WILEY FINANCE

Economics and finance of risk and of the future

ROBERT KAST
ANDRÉ LAPIED

Economics and Finance of Risk and of the Future

Robert Kast and André Lapied

John Wiley & Sons, Ltd

Economics and Finance of Risk and of the Future

For other titles in the Wiley Finance Series
please see www.wiley.com/finance

Economics and Finance of Risk and of the Future

Robert Kast and André Lapied

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging in Publication Data

Kast, Robert.
Economics and finance of risk and of the future / Robert Kast and André Lapied.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-470-01577-3
ISBN-10: 0-470-01577-2
1. Risk management. 2. Risk assessment. 3. Decision making. I. Lapied, André. II. Title.
HD61.K37 2006
368—dc22

2006002083

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13 978-0-470-01577-3 (HB)
ISBN-10 0-470-01577-2 (HB)

Typeset in 10/12pt Times by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Acknowledgements

As always in the case of a research synthesis, our thanks go to all our colleagues who brought up constructive criticisms during congresses and seminars, particularly to the participants to the Risk Utility and Decisions (RUD) and the Foundations of Utility and Risk (FUR) conferences. We'd like to mention specially our colleagues Marie-Cécile Fagart and Meglena Jeleva who were kind enough to read through and make us correct chapters concerning insurance and Alain Chateauneuf for his collaboration on previous works.

The idea of the book came from the difficulties we encountered to make understand the notions developed by the theories, at the application level: Our thanks go to our collaborators in applied studies and particularly to our then Ph. D. student Sophie Pardo. Last but not least, we are very grateful to Professor of Finance Eric Briys, invited professor at Université des Antilles et de la Guyanne and co-founder of Cyberlibris, for his always renewed encouragements in our work, which converged toward writing this book.

General Introduction

This book aims to reconsider the analysis of risks with the objective of proposing new or adapted management instruments for problems we face today. In the first section of this General Introduction we shall recall some historical facts about financial assets. The second section introduces risk, with all its different meanings, and the way it relates to financial assets and other management instruments. The third section addresses the perception of the future in a general way and introduces the new problems in risk management. The last section presents the solutions that have been found for some of them and dispatches them among the three parts of the book.

FINANCIAL ASSETS

Financial assets very likely appeared with the first historical civilisation in Sumer (South of Iraq), some 3000 AD, to promote production plans. Much later (1000 AD) but still in Mesopotamia, clay tablets in cuneiform writing from Babylon and Assur relate contracts between a borrower and a lender: they indicate the money amount (or the commodity), the warrantee, the term (expiration date) and the payoffs. Interest rates (around 20% in general) were mentioned or can be deduced from the written payoffs. They include some contingencies and ways out in case of default or of difficulties, and the warrantees were rather important.[1] Such contracts could be signed and honoured in a well-established legal and political system that enforced them. A key to this system was the official religion. A contract was placed under the authority of a god, the priests and the political system relying on it were there as witnesses. Faith and confidence in the god(s) assured that the terms of the contract would be honoured: the lender was "certain" that the borrower would do the necessary to fulfil the contract's clauses.

A financial contract was (and still is) a claim that promises that some payoffs will be paid by the borrower to the lender, under some conditions that are spelled out in the contract's terms. In Mesopotamia and during the next 5000 years or so, conditions mentioned some events that could happen, independently of the willingness of the contracting parties, but they mainly related to time periods that represented the future to both parties. Time periods were generally months or years.

In modern terms, such financial assets are considered as "riskless", in the sense that the borrower was assured that the payoffs will be honoured at the times set by the contract. Such an assurance does not mean that the borrower was not aware of taking some risk (in the modern sense), it only means that he or she relied on the respect of religious oath and the political and juridical power that went with it. There were still some uncertainties about

[1] Crawford (1991).

the future, even though neither the word nor the concept existed, and the future was only perceived in terms of time, as measured by the calendar.

RISKS

"Risk" was the first word to appear in European languages to address uncertainty. That was around 1200 in Venice (*rischio*).[2] Other words, such as "uncertainty", only appeared much later. The word was used to name a concept: a cargo, which had been shipped in the hope of trading it at some profit, could be destroyed if the ship met a reef.[3] There was nothing new about the fact that valuable cargos may not arrive safely. What was new was to say it in one word, or to have a relevant name for a new concept and for instruments to manage it. The problem facing a risk investor was not related to the dangers faced in shipping: sailors knew how to manage these risks (hazards) in their best interests. It wasn't commercial either: traders knew how to buy and sell at the best prices. Neither could the problem be solved by bankers: at that time, their job consisted of lending, borrowing or holding money at interest, i.e. what we would call "riskless" investments today. However, it was this group, the money holders, who needed to learn a way to deal with investments in cargos at risk. The problem they faced originated in religion. Indeed, fickleness of fortune, caprices of the sea, etc. depended on the ancient gods, and on the unique God at that time. The three religions based on the Good Book were present all around the Mediterranean and all of them condemned magic skills. Hence, it was blasphemy to bet against what was considered as God's will. In order to escape the religious domain, it was necessary to create a new range of skills based on observed facts. The intuition was to designate risk by something concrete in order to be free from any magic or religious suspicion.

The facts are the following:

A money amount (certain) is transformed into a list of possible future payments.

This is what we call a financial risk.

To make the concept concrete, contracts were written so that they could be directly traded if their holder could not face the risk any more for some reason.[4] Three types of (risky) contracts can be distinguished nowadays:

1. Risk sharing is achieved by contracts saying that payments are proportional to investments (e.g. securities).
2. Lending or borrowing at an interest rate above the usual term interest rate, because it integrates a "risk premium".
3. Insurance contracts, which transfer the risk from one economic agent (the insuree or insured agent) to another (the insurer).

Each type of contract requires particular expertise skills. In fact, the first contracts to appear combined all the sides of risk and it is only much later that finance and insurance split

[2] A very interesting story of risk, insurance and financial markets can be found in Briys and de Varenne (2000). See also: Bernstein (1998).

[3] The word risk comes from *rhiz ikon* "image of a reef" in Constantinople's Greek spoken by sailors. Venice and Constantinople were the main trade centres in the Mediterranean and commercial partners/competitors.

[4] David Schmeidler told us that tracks of similar contracts have been found in Tyre (actual Lebanon) for shipping to Carthage (actual Tunisia), some 1500 years before, but the notion of risk did not arise.

apart into distinct disciplines and businesses. Today, extensions of risk coverage and risk consciousness have led to the remerger of these areas, and we can find financial contracts for insurers which are as complex as the first risky contracts in the Middle Ages. This is due to managing problems caused by new situations that introduce risk levels higher than usual into the insurers' portfolios. Some natural and industrial catastrophes cause much worse casualties than they had used to, "new risk problems" such as "mad cow disease" or the Aids endemic disease are not sufficiently well known by scientists to forecast or to prevent them, etc.

Health and environmental problems have opened many questions regarding the usual way of thinking about risks. We face a problem with respect to managing human activities because they can harm our environment and our health. As humans, we also face hazards caused by our natural and social environments. Hence, we face a disequilibrium of an ecological type between the sphere of human activities on the one side, and nature on the other.

Such a disequilibrium opens a management problem, because it threatens each party with some risks. Natural and industrial catastrophes generate risks that have to be managed when natural sciences are not able to offer satisfactory technical solutions to the reduction of such hazards. Managing risks means investing in preventive devices, insurance contracting and forecasting future investments with unwarranted payoffs. In turn, the management problem opens one in economics, in which lies the origin of the risk concept.

If one looks at the facts, "new risk problems" are not newer than "old" ones were when the word risk merged into the language. Catastrophes, hurricanes and environmental damages have always threatened humankind. But the situation has changed: risk covers a huge range of meanings, the place then held by religion now belongs to science and people often think of engineering as a kind of magic skill able to protect them from any hazard. Above all, the consciousness of risks, which used to be confined to individual responsibility, then rose to the public sphere when methods of prevention became available. Nowadays the limits of these methods are reached when:

- The public sphere outranges any collective organisation and/or its ability to manage risks (international level).
- Controversies arise over which scientific theory applies to a given risky phenomenon.
- Science is at loss in the face of some new hazard.

Economics is not concerned with looking for means to prevent risks, reduce them or turn them away. Neither is economics in charge of understanding how risks are perceived, how much importance is given to them or what regulations of responsibility are acceptable. Risk problems enter into the economic sphere and can be studied in financial terms only when several conditions are satisfied:

- The phenomenon behind the risk is well described.
- The society and individuals who are concerned by taking the risk, want to avoid it or reduce it, or are willing to share it, are well identified.
- The time horizons at which expected outcomes will be observable are known.
- The law is able to enforce clauses settled after economics and finance have helped to design the contracts.

Risk management problems are considered as "new" if at least one of the above conditions is not satisfied. For example, "mad cow" disease didn't satisfy any of them when it first appeared. Environmental risks also, even though some of them only question the first and the last condition. The first one because the phenomena are complex and rarely belong to one scientific discipline, while the last one would require an international power able to enforce laws.

When a risk is well defined, economic calculus can investigate money amounts to invest and expected payoffs from this investment to be compared with future losses if it were not done. Also, it looks for the best method to apply to figure out the amounts and to compare them. Economics is concerned with interactions between economic agents, as well as with the political implications of regulations and economic incentives. Economic theories have been developed and are referred to by economic calculus and by finance in order to propose methods and instruments with which risks can be managed.

UNCERTAINTY AND PRECAUTION

Economic calculus and financial market theory have developed considerably since both appeared formally during the 1950s.[5] A difference in the ways of thinking about risk between the beginning and the end of the 20th century is related to the consciousness of science's limitations. In particular, it is commonly admitted today that there may be some scientific uncertainties and that these may hold, whereas the typical behaviour during the 1950s was to turn to scientists to erase any uncertainty. This behaviour was inherited from the 19th century, during which the incredible rate of development of all sciences gave the public an impression that everything would be understood some day and become ultimately under control.

What is the difference between common uncertainty and scientific uncertainty? Industrial societies were built on technologies based on scientific knowledge. Science became society's reference during the 18th century, when new technical knowledge was developed from scientific findings (instead of know-how, i.e. particular skills developed on the basis of pragmatic experiences). For more than a century, industrial countries believed that science could explain every phenomenon and that methods would be found to master or to control these. Even though many physicists had announced and proved as from the end of the 19th century that this was not possible, this faith was maintained among people. What is true, though, is that science and techniques have developed at an increasing rate during the last two centuries. It is indeed true that many phenomena are understood and under control, reinforcing the popular confidence that the sciences are in charge with most risks eliminated and the prospect that we will know everything with certainty some day.

However, what is certain? From the Latin *certus* we understand that something is certain if it is "certified". Without some certitude, no one would undertake anything: certainty settles, secures and makes one find the courage to engage in enterprises. However, certitudes may induce actions that could turn out to generate irreversible damages. Before the development of the sciences, certitudes were obtained from the gods who spoke through oracles. Since the 18th century, we have turned to the sciences to obtain some certitudes about the phenomena they have studied. Today, something is certain if it is certified as such by a group of well-known scientists. Now, what do the sciences mean by "certain"?

---

[5] Economic calculus was initiated by Colbert, a French Prime Minister during the 17th century, but not many uncertainties were taken into account then.

Science elaborates theories able to integrate observed facts. Theoretical predictions are tested, if possible, through repeated experiences, or by observations. Observed results are never identical to predictions. For any theory, there is a lower precision level for which it is true, and a higher one for which it doesn't pass the test. At the given precision level, generally chosen according to the capacities of the instruments, the theory is either verified or not. If it is, it can be relied upon for applications, which will produce results that are certain *at that precision level.*

In experimental sciences, observations yield statistics, from which estimates of the resulting frequencies are computed. A result verifies a theory if it is the empirical mean of the experimental results. The precision level can be defined by the variance of the sample, or any other dispersion measure. A technical device can be derived from the theory if, and only if, it is supported at this precision level.[6] Some sciences study matters that are difficult to put into experiments, notably the social sciences. However, observed facts yield statistics as well. For example, econometrics was developed to study such observations. Mathematical statistics have shown that such statistics do not allow one to obtain results as precise as those derived from experiments. Still, a theoretical result must be close to the empirical mean of the observed results.

Scientific certainty is defined by a mean value, or, more precisely, by a probability distribution.

Probability calculus was invented by the French scientist and philosopher Pascal, in the 17th century, and the Dutch astrophysicist Huygens wrote the first work on probability theory during the same period. Both were interested in games of chance and their findings are related to the kind of uncertainty that can be obtained through a mechanical device, e.g. a roulette wheel. Such devices make it feasible to prove[7] that an event will occur with a given frequency.

Chance comes from the Latin *cadentia*: the thing that falls or happens regularly. It addresses the type of uncertainty that can be found in games where the outcomes occur regularly. Humans have always been fascinated by games of chance, particularly those who need to train themselves to face dangers (soldiers and sailors) or those who get bored when they have no dangers to face! Most chance games originated in magic practices to learn something about the gods' will.

Games are a concrete formalisation of situations where a decision with uncertain consequences has to be taken. In 1933, Russian mathematician Kolmogorow proposed to formalise uncertainty by a space of states of nature, Ω, together with a set of events F satisfying some properties (algebra or σ-algebra) and a probability distribution P assigning a number to each event. The "probability" space (Ω, F, P) is totally unknown and not observable. Ω is such that any random variable can be considered as a mathematical function from this set to the set of numbers. Kolmogorow's probability space represents all possible circumstances, past, present or future, which can be imagined (or not!) to be a cause for a particular outcome of any chance device. It is easy to represent with such a probability space uncertain situations, particularly those that are similar to chance games.

---

[6] For example, in most European countries, electrical power is certified at 220 V (110 V in America). This means that it varies between 180 and 250 V, only outpassing these limits with a very low frequency. With this certitude, electrical devices are built in a way to bear any power varying around 220 V and are certified not to break down.

[7] Probable used to mean: that can be proved.

Probability theory is the mathematical theory underlying mathematical statistics. Probability and statistics were developed as a foundation for sampling, tests and estimation methods, and as such they also underlie the "theory of errors" in physics. In this last application, probability became *the* measure of scientific uncertainty.

