrable random variables and $L^2(\Omega, \mathcal{A}, P)$ the space of equivalence classes for the binary relation "almost-surely equal" defined on $\mathcal{L}^2(\Omega, \mathcal{A}, P)$. We get the following proposition.

Proposition 61 *1) $L^2(\Omega, \mathcal{A}, P)$ is a vector subspace of $L^1(\Omega, \mathcal{A}, P)$*

2) Let X and Y be 2 elements of $L^2(\Omega, \mathcal{A}, P)$; the product XY is in $L^1(\Omega, \mathcal{A}, P)$.

Proof. 1) It is a direct consequence of:

a) the linearity of the expectation operator;

b) the fact that a square-integrable variable is also integrable.

2) We are going to show that

$$E(XY)^2 \leq E(X^2)E(Y^2) \tag{2.16}$$

Let $Z = X + tY$ with $t \in \mathbb{R}$:

$$\begin{aligned} E\left(Z^2\right) &= E\left(X^2 + 2tXY + t^2Y^2\right) \geq 0 \\ &= E\left(X^2\right) + 2tE\left(XY\right) + t^2E\left(Y^2\right) \end{aligned}$$

The second line is a polynomial of degree 2 in t. It is always positive or equal to 0, therefore its reduced discriminant Δ' is negative or equal to 0. But Δ' is equal to:

$$\Delta' = E\left(XY\right)^2 - E\left(X^2\right)E\left(Y^2\right)$$

It shows inequality 2.16. The RHS of this inequality is finite since X and Y belong to L^2. It implies that the LHS is also finite and shows that XY is integrable. ■

This proposition allows to define an inner product on L^2.

Proposition 62 *The mapping from $L^2 \times L^2$ to $\mathbb{R}$ denoted by $\langle .,. \rangle$ and defined by:*

$$(X,Y) \to \langle X,Y \rangle = E(XY) \tag{2.17}$$

is an ***inner product*** *on L^2.*

The induced norm is defined by:

$$\|X\|_2 = \sqrt{\langle X,X \rangle} = \sqrt{E(X^2)} \tag{2.18}$$

and the induced metric d_2 is defined by $d_2(X,Y) = \|X-Y\|_2$.

Proof. The mapping $\langle .,. \rangle$ is positive because $\langle X,X \rangle = E(X^2) > 0$ if X is not (P-a.s) equal to 0. Bilinearity is obvious since the expectation operator is linear. ■

As it was the case for L^1, we can define a convergence on L^2 simply called L^2-convergence or convergence in quadratic mean.

Definition 63 *A sequence $(X_n, n \in \mathbb{N}^*)$ converges in L^2 with a limit $X \in L^2$ if and only if :*

$$\lim_{n \to +\infty} E\left[(X_n - X)^2\right] = 0$$

L^2 is in fact a Hilbert space[11]. At a first glance this distinction appears purely technical but it has important consequences on the properties of vectors in L^2. There are two important well known theorems, valid on $\mathbb{R}^n$, that are still true on Hilbert spaces and especially on L^2. These are the projection theorem and the Riesz representation theorem.

The Riesz representation theorem

First remember that, on $\mathbb{R}^2$, a linear mapping $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by:

$$\forall x \in \mathbb{R}^2, f(x) = a_1 x_1 + a_2 x_2 \tag{2.19}$$

where a_1 and a_2 are real numbers and $x' = (x_1, x_2)$. It means that the numbers (a_1, a_2) represent the mapping f. Moreover, the vector $a' = (a_1, a_2)$ has the same dimension than the vector $x \in \mathbb{R}^2$ and $f(x)$ can be written as the inner

[11] A Hilbert space H is a normed vector space where the norm comes from an inner product, and the metric induced by the norm makes H complete as a metric space (any Cauchy sequence in H is convergent).

product[12] of a and x. To summarize, we could say that any mapping f from $\mathbb{R}^2$ to $\mathbb{R}$ is represented by a vector $a \in \mathbb{R}^2$. It is the Riesz representation theorem in the two-dimensional space. Fortunately, this result is still valid in $L^2(\Omega, \mathcal{A}, P)$.

Theorem 64 *Let f denote a continuous linear mapping from L^2 to $\mathbb{R}$; there exists $Y_f \in L^2$ such that for any $X \in L^2$:*

$$f(X) = \langle X, Y_f \rangle = E(XY_f)$$

Suppose that X is the payoff of a financial asset and $f(X)$ denotes the initial price of this asset. The mapping $X \rightarrow f(X)$ is called a valuation operator or a pricing kernel. Assume for example that $Card(\Omega) = N$. The preceding theorem says that there exists a vector Y_f such that

$$f(X) = \langle X, Y_f \rangle = E(XY_f) = \sum_{i=1}^{N} X(\omega_i) Y_f(\omega_i) P(\omega_i) \tag{2.20}$$

Consider the simplest case where $X = e_i = \mathbf{1}_{\{\omega_i\}}$, the Arrow-Debreu security contingent on ω_i.

$$f(e_i) = \langle e_i, Y_f \rangle = P(\omega_i) Y_f(\omega_i) \tag{2.21}$$

$f(e_i)$ is the market price at date 0 of a security which pays 1 at date 1 if the true state of nature is ω_i. It is equal to the product of the probability $P(\omega_i)$ and $Y_f(\omega_i)$. In fact $Y_f(\omega_i)$ depends on two elements. First, the risk-free rate because you pay today the price $f(e_i)$ to receive the contingent payoff at a future date. The second element influencing $Y_f(\omega_i)$ is risk aversion. If you are highly risk-averse, $Y_f(\omega_i)$ will be lower because you are not ready to pay much to receive the contingent payoff.

Consequently, $Y_f(\omega_i)$ is interpreted as the risk-adjusted discount factor for the state ω_i. But any general financial asset X, defined by its payoffs, is a portfolio of such Arrow-Debreu securities.

$$X = \sum_{i=1}^{N} x_i e_i$$

with $X(\omega_i) = x_i$. It follows that:

$$f(X) = \langle X, Y_f \rangle = \sum_{i=1}^{N} x_i f(e_i) = \sum_{i=1}^{N} x_i Y_f(\omega_i) P(\omega_i) \tag{2.22}$$

[12] Remember that on $\mathbb{R}^n$, the usual inner product of two vectors x and y is defined by $< x, y >= \sum_{i=1,..n} x_i y_i$.

The projection theorem

The second important result in L^2 is the projection theorem. To introduce the result, consider one more time a vector x in $\mathbb{R}^2$ and a convex[13] subset $C \subset \mathbb{R}^2$. On figure 2.2, C is the grey ellipse. z is the point in C which is closest to x. In other words, z is the orthogonal projection of x on C. Consider any vector y whose extremity is in y. It is easy to see that the angle between $x - z$ and $y - z$ will be between 90° and 270°. The cosine of this angle is then negative. But the cosine of an angle between two vectors is proportional to the inner product of these vectors[14]. We then have:

$$< x - z, y - z > \quad \leq 0 \tag{2.23}$$

Here the convexity hypothesis is fundamental because it allows to say that all points in C are on the same side of the tangent to C (on which lies z).

[13] A subset A of a vector space is convex if:
$\forall \lambda \in [0;1], \forall (x,y) \in A \times A, \lambda x + (1-\lambda) y \in A$.

[14] In $\mathbb{R}^n$, the cosine of two vectors x and y is defined by $< x, y > / \|x\| . \|y\|$.

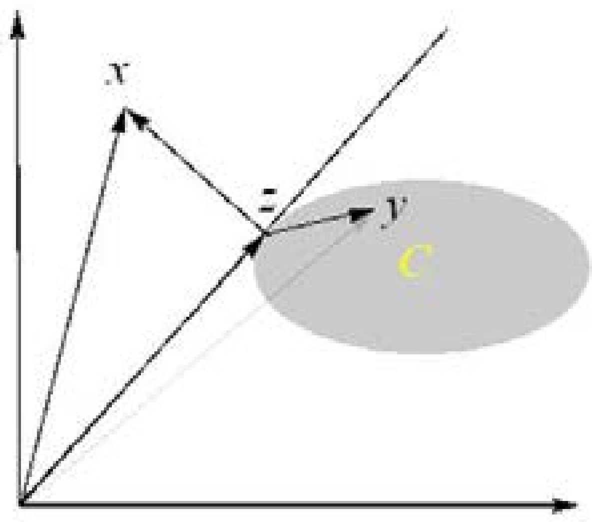

Figure 2.2: Projection theorem in $\mathbb{R}^2$

This result is still valid for square-integrable random variables and also called the projection theorem.

Theorem 65 *Let C denote a non-empty convex subset of L^2 and $X \in L^2$. There exists $Z \in C$ such that:*

$$\langle X - Z, Y - Z \rangle \leq 0 \textit{ for any } Y \in C$$

Z is the **orthogonal projection** of X on C. In chapter 4, we will introduce conditional expectations and show that they can also be interpreted in terms of orthogonal projections.

2.3.4 Covariance and correlation

An economic agent investing in a portfolio of financial securities has to take into account the relationships between assets' returns to make an optimal choice. For example, a portfolio containing several stocks of firms in a given industrial sector is more sensitive to news concerning this specific sector than a portfolio diversified across industries. Covariance is used to measure these relationships between returns.

Definition 66 *Let X and Y denote two random variables in $L^2(\Omega, \mathcal{A}, P)$; the **covariance** between X and Y, denoted $Cov(X,Y)$ (or simply σ_{XY}) is defined as:*

$$cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

Example 67 *Let X and Y be defined on a 4-state space as in table 2.1:*

State	$X(\omega)$	$Y(\omega)$
ω_1	1	3
ω_2	0	1
ω_3	3	1
ω_4	4	3

Table 2.1: Definition of X and Y

We assume that all states are equally likely, so $E(X) = E(Y) = 2$. The corresponding centered variables are given in table 2.2.

State	$X(\omega) - E(X)$	$Y(\omega) - E(Y)$
ω_1	-1	1
ω_2	-2	-1
ω_3	1	-1
ω_4	2	1

Table 2.2: Definition of the centered variables

The covariance is then calculated as follows:

$$cov(X,Y) = \frac{1}{4}\left(-1 \times 1 + (-2) \times (-1) + 1 \times (-1) + 2 \times 1\right) = 0.5$$

This quantity depends, like expectation and variance, on the probability measure P. Covariance gives a quantitative measure of the (linear) relationship between X and Y. Moreover, covariance is bilinear; in other words, for any quadruple a, b, c, d of real numbers and any quadruple of square-integrable random variables X, Y, Z, W we have:

$$Cov(aX + bY, cZ + dW) = ac \times \sigma_{XZ} + ad \times \sigma_{XW} + bc \times \sigma_{YZ} + bd \times \sigma_{YW} \quad (2.24)$$

It means in particular that $Cov(aX, Y) = aCov(X, Y)$.

Remark 68 *Relation 2.24 also allows to write:*

$$V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$$

because $Cov(X,X) = V(X)$.

The magnitude of $Cov(X,Y)$ depends on the magnitude of the values of X and Y. It is then difficult to compare two covariances and to give an economic interpretation in terms of the intensity of the relationship linking two variables. To overcome this problem, we refer to the correlation coefficient which uses standardized variables.

Definition 69 *Let X and Y denote two variables in L^2; the **correlation coefficient** between X and Y, denoted as ρ_{XY}, is defined as:*

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

where $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of X and Y.

ρ_{XY} can also be written $Cov(\frac{X}{\sigma(X)}, \frac{Y}{\sigma(Y)})$, which is the covariance of two variables with unit variance. It is the meaning we give to the word "standardization". In fact, when X and Y are zero-mean random variables, we get:

$$\rho_{XY} = \frac{\langle X, Y \rangle}{\|X\|_2 \|Y\|_2}$$

Here, ρ_{XY} may be interpreted as the cosine of the angle between the two vectors X and Y. It appears that the length of the vectors, which is in fact their standard deviation, is neutralized when measuring a correlation coefficient. The reader must be warned about the fact that a high correlation means a strong linear relationship between two variables, but a low correlation does not always mean a weak relationship. It only means that a strong **linear** relationship does not exist.

In the example of table 2.1, we get:

$$\begin{aligned}
\sigma(X) &= \sqrt{\frac{1}{4}\left((-1)^2 + (-2)^2 + (1)^2 + (2)^2\right)} = \sqrt{2.5} = 1.58 \\
\sigma(Y) &= \sqrt{\frac{1}{4}\left((1)^2 + (-1)^2 + (-1)^2 + (1)^2\right)} = 1 \\
\rho_{XY} &= \frac{0.5}{1.58} = 0.316
\end{aligned}$$

The correlation coefficient is positive but far from 1. We can remark that if X and Y are multiplied by a constant, the correlation coefficient does not change, even if covariance is different. The essential properties of ρ are summarized in the next proposition.

Proposition 70 *Let X and Z denote two random variables in L^2 and a, b, c, d four real numbers:*

$$\begin{aligned} Cov(aX+b, cZ+d) &= ac \times Cov(X,Z) \\ \rho_{aX+b,cZ+d} &= sign(ac) \times \rho_{XZ} \end{aligned}$$

Proof. The first equality comes directly from equality 2.24 by choosing Y and W identically equal to 1. The second equality is obtained by noting that $\sigma(aX+b) = |a|\,\sigma(X)$ and $\sigma(cY+d) = |c|\,\sigma(Y)$. ■

An important property of correlations (and covariances) is linked to independence .

Proposition 71 *If two variables X and Y in L^2 are independent, their correlation (covariance) is equal to zero.*

It is important to note that the proposition is an implication, not an equivalence. We could build counter examples of uncorrelated but non independent variables.