More recently, mathematics was able to show that probability can measure individual likelihood perceptions (subjective probability: de Finetti (1930) and Savage (1954), see Chapter 2). Probability measures are derived from a preference order ranking the results. Likelihood, statistical frequencies, chances and errors, are all measured by probability distributions, although these distributions have different meanings in each case. A scientific theory relies on one or the other of these definitions in order to predict a result's occurrence, and this is what makes the result "scientifically certain". We will see in Chapter 2 that it so happens that decision theory and economics restricted the original interpretation of risk to this type of uncertainty, which was then called risk or "situations of risk"!

Is uncertainty the opposite of certainty? If so, for instance in scientific cultures, then uncertainty is scientific uncertainty. This has at least two meanings:

- There are several theories in competition, all of them able to describe the same phenomenon. Hence, several scientists may certify different results.
- No scientist is able or willing to certify any consequence of some phenomena.

Scientific uncertainty is one of the characteristics of new risk problems that are not easy to deal with. The different formulations of the precautionary principle mention it. The principle appeared in discussions by several international committees which were concerned with environmental problems. Since it was signed by some states, it is invoked for many risk management problems, including pollution, wastes, public health and industrial incidents. Besides scientific uncertainty, i.e. the fact that uncontroversial probability cannot be ascertained for some event, other elements are always present in the precautionary principle statements:

- Precautions should be taken without delay. More generally, given that information will come along with time, it is necessary to take it into account. Uncertainty can be reduced (in some sense) in the future, hence present actions should not stop future ones from using information, and processes should avoid irreversible situations. Information can also increase uncertainty in the sense that a theory accepted today can be proved wrong later. For instance, a probability zero event may happen, or become highly likely to do so.
- Precautions have a cost. This cost should be economically and socially acceptable, and costs are to be compared with expected benefits. Individual responsibility is not the only one at stake if the consequences of someone's actions do not concern that person only. This is true already, when prevention is in order, and in this case costs are compared with the benefits of avoiding hazards. However, when precautions are taken, there is not enough knowledge about consequences to enforce them by invoking solidarity only. This is particularly true when the collective choice concerns countries with different cultures, interests, political and legal powers, etc. Precautions require some coordination of means and this is hard to achieve when no uncontroversial benefit estimation is done.

PROBLEMS: NEW METHODS AND NEW INSTRUMENTS

The evolution of risk economic analysis has proceeded in two directions that, in spite of many formal features of resemblance, prove to be opposite from a methodological point of view. One is given by the developments in individual decision theory and the other one follows the study of markets. Each one has some relevance, provided that one takes care to apply them in their own field. The decision theory range of applications is individual risks. Markets are mechanisms that make agent behaviours compatible and that yield aggregated risk measurements. Confusion between individual value and market value is to be avoided absolutely. These two measurements would be equal only if all agents were identical, but then we would be facing a paradox: these agents would not have any incentive to exchange and this would empty the concept of market from any significance. Attempts to justify the confusion of the two values by the fiction of a representative agent is not convincing, because it only holds if the heterogeneity of individuals is ignored. It has been known for a long time that properties of individual behaviours are not necessarily found at the aggregate level. To choose a simple example, if each agent consumes in proportion with its income, national consumption would be proportional to the national income only insofar as all agents are identical (in this case the proportionality factor is the same). Conversely, a property at the aggregate level is not always essential at the individual level. An invariance in a sum of behaviours can hide increases and reductions among its terms.

Individual decision theory experienced many developments because of the questions concerning the dominant paradigm of expected utility. This scientific evolution has an empirical origin: the standard theory is not able to explain many individual behaviours in the face of risk. It then progressed simultaneously in two directions: generalisations based on less demanding axioms, on the one hand, and research for a coherence with empirical results obtained in experimental economics, on the other hand.

Such theoretical extensions led to the enlargement of the scope of individual attitudes in the face of risk, and the renewal of the traditional approach to demand for insurance. This last field also profited from the analysis of risk selection by insurers (i.e. ways to distinguish "bad" risks from "good" ones) and from the study of methods to encourage policy-holders to adopt a careful behaviour. More recently, the panoply of instruments available to economists and experts has been enlarged by taking into account background risk: i.e. an exogenous risk that can influence decisions in the face of a specific risk because of hedging this risk or increasing it.

The search for a better knowledge of the risks undertaken by the credit institutions remained, for a great part, within the framework of modelling individual behaviours. What is at stake is the checking that the probability of ruin is contained within acceptable limits. The reasoning extends then to a more general analysis of solvency. The results obtained in this field greatly influenced the management of specialised establishments. They involve, indeed, a reflection on the allocation of capital to various activities. The respect of the equity ratio requires that capital is available to warrant their feasibility. Given the resource is not inexhaustible, it becomes crucial to select its uses, admittedly according to their expected profitability but also according to capital consumption, and thus to their risk. One then ends up with the traditional behaviour in finance: arbitrage between expected return and risk.

With regard to non-commercial goods, the absence of price must be circumvented to define a value common with commercial goods. This will then allow cost–benefit analysis to be performed. A market approach can be relevant for non-commercial goods, even if that can seem paradoxical. In this case, prices will yield an indirect measurement extracted from traded good prices: for instance, the value assigned to one of the characteristics of a private property will price the same characteristic in a public good. One of the major ways in this field of research led to the building of investigations in which agents are made to reveal the individual value they grant to public goods (i.e. their willingness to pay). This approach profited from enriching collaborations between specialists in various social sciences (sociology and psychology). However, as for any method founded on individual behaviour, its application to collective choices comes up against the aggregation problem. This question could find a solution in the use of market games, which try to replicate the market aggregation mechanisms through experimentation.

The financial theory of valuation developed general principles that led to fruitful advances, both in finance and in economics. They rely on the basic structure of the markets, and they connect prices of various assets. This method yields a characterisation of the market price of risk and, consequently, of a decomposition of an asset's expected returns in two parts: the riskless return representing the price of time, and the risk premium corresponding to the value of the asset's risk. These results have been brought close to traditional economic models of equilibrium so as to consolidate their theoretical basis. Applications of these principles to the valuation of a great number of new financial instruments (options, futures, swaps, etc.) contributed to their successes. Their merits also extend to the field of investment choice, with the theory of "real options".

Market approaches also found a justification in the development of securitisation, which transforms individual risks into tradable credits. These evolutions were followed in several fields. For example, the field of insurance saw the development of "cat-bonds" (catastrophe bonds, i.e. bonds linked to a catastrophe index) and "weather derivatives", while mortgage claims and automobile loans were introduced in the field of finance by securitisation.

Among the above problems, most have found satisfying solutions that allow us to manage risks with new instruments, based on solid theoretical grounds. These instruments prove both handy and efficient in practice, or promising when they have not been fully tested.

PRESENTATION OF THE BOOK

The structure of this work is organised according to three topics treated in separate parts. This choice does not claim to avoid any overlap or offer an equitable treatment to all the questions posed in this field. However, it makes it possible to put forward the problems often left in the shade by applications in economics as well as in management science, and which seem to us particularly important from the theoretical point of view on which they rely. Each one of these topics is presented in the form of an opposition between two alternative modes of representation. The goal is not to promote a universally preferable approach and to stigmatise any other way, but to show the importance of the choices usually carried out in these fields and to propose to bring the approaches closer to their relevant context.

Several situations may be faced that will yield different solutions to the two valuation problems: monetary valuation of non-marketed consequences, and present valuation of monetary consequences. In each part of the book we distinguish situations where these two valuation problems may be confused. Depending on the situation, the solutions to the decision and valuation problems will be deduced from different theories. The managing instruments suggested by the theories must be clearly distinguished depending on the situation to face. The three situations we focus on are:

- Who is making the decision?
- How is the uncertainty about the outcomes represented?
- What is the timing relevant for decisions to be made?

The first part of the book makes the distinction between collective and individual choices. The reference example is the Millau viaduct project in Southern France. It is a public project, and as such corresponds to collective choice. However, the project has been realised by a private company, hence most decisions are taken by an individual (in the legal sense).

The second part of the book distinguishes between two different uncertain situations: there is a known probability distribution over the decision's consequences, or there are no uncontroversial ones. The first situation is relevant for insurable risks: insurance companies rely on a probability distribution that they know. The second type of uncertain situation presents risks that are not easily insurable by classical insurance methods, instead they rely on financial methods.

The third part of the book opposes two ways to take time into account: a static approach differs from a dynamic one. The same risk problem may be considered from one or the other point of view, however the managing methods and instruments may differ greatly if information arrivals are not taken into account.

In a real case study, the three distinctions on which each part is constructed can overlap and the choice made according to one domain may interfere with the ones made in others. Most of the technical studies for the Millau viaduct project were based on well-known phenomena and probabilities of hazards were available. The eventuality of never-observed geological problems (e.g. earthquake) would have led us to face a situation of uncertainty. Without objective probabilities measuring hazards, we are bound to face controversies over the likelihood of a catastrophe, making the public choice much more complex.

Taking dynamic aspects into account and choosing a flexible project so as to take a range of options into account, makes it easier to accept a public investment because it opens the door to modifications of the initial decision.

Individual decision-making under uncertainty in a dynamic setting presents particular problems. If agents' behaviours do not satisfy conditions under which their choices can be represented with a subjective probability (as is the most common feature observed in experiments), dynamic consistency cannot be taken for granted. This would lead to *ex post* rejections of a plan previously accepted.

Each of the three parts we have just presented above is separated into four chapters. The first one concentrates on examples. The goal is to give evidence to problems that are opened in the corresponding domain and to enlighten the importance of the chosen distinction. The second and third chapters are devoted to economic and financial theories relevant to the analysis of the questions at stake. Each of the chapters presents one of the theoretical facets of the chosen distinction. The last chapters propose managing instruments. These tools are adapted to the problems presented in the first chapter's examples and they are founded on the relevant theories exposed in the second and third chapters.

Part I

Individual vs Collective Choice

INTRODUCTION TO PART I

This part of the book is built according to the distinction between individual and collective choices. The theoretical importance of these differences has already been explained in the General Introduction, and we will limit ourselves to illustrate it through the example presented in Chapter 1:

- Is the construction of the Millau viaduct concerned with collective choice, insofar as the state is responsible for the road infrastructures, or individual choice, since the manufacturer and delegate manager of this work is a private company?
- Is the public–private distinction relevant for this topic?

It is not the mode of property (public or private) that is important here, but the selected objectives. Indeed, the study of a possible investment in a purely productive logic by a public company cannot be restricted to the scope of collective choice. Individual decisions optimise the objective of an agent, while collective decisions take into account the objectives of several agents.

For the majority of economic questions, states turn to markets to organise exchanges and to determine values (through prices). This basic tendency, backed by arguments concerning the economic efficiency of competitive markets, is also supported by the bankruptcy of systems having adopted other structures of exchange (imperative planning). It follows that the sphere of public economic decision is gradually reduced to public commodities (commodities the consumption of which by an individual does not decrease the possible consumption of the other agents) and to externalities (non-market effects). Public economic calculation (known as cost–benefit analysis) is thus primarily distinct from private investments when they are deprived of the integration of externalities. If they are positive, externalities increase the value of the project. For example, there are positive effects on the regional economic activity due to the viaduct construction. In the opposite case, the investment value must be lessened, e.g. because of environmental damages. The private decision-maker does not have any reason to take into account these variations of value, insofar as, in the absence of internalising mechanisms (such as pollutant pays principle), external effects do not yield any profits from favourable consequences and do not pay for the unfavourable repercussions. In contrast, the public decision-maker who is supposed to take into account the whole population's interests, will retain these additional advantages and disadvantages. Thus public choice may differ from private decision and an economically unprofitable project may be accepted or a project with positive returns may be turned down. A public decision investment system related to granting a concession to a private company can cause dysfunctions. In the case of the Channel tunnel, the states tardily added constraints to construction in order to internalise external effects, reducing the financial profitability of the project, with the well-known consequences observed on financial markets for the tunnel company's shares. This exemplifies the differences between public and private valuations.

In this first part, Chapter 1 describes the Millau viaduct project and the problem that had to be solved before the decision was taken and the bridge achieved (inaugurated in December 2004). Most of the problems to analyse in order to manage risks are present in this project. In particular, the opposition between individual and collective choices is underlined.

Chapter 2 concentrates on individual decision theory: classical models, which are the most referred to, are obtained as special cases of a general theory developed since the 1980s. The assumptions are underlined because they are crucial for the relevance of the model used to value and manage a risky project.

Chapter 3 introduces difficulties met to aggregate individual decisions in a collective choice. Then a first theory founding a collective valuation process is presented: competitive markets. When available, competitive markets yield valuation for risks that is an aggregate of agents' behaviours.

Both individual and collective risk analyses and valuation propose management methods and instruments for risks, as presented in Chapter 4.

1
Risks in a Public Project: The Millau Viaduct

The project of constructing a huge bridge for the central North–South freeway axis in France, opened many fundamental problems in public decision processes and risk management. We focus on this example to illustrate risk-taking problems because of its width, its strategic importance for managing a region and because of its socio-economical as well as ecological impacts. The project was introduced by several individual propositions: they were then shared between groups of decision-makers, discussed among meeting participants, violently defended on some occasions or abandoned for new ones when they were more easily agreed on. In the end, the choice had to be collective in some sense we shall investigate, even though there is only one person (the Minister of Transportation in this case) responsible to sign up the final decision.

The choice process is complex, more or less organised according to state regulations, and its characteristics vary depending on the level at which partial decisions are taken. We shall mention some of the process features that we need to understand, but the process itself is outside the scope of economics and of this book. A main trait we are particularly interested in for this part of the book, is the distinction between individual and collective choice at different stages of the process.

The choice is individual if a unique person or a group speaking in one voice makes a decision, it is collective when an assembly uses some process to choose among alternative decisions. In both cases, the main arguments for a project's pros and cons rely on previous studies of its possible impacts. There are many of them, at all stages of the construction and even during the decision process itself. Some impacts concern individual agents,[1] others are collectively perceived. The way impacts are perceived results from reactions to previous, similar ones. Indeed, individual behaviours about the construction project may modify its consequences. For instance, the announcement of the project has induced some speculation about the concerned lands, hence increasing costs. During the construction works, thousands of visitors invaded the plant so that Millau township had to organise them in a way not to perturb works. Furthermore, overseers had to be employed to prevent some visitors taking and causing risks, because of their hazardous behaviours. On the other hand, constructions induced many lateral activities in the area, generating incomes and attracting workers.