2.4 Equivalent probabilities and Radon-Nikodym derivatives

2.4.1 Intuition

To buy a financial asset, risk averse economic agents are not ready to pay the expected present value of future cash-flows discounted at the risk-free rate. Considering the randomness of cash-flows, they require a risk premium and are ready to pay only a lower amount.

To translate this problem in simple terms using lotteries and keeping the risk-free rate equal to 0, we use the following example. Let X_1 denote the terminal payoff of a lottery, X_1 being 200 or 0 with equal probabilities. Assume that a risk averse agent is ready to pay only $X_0 = 90$ to play the lottery knowing that

$$E(X_1) = \frac{1}{2}[0 + 200] = 100 \tag{2.25}$$

In financial theory, there are two ways to link the terminal expected payoff to the initial price. In a CAPM[15]-like approach, the expectation is discounted by a risk premium such that

$$X_0 = \frac{E(X_1)}{1 + Risk\_premium} = 90 \tag{2.26}$$

The difficulty of this approach is obviously the determination of the risk premium. If X_0 is the equilibrium market price, the risk premium is a complicated function of the utility functions of all agents.

The other approach, mainly used to value derivative securities, is to discount future cash-flows at the risk-free rate (equal to 0 in our example) but to change the probabilities coming in the calculation of the expectation.

[15] CAPM = Capital Asset Pricing Model

In the above example, we are looking for an alternative probability measure Q (different from P) such that

$$X_0 = E_Q(X_1) \tag{2.27}$$

The probabilistic framework is very simple since there are only two states of nature. In fact, $\Omega = \{\omega_1, \omega_2\}$, $X_1(\omega_1) = 200$ and $X_1(\omega_2) = 0$.

Let Q be defined by:

$$\begin{aligned} Q(\omega_1) &= q_1 = 0.45 \\ Q(\omega_2) &= q_2 = 1 - q_1 = 0.55 \end{aligned}$$

It is obvious that we get the desired result $E_Q(X_1) = 90 = X_0$. The probability Q is easily obtained because we have only to solve a system of two equations with two unknowns:

$$q_1 \times 200 + q_2 \times 0 = 90 \tag{2.28}$$
$$q_1 + q_2 = 1 \tag{2.29}$$

This technique is now very common in finance, especially in the valuation of options. However, the idea seems artificial because one can have the feeling that Q depends on the risky asset considered in the calculation. Nevertheless it is very powerful when associated with the no arbitrage assumption, as shown in the following example.

Example 72 *Consider a two-state one-period economy with one risk-free asset (the risk-free rate is still assumed to be 0) and two risky assets defined by:*

$$X_1(\omega_1) = 200 \quad X_1(\omega_2) = 100 \quad X_0 = 130 \tag{2.30}$$
$$Y_1(\omega_1) = 150 \quad Y_1(\omega_2) = 110 \quad Y_0 = 120 \tag{2.31}$$

We first look for a probability Q such that $X_0 = E_Q(X_1)$. We have to solve:

$$130 = 200Q(\omega_1) + 100\,(1 - Q\,(\omega_1))$$

The solution is $Q(\omega_1) = 0.3$.

Look now for a probability Q' such that $Y_0 = E_{Q'}(Y_1)$. The equation to be solved is:

$$150Q'(\omega_1) + 110\,(1 - Q'\,(\omega_1)) = 120$$

and the solution is $Q'(\omega_1) = 0.25$.

Q and Q′ are different. It then seems that changing the probability measure is useless since we get a different probability measure, depending on the risky asset chosen for the calculation.

Fortunately, it is not the case if there are no arbitrage opportunities. An arbitrage opportunity is a portfolio which costs nothing at date 0 (or has a negative cost) and pays positive (at least zero) amounts in all states of nature at date 1. If such opportunities exist, current prices cannot be equilibrium prices.

We show hereafter that if Q *and* Q' *are different, there must be an arbitrage opportunity or, equivalently,* (X_0, Y_0) *cannot be a vector of equilibrium prices.*

Let us build a zero-cost portfolio θ *satisfying :*

$$\theta_X \begin{bmatrix} 200 \\ 100 \end{bmatrix} + \theta_Y \begin{bmatrix} 150 \\ 110 \end{bmatrix} + \theta_Z \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where θ_Z *is the number of units of the risk-free asset whose price and payoffs are equal to 1 (because the risk-free rate is zero). There are infinitely many portfolios satisfying these equations. The question is to know if it is possible to find such a portfolio with a negative cost (it would be an arbitrage opportunity).*

Consider $\theta_X = -2; \theta_Y = 5; \theta_Z = -350$*; we get :*

$$-2 \begin{bmatrix} 200 \\ 100 \end{bmatrix} + 5 \begin{bmatrix} 150 \\ 110 \end{bmatrix} - 350 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The cost of this portfolio is given by: $-2 \times 130 + 5 \times 120 - 350 = -10$

It happens that an agent characterized by a strictly increasing utility function is ready to "buy" (but the price is negative!) an infinite quantity of this portfolio. Consequently (X_0, Y_0) *cannot be a vector of equilibrium prices.*

As the arbitrage portfolio requires a short position on X_1 *and a long position on* Y_1*, the price* X_0 *is too high and* Y_0 *is too low (in relative terms). To get an equilibrium, it is necessary that* $-2X_0 + 5Y_0 = 350$*. As long as this equality is not satisfied, it is possible to build an arbitrage portfolio. To keep things simple, suppose that the price adjustment is realized only on* Y_1*. The equilibrium price is then* $Y_0 = 122$*.*

Recalculate now the probability Q" with this new price.

$$122 = 150Q"(\omega_1) + 110\,(1 - Q"(\omega_1))$$

implies $Q"(\omega_1) = \frac{12}{40} = 0.3$.

A kind of miracle appears in this example! $Q"$ is exactly the probability measure Q when the market is free from arbitrage opportunities. In other words, the probability we build does not depend on the security we use to calculate it, if prices are equilibrium prices. Q characterizes the whole market, not a specific security. It is the reason why the change of probability technique is so powerful for valuing derivative securities.

We let the reader check that when the risk-free rate r is not zero, that is when the price of the risk-free asset is $\frac{1}{1+r}$, the price of a contract X_1 is written;

$$X_0 = \frac{1}{1+r} E_Q(X_1) \qquad (2.32)$$

The economic interpretation of this formula is also very interesting. If the price of any asset is obtained by discounting the expected cash-flows at the risk-free rate, it is as if agents were risk-neutral because they don't require any risk premium. It explains why the probability measure Q is usually labelled "risk-neutral probability measure".

2.4.2 Radon Nikodym derivatives

To justify the not so intuitive assumptions of the current section, we mention an important point relative to the prices of Arrow-Debreu securities. An Arrow-Debreu security contingent on a state ω is a financial asset which pays one monetary unit if ω occurs and nothing if another state occurs. It can be written as the indicator function $\mathbf{1}_{\{\omega\}}$.

Consider an Arrow-Debreu security contingent on a state ω_1, denoted as A_1^1, in a one-period model with a zero risk-free rate. If $P(\omega_1) > 0$, the initial price A_0^1 of this security is strictly positive but possibly lower than $P(\omega_1)$, due to risk aversion. When changing the probability to write $A_0^1 = E_Q\left(A_1^1\right)$, one remarks that $E_Q\left(A_1^1\right) = Q(\omega_1)$. Consequently, the probability change works (on the economic point of view) only if events having a positive (zero) probability under P also have a positive (zero) probability under Q. The reason is that the price of an Arrow-Debreu security is positive (zero) when the probability of getting one monetary unit at the terminal date is positive (zero).

This remark justifies the definitions hereafter.

Definition 73 *a) A probability measure Q is absolutely continuous with respect to (w.r.t) P if:*

$$\forall B \in \mathcal{A},\ P(B) = 0 \Rightarrow Q(B) = 0$$

and we note $Q << P$.

b) Two probability measures P and Q are equivalent if:

$$\forall B \in \mathcal{A},\ P(B) = 0 \Leftrightarrow Q(B) = 0$$

Point (b) also means that $Q << P$ and $P << Q$.

The essential result allowing to use the change of probability method in a rigorous way is the Radon-Nikodym theorem.

Proposition 74 *$Q << P$ if and only if there exists a positive $\mathcal{A}$-measurable function ϕ such that:*

$$\forall B \in \mathcal{A},\ Q(B) = \int_B \phi dP$$

Remarks : It is clear that $P(B) = 0 \Rightarrow Q(B) = 0$ since the integral is 0. Moreover, ϕ, which is a random variable, is a.s strictly positive. If we note that $Q(B) = \int_B \phi dP$, it allows to denote $\phi = \frac{dQ}{dP}$ by analogy with usual differential calculus. ϕ is then called the Radon-Nikodym derivative of Q with respect to P. Finally, if P and Q are equivalent, $\frac{dQ}{dP}$ and $\frac{dP}{dQ}$ exist and:

$$\frac{dQ}{dP} = 1/\frac{dP}{dQ}$$

Proposition 75 *Let P and Q two equivalent probability measures on $(\Omega, \mathcal{A})$ and $\phi = \frac{dQ}{dP}$. We have the following equality:*

$$E_Q(X_1) = E(\phi X_1)$$

In the abovementioned financial interpretation involving a simple lottery X_1, $E_Q(X_1)$ is the price of X_1. It can also be expressed as the expectation under P of a transformed payoff[16] ϕX_1. We can also write $E(\phi X_1) = \langle \phi, X_1 \rangle$ where the inner product is the one defined in $L^2(\Omega, \mathcal{A}, P)$. ϕ then represents (in the sense of the Riesz theorem) the linear valuation operator.

Specific case : Assume $Card(\Omega) = N$, $\mathcal{A} = P(\Omega)$ and $P(\omega) > 0$ for any ω; each state of nature is an event and we get:

$$Q(\{\omega\}) = \int_{\{\omega\}} \phi dP = \phi(\omega)P(\omega)$$

ϕ is then defined by:

$$\phi(\omega) = \frac{Q(\omega)}{P(\omega)}$$

[16] In the usual microeconomic one-period model, ϕ is proportional to the ratio of the date-1 marginal utility of consumption and of the date-0 marginal utility of consumption (see Dothan, 1990).

2.5 Random vectors

2.5.1 Definitions

In portfolio management, it is common to deal with a large number of stocks whose returns are random variables. It is then more efficient to use vector notations and matrices to present the calculations.

Definition 76 *A n-dimensional* ***random vector*** *is a random variable defined on* $(\Omega, \mathcal{A}, P)$ *and taking values in* $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$. *We then write* $X = (X_1, ..., X_n)'$ *where*[17] *the* X_i *are real random variables.*

The joint distribution of a random vector is defined by its cumulative distribution function or its density (when it is relevant).

[17] The "prime" denotes transposition as usual. Without other indications, vectors are column-vectors.

Definition 77 *a) The CDF of a vector* $X = (X_1, ..., X_n)'$ *is the function* F_X *from* $\mathbb{R}^n$ *to* $[0;1]$ *defined by:*

$$F_X(x) = P\left(\cap_{i=1}^n \{X_i \leq x_i\}\right)$$

where $x \in \mathbb{R}^n$ *is equal to* $(x_1, x_2.., x_n)'$.

b) If the X_i *are continuous random variables, the joint density of* X *is a positive function* f_X *from* $\mathbb{R}^n$ *to* $\mathbb{R}$ *satisfying :*

$$F_X(x) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} ... \int_{-\infty}^{x_n} f_X(x)dx_1...dx_n$$

The vocabulary defined for random variables is still valid for random vectors; in particular a random vector is integrable or square integrable if all its components satisfy this property. In this case, $E(X)$ denotes the vector of expectations of the X_i and $\mathbf{V}_X$ stands for the covariance matrix defined by:

$$\mathbf{V}_X = \begin{bmatrix} V(X_1) & ... & Cov(X_1, X_j) & Cov(X_1, X_n) \\ & & & \\ Cov(X_j, X_1) & & V(X_j) & \\ Cov(X_n, X_1) & & & V(X_n) \end{bmatrix}$$

A simplified and common notation for $\mathbf{V}_X$ is the following:

$$\mathbf{V}_X = \begin{bmatrix} \sigma_1^2 & ... & \sigma_{1j} & \sigma_{1n} \\ & & & \\ \sigma_{j1} & & \sigma_j^2 & \\ \sigma_{n1} & & & \sigma_n^2 \end{bmatrix}$$

Using random vectors is especially interesting because the rules governing matrix calculations can be used. The following proposition summarizes these essential calculation rules.

Proposition 78 *Let* X *denote a square-integrable* n*-dimensional random vector and* U, W *denote two* n*-dimensional vectors in* $\mathbb{R}^n$.

1) $E(U'X) = U'E(X)$

2)$E(U'X, W'X) = U'E(XX')W$

3)$V(U'X) = U'\mathbf{V}_X U$

4)$CoV(U'X, W'X) = U'\mathbf{V}_X W$

These notations may appear confusing at a first glance because $U'X = \sum_{i=1}^{n} U_i X_i$ is a random variable, so $V(U'X)$ is a number and $\mathbf{V}_X$ is a (n,n) matrix. Moreover, XX' is a $n \times n$ matrix so $E(XX')$ is also a $n \times n$ matrix with generic element $E(X_i X_j)$. It is important for a student in finance to be comfortable with these notations because they are very common in portfolio choice problems.

2.5.2 Application to portfolio choice

Consider a financial market with n traded stocks; X denotes the vector of returns and $U \in \mathbb{R}^n$ stands for the vector of proportions invested in the n stocks. The random return of portfolio U, denoted R, can be written as:

$$R = U'X = \sum_{i=1}^{n} U_i X_i$$

By proposition 78, the expected return and the variance of the portfolio return are:

$$\begin{aligned} E(R) &= U'E(X) \\ V(R) &= U'\mathbf{V}_X U \end{aligned}$$

Assume that $E(X)$ has at least two components which differ in value. If it was not the case, all portfolios would have the same expected return and the problem would be trivial.

For U to be a portfolio, it is necessary that:

$$\sum_{i=1}^{n} U_i = 1$$

which may be written $U'\mathbf{1}$ =1 where $\mathbf{1}$ is a vector in $\mathbb{R}^n$ whose components are all equal to 1.

The standard portfolio choice problem consists in finding the minimal variance of return, constrained by a given level of expected return, say e. $\mathbf{V}_X$ is assumed positive definite[18], meaning that it is not possible to build a risk-free portfolio with risky assets. It is not really a restrictive assumption because a risk-free asset can be separately introduced in the model.

[18] A (n,n) matrix M is positive definite if anf only if $\forall x \in \mathbb{R}^n, x \neq 0 \Leftrightarrow x'Mx > 0$.

The optimization problem is then written:

$$\min \frac{1}{2}U'\mathbf{V}_X U$$
with the constraints
$$U'E(X) = e$$
$$U'\mathbf{1} = 1$$

The coefficient $\frac{1}{2}$ does not change the optimal solution. It is purely conventional and avoids to keep coefficients 2 in the expression of the derivative.

The Lagrangian is given by:

$$\mathcal{L}(U, \lambda, \mu) = \frac{1}{2}U'\mathbf{V}_X U + \lambda (e - U'E(X)) + \mu (1 - U'\mathbf{1})$$

To simplify notations in the following, let us denote $\mathbf{V}_X = \mathbf{V}$ and $E(X) = \mathbf{r}$; the first-order conditions of the problem are:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial U} &= \mathbf{V}U - \lambda \mathbf{r} - \mu \mathbf{1} = \mathbf{0} \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= e - U'\mathbf{r} = \mathbf{0} \\ \frac{\partial \mathcal{L}}{\partial \mu} &= 1 - U'\mathbf{1} = \mathbf{0} \end{aligned}$$

As $\mathbf{V}$ is invertible, the first condition gives:

$$U = \lambda \mathbf{V}^{-1}\mathbf{r} + \mu \mathbf{V}^{-1}\mathbf{1}$$

Using the two other conditions leads to:

$$\begin{aligned} e &= \lambda \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} + \mu \mathbf{r}'\mathbf{V}^{-1}\mathbf{1} \\ 1 &= \lambda \mathbf{1}'\mathbf{V}^{-1}\mathbf{r} + \mu \mathbf{1}'\mathbf{V}^{-1}\mathbf{1} \end{aligned}$$

After a few calculations, we get:

$$U = \frac{1}{D}\left[(eC - A)\mathbf{V}^{-1}\mathbf{r} + (b - eA)\mathbf{V}^{-1}\mathbf{1}\right]$$

where

$$\begin{aligned} A &= \mathbf{r}'\mathbf{V}^{-1}\mathbf{1} \\ B &= \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} \\ C &= \mathbf{1}\mathbf{V}^{-1}\mathbf{1} \\ D &= BC - A^2 \end{aligned}$$

As an exercise, the reader can check that the mapping $x \to \sqrt{x'\mathbf{V}^{-1}x}$ defines a norm on $\mathbb{R}^n$ induced by the inner product $\langle x, y \rangle = x'\mathbf{V}^{-1}y$. Deduce from this result that D is strictly positive.

Chapter 3

Usual probability distributions in financial models

This chapter covers only the most popular probability distributions encountered in financial models, and especially in finance textbooks. Obviously, it is not a complete tour of probability distributions and, even in finance, many others can be found in scientific papers in the field. However, the few distributions developed hereafter largely cover most of what appeared in financial models in the 50 last years. It is then a good starting point for students in finance.

This chapter is divided in two parts. The first one describes the properties of discrete distributions, essential the Bernoulli, binomial and Poisson distributions. The second part deals with the most common continuous distributions, namely the uniform, Gaussian and Log-normal distributions. At the end, we also shortly present some other useful distributions appearing in statistical tests. These are the χ^2, the Student-t and the Fisher-Snedecor distributions. They are deduced from the Gaussian distribution.

3.1 Discrete distributions

3.1.1 Bernoulli distribution

Definition and example

The most simple probability distribution is the so called Bernoulli distribution.

Definition 79 *A random variable X follows a **Bernoulli distribution** with parameter p if X takes values 1 and 0 with probabilities p and $1-p$.*

We can observe that if $B \in \mathcal{A}$ and $P(B) = p$, the indicator function of B follows a Bernoulli distribution with parameter p. It is denoted as

$$\mathbf{1}_B \sim \mathcal{B}(p) \tag{3.1}$$

As a natural extension, any variable taking two different values is often called a Bernoulli distribution. In fact, if X takes values a and b $(a > b)$ with probabilities p and $1-p$, the variable $Y = \frac{1}{a-b}(X - b)$ takes values 1 and 0 with probabilities p and $1-p$. Y is then $\mathcal{B}(p)$. The use of this distribution in finance is, in many cases, pedagogical. For example, in a one-period model, it is convenient to modelize variations of the logarithm of a stock price by a Bernoulli distribution. In this case, the two possible values are denoted $\ln(u)$ and $\ln(d)$ (u for *up* and d for *down*). If the initial (date-0) price is denoted S_0, the date-1 price, denoted S_1, takes two possible values uS_0 and dS_0. We observe that :

$$\ln(S_1) = \ln(S_0) + X$$

where X is a Bernoulli variable taking values $\ln(u)$ and $\ln(d)$. This simple model, extended to the multi-period case, has given rise to the so called binomial model (see next section).

In chapter 1, we described binary options (example 36) traded on the Chicago Board Options Exchange. These options pay \$100 or 0 at the maturity date, depending on the occurrence of an event $B = \{SP_T \geq K\}$ where SP_T is the value of the S&P500 index at the maturity date of the contract, T and K is the exercise price. Definition 79 shows that the payoff of this type of option follows a Bernoulli distribution. Obviously, we can expect that the price quoted at date 0 is a function of $P(B)$.

Expectation and variance

Proposition 80 *If* $X \sim \mathcal{B}(p)$, *then* $E(X) = p$ *and* $\sigma^2(X) = p(1-p)$

Proof. If a random variable X follows a Bernoulli distribution with parameter p, the expectation $E(X)$ is given by :

$$E(X) = p \times 1 + (1-p) \times 0 = p$$

The variance of X, denoted as $\sigma^2(X)$ is obtained in the same way by using the formula[1]:

$$\sigma^2(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$$

■

[1] Remark that a Bernoulli variable taking only values 0 and 1, it satisfies $X = X^2$.

Example 81 *Assume that a random variable Y takes two values y_1 and y_2 with probabilities p and $(1-p)$. Then, we get:*

$$\sigma^2(Y) = p(1-p)(y_1 - y_2)^2 \tag{3.2}$$

To show this equality, note that $X = \frac{1}{y_1 - y_2}(Y - y_2)$ follows $\mathcal{B}(p)$. Writing $Y = (y_1 - y_2)X + y_2$ allows to directly deduce:

$$E(Y) = (y_1 - y_2)E(X) + y_2 = py_1 + (1-p)y_2 \tag{3.3}$$
$$\sigma^2(Y) = (y_1 - y_2)^2\sigma^2(X) = p(1-p)(y_1 - y_2)^2 \tag{3.4}$$

In the specific case of the binomial model of stock prices, introduced in chapter 1 (example 9), and if Y denotes the logarithmic return of the stock, we have $y_1 = \ln(u)$ and $y_2 = \ln(d)$. Consequently, the variance of the stock return is:

$$\sigma^2(Y) = p(1-p)\ln\left(\frac{u}{d}\right)^2 \tag{3.5}$$

3.1.2 Binomial distribution

Definition and example

Stock returns in a one-period model were represented by a Bernoulli distribution. Consider now a multi-period model and assume that successive returns are independent Bernoulli random variables. Implicitly, independence of successive returns refers to the efficient market hypothesis, meaning that all information is instantaneaously reflected in prices. If, in addition, we assume that the parameters u and d are constant over time (constant volatility), the log-price variations on given horizons are driven by a binomial distribution defined as follows.

Definition 82 *A variable X follows a **binomial distribution** with parameters n and p if X is written as the sum of n independent Bernoulli variables X_i, $i = 1, ..., n$, each of them following $\mathcal{B}(p)$. We then have:*

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of combinations of k objects among n. The distribution of X is denoted $\mathcal{B}(n, p)$.

The binomial distribution is very popular in finance because it is the foundation of the famous option valuation model developed by Cox-Ross-Rubinstein (1979). This model describes the evolution of a stock price S in discrete-time; if S_t is the date-t stock price, the date-$(t+1)$ price is defined by:

$$S_{t+1} = S_t \times X_{t+1}$$

where X_{t+1} takes values u and d with probabilities p and $1-p$. The variables X_t are assumed independent. S_t is then equivalently defined by :

$$S_t = S_0 \times \prod_{s=1}^{t} X_s$$

from which we get :

$$\ln\left(\frac{S_t}{S_0}\right) = \sum_{s=1}^{t} \ln(X_s)$$

The left hand side (LHS) is the stock return between $s = 0$ and $s = t$ and the right hand side (RHS) is the sum of t independent Bernoulli variables taking values $\ln(u)$ and $\ln(d)$ with probabilities p and $1-p$.

The probability distribution of $\ln(S_t)$ is $\mathcal{B}(n, p)$ which implies :

$$P\left(\ln(S_t) = \ln(S_0) + k \times u\right) = \binom{n}{k} p^k (1-p)^{t-k}$$

where $\binom{n}{k}$ is the binomial coefficient counting the number of price paths containing k up moves (and then $t-k$ down moves).

Expectation and variance

The expectation and variance of the binomial distribution come immediately from the moments of the Bernoulli distribution.

Proposition 83 *If* $X \sim \mathcal{B}(n, p)$ *then* $E(X) = np$ *and* $\sigma^2(X) = np(1-p)$

Proof. The binomial distribution $\mathcal{B}(n, p)$ is defined as the sum of n independent and identically distributed variables (obeying $\mathcal{B}(p)$).

It follows immediately that if $X \sim \mathcal{B}(n, p)$:

$$E(X) = np \text{ and } \sigma^2(X) = np(1-p)$$

because expectations and variances of the n independent Bernoulli variables entering the binomial distribution can be added[2]. ■

In the abovementioned Cox-Ross-Rubinstein model, we get the moments of log returns on t periods of time as:

$$\begin{aligned} E\left[\ln\left(\frac{S_t}{S_0}\right)\right] &= t\left(p\ln(u) + (1-p)\ln(d)\right) \\ \sigma^2\left[\ln\left(\frac{S_t}{S_0}\right)\right] &= tp(1-p)\ln\left(\frac{u}{d}\right)^2 \end{aligned}$$

These expressions put to light the advantage of logarithmic returns. The first two moments of returns are simply the sum of the same moments on subintervals. For example, the variance of weekly returns is the sum of the variances of daily returns in the week under consideration. This property is not true when returns are calculated linearly as $\frac{S_{t+1}-S_t}{S_t}$.

[2]Remember that independent variables are uncorrelated.

3.1.3 Poisson distribution

Definition

This probability distribution is often used in insurance to describe the arrival of damages or in microstructure theory to modelize the flows of buy and sell orders on financial markets.

Definition 84 *A variable X follows a* ***Poisson distribution*** *with parameter λ if X takes positive integer values and is defined by:*

$$\forall k \in \mathbb{N},\ P(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$$

We then note $X \sim \mathcal{P}(\lambda)$.

Figure 3.1 shows the CDF of $\mathcal{P}(2)$. It appears like a step-function because the Poisson distribution is discrete.