Public investment decisions follow a regulated procedure, which is codified in France[2] and in many countries; it may be based on jurisprudence in other countries. In all cases, except

[1] By agents, we mean economic agents, i.e. individuals as well as firms or official organisations.

[2] Concerning transport, a law article makes precise the internal transportation regulations (LOTI, 30 décembre 1982). Furthermore, the state-owned consulting group edited a Recommendation Guide on Economic and Social Valuation for inter-cities road projects: "Évaluation et économique et sociale des projets routiers interurbains", SETRA.

maybe in a dictatorship, we can find some common elements in the preparatory studies. The main ones concern a study of impacts and a cost–benefit analysis: they serve as a basis for the political decision, which is taken in the last rounds of the choice process. In France, the political system is founded on the people's representation by assemblies, which sit at different levels: town, county, regional councils and finally the national assembly. In the case of the viaduct project, all of these councils have been consulted and they presented alternative issues or chose among the proposed ones. In the last round, however, the minister in charge of public transportation ended the process. This last decision was driven by technocratic civil servants, who were in charge of studies and reports of the different assemblies' decisions. Studies were done by consulting groups in engineering, by experts in geology and by some economists,[3] at a rather theoretical level for the latter. Indeed, implementing practical impact valuations according to the theories and tools presented in this book would have required a number of workers and a lot of time, well beyond the willingness to pay of the Ministry. In fact, the idea of this book and some of its materials came partly from our amazement at the ignorance of engineers and technocratic civil servants about the existence of managing tools and theoretical means of study, for risky investments.

A study consists of producing a complete list of possible impacts, having several alternative solutions to propose and finally following a comprehensible comparison process between impact valuations and solutions, in order to guide decisions that will be made at a political level. Impact valuation and analysis present many difficulties. Identifying concerned populations and environments can take a long time and one must be prompt to add to the list during the study. Valuation methods based on inquiries require the elaboration of consistent and credible scenarios to be presented to the population sample. Uncertainty opens up a whole new problem in itself. First of all, it is generally controversial. Even at a technical level, architects do not agree. Engineers with the same background disagree about some risks and their management during the construction process. At the population level, unreliable information, ethical values and cultural prejudice interfere with an objective description of uncertainty and its valuation. Furthermore, too much focus on hazards, damages and inconveniences among the population, as well as by people who could be held responsible, biases judgement on the balance that must be found between possible benefits and losses. Last but not least, uncertainty quantification, i.e. assigning probability measures or at least a ranking among likelihoods, is highly controversial in general. Add to this that any retained possible event requires a new scenario in the inquiries and a partition of the population sample, given that each subsample can be presented with only one credible situation.

The study of alternative solutions is more technical, but raises other types of uncertainty and controversy on technological choices and on risk valuation and control. These uncertainties regard technical means as well as financial managing instruments. Whatever the approach chosen for the study, a method must be elected to compare solutions. Two extreme cases can be considered: the first one would try and take all the complexities into account, while the second option relies on the analysis of solutions only.

In a centralised procedure such as the one here, the first case is not convenient because complexity makes choices rather obscure and prevents a clear justification of opposable solutions. The second one is flawed by its arbitrary simplifications and reductions, which induce us to neglect some relevant elements that were present in the studies.

[3] Among them, the authors of this book who presented some of the material contained in this book.

Economic analysis is considered as belonging to the second category. Cost–benefit analysis is typically presented in a particularly simplified and reducing manner. For instance, it would be very reducing indeed to consider net benefit as a decision criterion: it is only an indication, summarising the economic and financial analysis in order to guide political decisions. Economists, as well as the other experts called in the study, should be careful to present the assumptions and their limitations together with the results they obtained. Then, decision-makers should take these limitations into account to ponder the results and make their choices in accordance with the importance they attach to them.

The viaduct in question relies on two high plains in the mountainous relief of central South France. This region has experienced many changes during the last two centuries. From highly populated it became one of the least populated, most inhabitants having migrated towards northern industrial towns. The land is poor but it used to be self-sufficient, today however, the main agricultural activity consistent with modern production criteria is raising sheep for milk and cheese.[4] Hence, many changes have prepared the population to adapt and organise particularly well in front of risks and mutations. For instance, in this region, there are a huge number of friendly societies and tiny mutual insurance companies. Many young people from all over Europe came to live on the high plains during the 1970s; the land was inexpensive and returning to country life was the fashion. Most communities collapsed, but many families stayed and are now living on their land; they are very sensitive to environmental, biological, food and social change. The native population is close to the land and aware of the quality of life, food, natural things, etc. It doesn't share the same values with the immigrants, but both groups have the same kind of concerns. The region used to be isolated: rather poor, wild, with a harsh climate and few roads. The construction of a central freeway connecting Amsterdam to Barcelona had, among its collateral goals, breaking the isolation of the area. The axis was nearly completed in 2000, but a last difficulty was still unsolved: crossing the River Tarn's deep canyon. This is what the viaduct achieved in 2005, after many discussions before deciding on: the project, the location of the bridge, the type of construction and, finally, the company to do the work. The population is and has always been concerned and consulted through its elected representatives, trade unions and non-governmental organisations. Its main questions were about:

- Impacts on environment (pollution, noise nuisance, landscape).
- Economic impacts on the town of Millau and on the region.
- New openings and closures of access routes to some valleys distributed by the freeway exits.
- Changes in tourist activities, notably the bridge attractiveness as an art and technological object.

Several studies have been conducted by national organisations as well as by local councils and associations. When one looks at the project's long history, political difficulties are evident. It took 25 years between the first decision on the freeway construction and the last decision on the Millau viaduct. Studies on the crossing of the river and on the type of bridge started in 1988. There were three possible locations: East, Median, West. The Median one was retained after one year. Then, there were two types of project: High or Low. The High one took one more year to decide on, but this option was battled against by several

[4] The blue cheese "Roquefort" is the most famous among them, and Roquefort is a village close to the viaduct.

organisations on the basis of various arguments. This option imposed a gigantic viaduct (the actual one is 2500 m long, with a platform at 200 m high and central piles of more than 300 m high). Finally, there was a competition among architects and Norman Foster's project won. That was in 1996, the works started in 2002!

One among the many reasons why there was so much delay in the decision-making process is that, at each stage of the discussions, controversies arose about the studies' conclusions and the relative importance of impacts.

As a matter of fact, study conclusions have not always been retained. It began with the first one – performed by a state-dependent consulting company road and transportation specialist, CETE,[5] in Aix-en-Provence, it compared the two types of construction: High or Low. It concluded that the High solution had to be withdrawn because it was geologically too hazardous. Obviously, someone in the Ministry of Transport thought differently and put forward the visibility of the High solution, in contrast with the usual Low solution type of bridge, and that was the one which has been worked on and eventually achieved. This first study concentrated on the technical problems and on geological risks, and it only made a general overview of impacts. It could have been expected that such a technical point of view would have been less controversial and indeed, the Ministry decision took for granted that no one would question its technical expertise. It happened to be wrong, as a "Comité de Propositions pour l'A75"[6] was formed in Millau, under the leadership of a retired engineer (from the same Ministry!), which performed a counter expertise. The conclusion was the same as CETE's, only it went further in the studies, particularly about geological hazards. These are of two kinds: one is relative to the stability of the bridge and had already been put forward by the first study; the other one is relative to ecological impacts. This region is made up of a karst (i.e. soft and hollowed limestone), and serves as a water reservoir and filter to the whole South of France water alimentation. The hazards induced by a heavy bridge and an expectedly important traffic, are relative to infiltrations of polluted matters. The construction weight crackles the rocks and rain water charged with oil, tyre tracks and all kinds of dirt carried by automobile wheels, pours into the cracks and gets mixed up with deep fresh water. Mastery of these hazards imposed more constructions: foundations, consolidations, drains, sewers, etc. and hence higher costs. About costs, the total proposed amount for the construction was €225 m, which was considered highly underestimated by most consultants and opponents to the project. Indeed, another huge bridge above the River Seine (Tancarville), much shorter and lower than this one, cost four times as much, and that was 10 years before! The argument for a lower cost for a bigger bridge was that construction techniques for such bridges are better known, require less qualified workers and some materials are less expensive. In the end, the final estimated cost was €396 m and not only did the contractor accept it, but actually did better (€394m). We'll come back to this point later on.

It is much more difficult to oppose arguments and compare costs when economic and ecological impacts are studied. This is due, in part, to separations that are usually followed in studies between some types of decision consequences: technical, socio-economic and ecological.

In this book, we'll argue that, up to certain limits, all impacts can be approximately valued, whether by market prices or by individual valuation. Obviously, some valuations may cost

[5] Centre d'Etudes Techniques et de l'Equipement.

[6] Comittee for proposals about the A75 (A75 is the name of the freeway).

more to obtain than they are worth to help decision-making. This is one of two reasons why consultants distinguish between types of consequences. Technical implications are accounted for in the range of costs, including future options and additional works. Costs are well known if the delay is not too far away and purveyors are known in advance, they are market values.[7] Economic impacts are easier to value than social issues. However, expertise has set some standards and is able to give a rough idea of the nuisance values, for example. Hence this second type of impact can be quantified and summarised in net monetary benefits. In contrast, ecological impacts, including environmental damages as well as hazards to human activities caused by nature, are much less easy to compute. Furthermore, they are the most likely to be controversial, given that they do not rest on sufficiently reliable scientific studies and that ethical values are different depending on social classes or cultural backgrounds. This is one of the reasons why studies are content with a purely qualitative valuation for this type of consequences. Another reason is more positive: qualitative arguments are much easier to defend in front of an assembly where few experts are present (e.g. elected representatives) than quantitative ones, which are bound to be confused with ethical values. In fact, preparing arguments to be presented in front of representative assemblies is one of the most important tasks of a consultant.[8] Obviously, such a task is based more on political science than on economics, and explains why it is often in contrast to the economic impact valuation. However, missing valuations of such risks deprive the decision-makers of objective criterions for managing them, at least *ex post*, such as insurance or financial instruments able to hedge or compensate possible negative outcomes. In order to apply some of the methods proposed in this book, it is necessary to collect data and generalise enquiries on the public willingness to pay, which are becoming available and gathered by specialised agencies.

In the case of the Millau viaduct example, the best studied impacts were those on the town of Millau and surrounding villages. They address urbanism, regional tourism and environmental issues. Urbanism meant to forecast the possible evolution of industrial plants and surrounding housing and shops implantation, in order to be prepared to open new access roads, streets, sewers and the like. Indeed, access to the freeway and to the bridge implied moving activities from the town centre and from some villages around, towards the access routes. In order to be prepared, it was decided by the town council and other local authorities to modify the road system, refuse dumps and other infrastructures so as to facilitate the probable extensions towards the freeway's access roads. Another bet was on the attraction of the old town centre for tourists who, up to now, were much too bored by traffic jams during the crossing of Millau to consider stopping. This implied advertising and enlarging access roads and parking lots. A third priority was to convince the state and the bridge construction contractor, to leave open most of the construction working roads made for trucks, so that they would provide access to the site, for use as a visiting sight-seeing tour for tourists interested in this major architectural realisation. Tourism is, indeed, a major development factor for the region. Controversies arose about this issue, too. Before the region was opened, tourists were mainly attracted by wildness, old-fashioned farming and sports of nature (canoeing, rock climbing, hiking and more risky sports). Obviously, a technological achievement changes the

[7] Except in centralised countries where arbitrage is flawed by state priorities. Notice that most European countries are still highly centralised as far as infrastructures are concerned.

[8] This was explained to the authors during an interview with one of the chief engineers in the CETE of Aix en Provence in charge of the first study, André Mérieux. The interview was published under the title: "Prévoir l'impact d'un projet d'équipement" (forecast an equipment project), together with a paper of the authors: "L'évaluation des risques dans les projets publics", (risk valuation in public projects), *Economie Publique*, **10** (2002) 9–49.

landscape (the bridge can be seen from miles away) and the general image of the region. The bet, this time, was that this new feature would attract another category of tourists, and would facilitate accessibility to them without repulsing the previous ones too much. There were harsh controversies among the tourist branch about this bet, in particular on the duration of the stays of the waves of Northern Europeans attracted by the Mediterranean coast during the summer and passing over the bridge instead of being delayed by traffic jams. Simulations of different scenarios were implemented by the local tourism organisations, based on past data and on regions that had been through similar modifications. Prospective studies were ordered from consulting firms, but from our knowledge they didn't proceed to enquiries. That made it possible to sketch risks, valuate them grossly, and help hotels and other tourist activities to orient their investments. Here again, the need for more data has been felt; it would have made it easier to analyse, forecast and get more precise conclusions than could be done with those available. Some could be found in similar regions, but gathering them would require so much work that, again, it necessitates the creation of a special agency.

However, environmental issues were left to expert appreciation and little was done with respect to public perceptions, without opportunities to open discussions about possible alternative scenarios. Therefore, it is not a risk (with both possible gains and losses) but a particular possible damage that is valued. Furthermore, most of the time, impacts are understood in the biased meaning developed for marketing, i.e. positive impacts able to attract customers, that say very little in terms of expected risks! In any case, without monetary valuation of the environmental and visual issues, risk is only appreciated in subjective terms and this leaves little material to think about its management.

It is likely that most great public projects are treated the same way. The discussions have been hard but, deprived of objective arguments about the financial part (except for the technological side), controversies and protests could not lead to much improvement. In the end, it seems that decisions were taken at a high Ministry level, a technocratic level, not a political one, even though the minister signed the decision. Indeed, ministers from all political parties changed over during the long process, and it so happened that it was a Communist minister who signed the most liberal decision! A contract was signed with a private company to build the bridge and a 78-year concession to manage it, including an option to sell!