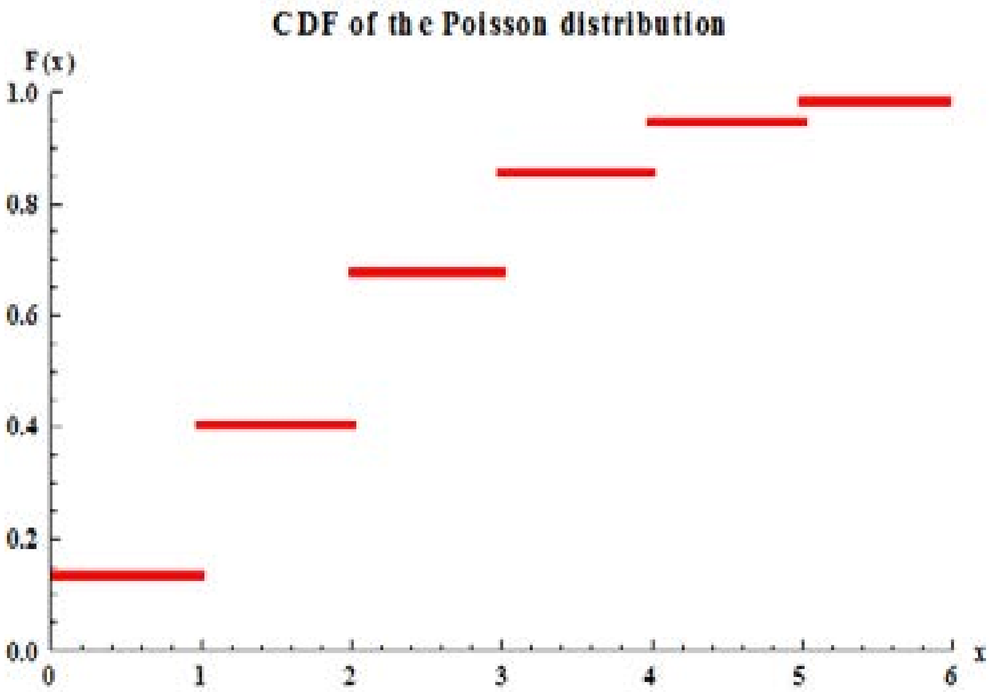

Figure 3.1: CDF of $\mathcal{P}(2)$

Expectation and variance

Proposition 85 *If* $X \sim \mathcal{P}(\lambda)$ *then* $E(X) = \lambda$ *and* $\sigma^2(X) = \lambda$

Proof. The expectation is deduced from the definition of the exponential function as the sum of an infinite series, $e^x = \sum_{k=0}^{+\infty} \frac{x^k}{k!}$.

$$\begin{aligned} E(X) &= \sum_{k=0}^{+\infty} kP(X=k) = \sum_{k=0}^{+\infty} k \exp(-\lambda)\frac{\lambda^k}{k!} = \exp(-\lambda)\sum_{k=1}^{+\infty} k\frac{\lambda^k}{k!} \\ &= \lambda \exp(-\lambda)\sum_{k=1}^{+\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \exp(-\lambda)\sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = \lambda \exp(-\lambda)\exp(\lambda) = \lambda \end{aligned}$$

To get the variance, the calculation is a little bit more involved:

$$\sigma^2(X) = E(X^2) - E(X)^2 = \exp(-\lambda)\sum_{k=0}^{+\infty} k^2\frac{\lambda^k}{k!} - \lambda^2 \qquad (3.6)$$

We can write :

$$\begin{aligned}
\sum_{k=0}^{+\infty} k^2 \frac{\lambda^k}{k!} &= \sum_{k=1}^{+\infty} k^2 \frac{\lambda^k}{k!} = \lambda \sum_{k=1}^{+\infty} k \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda \sum_{k=1}^{+\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} + \lambda \sum_{k=1}^{+\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda^2 \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} + \lambda \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} \\
&= \left(\lambda^2 + \lambda\right) \exp(\lambda)
\end{aligned}$$

Replacing in equation 3.6 leads to $\sigma^2(X) = \lambda$. ■

A specificity of $\mathcal{P}(\lambda)$ is the equality of expectation and variance. This property is useful when one wants to test if a given random variable follows this distribution.

The Poisson distribution is also commonly used to approximate a binomial distribution $\mathcal{B}(n, p)$ when n is large and p close to 0. For example, if you analyze the number of kjackpot winners in a 6/49 lotto games, the probability of winning the jackpot is around a chance over 15 millions. It is usual that 20 or 30 millions tickets are bought by players. The number of jackpot winners then follows a binomial distribution with n equal to the number of tickets and p is the probability of winning the jackpot. The reader can easily check that the expected number of winners is np and the variance $np(1-p)$ according to the properties of the binomial distribution. But $np(1-p) \simeq np$ since p is almost 0. Consequently, expectation and variance are almost equal and the distribution of the number of winners can be approximated by a Poisson distribution with parameter $\lambda = np$.

The three distributions presented in this section are discrete, the two first taking a finite number of values and the last one, $\mathcal{P}(\lambda)$, has a countable support, that is, the set of all integers. The following section is devoted to the most common continuous distributions appearing in financial models, especially to modelize prices and returns.

3.2 Continuous distributions

3.2.1 Uniform distribution

Definition

Definition 86 *X follows a* ***uniform distribution*** *on the interval* $[a;b]$, $a<b$, *if its density* f_X *is given by:*

$$f_X(x)=\begin{cases}\frac{1}{b-a} & \text{if } x\in[a;b]\\ 0 & \text{elsewhere}\end{cases}$$

We then denote $X\sim\mathcal{U}([a;b])$.

The CDF (F_X) of X is obtained by integrating the density.

$$F_X(x)=\begin{cases}\frac{x-a}{b-a} \text{ if } x\in[a;b]\\ 0 \text{ if } x<a\\ 1 \text{ if } x>b\end{cases}$$

Figure 3.2 shows the CDF of the uniform distribution on the interval $[0;1]$.

From this definition we deduce that, on any interval $[c;d]$ included in $[a;b]$:

$$P_X([c;d])=P_X(]c;d])=\frac{d-c}{b-a}=F_X(d)-F_X(c)$$

The probability of a given subinterval is proportional to its length and all subintervals of $[a;b]$ with a given length have the same probability (this explains the name "uniform distribution"). We will see hereafter that the moments of a uniform distribution are simple functions of a and b.

Expectation and variance

Proposition 87 *If* $X\sim\mathcal{U}([a;b])$ *then* $E(X)=\frac{b+a}{2}$ *and* $\sigma^2(X)=\frac{(b-a)^2}{12}$

Proof. If X follows a uniform distribution on $[a;b]$, the expectation of X is given by:

$$\begin{aligned}E(X) &= \int_{-\infty}^{+\infty}xf_X(x)dx=\frac{1}{b-a}\int_a^b xdx=\frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b\\ &= \frac{1}{2}\frac{(b^2-a^2)}{b-a}=\frac{b+a}{2}\end{aligned}$$

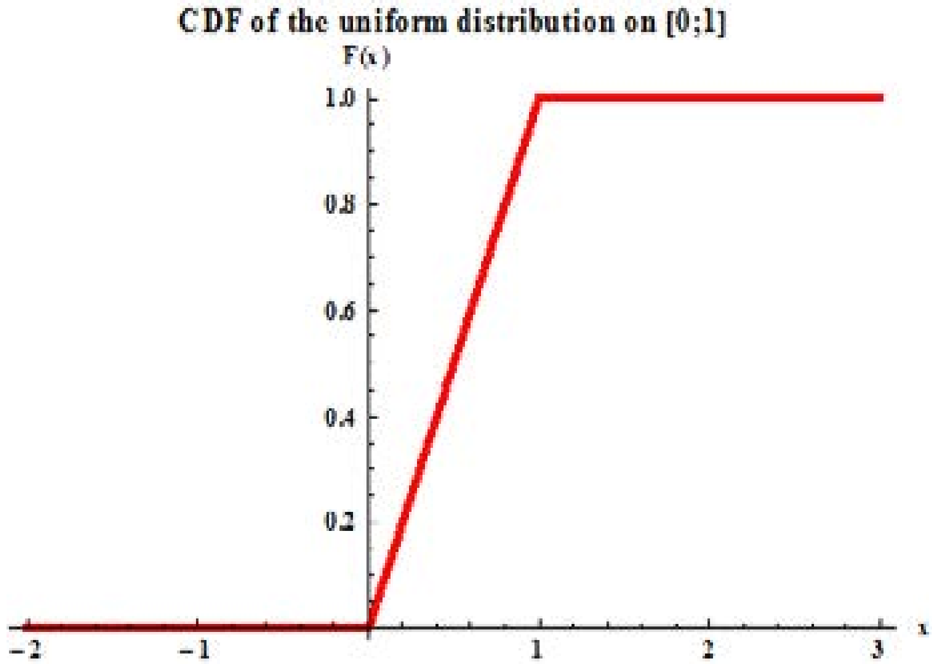

Figure 3.2: CDF of the uniform distribution on $[0;1]$

In the same way, the variance is given by:

$$\begin{aligned}\sigma^2(X) &= \int_{-\infty}^{+\infty} x^2 f_X(x)dx - \left(\frac{b+a}{2}\right)^2 = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{1}{3}\frac{(b^3-a^3)}{b-a} - \frac{1}{4}(a^2+2ab+b^2) \\ &= \frac{1}{3}(a^2+ab+b^2) - \frac{1}{4}(a^2+2ab+b^2) \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

■

3.2.2 Gaussian (normal) distribution

Definition

The normal distribution is the most common probability distribution in all sciences. It is due to a mathematical result called "central limit theorem" we will present in the next chapter.Without specifying the assumptions for the moment, this theorem states that the sum of a large number of independent and identically distributed (i.i.d) random variables is approximately driven by a Gaussian distribution. As a consequence, a number of statistical tests are based on the Gaussian distribution. It allows to modelize stock returns with a reasonable accuracy.

Nevertheless, one has to remember that it is not the optimal fit to the distribution of stock returns.Some empirical works show that alternative distributions (like Levy distributions) are better choices to take into account large variations (especially crashes) regularly encountered on financial markets. However, these distributions do not have nice statistical properties, so they are not often used in standard models. An important question nowadays (especially after the 2008 financial crisis) is that the "Gaussian world" may be a dangerous assumption when defining risk measures such as Value at Risk. This measure doesn't take into account crashes, liquidity crises and other extreme movements.

Definition 88 *X follows a **Gaussian distribution** with parameters m and σ (also written $X \sim \mathcal{N}(m, \sigma)$) if its density f_X is defined by:*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$$

f_X is the famous bell curve encountered in all statistical textbooks. It is symmetric with respect to the line $x = m$. Without entering the details, it is worth to recall that around 2/3 of the outcomes of a Gaussian distribution lie in the interval $[m - \sigma; m + \sigma]$ and 95% of the outcomes lie in the interval $[m - 2\sigma; m + 2\sigma]$.

Figure 3.3 shows the density of the Gaussian distribution $\mathcal{N}(0, 1)$ also called the standard Gaussian distribution.

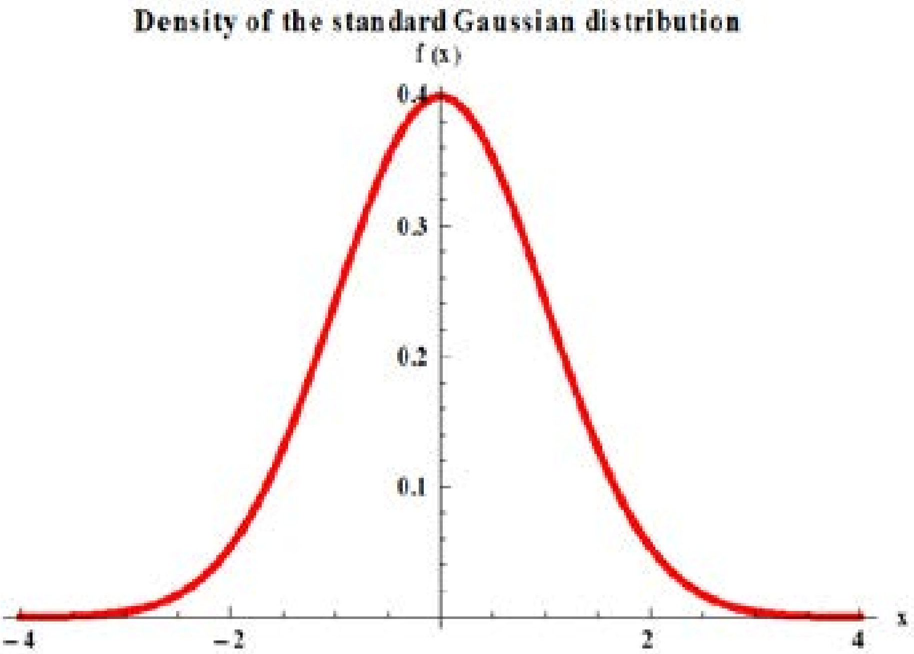

Figure 3.3: Density of $\mathcal{N}(0, 1)$

Moreover, most of the probability distributions used for statistical tests are functions of the Gaussian distribution. It is the case for the χ^2 distribution, the Student or the Fisher-Snedecor distributions presented in the next section.

Expectation and variance

Proposition 89 *If* $X \sim \mathcal{N}(m, \sigma^2), E(X) = m$ *and* $\sigma^2(X) = \sigma^2$

Proof. The definition of the density gives, for the Gaussian distribution:

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} x \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right) dx$$

Denoting $y = \frac{x-m}{\sigma}$, allows to write:

$$\begin{aligned} E(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma y + m) \exp\left(-\frac{1}{2}y^2\right) dy \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y \exp\left(-\frac{1}{2}y^2\right) dy + \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}y^2\right) dy \\ &= \left[-\frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)\right]_{-\infty}^{+\infty} + m = m \end{aligned}$$

Finally, $E(X) = m$.The last equality comes from the fact that $\exp\left(-\frac{1}{2}y^2\right)$ is the density of a standardized normal variable.

The same change $y = \frac{x-m}{\sigma}$ leads to calculate the variance as follows:

$$\begin{aligned} \sigma^2(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma y + m)^2 \exp\left(-\frac{1}{2}y^2\right) dy - m^2 \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y^2 \exp\left(-\frac{1}{2}y^2\right) dx + \frac{2m\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y \exp\left(-\frac{1}{2}y^2\right) dx \end{aligned}$$

The second term of the RHS is equal to 0 as the expectation of a zero-mean gaussian variable (times $2m\sigma$) ; the first term is integrated by parts.

$$\begin{aligned} &\frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y \times y \exp\left(-\frac{1}{2}y^2\right) dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left(\left[y \exp\left(-\frac{1}{2}y^2\right)\right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} - \exp\left(-\frac{1}{2}y^2\right) dx\right) \\ &= \sigma^2 \left[\left[\frac{1}{\sqrt{2\pi}} y \exp\left(-\frac{1}{2}y^2\right)\right]_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}y^2\right) dx\right] \end{aligned}$$

The first term between brackets is zero and the second term is equal to 1. It proves that $\sigma^2(X) = \sigma^2$. ■

As can be seen in the density, the Gaussian distribution is entirely determined by its two first moments. This property is especially interesting when stock returns are gaussian. It means that portfolio choice is uniquely guided by these two first moments. One doesn't need to assume quadratic utility functions to manage the portfolio problem in the mean-variance world of Markowitz (1952).

3.2.3 Log-normal distribution

Definition

The continuous return of a financial security between dates 0 and t is given by $r = \ln\left(\frac{S_t}{S_0}\right)$ if S_t denotes the date-t price ($t > 0$). Consequently, getting a price when starting from a return, needs an exponential transformation by writing $S_t = S_0 e^r$. Using proposition 38 of chapter 1 leads to the characterization of a Log-normal random variable.

Definition 90 *X follows a* ***Log-normal distribution*** *with parameters m and* σ^2 *if* $\ln(X) \sim \mathcal{N}(m, \sigma^2)$. *The density of X is given by:*

$$f_X(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(x)-m}{\sigma}\right)^2\right) & \text{if } x > 0 \\ 0 \text{ elsewhere} \end{cases}$$

We note $X \sim LN(m, \sigma^2)$.

Figure 3.4 shows the density of the Log-normal distribution with parameters $m = 0$ and $\sigma = 1$.

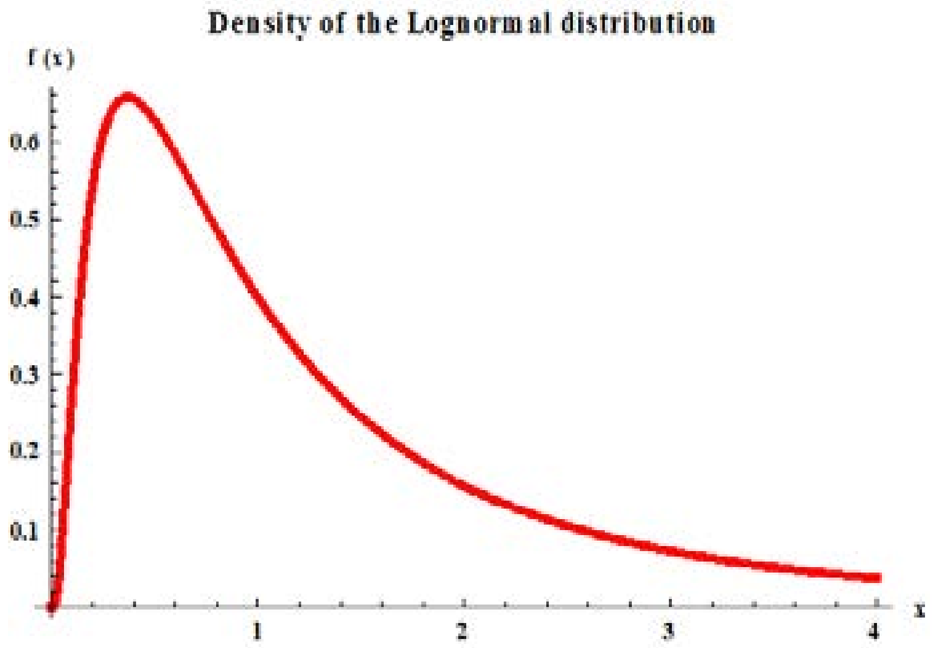

Figure 3.4: Density of the Lognormal distribution

Contrary to what was observed for the Gaussian distribution, the density is not symmetric. It explains some surprising results like the one illustrated in example 93 at the end of this chapter.

A Log-normal distribution is the usual assumption for stock prices, especially in the Black-Scholes option valuation model. It is worth to notice that it takes only positive values, a consistent characteristic for stock prices, due to the limited liability of shareholders.

Expectation and variance

Proposition 91 *If* $X \sim LN(m, \sigma^2)$, $E(X) = \exp\left(m + \frac{\sigma^2}{2}\right)$ *and* $\sigma^2(X) = \exp\left(2m + \sigma^2\right)\left(\exp(\sigma^2) - 1\right)$

Proof. The expectation is calculated as follows:

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{+\infty} \exp\left(-\frac{1}{2}\left(\frac{\ln(x) - m}{\sigma}\right)^2\right) dx$$

We use $y = \ln(x)$, and rewrite the expectation as:

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(y) \exp\left(-\frac{1}{2}\left(\frac{y - m}{\sigma}\right)^2\right) dy$$

Rearranging the terms in the exponentials leads to:

$$\begin{aligned} E(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}\left(\frac{(y - (m + \sigma^2))^2}{\sigma^2}\right)\right) \exp\left(m + \frac{\sigma^2}{2}\right) dy \\ &= \exp\left(m + \frac{\sigma^2}{2}\right) \end{aligned}$$

The integral is equal to 1 because it is the integral of the density of a normal random variable with mean $(m + \sigma^2)$ and variance σ^2.

The same method is used to calculate $V(X)$. We first show that $E(X^2) = \exp\left(2(m + \sigma^2)\right)$ and we finally get $V(X) = \exp\left(2m + \sigma^2\right)\left(\exp(\sigma^2) - 1\right)$ ■

Example 92 *Let* $Y \sim \mathcal{N}(0, 1)$ *and* X *the random variable defined by:*

$$X = exp\left(\left(m - \frac{\sigma^2}{2}\right) + \sigma Y\right)$$

where m *and* σ *are real numbers,* $\sigma > 0$. X *represents the date-*1 *price of a stock whose return is Gaussian with parameters* m *and* σ *when the date-0 price is equal to 1.*

A call option contract on X *with exercise price* K *and maturity* 1 *is a financial security paying* $\max(X - K; 0)$ *at the maturity date.*

What is the expected value of this final payoff?

We need to calculate $E\left[(X-K)_+\right]$ *where* $(x)_+ = x$ *if* $x > 0$ *and* $(x)_+ = 0$ *elsewhere. Denoting* f_X *the density of* X*, we get:*

$$E\left[(X-K)_+\right] = \int_0^{+\infty} f_X(x)\max(x-K;0)dx = \int_K^{+\infty} f_X(x)(x-K)dx \quad (3.7)$$

$$= \int_K^{+\infty} x f_X(x)dx - K\int_K^{+\infty} f_X(x)dx \quad (3.8)$$

$$= \int_K^{+\infty} x f_X(x)dx - KP(X \geq K) \quad (3.9)$$

$$= \int_K^{+\infty} x f_X(x)dx - KP(\ln(X) \geq \ln(K)) \quad (3.10)$$

We know by definition of X *that*

$$P(\ln(X) \geq \ln(K)) = P\left[\left(m - \frac{\sigma^2}{2}\right) + \sigma Y \geq \ln(K)\right] \quad (3.11)$$

$$= P\left[Y \geq \frac{\ln(K) - \left(m - \frac{\sigma^2}{2}\right)}{\sigma}\right] \quad (3.12)$$

If we denote $N(x)$ *the CDF of a standard Gaussian variable, we obtain:*

$$P(X \geq K) = 1 - N\left(\frac{\ln(K) - \left(m - \frac{\sigma^2}{2}\right)}{\sigma}\right) \quad (3.13)$$

$$= N\left(\frac{-\ln(K) + \left(m - \frac{\sigma^2}{2}\right)}{\sigma}\right) \quad (3.14)$$

The last equality is coming from the symmetry of the density of a Gaussian variable.

Using the technique in proposition 91, the first term in equation 3.10 may be written as:

$$\int_K^{+\infty} x f_X(x)dx = e^m N\left(\frac{-\ln(K) + \left(m + \frac{\sigma^2}{2}\right)}{\sigma}\right) \quad (3.15)$$

These formulas are the basis of the famous option valuation model developed in the seventies by Fisher Black and Myron Scholes (Black-Scholes, 1973).

Example 93 *Assume that the logarithmic return of the S&P500 is driven by a normal distribution with parameters* $m = 3\%$ *and* $\sigma = 20\%$. *The current value of the index is 1000 points. What is the price a risk-neutral investor is ready to pay to buy a contract delivering $100 if the S&P500 value in one year is in the interval* $[900; 1000]$. *What price is he ready to pay if the interval is* $[1000; 1100]$? *Why may it be a surprising result?*

3.3 Some other useful distributions

When using large samples of i.i.d variables, statistics like the mean follow a Gaussian distribution according to the central limit theorem (see chapter 4). The variance is then written as a function of Gaussian random variables. Moreover, when standardizing variables (transforming to get a zero-mean and a unit variance) leads to use more or less complicated functions of Gaussian variables. It is the reason why a number of useful distributions in statistics are "derived" from the Gaussian distribution. We shortly present hereafter the χ^2 distribution, the Student-t and the Fisher-Snedecor distributions.

3.3.1 The χ^2 distribution

Definition 94 *A random variable Y follows a χ^2* ***distribution*** *with n degrees of freedom if Y can be written as:*

$$Y = \sum_{i=1}^{n} X_i^2 \tag{3.16}$$

where the X_i are independent standard Gaussian distributions, that is, $\forall i$, $X_i \sim \mathcal{N}(0,1)$.

This distribution is useful when one wants to perform a statistical test for the variance σ^2 of a random variable. If $(X_1, ...X_n)$ are identically distributed Gaussian random variables with parameters (m, σ_0^2), the variable Y defined by:

$$Y = \sum_{j=1}^{n} \left(\frac{X_i - m}{\sigma_0} \right)^2 \tag{3.17}$$

follows a χ^2 distribution with n degrees of freedom. Equality 3.17 leads to:

$$\frac{\sigma_0^2 Y}{n} = \frac{1}{n} \sum_{j=1}^{n} (X_i - m)^2 \tag{3.18}$$

The expression on the right-hand side of equation 3.18 is the empirical variance. If m is unknown and estimated by $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the variable Y^*, obtained by replacing m by $\overline{X}$ in equation 3.17 follows a χ^2 distribution with $n-1$ degrees of freedom. In this case $\frac{1}{n-1}\sum_{j=1}^{n}\left(X_i - \overline{X}\right)^2$ is used as an unbiased estimator of the variance to perform the test.

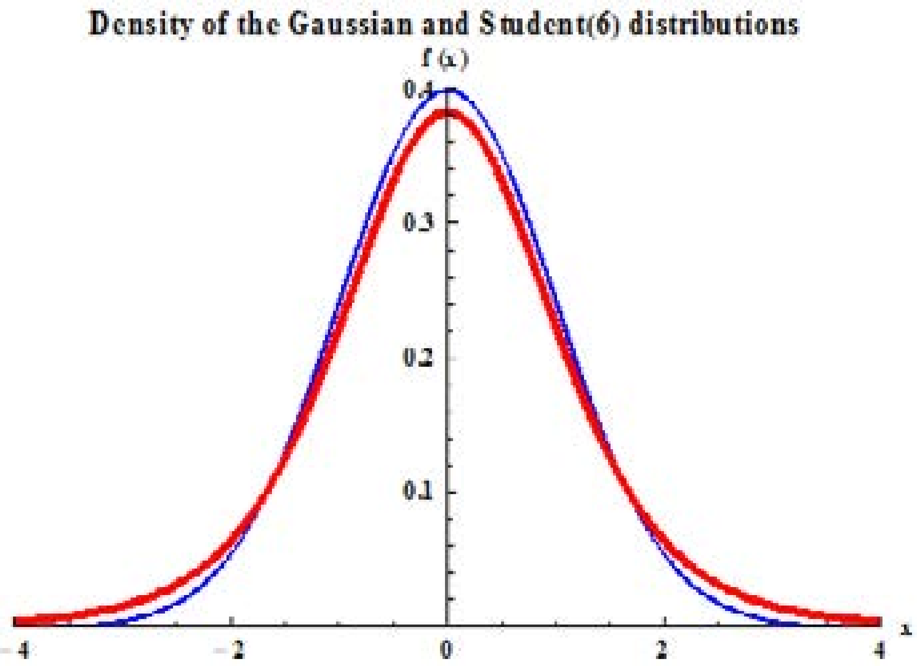

Figure 3.5: Comparison of the densities of a Student and a Gaussian density

The χ^2 distribution also appears in the well-known χ^2 test aimed at testing the independence of two distributions, and in the test comparing a empirical distribution to a theoretical distribution.

3.3.2 The Student-t distribution

Definition 95 *A random variable Y follows a **Student–t distribution** with n degrees of freedom if Y can be written:*

$$Y = \frac{Z}{\sqrt{\frac{X}{n}}} \tag{3.19}$$

where Z is a standard Gaussian distribution and X follows a χ^2 distribution with n degrees of freedom.

The *Student-t* distribution is used to test the equality of means in two populations, or to test regression coefficients. For example, they are common when testing the market model or the Capital Asset Pricing Model.

On figure 3.5, we can see than the Student density (bold line) has fatter tails than the Gaussian density (thin line) when the number of degrees of freedom is low (6 in the example). It is then sometimes used to represent stock returns when one wants to take into account that extreme returns are more frequent on real markets than what is predicted by a Gaussian distribution.

More generally, a Student(n) has a variance equal to $n/(n-2)$ and a kurtosis equal to $3(n-2)/(n-4)$. It is only defined when $n > 4$. We observe that for $n = 6$, the kurtosis is equal to 6, that is greater than the corresponding moment for the Gaussian distribution.

3.3.3 The Fisher-Snedecor distribution

In a multiple regression, beyond the significance of the individual regression coefficients, most softwares provide the so-called F of the regression. It comes from the Fisher-Snedecor distribution, defined as follows.

Definition 96 *A random variable Y follows a **Fisher-Snedecor distribution** if it writes:*

$$Y = \frac{\frac{X_1}{n_1}}{\frac{X_2}{n_2}} \tag{3.20}$$

where X_1 (X_2) follows a χ^2 distribution with n_1 (n_2) degrees of freedom.

It can be seen that a $F(n_1, n_2)$ is the inverse of a $F(n_2, n_1)$ variable. It is the reason why statistical tables of F variables only provide values greater than 1. When you get an observed value below 1 when testing the equality of two variances, take the inverse, reverse n_1 and n_2 and look at the corresponding position in the statistical table. Obviously, when testing the relevance of a regression model, the two variances are not equivalent. You just want to know if the variance explained by the model is significantly greater than the unexplained variance. In this case, a F statistic lower than 1 simply means that your model is not the right one.