To what extent is such a decision a collective one? What is the role of individual decisions in such a complex process? In fact, there are several levels at which decisions are taken. Even when an individual actually decides, collective features are taken into account. Take the ultimate stage: when the Minister of Transportation signed the enactment and the contract, he acted as a nation representative and in that sense his decision was the reflection of a collective one. Delegating the work to a private company was certainly not a representation of his own political preferences. He was convinced to do it, because state-owned companies or public services would not have been able to achieve the work for the proposed cost. State-managed entrepreneurships are constrained by many regulations: routine, trade union agreements and heavy administrative costs. Furthermore, they are not very efficient at managing financial risks (the state is its own insurer) and investment profitability. But, if no private companies had taken the challenge to meet the requirements imposed to enter the competition, the state would have had to actually do the job.

A private company is considered economically as an individual agent. Its decision, however, integrates collective choices. First, decisions are taken by a board of directors that represent the owners: they may have different objectives. Second, decisions integrate collective choice at the national (and even supranational) level, because they are taken under the constraints of regulations and responsibility in front of the law.

It is interesting to learn[9] how this private company analysed the problem. Several reasons explain why it was able to win the competition and achieve the works before the expiration date and at a lower cost than the proposed one. First of all, it had already built several bridges of the same type, although less huge, and it had learned to master the technology. Second, it owns a group of dependent companies who could deal with most of the underwriting tasks: notably a concrete company, a metal works company and several road work companies. That made it possible to lower costs or to integrate them into an overall financial plan. Third, it was able to conduct further studies about climatic and geological risks and make sure it could hedge them financially. Last, the long-term concession and the put option made it able to smooth costs with expected returns and have its image and capital valued by the stock exchange after it achieved this impressive work. A detail worth noticing is that within the interview the word “risk” appears in all the answers, when it is hardly mentioned explicitly in any of the previous studies’ reports. This is because, contrary to a state-dependent company or a consulting firm, such an important international company is not limited to its engineering ability, but is a financial company as well, with an important group of finance and insurance specialists, able to diversify and deal with hedging instruments, which are out of reach of the public ones. Let us conclude about collective choices included in the private decision to take the job: the choice is individual in the sense that it endorses the company responsibility, both financially and legally. However, it integrates collective valuation because it is priced on the stock exchange market, its costs are computed through market values and its financing (bonds and other borrowings) is made at a competitive rate.

In contrast, the first cost valuation was made under the assumption that the state would be in charge of the works, on the basis of materials at then present prices and very low discount rates. This is common practice for state-owned companies and decisions; it is justified in part by the state being its own insurer and investments being managed, *ex post*, by taxes, Treasury bills and discretionary tolls. As a result, its efficiency is hard to compare with private company standards, and the decision is fundamentally of an individual type, even though risks are collectively supported.

The next two chapters should make clear the profound differences in economic theory between individual decisions and collective ones.

[9] This is taken from an interview with Jean-François Roverato, Board President and Chief Executive of Eiffage, the contractor company, given to a local newspaper: *Le Midi Libre*, special issue, August 2004.

2

Individual Valuations and Economic Rationality

A choice is rational if it is founded on some formal reasoning, so that it can be justified in front of a discussant. This is easier to do if both parties agree on a common theory: we refer to individual decision theory in this chapter. An individual is not restricted to a person, here as in economic theory in general. It is to be understood in the broad sense of a socially well-defined entity, considered as responsible for the consequences of its decisions, for instance in front of a court. The purpose of (individual) decision theory is to define what a rational choice could be.

Starting from previous works by Ramsey (1926) and Keynes (1921), among the most prominent ones, decision theory blossomed during the 1939–1945 world conflict. It was being used to determine US military strategies. "Operations research" was the name given to this trend of research, including the representation of individual (in a broad sense) behaviour, and of strategic interactions. The name was kept when the theories extended their field of applications to political and economic problems such as industrial development strategies. Operations research includes some mathematics and computer sciences, which were just appearing at the time. Economic theory and finance theory foundations were then formulated and developed using these methods from the 1950s on.

In order to develop game theory, where decision-makers' interactions play a central role, von Neuman and Morgenstern (1944) proposed the first model of rational choice in the face of uncertainty: the Expected Utility (EU) theory. It extends the utilitarian approach in economics to decisions in the face of uncertainty. Difficulties to justify choices, most of the time come from uncertainties about the decisions' consequences, which depend on elements (events) that are not under the decision-maker's control. Among these, natural phenomena or other decision-makers' reactions are the most common ones. In game theory, both interact and the set of uncertain states was assumed to be measured by a probability distribution. The EU criterion was already known (it had been proposed in the 18th century by Daniel Bernoulli (1738)), but without axiomatic characterisation of the agent's preferences.

The model was soon extended to situations where probability is unknown, or even unknowable, by Savage (1954), following previous works by de Finetti (1930). More general models were developed later to take into account observed violations of axioms such as the Allais paradox (Allais, 1953; see Section 2.6 below). We shall present a model inspired by the one proposed by Schmeidler (1989), general enough to derive from it (nearly) all other models as particular cases. It encompasses the two types of uncertain situations: with or without a probability distribution on the decision's outcomes. The distinction between the two types of uncertain situations is the topic of Part II, where results presented here will be exploited in Chapter 6.

2.1 PREFERENCES ON CONSEQUENCES AND UTILITIES OF DECISIONS

Decision theory starts with a general formulation of a choice problem and then specifies its elements. If there is a decision *problem*, it must be the case that the decision-maker does not know how to order the decisions into consideration. The first rationalisation step consists of describing how to relate preferences on consequences and decisions to be taken. The second step makes precise the conditions under which the preference order can induce a choice criterion on decisions. The criterion will be the (mathematical) expression of rationality: to decide rationally is to choose the decision(s) that obtains the criterion's optimal value. In order to build such a criterion, some conditions must be satisfied by preferences. The criterion will order the decisions, it must be the case that this order is consistent with the preference order on consequences.

A preference order is not necessarily an order in the mathematical sense, i.e. satisfying the transitivity property: if a is preferred to b, and b to c, then a is preferred to c. In practice, we often observe cycles in a sequence of choices: a was chosen over b and b over c but c was chosen over a! For instance, some people prefer their sister/brother to their cousin, their cousin to their neighbour, but then prefer their neighbour to their sister/brother ... likely for different reasons in the first two cases than in the third one: this will not be considered as rational, here. A preference relation aims to aggregate the preference orderings on the different components of the objects of choice. The previous example fails to aggregate a preference based on the degree of proximity with relatives and a preference based on the proximity with neighbours. A lexicographic order is founded on an *a priori* hierarchy between the components (e.g. letters in a word) and follows then the preference order on each component (alphabetical order): thanks to this a dictionary is not a mess! Cycles are observed as soon as the aggregation on preferences over the several components fails to obtain transitivity. This is evidenced in Condorcet's Paradox (see Chapter 3): the majority vote procedure to aggregate individual (transitive) preferences does not satisfy transitivity.

Non-transitive preference orderings cannot be represented by a unique numerical criterion. Indeed, a criterion is a numbering of objects such that the numbers are in the same (natural) order as the preference order, e.g. the words in a dictionary can be numbered from 1 for the first entry (*A*) to whatever number it is for the last one (*Zymotic*). Then, as the natural order on numbers is transitive, it must be the case that the preference order be transitive as well. Transitivity is the first restriction (axiom) imposed on preference orders. More precisely, the preference order decomposes into an equivalence order (indifference) and a strict order on equivalence classes. This is called a pre-order in mathematics,[1] furthermore the pre-order

[1] A pre-order is a transitive binary relation (preferred or indifferent to) on a set that satisfies furthermore "reflexivity": a is preferred or indifferent to a. A pre-order defines two relations: an equivalence (or indifference) relation, and a strict order on equivalence classes defined by the equivalence relation. An equivalence relation is a transitive binary relation (indifferent to) on a set that satisfies reflexivity and "symmetry": if a is indifferent to b, then b is indifferent to a. It defines equivalence (indifference) classes, i.e. a partition of the set, into subsets formed of indifferent elements. A strict order is a transitive binary relation on a set that satisfies furthermore "anti-symmetry": if a is strictly preferred to b, then it is not true that b is strictly preferred to a.

must be complete (all objects can be compared), otherwise some objects could not be ranked. When the set of objects is finite or denumerable, the existence of a criterion is obvious: any one-to-one function from the set of objects to a set of numbers with the same cardinal is a criterion. However, when the set of objects is not denumerable, it is necessary that the preference order satisfy the structure of the real number set: it must be continuous.[2]

It must be clear that such a representation by a numerical function is not unique: the criterion is purely ordinal. A criterion representing preferences is usually called a "utility function", U, and it is characterised by: a is preferred to b if and only if $U(a) \geq U(b)$. It is straightforward to see that any increasing function of a utility function is a representation as well (e.g. the more complicated criterion: $2\exp[U(a)] \geq 2\exp[U(b)]$). We see that the order is represented by utility functions, but its value differences do not indicate a relative importance between the objects. If we want these differences to be taken into account, it is necessary that the criterion have a cardinal property: its differences must represent an order over the differences between objects, consistent with preferences. The problem is that such differences (qualitative ones in general) must make sense: this is the case if objects are numbers (or vectors of numbers), i.e. if they are represented by some quality measure (or list of quality measures). This remark is of importance for random consequences, for instance expected utility has a cardinal property as we shall see. We shall insist on the distinction between ordinal and cardinal utility in Chapter 3.

In the unusual case where there is no uncertainty about a decision's consequences, a utility function would be a decision criterion: the decision-maker would choose the decision corresponding to the consequence that maximises the utility function (or any increasing function of it). Consumer theory is an example where the decisions are identified with their consequences: choices are made on consumption commodities and the consequence of a choice is to consume the chosen commodities. The "mad cow" disease has shown that such a simplifying assumption cannot always be applied to consumption, even for a common commodity: the choice of which meat to buy can take into account possible future consequences (on health) of its consumption.

2.2 DECISIONS, ACTS AND CONTINGENT ASSETS

Three basic sets can describe a decision problem: a set of realisable actions or decisions, a set of possible consequences and a last set representing uncontrolled elements.

Let A be for the action set. Its elements are determined by the decision-maker depending on its capacity and its constraints. It is on this set that we need to define a criterion.

The set of possible consequences is defined, in part, by the decision-maker's imagination, but some of its elements come out of necessity and logic.[3] On the set of consequences, the decision-maker has to make his preference order precise. Let C stand for consequences.

A difficulty arises for the last step: representation of uncontrolled elements, i.e. uncertainties. This is because, in general, actions are not linked to consequences by a well-defined function. If it were the case, there would be no decision problem at all. Indeed, a well-defined link (say, $f : A \rightarrow C$) would in turn induce a direct order on the action set from

[2] Continuity means that there are no jumps among preferences: given a sequence of consequences, each of them preferred to A, then the sequence limit must be preferred to A. Gérard Debreu has shown the existence of such a representation by a numerical function (utility function) in a 1954 paper, a fundamental one for microeconomics theory.

[3] If one chooses to take a plane to New York, one possible consequence is that the plane lands at one of New York's airports!

the preference order on consequences (a is chosen over a' if and only if $f(a)$ is preferred to $f(a')$). For example, this is the case in consumer theory where an action is a commodity bundle, which is considered as its own consequence as well. But consumer theory extends to time and to uncertainty, and then preferences bear not only on consumption bundles, but also on commodities contingent on time, on uncertain states or on both. More generally, contingent assets valuation is a main concern of this book, and we introduce the topic in two examples: consumption contingent on time and games of chance.

2.2.1 Time Contingent Consumption

The classical consumer theory is a static one, but we can enlarge its scope to a dynamic framework. We can still consider consumption bundles as decisions, but then a consequence of a bundle, say a, is not unique, it depends on t, the time at which it is obtained: it is an asset or a contract contingent on time. Hence the whole consequence of decision a will be a process of commodity bundles $(c_t)_{t\in T}$, where T is the set of times. A function linking consequences to actions is well defined in this case:

$$c : A \times T \to C \quad \text{with} \quad c(a, t) = c_t$$

It is easy to understand that the preference order on consumption bundles is not sufficient to define a criterion on actions: we need a preference order on bundle processes, as well. For instance, coffee and cereals may be preferred to wine and cheese at breakfast, but the reverse may be true at lunch. What is needed to decide on a diet (a list of dishes for every meal during a week, say), is a preference order on meal processes. If the decision-maker has such a preference order, then its decision problem has a direct solution: indeed, $f = c(a, \cdot)$ is a well-defined function from A to C and induces a ranking on A. As we have mentioned, a utility function can represent the order of preferences on $C : c$ is preferred to c' if and only if $U(c) \geq U(c')$. As we shall see, such a representation is subject to some requirements (axioms). In the case of consumption theory in our dynamic framework, we can see how the functional link we have defined, between actions and consequences through the artefact of the time set, induces a criterion on actions from preferences on consequences:

$$\text{action } a \text{ will be chosen over action } a'\text{, if and only if } U[c(a, \cdot)] \geq U[c(a', \cdot)]$$

$V(a) = U[c(a, \cdot)]$ defines a criterion on A. We shall see below under which conditions a function such as U exists. Furthermore, under some more conditions on the preference order, U may be decomposed into a function, say u on C, and another function, say r on T. In this case we could obtain a decomposition into a "discounted" sum: if $T = \{1, \ldots, T\}$:

$$U[c(a, \cdot)] = r(1)u[c(a, 1)] + \cdots + r(T)u[c(a, T)]$$

2.2.2 Games of Chance

Games of chance are other examples of decision problems where consequences are contingent on some variables: uncertain outcomes. A lottery player faces the set of all authorised bets on each number (or event, e.g. combination of numbers), under its budget constraint. That is A. For each chosen bet, a set of consequences is defined by the lottery rules: the set of prizes associated with each event (contingent prize). If we call E the set of events on which players can bet, then the lottery's rules yield a well-defined function:

$$c : A \times E \to C \text{ where } c(a, e) \text{ is the prize obtained by betting } a \text{ if event } e \text{ is realised}$$

This formalisation is adapted to consumption commodities with uncertain qualities as well. The actions are the commodities, the consequences are the same commodities with the quality that is observed *ex post*. More generally, contingent assets correspond to this representation: their payoffs are contingent on some future events.