Chapter 4

Conditional expectations and Limit theorems

As already mentioned in chapter 1, learning a piece of information changes probabilities of events. It then changes expectations of random variables because these expectations are probability-weighted averages. Conditional expectations are a natural tool to address this issue. Moreover, conditional expectations play an important role in valuation models. One essential result in finance theory is the following: when there are no arbitrage opportunities, the date-t value of an asset is the discounted expected value[1] of date-$t+1$, conditional on the information known by date t. In the book to follow (Stochastic Processes for Finance, Roger, 2010), conditional expectations will play an even more important role. In muti-period models, the stochastic processes called martingales are fundamental. But their definition relies essentially on conditional expectations. It is the reason why we want to insist now on the importance of understanding this (maybe difficult) topic.

4.1 Conditional expectations

4.1.1 Introductive example

We start with a very simple framework allowing to carefully describe what is going on when an information is revealed. Consider a probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, $\mathcal{A} = \mathcal{P}(\Omega)$ and $P(\omega_i) = 0.25$ for all $i = 1, .., 4$. Two random variables X and Y are defined on Ω and their values appear in table 4.1.

[1] The expectation is calculated with respect to a specific probability measure called the risk-neutral measure. In this framework, the risk-free rate is used for discounting.

State	X	Y
ω_1	1	1
ω_2	2	1
ω_3	3	2
ω_4	4	2

Table 4.1: Definition of X and Y

We get immediately:

$$E(X) = \frac{1}{4}(1+2+3+4) = 2.5 \quad (4.1)$$

$$E(Y) = \frac{1}{4}(1+1+2+2) = 1.5 \quad (4.2)$$

Suppose now that the value of Y is observed before the value of X is revealed.

If $Y(\omega) = 1$, the true state ω may be either ω_1 or ω_2. In other words, the event $\{\omega_1, \omega_2\}$ occurs and it is equal to the event $\{Y = 1\}$. Probabilities of all states change and become conditional on the event $\{Y = 1\}$. The new (conditional) probabilities for the 4 states are now:

$$(P(\omega_i \,|\{Y = 1\}), i = 1, ..., 4) = \left(\frac{1}{2}; \frac{1}{2}; 0; 0\right) \quad (4.3)$$

The other consequence is the change in the expectation of X which becomes the conditional expectation denoted $E(X\,|\{Y = 1\})$ with

$$E(X\,|\{Y = 1\}) = \sum X(\omega_i)P(\omega_i\,|\{Y = 1\}) = \frac{1}{2}(1+2) = 1.5 \quad (4.4)$$

In other words, if $E(X)$ is the initial price of a stock, $\{Y = 1\}$ corresponds to bad news leading to a price decrease.

The important fact here is that the change of probabilities and expectations is not linked directly to the values of Y but to the information revealed by the observation of Y. If the values of Y, 1 and 2 in the example, had been replaced by 100 and 200, the result would have been the same. The conditional expectation of X would also be 1.5. Each event with respect to which we define conditional probabilities generates an other conditional expectation.

4.1.2 Conditional distributions

Discrete variables

Let us first consider two discrete variables X and Y whose supports are respectively $(x_i, i = 1, ..., n)$ and $(y_j, j = 1, ..., p)$.

Definition 97 *a) The* ***conditional probability distribution*** *of X knowing $\{Y = y_i\}$ is the mapping denoted as $P_{X|Y}(. |y_i)$ and defined by:*

$$P_{X|Y}(x|y_i) = P(X = x|Y = y_i) = \frac{P(\{X = x\} \cap \{Y = y_i\})}{P(\{Y = y_i\})}$$

In this definition it is assumed that $P(\{Y = y_i\}) \neq 0$, but it is in fact implicit in the definition of the support of Y. $P_{X|Y}(. |y_i)$ effectively induces a probability measure on the support of X.

Continuous variables

Denote f_{XY} the joint density[2] of a couple of continuous random variables, f_X and f_Y being the densities of X and Y.

Definition 98 *For any y satisfying $f_Y(y) > 0$, the* ***conditional density*** *of X w.r.t. $\{Y = y\}$ is the function $f_{X|Y}(.\,|y)$ defined by :*

$$f_{X|Y}(x\,|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

Remark 99 *When X is continuous with density f_X and B is an event such that $P(B) \neq 0$ the density of X conditional on B is defined by:*

$$f_X(x\,|B) = \begin{cases} \frac{f_X(x)}{P(B)} \text{ if } x \in X(B) \\ 0 \text{ elsewhere} \end{cases} \tag{4.5}$$

We can now characterize conditional expectations, starting with the most simple case of conditioning with respect to an event.

4.1.3 Conditional expectation with respect to an event

The introductive example shows how to define the conditional expectation of a random variable with respect to an event in $\mathcal{A}$.

Definition 100 *a) The conditional expectation of a discrete variable X , taking values $x_1, ..., x_N$, w.r.t an event B in $\mathcal{A}$, is the quantity $E(X\,|B)$ defined by :*

$$E(X\,|B) = \sum_{i=1}^{N} x_i P(\{X = x_i\}\,|B)$$

b) The conditional expectation of a continuous variable X with density f_X w.r.t an event B in $\mathcal{A}$, is the quantity $E(X\,|B)$ defined by :

$$E(X\,|B) = \frac{1}{P(B)} \int_{X(B)} x f_X(x)dx = \int_{-\infty}^{+\infty} x f_X(x\,|B)dx$$

[2]See chapter 2, definition 77.

In the introductive example at the beginning of this section, if $\{Y = 2\}$ we obtain:

$$
\begin{aligned}
E(X\,|\{Y=2\}) &= \sum_{i=1}^{N} x_i P(\{\omega_i\}\,|\{Y=2\}) && (4.6)\\
&= 3 \times P(\omega_3\,|\{Y=2\}) + 4 \times P(\omega_4\,|\{Y=2\}) && (4.7)\\
&= \frac{1}{2}(3+4) = 3.5 && (4.8)
\end{aligned}
$$

The first equality comes from $P(\omega_1\,|\{Y=2\}) = P(\omega_2\,|\{Y=2\}) = 0$.

We could have saved some place and notation if we had addressed the problem in a more general way. In fact, we could define directly the conditional expectation of X with respect to the random variable Y. Before knowing the value of Y, we already know how to calculate the conditional expectation if one of the two events occurs. This remark allows to propose a more general approach.

4.1.4 Conditional expectation with respect to a random variable

Discrete variables

Definition 101 *The conditional expectation of a discrete variable X, taking values $x_1, ..., x_N$, w.r.t. a discrete random variable Y, taking different values $y_1, ..., y_M$, denoted as $E(X\,|Y)$, is the random variable defined by:*

$$\forall \omega \in \{Y = y_j\}, \ \ E(X\,|Y)(\omega) = \sum_{i=1}^{N} x_i P(\{X = x_i\}\,|\{Y = y_j\}) \qquad (4.9)$$

It leads to characterize the conditional expectation of X w.r.t. Y as in table 4.2.

State	$E(X\,\vert Y)$
ω_1	1.5
ω_2	1.5
ω_3	3.5
ω_4	3.5

Table 4.2: Conditional expectation of X with respect to Y

Continuous variables

Suppose now that X and Y are continuous with densities f_X and f_Y, the conditional density being denoted $f_{X|Y}(x|y)$ as before. The conditional expectation of X w.r.t. $\{Y = y\}$ is written:

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

More generally, the conditional expectation of X w.r.t. Y is the random variable defined by:

$$\forall \omega \in \{Y = y\}, \quad E(X|Y)(\omega) = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx$$

Remark 102 *We observe that, when Y is discrete, the subsets $\{Y = y_j\}$ define a partition on Ω. Second, the value of the random variable $E(X|Y)$ is constant on each subset of the partition, and it is also true for Y, by construction. In other words, the information revealed by Y is the same as the information revealed by $E(X|Y)$. A key remark here is that $E(X|Y)$ is $\mathcal{B}_Y$-measurable. It leads to the general approach of conditional expectations.*

4.1.5 Conditional expectation with respect to a sub-tribe

The examples provided before for discrete variables showed that the key point when conditioning with respect to a random variable is not the values taken by this variable but the information these values reveal about the true events. Consequently, the more general way to define conditional expectations is to condition with respect to subsets of events or, more precisely, with respect to sub-tribes.

Definition 103 *The conditional expectation of an integrable random variable* $(that\ is\ X \in L^1(\Omega, \mathcal{A}, P))$*, w.r.t a sub-tribe* $\mathcal{B}$ *of* $\mathcal{A}$*, is any* $\mathcal{B}$*-measurable random variable* Z*, satisfying:*

$$\forall B \in \mathcal{B}, E\left(Z\mathbf{1}_B\right) = E\left(X\mathbf{1}_B\right) \tag{4.10}$$

This definition deserves several comments.

- As Z is defined by means of integrals, two variables Z and Z' can satisfy equality (4.10) if they differ only on negligible events. They are called **versions of the conditional expectation**. Any version used in calculations is in general denoted $E(X\,|\mathcal{B})$.

- The definition also means that a variable X and its conditional expectation $E(X\,|\mathcal{B})$ have the same mean on any event of the tribe $\mathcal{B}$. We let the reader check it was actually the case in the former example (see table 4.1).

- The equality 4.10 implies that if X is $\mathcal{B}$-measurable, $E(X\,|\mathcal{B}) = X$.

Example 104 *Let* $Card(\Omega) = 4, P(\omega_i) = p_i$ *for each* ω_i*, and* $\mathcal{B}$ *defined by:*

$$\mathcal{B} = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\}$$

Denote $B_1 = \{\omega_1, \omega_2\}$, $B_2 = \{\omega_3, \omega_4\}$ *and let* X *be defined by*[3] $X = (x_1; x_2; x_3; x_4)$.
The equality 4.10 implies:

$$\begin{aligned} p_1x_1 + p_2x_2 &= p_1z_1 + p_2z_2 && (4.11) \\ p_3x_3 + p_4x_4 &= p_3z_3 + p_4z_4 && (4.12) \end{aligned}$$

[3] As $Card(\Omega) = 4$, X is defined by the vector of values it takes on the four states of nature.

where the conditional expectation Z takes values $(z_1; z_2; z_3; z_4)$. Equation 4.11 refers to evant B_1 and equation 4.12 to event B_2. Moreover, Z is $\mathcal{B}$-measurable; it means that it is constant on B_1 and on B_2. It implies that:

$$\begin{aligned} z_1 &= z_2 \\ z_3 &= z_4 \end{aligned}$$

We finally get:

$$\begin{aligned} z_1 &= z_2 = \frac{1}{p_1 + p_2}\left[p_1 x_1 + p_2 x_2\right] = E\left(X\,|B_1\right) \\ z_3 &= z_4 = \frac{1}{p_3 + p_4}\left[p_3 x_3 + p_4 x_4\right] = E\left(X\,|B_2\right) \end{aligned}$$

The result is intuitive. The conditional expectation on B_1 (B_2) is the (conditional probability) weighted average of the values taken by X on this subset $B_1(B_2)$.

We also check here that if X was already $\mathcal{B}$-measurable, then $E\left(X\,|\mathcal{B}\right)$ would be equal to X.

4.2 Geometric interpretation in $L^2\left(\Omega, \mathcal{A}, P\right)$

Conditional expectations have a natural geometric interpretation when the analysis is restricted to square integrable random variables, that is to elements of the vector space $L^2\left(\Omega, \mathcal{A}, P\right)$. We then assume it is the case in this section.

4.2.1 Introductive example

To explain what is the "geometry" of conditional expectations, first consider a simple example in the two-dimensional space $\mathbb{R}^2$, endowed with the usual metric:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

where $x' = (x_1, x_2)$ et $y' = (y_1, y_2)$.

For a given $x \in \mathbb{R}^2$, suppose that we want to determine the point $z = (z_1, z_1)$ on the bisector of the positive orthant which is the closest to x. We have to solve:

$$min_{z_1}(x_1 - z_1)^2 + (x_2 - z_1)^2$$

because points on the bisector have the same coordinates $z_1 = z_2$.

We find immediately $z_1 = \frac{x_1+x_2}{2}$. In other words, z is the orthogonal projection[4] of $x \in \mathbb{R}^2$ on the one-dimensional subspace defined by the bisector.

The reason is that $z - x$ is orthogonal to z.

$$< z - x, z >= (z_1 - x_1)z_1 + (z_1 - x_2)z_1 \quad (4.13)$$
$$= \frac{x_2 - x_1}{2}z_1 + \frac{x_1 - x_2}{2}z_1 = 0 \quad (4.14)$$

Suppose now that $\mathbb{R}^2$ is endowed with the metric:

$$d^*(x, y) = \sqrt{p(x_1 - y_1)^2 + q\,(x_2 - y_2)^2}$$

with $p + q = 1, p > 0, q > 0$. It simply means that the two coordinates are not equally weighted.

Solving the same optimization problem leads to:

$$z_1 = px_1 + qx_2$$

z_1 is then a weighted average (we are tempted to write "an expectation") of the components of x.

4.2.2 Conditional expectation as a projection in L^2

The preceding approach can be applied almost without modifications to the vector space of square-integrable random variables. If X is an element of $L^2(\Omega, \mathcal{A}, P)$, the conditional expectation $E(X|\mathcal{B})$ is $\mathcal{B}$-measurable and so belongs to the subspace[5] of $\mathcal{B}$-measurable variables denoted $L^2(\Omega, \mathcal{B}, P)$.