As a general rule, a decision's consequences occur in the future. They are contingent on time and on other events, both phenomena that escape the decision-maker's control. An asset, in general, is contingent on such variables. Such uncontrolled phenomena may be due to:

- A random mechanism, such as a lottery.
- Other players' reactions, e.g. in a chess game.
- Both, in a card game for example.
- Error variables: when a mechanism is assumed to have deterministic outcomes. Several error sources can blur the results obtained: flaws in the deterministic theory, reading imprecision, errors in using the device, "collateral damages", etc.
- Social or physical factors considered by the decision-maker as impossible to forecast precisely: election results, earthquakes, etc.
- Limited ability of the decision-maker to dissociate consequences of a chosen action, as in shipping risks, venture capital, foreign country risk (see Chapter 9), etc.

In all cases, we shall be able to formalise the link between actions and consequences by introducing a set S that represents uncontrolled variables, whatever their origin:

$$c : A \times S \to C$$

Note that S is often an ill-defined set and is only an artefact introduced for formalisation purposes, in decision theory.[4] However, the same formalisation can be used to represent contingent contracts, and, in this case, the set S is obviously well defined by the contracts' clauses. S can always be "constructed" in the following way: the decision-maker knows the set of consequences associated with each decision a, say C_a. A set of "qualities" s may be associated with each element of $C_a : c(a, s)$ is the quality of a particular element of C_a. The set S is obtained as the union of the sets of $c(a\,,\ s)$'s. Qualities may be time, random events, uncertain events (the difference between random and uncertain is the topic of Part II) coming from endogenous and/or exogenous causes and/or other decision-makers' actions. In any case, the formalisation above is always relevant, even if the set S remains ill-defined, as long as the function c is well defined and induces a criterion on A from the preference order on C.

It is often convenient to decompose S into a product set, for instance the product of a set representing time, T, and another one, E, representing uncertain events: $S = E \times T$. What distinguishes a dynamic problem from a static one is the type of the decision-maker's preferences on consequences: the problem is dynamic if and only if preferences vary according to information arrivals in the future (the distinction static vs dynamic is the object of Part III).

Another distinction in decision problems formalisation refers to the type of uncertainties the decision-maker is facing. If the set has a known probability distribution, as in lottery betting, the problem is different from when there is not, as in horse race betting. There again, the distinction is relevant, only if the decision-maker preference order has different properties in one case from the other. In consumer theory under uncertainty (or Arrow–Debreu's General Equilibrium model, see Chapter 3), for example, preferences are general

[4] S plays the same role as the set Ω, Kolmogorow's probability it space we mentioned in the General Introduction.

enough to encompass both cases. Some versions of this model, however, restrain preferences to the case where they decompose into a utility on certain consequences and a probability on random events, in order to obtain particular results.

The general formulation of a decision problem allows us to forget about the set of actions in order to concentrate on the consequence function: $c : A \times S \to C$. The relevant object of study is the set, $\{c(a, \cdot); a \in A\}$, of functions associating an uncontrolled variable to a particular consequence. Savage (1954) called "acts" those functions that correspond to actions, and the term was kept in the literature; we shall follow this use throughout. Acts have the same representation as assets with payoffs contingent on future events, or contracts contingent on the realisation of some observable events. Assets and other contingent contracts are the objects of trades in financial markets.

In practice, solving a decision problem by decision theoretic tools relies heavily on the way the problem is described. In particular, the analysis of the set S of uncontrolled variables is crucial. We shall see in Chapter 4 how to use a decision tree to analyse and define the elements of a decision problem.

For normative purposes, decision theory provides a basis on which a criterion may be defined and constructed in practice. A criterion is a numerical function on the set of actions such that action a is chosen over action a' if and only if the consequences of a are preferred to the consequences of a'. In terms of acts, or assets, the criterion defines a value function. The economic literature calls "certainty equivalent", or more precisely, "present certainty equivalent", an act of valuation when future consequences are uncertain. This appellation clarifies that valuation erases the decision problem's temporal and uncertain aspects. Indeed, the criterion "averages" consequences over time and uncertain events. Obviously, this averaging may hide the limitations due to the conditions on preferences on which the criterion was constructed. For example, a simple criterion such as the discounted expected monetary value assumes that the decision-maker uses a probability distribution on the set of events and a discount factor, which should represent its preferences for present consumption.

2.3 CRITERION AND INDIVIDUAL VALUATION: AVERAGING

A decision criterion is a valuation of a decision consequence set. Consequences are contingent on time and on uncertain events. We have noticed already that consequence aggregation may lose some temporal aspects as well as some components of uncertainty. The task of decision theory is to put forward a set of conditions (axioms) under which the relevant aspects that the decision-maker wants to preserve are explicitly taken into account by the decision criteria. Obviously, for application purposes, such conditions should try to be expressed in such a way that they can be understood and scrutinised by the decision-maker. In practice, experimental economic procedures have been used to verify whether the decision-maker satisfies the conditions or not. There is always a percentage of subjects in an experiment who do not satisfy the conditions, but this does not necessarily mean the subjects are irrational: it may be that the axioms are not consistent with their rationality. For instance, transitivity of preferences is often violated. Another important axiom is not satisfied by about 60% of the subjects in most experiments: the so-called "independence axiom", which is basic for expected utility. For such subjects, relaxations of the axiom must be found so that they can be satisfied by the subjects' behaviours.

Under the set of conditions, the existence of a criterion (or value function) is established and yields a summary of the consequences of the decision. The simplest formula that summarises consequences contingent on a set of times and uncertain events (we assume both to be finite for simplification: $e = e_1, \ldots, e_N$ and $t = t_1, \ldots, t_T$) is:

$$V(a) = u[c(a,\ e_1,\ t_1)]p(e_1)r(t_1) + \cdots + u[c(a,\ e_N,\ t_T)]p(e_N)r(t_T) \tag{2.1}$$

In this formula (discounted expected utility), consequences are valued by their utilities, u, events are given weights measuring their relative importance (probability, in the general mathematical sense), p, and times are measured by discount factors, r.

However, it may be the case that relative importance weights depend on the time at which they occur: p_{ti}. Similarly, discount rates can be contingent on prevailing events: $r(e_1, \ldots, e_{t-1},\ t)$.[5] More generally, the utility function and both measures could depend on the previous history of the decision process, including past times, events and decisions. In general, all factors may be interdependent as they are obtained by a decomposition of the preference order into the different elements on which the consequences are contingent. Conditions on preferences make such a decomposition more or less restrictive, for instance the previous discounted expected utility criterion is obtained under very restrictive assumptions. For the remainder of this chapter, we shall concentrate on uncertainty, while we shall emphasise the role of time in Part III. In this simpler case, formula (2.1) becomes the expected utility:

$$V(a) = u[c(a,\ e_1)]p(e_1) + \cdots + u[c(a,\ e_N)]p(e_N) \tag{2.2}$$

Let us point out that the mean of a set of values, central to all aggregation formulas, may be written in three different ways. Each one of them offers a particular interpretation. The most common point of view is the one adopted in formulas (2.1) and (2.2) and has a geometrical interpretation: the mean is the centre of a set of points (utilities) weighted by a measure (probability). That is, the mean is a middle point and the formula expresses it directly.

The second point of view is related to the way the middle point is reached, it is based on distance increases between two points. Let us rank the points by ordering the events according to their utility values, or assume we have first numbered the events so that utility values are in the same order: $u[c(a,\ e_1)] \leq \cdots \leq u[c(a,\ e_N)]$.

There are two main ways to reach the middle point. One can start from an extremity and progress according to the natural order, with weights according to the length of the distance between two adjacent points:

$$\begin{aligned} V(a) = {} & u[c(a,\ e_1)]p(e_1, \ldots, e_N) + \{u[c(a,\ e_2)] - u[c(a,\ e_1)]\}p(e_2, \ldots, e_N) + \cdots \\ & + \{u[c(a,\ e_N)] - u[c(a,\ e_{N-1})]\}p(e_N) \end{aligned} \tag{2.2$'$}$$

Alternatively, in a way more similar to formula (2.2), one can weight each utility by the increases of respective weights:

$$\begin{aligned} V(a) = {} & u[c(a,\ e_1)][p(e_1, \ldots, e_N) - p(e_2, \ldots, e_N)] \\ & + u[c(a,\ e_2)][p(e_2, \ldots, e_N) - p(e_3, \ldots, e_N)] + \cdots + u[c(a,\ e_N)]p(e_N) \end{aligned} \tag{2.2$''$}$$

5 We shall come back to these difficulties in Part III.

It is straightforward to check that the two last formulas are equivalent to the previous ones only if the relative measure, p, is such that: $p(e_n, \ldots, e_N) - p(e_{n-1}, \ldots, e_N) = p(e_n)$ for all elementary events e_n. Otherwise stated, the measure must satisfy: for any disjoint events A and B, $p(A \cup B) = p(A) + p(B)$, the so-called additive property that a probability measure satisfies.

Such a property need not be satisfied by more general measures and we shall see that some non-additive measures are relevant for representing behaviours in the face of uncertainty. Let us consider the following example.

Take a national bingo (seven numbers drawn at random) where bets are €1 (for instance by buying a fixed price ticket) in the hope of winning a million euro (€1m) prize (and zero otherwise). A player buying a ticket must value it at €1, at least. If the value of the bet is the mean of its consequences (formula (2.2) where $u(x) = x$), the weight of the prize must be 1/1 000 000 as: $V(\text{one ticket}) = 1 = 1\,000\,000 \times 1/1\,000\,000$.

A first interpretation considers 1 millionth as the player's likelihood (subjective probability) of the ticket's number. However, everybody knows (or could compute) that there are about 86 million possible seven-digit numbers (85 900 584 exactly). As a consequence, this valuation does not satisfy additivity of the event measure. Were it the case, the player should be considered to be sure to win against the 85 million numbers he/she didn't play.

A second interpretation of this way to play, and hence of valuing the likelihood to have the winning number, takes the frequency of this number as its measure (86 millionth). However, the utility of winning €1 m is at least 86 times as much as the utility of €1, so that:

$$V[u(1\,\text{euro})] = u(1\,000\,000) \times 1/86\,000\,000 = 1$$

A third interpretation combines the two previous ones in the player's rationality. The utility of €1 m is more than a million times the utility of €1, (because the player's life would be changed) and the likelihood of the chosen number occurring is based on its frequency, but modified by a subjective factor (luck factor) (because the number is the player's loved one – a birthday date, for instance), so that neither the utility of 1 000 000 is 1 000 000, nor the weight assessed to winning is 1/86 000 000, and:

$$V[u(1\,\text{euro})] = u(1\,000\,000) \times \phi(1/86\,000\,000) = 1$$

This formula is a generalisation of the two previous ones in which the function, u, is on money amounts and ϕ is a deformation function on frequencies. If frequencies are not known or are not used as such, a similar formula expresses the weighting by a non-additive subjective measure, μ, of the winning event, say e^*:

$$V[u(1\,\text{euro})] = u(1\,000\,000) \times \mu(e^*)$$

In the first interpretation, the utility of €1 m was 1 000 000 and μ was the subjective probability to win: 1 millionth.

This example shows that the simple valuation of random consequences by their mean value (Huygens–Pascal criterion) would induce any one not to play bingo. As advisable as this conclusion may sound, it is in contrast with the behaviours of millions (even billions) of players,[6] not all of them dramatically lunatic and among them very rational investors and policy decision-makers.

---

[6] See Chapter 9.

2.4 A SIMPLE DECISION THEORETIC MODEL

Before entering into the main theory, which yields the most general criterion, we can understand its main features through a simpler theory and criterion. It is inspired at the same time by the Schmeidler (1989) and Yaari (1987) models that we shall present later, and is due to Chateauneuf (1991) who modestly calls it "simplified Yaari–Schmeidler model". In this model, consequence utilities are their monetary values and the set of acts is the set B of bounded continuous functions on the set of real numbers. The criterion is directly defined if preferences on acts satisfy a simple monotonic axiom:

An act is strictly less preferred than the same act to which is added a positive monetary amount in each of the consequences.

Under this condition, indeed, the value of an act, b, in the set B of acts, is simply the minimum certain (i.e. constant) monetary amount preferred to that act:

$$V(b) = \inf\{r \in R / r \text{ is preferred to } b\}$$

Any criterion obtained by a non-decreasing function of V represents the decision-maker's preferences as well. For example, a change of monetary units would do. In practice, such a criterion is parametrically estimated by a decreasing auction procedure for a given act and such that the decision-maker reveals the minimum amount of money preferred to the act. However, if preferences have a supplementary property, the criterion becomes much easier to use. Indeed, it is then only defined up to an affine (i.e. linear plus a constant) function, which makes it "cardinal" instead of simply "ordinal", and can be decomposed into an integral: a mean value of the previous type, formula (2.2).

Here is this fundamental property:

Additivity: *For all acts X, Y and Z, if act X is preferred to act Y and if they are both added to the same act Z, then the act $X + Z$ is still preferred to act $Y + Z$.*

If this property is satisfied, a simple exercise shows that the criterion is additive: $V(X + Z) = V(X) + V(Z)$. In the case where the set of uncertain events is finite: $\Omega = \{\omega_1, \ldots, \omega_n\}$, acts are column vectors of real numbers. Let μ be an additive non-negative and bounded by 1 measure on this set (a probability measure), then the preceding result is a well-known one from linear algebra: μ is defined by a row vector of positive numbers summing to 1. The criterion can be written as:

$$V(X) = X(\omega_1)\,\mu(\omega_1) + \cdots + X(\omega_n)\,\mu(\omega_n)$$

If, without loss of generality, we assume that $X(\omega_1) \leq \cdots \leq X(\omega_n)$, the criterion can be written according to formula (2.2′) as a mean of weighted increases:

$$V(X) = X(\omega_1)\,\mu(\{\omega_1, \ldots, \omega_n\}) + [X(\omega_2) - X(\omega_1)]\,\mu(\{\omega_2, \ldots, \omega_n\}) + \cdots$$
$$+ [X(\omega_n) - X(\omega_n - 1)]\,\mu(\omega_n)$$

Indeed, here

$$\mu(\{\omega_1, \ldots, \omega_n\}) = \mu(\{\omega_1\}) + \cdots + \mu(\{\omega_n\})$$

and

$$\mu(\{\omega_1, \ldots, \omega_n\}) - \mu(\{\omega_2, \ldots, \omega_n\}) = \mu(\{\omega_1\}), \text{ etc.}$$

In practice, submitting the decision-maker to an experiment will provide an estimation of the measure μ, and the criterion can be deduced. Notice that, even though μ is mathematically a probability measure, it is a subjective one and it need not be a known probability distribution (i.e. frequency).