In example 104, $L^2(\Omega, \mathcal{A}, P)$ could be identified to $\mathbb{R}^4$ and $L^2(\Omega, \mathcal{B}, P)$ to $\mathbb{R}^2$ since the variables in this subspace had only two different components. We are going to show that $E(X|\mathcal{B})$ is the **orthogonal projection** of X on $L^2(\Omega, \mathcal{B}, P)$. In other words, $E(X|\mathcal{B})$ solves the optimization problem:

$$min_{Z \in L^2(\Omega,\mathcal{B},P)} E\left[(X - Z)^2\right] = min_{Z \in L^2(\Omega,\mathcal{B},P)} d(X, Z)^2 = E\left[(X - E(X|\mathcal{B}))^2\right]$$

To keep things simple, we just show this property with the data of example 104. As $E(X|\mathcal{B})$ is $\mathcal{B}$-measurable, we know that

$$z_1 = z_2 \quad (4.15)$$
$$z_3 = z_4 \quad (4.16)$$

[4]Remember that two vectors are orthogonal when their inner product is zero.

[5]To be completely rigorous, we should adopt a different notation for P (for example $P_{\mathcal{B}}$ because it is defined on the sub-tribe $\mathcal{B}$ in $L^2(\Omega, \mathcal{B}, P)$. For the sake of simplicity, we keep P to denote the probability measure on $\mathcal{B}$.

Therefore,

$$E\left[(X-Z)^2\right] = p_1(x_1-z_1)^2+p_2(x_2-z_1)^2+p_3(x_3-z_3)^2+p_4(x_4-z_3)^2 \quad (4.17)$$

The partial derivatives with respect to z_1 and z_3 must be zero to obtain an optimum.

$$\frac{\partial E\left[(X-Z)^2\right]}{\partial z_1} = -2\left[p_1(x_1-z_1)+p_2(x_2-z_1)\right]=0 \quad (4.18)$$

$$\frac{\partial E\left[(X-Z)^2\right]}{\partial z_3} = -2\left[p_3(x_3-z_3)+p_4(x_4-z_3)\right]=0 \quad (4.19)$$

The minimum is obtained with:

$$z_1 = z_2 = \frac{1}{p_1+p_2}\left(p_1x_1+p_2x_2\right) = E\left(X\left|\mathcal{B}\right.\right)(\omega_1) = E\left(X\left|\mathcal{B}\right.\right)(\omega_2) \quad (4.20)$$

$$z_3 = z_4 = \frac{1}{p_3+p_4}\left(p_3x_3+p_4x_4\right) = E\left(X\left|\mathcal{B}\right.\right)(\omega_3) = E\left(X\left|\mathcal{B}\right.\right)(\omega_4) \quad (4.21)$$

The reader can check that the second partial derivatives are positive, ensuring that the stationary point is a minimum (also because the cross-derivatives are 0). We are done.

The properties of conditional expectations can now be presented in a more intuitive way, using this geometrical interpretation.

4.3 Properties of conditional expectations

Proposition 105 *Let* (X,Y) *be two random variables in* $L^2(\Omega,\mathcal{A},P)$ *and* $\mathcal{B},\mathcal{B}'$ *two sub-tribes of* $\mathcal{A}$ *satisfying* $\mathcal{B}\subset\mathcal{B}'$*:*

1) If X *is a constant* $c\in\mathbb{R}$*,* $E(X|\mathcal{B})=c$

2) $\forall(a,b)\in\mathbb{R}^2,\ E(aX+bY|\mathcal{B})=aE(X|\mathcal{B})+bE(Y|\mathcal{B})$

3) If $X\leq Y,\ E(X|\mathcal{B})\leq E(Y|\mathcal{B})$

4) $E(E(X|\mathcal{B}')|\mathcal{B})=E(X|\mathcal{B})$

5) If X *is* $\mathcal{B}$*-measurable* $E(XY|\mathcal{B})=X\,E(Y|\mathcal{B})$

6) If X *is independent of* $\mathcal{B}$*,* $E(X|\mathcal{B})=E(X)$

We do not provide all the details of the proof but it is worth to underline the intuitions leading to some of these results.

First, a constant c can also be written $c\mathbf{1}_\Omega$. It is then a random variable measurable with respect to any tribe, especially w.r.t. $\mathcal{B}$. The projection theorem then implies that c is its own projection on $L^2(\Omega,\mathcal{B},P)$. Remember that $L^2(\Omega,\mathcal{B},P)$ being a vector sub-space of $L^2(\Omega,\mathcal{A},P)$, it is a convex set.

Points (2) and (3) are direct consequences of the definition of conditional expectations.

Point (4), which doesn't seem obvious at first glance, may be easily understood using the geometric interpretation of the conditional expectation. $E(X|\mathcal{B}')$ is the orthogonal projection of X on $L^2(\Omega,\mathcal{B}',P)$. $E(E(X|\mathcal{B}')|\mathcal{B})$ is the projection on $L^2(\Omega,\mathcal{B},P)$ of $E(X|\mathcal{B}')$.

Point (4) simply says that projecting first on $L^2(\Omega,\mathcal{B}',P)$ and then on $L^2(\Omega,\mathcal{B},P)$ is equivalent to make directly the projection on the smallest space $L^2(\Omega,\mathcal{B},P)$. It is a well-known property of projections on finite-dimensional spaces. Moreover, it is worth to notice that if $\mathcal{B}=\{\varnothing,\Omega\}$, $E(X|\mathcal{B})=E(X)$ and then $E(E(X|\mathcal{B}'))=E(X)$ whatever $\mathcal{B}'$ is.

Point (6) could be written $E(X-E(X)|\mathcal{B})=0$ since $E(X)$ is a constant

(see point (1)). In other words, $X-E(X)$ independent of any variable Y in $L^2(\Omega,\mathcal{B},P)$ means.

$$E((X-E(X))\,Y)=E(X-E(X))\,E(Y)=0 \tag{4.22}$$

The first term on the left-hand side of equality 4.22 is the inner product of $X - E(X)$ and Y. Two vectors with an inner product equal to zero are said orthogonal. So, there is a close link between independence and orthogonality in the space of square-integrable variables.

4.3.1 The Gaussian vector case

To end this section on conditional expectations, we present the specific case of Gaussian vectors[6]. Such conditioning with Gaussian vectors is common either in the microstructure literature or in the "information" literature in which investors are supposed to receive private signals[7]. Equilibrium prices are more easily obtained when the private signals follow a joint Gaussian distribution.

Definition 106 *A random vector $X = (X_1, ..., X_n)$ is said Gaussian if every linear combination $\sum_{i=1}^{n} a_i X_i$ is a Gaussian variable.*

Denot $m' = (E(X_1), ..., E(X_n))$ the vector of expectations and $\mathbf{V}_X$ the covariance matrix of X. The density f_X of X is given by:

$$\forall x \in \mathbb{R}^n, f(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{Det(\mathbf{V}_X)}} \exp\left(-\frac{1}{2}(x-m)'\mathbf{V}_X^{-1}(x-m)\right) \tag{4.23}$$

where $Det(\mathbf{V}_X)$ is the determinant of the covariance matrix (assumed different from zero).

Proposition 107 *Let $X = (X_1, ..., X_n)$ be a Gaussian random vector with parameters m and $\mathbf{V}_X$; for $p < n$ let $Y_1 = (X_1, ..., X_p)$ and $Y_2 = (X_{p+1}, ..., X_n)$. Decompose $\mathbf{V}_X$ in the following way:*

$$\mathbf{V}_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where Σ_{ii} is the covariance matrix of Y_i and Σ_{ij} is the matrix containing covariances between the components of Y_i and Y_j for $i, j = 1, 2, i \neq j$. The probability distribution of Y_1 conditioned on $Y_2 = y_2 \in \mathbb{R}^{n-p}$ is Gaussian with the following two first moments:

$$E(Y_1 | Y_2 = y_2) = E(Y_1) + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - E(Y_2)) \tag{4.24}$$

$$\mathbf{V}_{Y_1|Y_2=y_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

[6] Random vectors have been presented in chapter 2, section 2.5.
[7] One of the seminal papers in the field is Grossman (1976).

Case $p = 1$ and $n = 2$

Applying the above proposition when $p = 1$ and $n = 2$ gives:

$$\begin{aligned} E\left(X_1 \left| X_2 = x_2\right.\right) &= m_1 + \frac{\sigma_{12}}{\sigma_2^2}\left(y_2 - m_2\right) \\ \mathbf{V}_{X_1|X_2=x_2} &= \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2} \end{aligned}$$

If ρ_{12} stands for the correlation between the two variables, we obtain:

$$\mathbf{V}_{X_1|X_2=x_2} = \sigma_1^2(1-\rho_{12}^2)$$

This result may be obtained by using the definition of the conditional density (with $x' = (x_1, x_2)$).

$$\begin{aligned} f_{X_1|X_2}\left(x_1 \left| x_2\right.\right) &= \frac{f_X(x_1,x_2)}{f_{X_2}(x_2)} = \frac{\frac{1}{(2\pi)\sqrt{|Det(\mathbf{V}_X)|}} exp\left(-\frac{1}{2}(x-m)'\mathbf{V}_X^{-1}(x-m)\right)}{\frac{1}{\sigma_2\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x_2-m_2}{\sigma_2}\right)^2\right)} \\ &= \frac{\sigma_2}{\sqrt{2\pi}\sqrt{\sigma_1^2\sigma_2^2-\sigma_{12}^2}} \frac{exp\left(-\frac{1}{2}(x-m)'\mathbf{V}_X^{-1}(x-m)\right)}{exp\left(-\frac{1}{2}\left(\frac{x_2-m_2}{\sigma_2}\right)^2\right)} \\ &= \frac{\sigma_2}{\sqrt{2\pi}\sqrt{\sigma_1^2\sigma_2^2-\sigma_{12}^2}} exp\left(-\frac{1}{2}\left((x-m)'\mathbf{V}_X^{-1}(x-m) - \left(\frac{x_2-m_2}{\sigma_2}\right)^2\right)\right) \end{aligned}$$

Calculating the distinct parts leads to:

$$\mathbf{V}_X^{-1} = \frac{1}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}\begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}$$

Denoting $A = (x-m)'\mathbf{V}_X^{-1}(x-m)$, we get:

$$\begin{aligned} A &= \frac{\sigma_2^2x_1^2 - 2\sigma_2^2x_1m_1 - 2x_1\sigma_{12}x_2 + 2x_1\sigma_{12}m_2}{\sigma_1^2\sigma_2^2-\sigma_{12}^2} + \\ & \frac{\sigma_2^2m_1^2 + 2m_1\sigma_{12}x_2 - 2m_1\sigma_{12}m_2 + \sigma_1^2x_2^2 - 2\sigma_1^2x_2m_2 + \sigma_1^2m_2^2}{\sigma_1^2\sigma_2^2-\sigma_{12}^2} \end{aligned}$$

Consequently, the conditional density can be written as:

$$\frac{f_X(x_1,x_2)}{f_{X_2}(x_2)} = \frac{\sigma_2}{\sqrt{2\pi}\sqrt{(\sigma_1^2\sigma_2^2-\sigma_{12}^2)}} \exp\left(-\frac{1}{2}\frac{\left(-\sigma_2^2x_1+\sigma_2^2m_1+\sigma_{12}x_2-\sigma_{12}m_2\right)^2}{\sigma_2^2\left(\sigma_1^2\sigma_2^2-\sigma_{12}^2\right)}\right)$$

If we look now to what is given in the proposition:

$$\begin{aligned} E\left(X_1 \left| X_2 = x_2\right.\right) &= m_1 + \frac{\sigma_{12}}{\sigma_2^2}\left(x_2 - m_2\right) \\ \mathbf{V}_{X_1|X_2=x_2} &= \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2} \end{aligned}$$

the corresponding density g is:

$$\begin{aligned} g(x_1) &= \frac{1}{\sqrt{\left(\sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right)}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_1 - m_1 - \frac{\sigma_{12}}{\sigma_2^2}\left(x_2 - m_2\right)}{\sqrt{\sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}}}\right)^2\right) \\ &= \frac{\sigma_2}{\sqrt{2\pi}\sqrt{\left(\sigma_1^2\sigma_2^2 - \sigma_{12}^2\right)}} \exp\left(-\frac{1}{2}\frac{\left(-\sigma_2^2 x_1 + \sigma_2^2 m_1 + \sigma_{12}x_2 - \sigma_{12}m_2\right)^2}{\sigma_2^2\left(\sigma_1^2\sigma_2^2 - \sigma_{12}^2\right)}\right) \end{aligned}$$

We then come to the desired result $g(x_1) = f_{X_1|X_2}\left(x_1 \left| x_2\right.\right)$.

The financial interpretation of $\mathbf{V}_{X_1|X_2=x_2} = \sigma_1^2(1 - \rho_{12}^2)$ when $X_2 = x_2$ is a signal received by an investor is quite natural. The variance of X_1 is lower after receiving the signal but the decrease depends on the correlation of the signal with the variable X_1 under consideration. Obviously, the sign of ρ_{12} does not matter because a negatively correlated signal brings as much information as a positively correlated signal.

4.4 The law of large numbers and the central limit theorem

In numerous real situations, we have to add a large number of random variables and the question is to know the behavior of the sum or of the average of these variables. For example, when studying the return of an equally weighted portfolio, we calculate an average return across a set of stocks.

The return on a long-term (say 20 years) investment in a given security is the sum of around 5000 daily returns. When successive returns are assumed independent, a large number of i.i.d random variables are added to get the long-term return. Can we say something about the probability distribution of this variable? This question is important in finance, for example when studying the *equity premium puzzle* initially presented by Mehra and Prescott (1985). Historical data show that stocks outperform bonds on the long run by around 5 to 6% in many countries. To analyze this premium, a first point is to assume a reasonable distribution for returns.

In other models like Ross' APT (Arbitrage Pricing Theory, 1976), the β coefficients on the different risk factors are obtained by building a (almost risk-free) arbitrage portfolio with a large number of assets. The specific risk is neglected because of the diversification provided by the large number of assets in the portfolio.