More generally, Chateauneuf (1991) assumed Ω to have a σ-algebra, F, of events and B to be the vector space of bounded continuous measurable functions on (Ω, F). Preferences on B are a complete continuous pre-order (i.e. any sequence of acts: $f_1, \ldots, f_n, \ldots$ all preferred to a given act f and converging towards a limit act f_0, then f_0 is preferred to f). Then V exists and is defined up to an affine real function (Debreu, 1954) by: $V(b) = \inf\{r \in R/r \text{ is preferred to } b\}$.

If, furthermore, preferences satisfy the additivity axiom for all acts in B, then the linear algebra result extends and there exists an additive, bounded, non-negative measure, μ such that V is the Lebesgue integral (expected value) of its consequences: $V(X) = \int X \, d\mu$.

This first step in the Chateauneuf (1991) model is the result obtained by De Finetti (1930) to define mathematically the notion of subjective probability and is a special case of expected utility (here the utility of a payoff is the payoff itself).

However, the additivity axiom, which yields the additive property to the criterion, is not often satisfied by decision-makers. Indeed, Chateauneuf shows that adding an act to another may modify its consequence variability (think of an act as a marketed asset and the sum as a portfolio formed by two assets). An extreme case is the following: take an act, X, preferred or indifferent to another act, Y, then combine them with $Z = -Y$, where $-Y$ is the act that yields minus the consequences of act Y (a short position on Y if it is a marketed asset or a perfect insurance contract if Y is thought of as a list of possible casualties). Then it is clear that act $X - Y$ is still an act with uncertain consequences, while $Y - Y = 0$ is certain (a perfectly hedged position). A decision-maker who is risk averse (in a general sense, i.e. doesn't like a financial position with uncertain consequences) is bound to prefer a riskless position, such as $Y - Y = 0$, to a risky one, such as $X - Y$. This extreme case shows that a risk-averse (uncertainty-averse, here) decision-maker may well invert the preference order after adding the same act to two previously ordered ones.

However, this hedging opportunity would not be available if act Z could not reduce the risk of X nor the risk of Y. This occurs if all acts vary in the same way. "Vary in the same way" is a loose expression but it can be given a precise meaning through the mathematical notion of comonotonicity:

Definition (comonotonicity) *Two measurable functions on Ω, X and Z, are comonotonic iff:*

$$\forall\ \omega, \omega' \in \Omega,\ X(\omega) \geq X(\omega') \Leftrightarrow Z(\omega) \geq Z(\omega')$$

Note that Z can be comonotonic with X and with Y but that X and Y need not be comonotonic with each other. In particular, a constant function is comonotonic with any measurable function.

The comonotonicity property defines a coverage of the set of acts (measurable functions) by subsets of comonotonic acts with non-empty intersections (the riskless act, or constant function, belongs to all of the subsets). For each of the subsets, the additivity axiom is satisfied and hence the criterion is additive for comonotonic acts. However, it is not additive on the whole set of acts, in general, because of the possible hedging effects of two non-comonotonic acts. More precisely, Chateauneuf (1991) shows that if preferences on acts satisfy the following axiom (comonotonic additivity), which is a relaxation of the previous additivity axiom, then there exists a measure, μ (a capacity), which is increasing (for inclusion), positive and bounded by 1, and a functional, V, additive on each comonotonic subset (but not on the whole set of acts), such that V can be decomposed as a (Choquet[7]) integral that we shall denote, for an act X, as: $V(X) = \oint X \, d\mu$.

In particular, in the case that Ω is finite as above, we obtain the same formula as in the additive case as a weighted mean of value increases:

$$V(X) = X(\omega_1)\ \mu(\{\omega_1, \ldots, \omega_n\}) + [X(\omega_2) - X(\omega_1)]\ \mu(\{\omega_2, \ldots, \omega_n\}) + \cdots + [X(\omega_n) - X(\omega_{n-1})]\ \mu(\omega_n)$$

However, here, we may have: $\mu(\{\omega_1, \ldots, \omega_n\}) - \mu(\{\omega_2, \ldots, \omega_n\}) \neq \mu(\{\omega_1\})$ so that the Choquet integral (weighted mean of value increases) does not collapse into a Lebesgue one (weighted mean of values) as in formula (2.2).

Theorem (Chateauneuf) *Preferences on the set B of bounded continuous measurable functions on a space* (Ω, *F*) *satisfy the following axioms:*

(1) They define a complete continuous pre-order on B.
(2) The pre-order is monotonic: for any act X and any non-negative constant act X_0, $X + X_0$ *is strictly preferred to X.*
(3) It satisfies comonotonic additivity: for any acts X, Y and Z in the same comonotonic subset of B, X is preferred or indifferent to Y iff $X + Z$ *is preferred to* $Y + Z$.

If and only if there exists a unique capacity, μ *(a monotonic, positive measure bounded by 1) on F such that preferences are represented by a functional V (Y is not strictly preferred to X iff* $V(Y) \leq V(X)$*), the Choquet integral of an act's payoffs with respect to capacity* μ: $V(X) = \oint X \, d\mu$, *for any act X.*

2.5 A GENERAL CRITERION OF INDIVIDUAL VALUATION

There are many criteria for valuing decisions with uncertain consequences. The oldest one is the mean value: arithmetic mean and weighted average when probabilities are known. The first one may be purely arbitrary if there are no reasons to believe that each value is equally likely to occur, but it may be justified by an argument of the type: "Given that there are no reasons why one value should be more important than the others, let's treat them equally". The second criteria required a definition of probability and was given independently by Huygens and by Pascal in the 16th century in order to obtain a mathematical criterion for deciding how to bet when playing card games. Neither of the authors, however, gave a justification for the criterion: it seemed natural, once the notion of probability weights (chance weights) was brought up. The question is: Why is it rational for an individual to decide according to the optimisation of this criterion?

The question was answered some three hundred years later by von Neumann and Morgenstern (1944) who needed an individual criterion to analyse rational economic behaviours within the mathematical framework of game theory. In a game, each player faces

[7] Choquet (1953).

a set of possible outcomes with known probability distribution (distribution on uncertain outcomes and on the strategies – mixed strategies, i.e. decisions played at random) of the other players. A set of conditions (axioms) on the players' behaviours yields the existence of a function on outcomes (utility) and a criterion: the utility's expectation with respect to the known probability distribution.

Other criteria exist and are used in practice for pragmatic reasons, although they cannot be considered as yielding a "rational" behaviour, even when they prove adequate and can be justified by subjective considerations. For instance, taking the minimum of the ranked outcomes if the player is pessimistic, the maximum if optimistic, or an arbitrary convex combination of both to ponder the two extreme attitudes. All three were used well before conditions under which behaviours would be consistent with their maximisation were provided by decision theorists. From the pioneering axiomatic theories of de Finetti (1930), von Neumann and Morgernstern (1944) and Savage (1954), many authors have contributed to the generalisations by relaxing and refining axioms, mainly since the 1980s. Obviously, we expect a condition (axiom) to restrict the set of admissible behaviours, however the one under which the last (and most commonly referred to) three criteria are obtained is *very* restrictive as regards behaviours in front of risks and/or uncertainties, as we shall see in Section 2.6 below.

Among the well axiomatised criteria, we shall present a very general one that was introduced by David Schmeidler (1986, 1989), and we extend it a little so that it encompasses other models as special cases. The model is written in the Anscombe–Aumann (1963) framework that we recall here.

We assume the decision-maker is able to produce a list of possible states of the world and events, (Ω, F), on which decision consequences are contingent. There is no need for a prior probability distribution. Furthermore, consequences may not be the ultimate ones: some of them are random, with known probability distribution. So, there may be two stages in the consequence process: once a decision is chosen, a state of nature occurs and yields the first stage consequence. It may be certain and then the ultimate consequence is obtained, but it may be random. In this case the consequence is a lottery ticket and the ultimate consequence will be the lottery's prize. The reason why this complexity is introduced is that the model can encompass all types of uncertain situations. Obviously, if the first stage consequence is not random, this is a special case of a lottery. Another special case is obtained if the probability distribution is known for the first uncertainty level. Anscombe and Aumann (1963) interpreted the two uncertainty levels as: a horse lottery for the first one, i.e. a ticket saying how much you bet on which horse number in a horse race, and a roulette lottery for the second one.

For an example close to this interpretation, assume you have a bet on a horse and that the prize you will obtain if you win provides an uncertain gain because you have already borrowed this money amount to buy a bingo ticket.[8] There is no objective probability for the horse winning event (even an expert's one is subjective), however you are well aware of the probability that the number you played on at bingo may come out (even though you may act as if you deformed it: you think it is 1 millionth instead of 86 millionth as in the example in Section 2.3). For another example, consider a dynamic choice: at the first stage, you decide whether to go on vacation to the mountain or to the beach. If the weather is nice and you have chosen the mountain, you will hike and you know how much this is worth to you. If the weather is bad, you will go to the bars and play a game at the local bingo. On

[8] We do not advise anyone, even the most risk-loving player, to act this way!

the other hand, if you go to the beach and if the weather is nice, you will enjoy swimming, otherwise, you will play roulette at the casino. Obviously, there are seasons when the first stage uncertainty is resolved, so your choice really depends upon two lotteries. In some cases, you can go to the casino or to the bar and not gamble so that your choice has an uncertain and non-probabilised outcome. We will come back to the distinction between these two types of uncertainty in the second part of the book. As far as rationality is concerned, however, both cases are encompassed in the following model.

A partial representation theory of preferences for each stage of uncertain consequences is invoked, then an axiom links the two criteria representing preferences at each stage to obtain a global valuation function of an act. For each decision there is a corresponding "act", here a measurable function from (Ω, F) to a set of lotteries, say M. To each of nature's outcomes corresponds a contingent lottery, i.e. a list of probabilities $\pi(\omega)$, contingent on the first uncertainty level outcome ω. An act is a measurable function, i.e. there exists a σ-algebra of events G, on the set of lotteries M, and let B be the set of acts $f:(\Omega, F) \to (M, G)$. In turn, lotteries will yield certain ultimate consequences in the set C.[9] The theory constructs a criterion representing the decision-maker's preferences on acts, and this criterion is itself decomposed into one representing preferences on lotteries and a final one (a utility function) on the certain consequences.

In practice, the process goes the reverse way: preferences on ultimate consequences are easy to reveal, then preferences on lotteries are questioned, and finally preferences on acts are obtained from other types of experiments.

If the different levels of preferences satisfy a set of axioms, then it can be proved that they are represented by some functions (criteria). We shall not present the axioms here, they are similar but more complete than in Chateauneuf's theory and the existence theorem much more involved. Let:

U stand for the utility on consequences ($U : C \to R$),
V for the criterion on lotteries ($V : M \to R$) and
W for the global one on acts ($W : B \to R$).

Each criterion is the integral of the previous one. A special case is when V, W or both are Lebesgue integrals (i.e. usual expectations) as in Anscombe and Aumann (1963). In Schmeidler (1989), W is a Choquet integral.

For a given act f:

$$W(f) = \oint_{\Omega} V[f(\cdot)]\mathrm{d}\mu \quad \text{Choquet expected utility}$$

In the special case where lotteries are degenerate: $f(\omega) = X(\omega) \in C$ and C are real monetary consequences ($f:\Omega \to R$ and $U(x) = x$) we are back to the Chateauneuf's criterion:

$$W(f) = V(f) = \oint_{\Omega} f\,\mathrm{d}\mu$$

In Schmeidler's model, $V[f(\cdot)]$ is the usual (Lebesgue) expected utility with respect to the given (objective) probability distribution, $f(\omega) = \pi_{\omega}$:

$$V[f(\omega)] = V(\pi_{\omega}) = \int_{C} U\,\mathrm{d}\pi_{\omega}$$

[9] More generally but still consistent with the model, consequences may be other lotteries, horse or roulette ones, but then more consistency axioms are required.

More generally, under conditions slightly different from those of Schmeidler,[10] the valuation may take into account a subjective deformation of probability, using a deformation function, ϕ of the cumulative distribution increments. Then preferences on lotteries (with distribution π) are represented by the Choquet expected utility with respect to the measure $\phi(\pi)$:

$$V(\pi) = \oint_C U \, \mathrm{d}\phi(\pi)$$

Such a deformation has been used to explain decision-makers' attitudes when they behave differently in the face of high and low probabilities. This was verified by experimentation (for instance Gonzales and Wu, 1999) and has been applied to estimate the value of statistical life (e.g. Kast and Luchini, 2004). It yields an inverse S shape function for ϕ, as depicted in Figure 2.1.

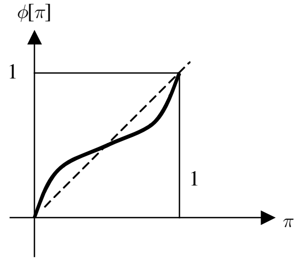


Figure 2.1 Deformation function

As we have seen above, the integral is written as a sum of increases on utility consequences ranked in non-decreasing order. Because the ranking is essential, V is called the "Rank Dependent Expected Utility" (RDEU), following the first model in which it appeared, developed by Quiggin (1982).

Finally, U, the utility function on consequences, is such that c is preferred to c' if and only if: $U(c) \geq U(c')$. In the special case where U is the identity function, I, and consequences are monetary values, V is called the "dual expected utility",[11] a criterion on lotteries that was proposed by Yaari (1987) (see Section 2.6):

$$V(\pi) = \oint_\Omega X \, \mathrm{d}\phi(\pi)$$

Let us go back to Schmeidler's more general model. If the criterion is additive, thanks to the restriction of one axiom, then the capacity measure μ is additive as well and is a probability measure. Such a subjective measure (revealed by the decision-maker's behaviour) is the one shown to exist by the Anscombe and Aumann (1963) model and is similar to the one put forward by Savage (1954) and, in the special case where $U = I$, by de Finetti (1930). The criterion is an expected utility with respect to a subjective probability distribution. For example, an expert's probability may be justified by such a theory. The criterion bears upon acts that are random variables from W to C in this case. Two specifications are well known. The first one we have seen in the previous paragraph: the de Finetti (1930)

[10] One could take Kast's (1991) model, replacing additivity with comonotonic additivity in the axioms at each uncertainty stage.
[11] Dual, here, refers to topopoly, in the sense that measures are in the dual space of functions.

model introducing probabilities as a representation of an order, we have obtained it as a restriction of Chateauneuf's criterion to the additive case. The second specification considers that a probability distribution (p) is known on (W, F) and is not deformed ($\phi = I$). Such a probability distribution is called "objective" or objective prior, in contrast with the subjective case. Then, we return to the first expected utility model proposed by Bernoulli (1738) and theoretically first founded by von Neumann and Morgenstern (1944):

$$V(\pi) = \int_{\Omega} U \, \mathrm{d}\pi$$

For this reason, V is often referred to as the "Bernoulli criterion" and U as the "von Neumann–Morgenstern utility function".

In the expected utility model, the decision-maker's behaviour is completely represented by the utility function on consequences. In the Yaari model, it is the deformation function ϕ that represents the behaviour. Last of all, if both U and ϕ are identity functions, the model erases attitudes towards both consequences and probabilities and is reduced to the Huygens–Pascal criterion, i.e. the mean value:

$$V(\pi) = \int_{\Omega} I \, \mathrm{d}\pi$$

Note that the general criterion on acts, W, is defined up to a non-decreasing function only (it is an ordinal criterion). However, the way it is decomposed into an integral defines a unique measure μ, and the utility function U is defined up to an affine transformation (cardinal utility). This cardinal property gives a meaning to increases so that, if U' is another function representing the same preferences, then: $U' = aU + b$. Then, for any consequences c and c':

$$U'(c) - U'(c') = a[U'(c) - U(c')]$$

The cardinal property is induced on the two upper criteria, V and W, because Choquet integrals are defined by increases.

In order to summarise these results, let us simplify a little bit: skip the most general theory (with acts as functions of the states set into a lottery set), and consider monetary consequences. Then uncertainty (or the future) is described by an unknown probability space (Kolmogorow) (Ω, F, Π) and acts as real measurable functions $X : (\Omega, F) \to (R, B)$. Preferences on a (suitable) set of acts are defined by a complete and continuous pre-order and are represented by a continuous functional V: an *ordinal* utility of acts. If we add conditions of preferences, i.e. on the functional, we get Table 2.1 distinguishing situations of uncertainty (the probability induced by act $X : \Pi_X$ is unknown, controversial or irrelevant) and situations of risk (Π_X is objectively known). In both cases, conditions decompose V into a *cardinal* utility U on consequences and a measure on (R, B). Names in Table 2.1 are those of the first authors and we go from the more general (higher) to the least one (lower) and conditions add restrictions on the utility (left) or the measure (right).

Note that Table 2.1 is only a summary of what was presented above and is not an exhaustive representation of all the decision models. There are many derivative versions of the ones there. Furthermore, three important models are not cited, the first two are very close to Choquet expected utility, the last one has very different foundations:

- The Jaffray (1989) linear utility theory.
- The Kahneman and Tversky (1979) and Tversky and Kahneman (1992) prospect theory.
- The Gilboa–Schmeidler (1995) case-based decision theory.

Table 2.1 Decision theories

Decision-making under uncertainty
No objective (and/or uncontroversial) probabilities on (Ω, F)
(Choquet expected utility, CEU: Schmeidler, 1989)
$V(X) = \oint_\Omega U(X)\ \mathrm{d}\mu$ (Choquet integral)
$U: R \to R$ cardinal utility, μ capacity

(Subjective EU: Savage, 1954)
$V(X) = \int_\Omega U(X)\ \mathrm{d}\mu$ (Lebesgue integral)
U cardinal, μ probability

(Subjective CE: Chateauneuf, 1991)
$V(X) = \oint_\Omega X\ \mathrm{d}\mu$(Choquet integral)
$U = Id_R$, μ capacity

(Subjective expectation: de Finetti, 1930)
$V(X) = \int_\Omega X\ \mathrm{d}\mu$ (Lebesgue integral)
$U = Id_R$, μ probability

Decision-making under risk
Known probability Π on the space (Ω, F)

(Rank dependent expected utility, RDEU: Quiggin, 1982)
$V(X) = \int_\Omega U(X)\ \mathrm{d}\phi(\Pi)$ Choquet integral
U cardinal, $\phi(\Pi)$ capacity

(EU: von Neumann–Morgenstern, 1944)
$V(X) = \int_\Omega U(X)\,\mathrm{d}\Pi$ (Lebesgue integral)
U cardinal, $\phi = Id_{[0,1]}$

(Dual theory: Yaari, 1987)
$V(X) = \oint_\Omega X\,\mathrm{d}\phi(\Pi)$ (Choquet integral)
$U = Id_R$, $\phi(\Pi)$ capacity

(Expectation: Huygens–Pascal, 1660)
$V(X) = \int_\Omega X\mathrm{d}\Pi$ (Lebesgue integral)
$U = Id_R, \phi = Id_{[0,1]}$

2.6 THE TWO MAIN CRITERIA FOR DECISION-MAKING IN FRONT OF A KNOWN PROBABILITY DISTRIBUTION: PARADOXES AND LIMITATIONS

The most well known theories for economic decisions under uncertainty are: expected utility and dual expected utility. Both deal with a particular kind of uncertainty, cases where the probability distribution of outcomes is known (situations of risk, see Part II).

From now on in this section, let us consider that uncertainty is represented by a probability space: (S, B, P) where P is known. If d is the chosen decision, the criterion is the expected utility of its consequences. Consequences depend on which state of the world occurs, so if the state s is realised, the utility of the decision will be: $U(d, s)$. The expected utility criterion can be written as: $V(d) = \int_S U(d, \cdot)\ \mathrm{d}P$.

In this formula, two utilities are present, they can be interpreted as: an *ex ante* global utility V, and an *ex post* utility U. If d is a quantity of some consumption commodity, as it is in microeconomic analysis for example, then assume we need to study the *ex ante* marginal utility of this commodity, $V'(d)$, or the marginal rate of substitution of commodity d_1 by commodity d_2: $V'(d_1)/V'(d_2)$, assuming both exist. Given that: $V'(d) = \int_S U'(d, \cdot)\ \mathrm{d}p$, we end up studying the *ex post* marginal utility, $U'(d, s)$, of the observed commodity quantity when we know which state s has occurred.

A question arises, however, as to whether the *ex ante* and the *ex post* decision problems are the same? A way to introduce Allais's paradox (below) is to answer: No, they are not! The first problem, *ex ante*, involves decision-making in the presence of risk: we don't know what commodity will be available for consumption. The second one is completely different and is much easier to solve: a consumption decision problem for a known commodity.

For example, take a restaurant in a tourist place you have never been in. You don't want to eat the stern international and expensive food at your hotel. In the rundown and poor surroundings, you've been told there is a 50/50 chance that you will find a genuinely very good and inexpensive restaurant. But you may also find yourself in one where hygiene is ignored and you will end up sick if you eat in it.

Once you've entered a restaurant, you face an *ex ante* risky choice (i.e. before you are too hungry to choose rationally!): order a safe dish, like boiled rice with vegetable sauce, but then you could have eaten at your hotel as well! Or order something that you're really fond of but is highly hazardous, like sea food.

Ex post, i.e. when you get some more information about which kind of restaurant you are in (e.g. by ordering some little entrées to nibble on while you think and drink a highly alcoholic[12] aperitif), you have a consumption choice to make: whether you have enough clues that this restaurant is clean to order what you are tempted to, or the converse and you stick to a safe choice tasteless dish.

Many economists, however, neglect this remark and refer to expected utility theory as a reliable analytic tool for decision-making in the face of uncertainty. For instance, risk attitude is characterised by a criterion (*ex post*): $-U''/U'$ (we'll come back to this in Chapter 6). It does make sense within expected utility theory, where the criterion for risky decisions, V, is completely characterised by U. However it relies on derivatives of the utility function on consumption, which addresses a riskless decision problem. This point was made as from 1953 by M. Allais, who argued that expected utility could not represent attitudes toward risks, in general. Indeed, it has been shown, by Allais and others, that about 60% of subjects (in good cases and up to 90% in some other cases) did not behave in accordance with this criterion if faced with an experimental risky prospect. Allais's argument was based on the *ex ante/ex post* effect we mentioned. It pointed out a particular axiom yielding additivity of the criterion V (and hence its integral form). There are many ways to present this axiom, the one chosen in Chateauneuf's model (Section 2.4) was in terms of random variables. It showed that hedging possibilities could be denied by the axiom while they could be appreciated by a decision-maker. Then the axiom was relaxed to encompass such behaviours. The axiom is often expressed in terms of lotteries, which is mathematically equivalent but suggests other interpretations. Allais's was that the axiom denies any difference a decision-maker could make between *ex ante* and *ex post* valuations. How does this work?

Consider decisions d and d', which yield money amounts $d(s)$ and $d'(s)$ if state s occurs. Then consider a given non-random money amount, c, and decisions $d+c$ and $d'+c$ yielding $d(s)+c$ and $d'(s)+c$ in state s. The axiom is known as the "independence axiom". "Independence" relates to the assumption that the preference order should be independent of the value levels obtained: adding (or subtracting) the same amount to two choices should not alter their ranking.

Independence axiom *d is preferred to d' if and only if $d+c$ is preferred to $d'+c$.*

This condition seems innocuous at first glance and many people have a tendency to accept it. However, many experimental examples show that, although subjects agree on the axiom, they

[12] High alcoholic degrees make it safe to drink, but much less to think! Maybe it is safer to have tea, that would make you less hungry anyway.

violate it in their observed choices. Allais (1953) was the first to present such an example. This phenomenon is called a paradox because most people, including some founders of expected utility such as Samuelson and Savage, found it paradoxical that people would violate an axiom they found "natural". The supporters of the axiom against Allais's example argued that people needed a little time to understand the choices and rationally correct their initial "mistakes". The (first) example goes as follow (the currency is the euro, although Allais's original example was written in terms of French francs).

Uncertainty is generated by a ball drawn from an urn containing 100 balls numbered 0, 1,…, 99. Bets are laid on three events, F, G and H, with different, easy to compute probabilities: event F occurs when the number of the ball drawn is 0, event G refers to numbers between 1 and 10, and the drawing of a ball with a number between 11 and 99 is the event H.

The following choices have to be ranked by preference order:

X yields €1 m with certainty.
Y yields €0 if F occurs, €50 m if G, €1 m if H.
Most people agree that Y is preferred to X.

Now two other choices have to be made:

A yields €1 m if F occurs, €1 m if G, €0 if H.
B yields €0 if F, €50 m if G, €0 if H.
Most people prefer A to B.

If they respect the independence axiom, however, their choices should be consistent with expected utility. Assume then that there exists a utility function U such that:

$$V(X) = U(1)$$

$$V(Y) = 0.01\ U(0) + 0.1\ U(50) + 0.89\ U(1)$$

$$V(A) = 0.01\ U(1) + 0.1\ U(1) + 0.89\ U(0)$$

$$V(B) = 0.01\ U(0) + 0.1\ U(50) + 0.89\ U(0)$$

Without loss of generality (U is defined up to an affine transform) assume that $U(0) = 0$ and $U(50) = 1$, leaving only $U(1)$ to be defined.

If A is preferred to B it must be that: $V(A) > V(B)$, i.e. $0.11U(1) > 0.1$, implying $U(1) > 1/11$. If Y is preferred to X it must also be that: $V(Y) > V(X)$, i.e $0.1 + 0.89U(1) > U(1)$, implying $0.11U(1) < 0.1$ and $U(1) < 1/11$. Here lies the contradiction: if Y is preferred to X, then it should be the case that B is preferred to A, if one agrees with the independence axiom.

In fact, if we let $X = d$ and $A = d + c$, we see that c is not a constant as in the previous axiom statement. Indeed, c yields 0, 0 and -1 respectively for events F, G and H. However, it is easy to show that the statement is equivalent[13] to the same expression, but where the constant c is replaced by a random outcome yielding $c(s)$ if s occurs. In this form, we can recognise the additivity axiom of Chateauneuf's model, in Section 2.4 above.

The expected utility criterion had itself been proposed to solve another paradox. It is called the St Petersburg paradox because it was proposed in this town, where the Tsar's

[13] Indeed, the second implies the first straightforwardly, the reciprocal is obtained by repeating the condition for each state s.

court stayed in the 18th century. At that time the only known criterion was the recent Huygens–Pascal mean (or expected) value. The paradox appears in a gamble in which the mean value is not finite. The mean value criterion would induce someone to bet whatever fortune is available, but few would do that in the following heads and tails gamble: a coin is flipped, if tails, stop. If heads, flip it again, and again until tails occurs, then stop. There is an infinite set of possible outcomes, however everybody is aware that going further than a hundred flips is highly unlikely (unless the coin is rigged!). However, the gains compensate for the low probabilities: the gain if tails obtains at the nth flip is 2^n roubles, and the probability is: $1/2^n$. Hence:

$$\sum_{n=1}^{\infty} 2^n \times \left(\frac{1}{2}\right)^n = \infty.$$

Bernoulli's solution was that, in fact, gamblers do not take 2^n roubles as it is, at least for n large enough. Most people, indeed, have a hard time distinguishing between €1 m and €1 000 850, although they do make a big difference between €850 and nothing. An interesting numerical function (the logarithm) had just been invented which performs the trick: transforming a number into a number less and less close to it when the number increases above 1. If, instead of comparing 2^n roubles with 1 rouble people compared $\ln 2^n$ to zero, then the sequence $(\ln 2^n)/2^n$ converges towards zero and the sum converges towards (logarithm of) the bet equivalent for people acting this way. Otherwise stated, the utility of money decreases with increased wealth (decreasing marginal utility) and the expected utility is the gamble valuation.