What is the mathematical result allowing to neglect the specific risk in a portfolio containing a large number of stocks? On the mathematical point of view, the tools used in these models are the law of large numbers and the central limit theorem. They are based on convergence of sequences of random variables. We already saw convergence in L^1 and L^2. We start this section by presenting three other types of stochastic convergence and then address the two essential theorems. It is worth to mention that a version of the central limit theorem shows the convergence of an option price in the Cox-Ross-Rubinstein (1979) model to the one obtained in the Black-Scholes model (1973).

4.4.1 Stochastic Convergences

Definition 108 *Let* $(X_n, n \in \mathbb{N})$ *be a sequence of random variables and* X *a random variable defined on a probability space* $(\Omega, \mathcal{A}, P)$;

1) $(X_n, n \in \mathbb{N})$ ***converges to*** X ***in probability*** *(denoted as* $X_n \xrightarrow{P} X$*) if for any* $\varepsilon > 0$*:*

$$\lim_{n\to+\infty} P\left(|X_n - X| > \varepsilon\right) = 0$$

2) $(X_n, n \in \mathbb{N})$ ***converges to*** X ***almost surely*** *(denoted as* $X_n \overset{a.s}{\to} X$*) if there exists a set* $\Omega_0 \subset \Omega$ *with* $P(\Omega_0) = 1$ *such that:*

$$\forall \omega \in \Omega_0, \lim_{n \to +\infty} X_n(\omega) = X(\omega)$$

3) Let P_{X_n} (P_X) *be the probability distribution of* X_n (X) *;* $(X_n, n \in \mathbb{N})$ ***converges to*** X ***in distribution*** *(denoted as* $X_n \overset{\mathcal{L}}{\to} X$*) if, for any bounded continuous function* f*:*

$$\lim_{n \to +\infty} \int_{\mathbb{R}} f(x).dP_{X_n}(x) = \int_{\mathbb{R}} f(x).dP_X(x)$$

This convergence is also called ***weak convergence****.*

These three notions of convergence appear in the limit theorems of the next section.

4.4.2 Law of large numbers

"It[8] *is difficult to understand why statisticians commonly limit their inquiries to averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once".* F. Galton

We study here the behavior of the average of a large number of random variables, by characterizing the expectation and the variance of the mean. Preliminary results will be useful to get laws of large numbers.

Proposition 109 ***Markov inequality***

Let X *be a random variable taking positive values, being integrable with* $E(X) = \mu$*. For any* $A > 0$ *the following inequality is satisfied:*

$$P(X \geq \mu A) \leq \frac{1}{A}$$

Obviously this result is interesting only if $A > 1$. Remark that no assumption is made on the type of probability distribution followed by X. It gives a bound for the probability that a given random variable goes above a given multiple of its own expectation. Markov inequality is valid in a very general framework. In particular, it is not assumed that X has a finite variance. But if it is the case, a more specific result is obtained as follows.

[8] I borrowed this citation in Koch-Medina and Merino (2003), p221.

Proposition 110 ***Byenaimé-Tchebychev inequality***

Let $X \in L^2(\Omega, \mathcal{A}, P)$ *such that* $E(X) = m$ *and* $V(X) = \sigma^2$*; for any* $B > 0$ *we have:*

$$P(|X - \mu| \geq B) \leq \frac{\sigma^2}{B^2}$$

A financial illustration of this result is given by Jorion (2000) in his book *Value at Risk* (see chapter 1, example 31). The Basel committee requires a 99% level and a 10-days horizon to calculate the VaR. The amount of required capital obtained by this calculation is then multiplied by a security coefficient equal to 3. The preceding inequality can also be written as:

$$P(|X - \mu| \geq A\sigma) \leq \frac{1}{A^2}$$

where A is a positive constant. If X has a symmetrical distribution, we get:

$$P(X - \mu < -A\sigma) \leq \frac{1}{2A^2}$$

For the RHS to be equal to 0.01, we need $A = \sqrt{\frac{1}{2\times 0.01}} = 7.0711$. However, banks often assume Gaussian returns. With Gaussian returns $A = 2.32$, that is 3 times less than the number obtained without this assumption.

Proposition 111 ***Weak law of large numbers***
Let $(X_n, n \in \mathbb{N})$ be a sequence of square integrable and identically distributed random variables (with expectation μ and variance σ), pairwise independent. Let $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$; $(Z_n, n \in \mathbb{N})$ converges in probability to the (constant) random variable μ. Moreover, for any $\varepsilon > 0$:

$$P\left(|Z_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Remark : The second part of the proposition is obtained by applying proposition 110.

Convergence in probability is not very intuitive. A sufficient condition can be obtained when square integrable variables are considered.

Proposition 112 *Let $(X_n, n \in \mathbb{N})$ be a sequence of square integrable random variables; X_n converges to X in probability (X is also assumed in L^2) if the two following conditions are satisfied:*

a) $\lim_{n\to+\infty} E(X_n) = E(X)$

b) $\lim_{n\to+\infty} V(X_n - X) = 0$

Proposition 113 ***Strong law of large numbers***
Let $(X_n, n \in \mathbb{N})$ be a sequence of square integrable i.i.d random variables and $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$(Z_n, n \in \mathbb{N})$ converges almost surely to μ.

On the contrary, if $E(|X_n|) = +\infty$, the sequence Z_n is almost surely unbounded.

Laws of large numbers insure the dividends of insurance companies shareholders. A large number of identical but independent policies reduces the variance of future liabilities. Taking into account risk aversion of agents, companies are able to require more than the expected damage (pure premium) to clients. The diversification of the portfolio of liabilities of a company allows to reduce their dispersion, generating, in most cases, a profit.

In the APT model, a multifactor structure of returns is assumed as follows.

$$r_i = E(r_i) + \sum_{k=1}^{K} \beta_{ik} F_k + \varepsilon_i \tag{4.25}$$

where r_i is the random return of asset i, $F_1, ..., F_K$ are random variables representing common factors. β_{ik} is the sensitivity of stock i to the variations of factor k and, finally, ε_i is a random variable representing the specific risk of firm i. The common factors are assumed uncorrelated ($Cov(F_k, F_j) = 0$ for $j \neq k$) and uncorrelated with specific risks ($Cov(F_k, \varepsilon_i) = 0$). Finally, specific risks are uncorrelated ($Cov(\varepsilon_i, \varepsilon_m) = 0$ for $i \neq m$).

The return on a given asset is then divided in two parts. The first one is linked to the common risk factors and the second to a specific factor. More precisely, consider a large number of stocks N; the return of an equally weighted portfolio is written as:

$$\frac{1}{N}\sum_{i=1}^{N} r_i = \frac{1}{N}\sum_{i=1}^{N} E(r_i) + \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \beta_{ik}F_k + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \quad (4.26)$$

$$= \frac{1}{N}\sum_{i=1}^{N} E(r_i) + \sum_{k=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}\beta_{ik}\right)F_k + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \quad (4.27)$$

Large portfolios allow to diversify away the specific risk, because of the law of large numbers. In other words the variance of $\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i$ tends to 0 when the number of stocks in the portfolio tends to infinity.

4.4.3 Central limit theorem

The central limit theorem explains why the Gaussian distribution is so important in all scientific fields. We provide hereafter two versions of the theorem. The first one assumes that the variables entering the mean are distributed according to the Bernoulli distribution and gives the intuition of why the Cox-Ross-Rubinstein model converges to the Black-Scholes model.

Proposition 114 ***Central limit theorem (CLT)***

Let $(X_n, n \in \mathbb{N})$ be a sequence of i.i.d Bernoulli random variables with parameter p; the sequence T_n defined by:

$$T_n = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$$

converges weakly to the standard Gaussian distribution.

This version of the CLT is not sufficient to obtain the convergence of the binomial model to the Black-Scholes model because the parameter p doesn't depend on n. In the Cox-Ross-Rubinstein model, parameters u and d depend on the delay between two trading dates and this delay tends to 0 when the number of sub-periods increases. The probability of an *up*-state also depends on the number of sub-periods. We then need a "dynamic" version of the theorem.

Definition 115 *Let*[9] $Y = \left(Y_1^n, ..., Y_{k(n)}^n, n \geq 1\right)$ *a triangular array of zero mean random variables. For any* n, *let* $s_n^2 = V\left(\sum_{i=1}^{k(n)} Y_i^n\right)$. Y *satisfies the* ***Lindeberg condition*** *if, for any* $\varepsilon > 0$, *the sequence* $U = \left(U_1^n, ..., U_{k(n)}^n, n \geq 1\right)$ *defined by:*

$$\begin{aligned} U_i^n &= Y_i^n \text{ si } |Y_i^n| \leq \varepsilon s_n \\ &= 0 \text{ sinon} \end{aligned}$$

satisfies :

$$\lim_{n \to +\infty} \frac{V\left(\sum_{i=1}^{k(n)} Y_i^n\right)}{s_n^2} = 1$$

The following proposition provides the right version of the CLT to study the convergence of the discrete-time option pricing model to the continuous-time model.

Proposition 116 *Let* $Y = \left(Y_1^n, ..., Y_{k(n)}^n, n \geq 1\right)$ *a triangular array of random variables such that the zero mean sequence* $\left(Y_1^n - E\left(Y_1^n\right), ..., Y_{k(n)}^n - E\left(Y_{k(n)}^n\right), n \geq 1\right)$ *satisfies the Lindeberg condition. For any integer* $n \geq 1$, *let* $Z_n = \sum_{i=1}^{k(n)} Y_i^n$. *If* $E\left(Z_n\right) \to \mu$ *and* $V\left(Z_n\right) \to \sigma^2 \neq 0$, *the sequence* Z_n *weakly converges to a standard Gaussian variable* Z.

This proposition is useful in calibrating the discrete-time model of Cox-Ross-Rubinstein where you need to define the parameters u and d characterizing the stock price process. u and d are chosen to keep constant the yearly expected return and variance, independently of the duration of sub-periods (see Hull, 2009, p 248-249).

[9] This definition and the following proposition can be found in Duffie, 1988, p244-246.

Bibliography

[1] Bernoulli, D. (1738), Specimen theoriae novae de mensura sortis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae,* 5, 175-92.

[2] Bernoulli, D. (1954), Exposition of a new theory of the measurement of risk, *Econometrica*, 22, 123-36

[3] Black.F-Scholes.M (1973), The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*, 81, 637-54.

[4] Cox.J.,Ross.S, Rubinstein.M (1979), Option Pricing : a Simplified Approach, *Journal of Financial Economics* **7,** 229-263.

[5] Dothan. M. (1990), *Prices in Financial Markets*, Oxford university Press

[6] Duffie.D (1988), *Security Markets, Stochastic Models,* Academic Press.

[7] Fama, E. (1970), Efficient markets : a review of theory and empirical work, *Journal of finance, 25, n*°2, 383-417

[8] Grossman, S.J. (1976), On the Efficiency of Competitive Stock Markets when Traders Have Diverse Information, *Journal of Economic Theory*, 31, 573-585.

[9] Harrison.J.M-Kreps.D.M (1979), Martingales and Arbitrage in Multiperiod Securities Markets, *Journal of Economic Theory,* 20, p381-408.

[10] Harrison.J.M-Pliska.S (1981), Martingales and Stochastic Integrals in the Theory of Continuous Trading, *Stochastic Processes and their Applications,* 11, 215-60.

[11] Huang.C.F-Litzenberger.R (1988), *Foundations for Financial Economics,* North-Holland.

[12] Hull, J.C. (2009), *Options, Futures, and other Derivatives,* Pearson, Prentice Hall.

[13] Koch-Medina.P.-Merino.S. (2003), *Mathematical Finance and Probability*, Birkhauser Verlag.

[14] Lintner.J (1965), The Valuation of Risky Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, *Review of Economics and Statistics,* 47, 13-37.

[15] Malkiel, B. (2003), *A Random Walk Down Wall Street*, W.W. Norton & Co. (original edition in 1973).

[16] Markowitz.H (1952), Portfolio Selection, *Journal of Finance,* 7, 77-91.

[17] Markowitz.H (1959), *Portfolio Selection : Efficient Diversification of Investment,* Wiley New York.

[18] Mehra, R., and Prescott E.C. (1985), The equity premium: a puzzle, *Journal of Monetary Economics,* 15, 145-161.

[19] Merton, R., C. (1969), Lifetime Portfolio Selection under Uncertainty: The Continuous Time Case, *Review of Economics and Statistics*, 50, 247–257.

[20] Merton, R., C. (1971), Optimal Consumption and Portfolio Rules in a Continuous-Time Model, *Journal of Economic Theory*, 3, 373-413.

[21] Merton, R., C. (1973) Theory of rational option pricing, *Bell Journal of Economics and Management Science,* 4, 141-83.

[22] Mossin.J (1966), Equilibrium in a Capital Asset Market, *Econometrica,* 34, 261-76.

[23] Quiggin.J (1982), A Theory of Anticipated Utility, *Journal of Economic Behavior and Organization,* 3*,* 324-343.

[24] Roger. P. (2010), *Stochastic Processes for Finance*, Bookboon.com.

[25] Roll.R-Ross.S (1980), An Empirical Investigation of the Arbitrage Pricing Theory, *Journal of Finance,* 35, 1073-1103.

[26] Ross.S (1976), Options and Efficiency, *Quaterly Journal of Economics,* 90. 75-89.

[27] Ross.S (1976), The Arbitrage Theory of Capital Asset Pricing, *Journal of Economic Theory,* 13, 341-60.

[28] Sharpe.W (1964), Capital Asset Prices : a Theory of Market Equilibrium under Conditions of Risk, *Journal of Finance,* 19, 425-42.