There was another way out of the paradox, but it was only discovered after the paper written by Yaari (1987), although the intuition is similar. Assume that gamblers know what the value of money is, even when millions are compared to units. However, they may face some difficulties in comparing $½^{85}$ and $½^{91}$ chances to lose or to win. Indeed, misperception of very low probabilities is a common observation, and it can be formalised by a function that modifies probabilities: an inverse S shaped one, for instance, as seen in the above paragraph (Figure 2.1). Then, with an appropriate function f, $2^n f(1/2^n)$ may converge towards zero and the sum may be finite.[14]

In Yaari's model, the independence axiom is replaced by a comonotonic independence axiom, and furthermore the sum differs because the weights are deformations of increases in probability.

As a conclusion, we have seen a series of theories defining a criterion on a mathematical basis and corresponding to the different uncertain situations that a decision-maker may face. All these theories, from the more general to the more restricted ones, are based on assumptions on agents' behaviours. Moreover, they are relevant for different decision problems, depending on the type of uncertainty that is faced. The limitations of their use in a decision problem are clearly indicated by the axioms they rely on and by the relevance of the assumptions on the representation of uncertainty. From the tableau in the previous section, we can see that the two models we have presented here are very low in the general setting. Indeed, they are close to the simple mean value criterion that is still commonly used in practice to summarise uncertain consequences. For economic analysis, however,

[14] Notice that the utility function must be concave and/or the deformation function convex to obtain convergence. Both are characterisations of "risk aversion", the first one within the expected utility model and the second one within the dual theory, as we shall see and discuss in Chapter 6.

these two models are interesting and help to define and understand concepts such as “risk aversion”, as we shall see in Chapter 6. This must not be confused with their relevance for most decision problems in the face of uncertainty. Relevance has to be questioned for each particular problem: known or unknown probabilities, hedging possibilities or independence of prizes, notably.

We have been concerned up to now with a single decision-maker able to rank consequences of the available actions. Furthermore, to specify preferences over acts, ranking consequences is not sufficient as the complete behaviour in the face of an uncertain outcome has to be revealed. This includes psychological features and, indeed, experiments necessitated cooperation between economists and psychologists. As a result, the theory can hardly be invoked to represent collective behaviours. Furthermore, an individual decision-maker is alone to face the consequences of its actions, which is not the case for a public decision, even if an individual decides in place of others.

To what extent, then, is it possible to refer to such theories to give guidance in public choices and produce management risk instruments? In the next chapter we present difficulties arising from tentative theories to aggregate individual preferences, and then we shall come back to instruments derived from individual decision theory in Chapter 4.

3

Aggregation of Individual Choices

Outside of a dictatorial framework, carrying out a collective choice requires that we take individual preferences into account. Two approaches are available for this task. The first one is given by public economics, which studies methods of direct individual preferences integration using the public decision-maker criterion. The second one uses contract prices, which reflect agent preferences indirectly, through the choices they have made.

3.1 PUBLIC CHOICES

A very simple and rather old idea, attributed to Bergson (1938), consists of transposing the formalism of individual decision theory to group decisions. This approach results in defining the so-called Bergson–Samuelson social welfare functions, which are the equivalent of individual utility functions in decision theory. In the same way that utility functions represent an individual agent's preferences, the social welfare function represents social preferences. We shall wonder about the meaning it is advisable to give to the concept of "social preferences", while seeking to characterise the social welfare functions.

A social welfare function connects a state of the world to the social utility that results from it. According to ideas of the 18th century philosophers, the state is benevolent, so that its single concern should be to best achieve the happiness of the greatest number of agents. An enlightened despot or an elected representative assembly will then have the same motivations. On the other hand, utilitarians taught us that an individual's satisfaction could be measured by a utility function. Using both points of view, the social utility function must depend on the agent utility levels and on these alone. In this way, the social welfare function will be considered as neutral with respect to the state's characteristics, which are not integrated into the individuals' utility functions. As agent utility levels themselves depend on the states of the economy, this design is completely compatible with the first definition given for social welfare functions.

This method, however, poses important problems which we find it advisable to tackle now.

3.1.1 Comparability of Individual Utility Functions

The first problem arising within this approach involves the way in which it is possible to take different agent utility levels into account. It takes us back to the comparability of utilities question, and is often juxtaposed with the cardinal utility–ordinal utility distinction,[1] although such a confusion is not entirely justified. It is thus necessary to accurately specify

[1] Let us recall that an ordinal utility only reflects a preference order, while for a cardinal utility, increments are also meaningful for a decision-maker.

these concepts, in order to clarify where stands the limit between what is possible and what is not, as regards social utility.

Public decisions will often be advantageous for some individuals and a nuisance for others. Choosing to carry out a project or not, will result in taking into account various agent utility profits and losses, simultaneously (according to both the benevolent state and the utilitarian designs).

If one wishes to characterize a social optimum by maximising the social utility function under a whole set of constraints, it is necessary to be able to compare the agents' utility levels for each realisable state of the economy. For example: one must be able to say which one out of two states s_1 and s_2 is better from the social point of view if state s_1 provides consumption $x_a(s_1)$ and thus utility $u_a(x_a(s_1))$ to agent a and consumption $x_b(s_1)$ and utility $u_b(x_b(s_1))$ to agent b, while state s_2 provides consumption $x_a(s_2)$ and thus utility $u_a(x_a(s_2))$ to agent a and consumption $x_b(s_2)$ and utility $u_b(x_b(s_2))$ to agent b. The answer will be obtained from a comparison of utility levels.

If the objective is to evaluate a public project, one will have to compare the agents' utility variations with respect to an initial situation. It is necessary, for example, to know if the project, P, which yields a utility variation $\Delta u_a(P)$ for agent a and a utility variation $\Delta u_b(P)$ for agent b is preferable or not to the status quo. Such a comparison between utility level variations forms part of the basis of cost–benefit analysis, which we shall present later on.

In the utilitarian framework of the 19th century, such comparisons do not pose a problem, since the utility function is assumed to be perfectly measurable. However, this conception of utility was strongly criticised from the very start of the 20th century. Criticism related to the fact that, in practice, it appears to be difficult to give a comparable measurement of different agent utilities, even if they were cooperative and honest, which is not necessarily compatible with their own interests. The move from a cardinal utility concept towards an ordinal utility[2] one was all the easier because of a major result of utility theory: the neoclassical approach to consumer theory can be founded on either concept. However, as we shall be able to demonstrate, this is unhappily not true any more for collective decisions. Because of this, in the second part of the 20th century, some economists pleaded in favour of the comparability of utilities but they did not necessarily favour cardinalism. This debate showed that it is possible to distinguish several levels for comparison, some levels being less demanding than others. On the whole, one can distinguish six different utility concepts with regard to comparability and the cardinality–ordinality distinction.

(1) Ordinality and non-comparability. Let us start from a list of utility functions, each one representing the corresponding agent's preferences. If we transform the utility functions by unspecified, increasing, monotonic functions, each one independent of the others, we still obtain a list of utility functions representing the agents' preferences. If agents a and b's preferences are represented by two utility functions, u_a and u_b, and if f and g are two increasing numerical functions, then: $v_a = f_a(u_a)$ and $v_b = f_b(u_b)$ are two utility functions representing the same preferences. Non-comparability is due to the fact that acceptable transformations of the utility functions are independent. Assume $u_a(x_a) > u_b(x_b)$, for some allocation x_a to agent a and x_b to agent b. There will exist some increasing functions f and g such that: $v_a(x_a) < v_b(x_b)$. It is then impossible to compare the agents' utility levels, since measurements u and v arrange the two agents' satisfactions in a different

[2] Ordinal utility: $U(x) > U(x')$ if and only if x is strictly preferred to x', U is defined up to an increasing function. Cardinal utility: U is ordinal and $U(x) - U(x')$ measures the difference between x and x', U is defined up to an increasing affine function: $aU + b$.

order. Now, if we consider an allocation transformation from x to y as a public project consequence, with $u_a(x_a) - u_a(y_a) > u_b(x_b) - u_b(y_b)$, some choices of f_a and g_b will be such that: $v_a(x_a) - v_a(y_a) < v_b(x_b) - v_b(y_b)$. Hence a comparison of utility level variations between agents is meaningless.

(2) Cardinality and non-comparability. If we transform agents a and b's utility functions, u_a and u_b, by unspecified, positive, independent affine functions: $v_a(\cdot) = \alpha_a u_a(\cdot) + \beta_a$ and $v_b(\cdot) = \alpha_b u_b(\cdot) + \beta_b$, with $\alpha_a > 0$ and $\alpha_b > 0$, we still obtain utility functions representing the two agents' preferences. For example, the von Neumann–Morgernstern's utility functions satisfy this. Here again, it is not possible to compare the agents' utility levels, since different affine transformations can arrange the two agents' satisfactions in different orders for functions u and v. In the same way, a comparison of utility level variations between agents is impossible. This shows that cardinality is not sufficient to ensure preference comparability.

(3) Ordinality and level-wise comparability. Let us suppose now that any transformation of utility functions by a single, increasing, monotonic function preserves the agents' preferences ($f_a = f_b = f$ in the first case). It then becomes possible to compare the individuals' utility levels because an agent with a higher utility than another will remain favoured for whatever transformation of utility functions: $u_a(x_a) > u_b(x_b) \Rightarrow v_a(x_a) > v_b(x_b)$. In contrast, it is still not possible to compare the utility levels' changes because the transformation does not necessarily preserve differences in values. Thus, cardinality is not necessary to ensure some form of preference comparability.

(4) Cardinality and level variations comparability. In this case, strictly positive affine function transformations, which only differ from one another by their constants ($\alpha_a = \alpha_b = \alpha$), preserve agent preferences. Contrary to the previous situation, level-wise comparisons are not possible, while comparisons by differences (gains or losses of utility) are: $u_a(x_a) - u_a(y_a) > u_b(x_b) - u_b(y_b) \Rightarrow v_a(x_a) - v_a(y_a) > v_b(x_b) - v_b(y_b)$. Indeed, because they differ, the constants can reverse the agent utility levels' hierarchy but they do not have an influence on the utility variations between levels.

(5) Cardinality and full comparability. Only an identical strictly positive affine transformation of all utility functions preserves the agents' preference order ($\alpha_a = \alpha_b = \alpha$ and $\beta_a = \beta_b = \beta$). Comparisons by levels, or level variations, are both possible. Thus, full comparability requires cardinality, otherwise stated: cardinality is necessary but not sufficient for full comparability.

(6) Cardinality and absolute measurement. We restrict transformations preserving the agents' preference orders to the same strictly positive linear functions for all utility functions ($\beta = 0$). This additional assumption is equivalent to introducing a zero value and yields an absolute value for utility functions: $u_a(o_a) = 0 \Rightarrow v_a(o_a) = 0$.

Only the first two formalisations (ordinality or cardinality without comparability) are supported by preference representation theorems. Comparability is only obtainable by adding an ad hoc condition on utility functions and cannot be justified by an axiom on preferences. As we announced, the implementation of social utility functions requires a comparison between agent satisfactions, given that they yield the social welfare as a function of the various agents' utilities. Depending on the use, a comparability by levels or by level variations is necessary.

3.1.2 Arrow's Impossibility Theorem

The second problem encountered with the social welfare function approach is due to the possibility of aggregating individual preferences with certain properties into social preferences with characteristics that are considered as desirable. The famous impossibility theorem, due to Arrow (1950), imposes drastic limitations on the hopes that had been formed in this field.

The starting point is a set of individual agent preferences ($i = 1, \ldots, I$) on a state space S, which are represented by complete pre-orders. The requirements for social preferences are as follows:

(1) Unrestricted field. Social preferences must be built upon any possible individual preference (complete) pre-order.

(2) Independence from irrelevant states. The order of two states given by social preferences depends only on individual orders of these two states.

(3) Weak Pareto condition. If each agent strictly prefers a state to another, the state will also be strictly preferred by the social pre-order.

(4) Non-dictatorship. There is no individual such that, if he strictly prefers a state to another, it will also be strictly preferred by the social pre-ordering, whatever the other agents' preferences.

It is then possible to state the following result.

Theorem (Arrow) *If the number of agents, I is finite and if the states space, S contains at least three elements, there is no pre-order of social preferences complete on S satisfying conditions (1) to (4).*

Thus, Arrow's theorem excludes any possibility of giving a coherent selection criterion which incorporates the whole set of individual preferences. It played the part of a relentless judge for appreciating a great part of the public economic literature.

It is not useless for intuitive comprehension, to parallel Condorcet's paradox with Arrow's impossibility theorem, even though the latter is definitely more powerful (and thus more hope-destroying). Let us suppose that society consists of three individuals: a, b and c, who have to choose between three possible states, s_1, s_2 and s_3, and that their (transitive) preferences on these states are such that:

- For individual a: s_1 is strictly preferred to s_2
  s_2 is strictly preferred to s_3
  s_1 is strictly preferred to s_3
- For individual b: s_3 is strictly preferred to s_1
  s_1 is strictly preferred to s_2
  s_3 is strictly preferred to s_2
- For the individual c: s_2 is strictly preferred to s_3
  s_3 is strictly preferred to s_1
  s_2 is strictly preferred to s_1

Let us suppose that group decisions are made according to the majority rule (which satisfies conditions (2) to (4) in the impossibility theorem). One can then easily check that the social preferences produced by the majority rule are such that:

- s_1 is strictly preferred to s_2,
- s_2 is strictly preferred to s_3, and
- s_3 is strictly preferred to s_1.

Hence, social preferences are not transitive, while all individual preferences from which they are constructed satisfy this coherence axiom. Obviously, one could obtain transitive social preferences from other individual preferences than those appearing above (for example, if the three individuals have the same preference order on the three states). However, Condorcet's paradox excludes one possible preference order configuration and violates condition (1) in Arrow's theorem as a consequence. This demonstrates the importance of the requirement that all the possible preferences be taken into account in Arrow's theorem.

One way to dodge the impossibility result is to try and weaken conditions (1) to (4). In particular, this is the case with condition (1): unrestricted field. Indeed, one can affirm that the social welfare function is built from the agents' preferences observable in an economy and not on all acceptable preferences, this relaxes condition (1). Social preferences can then be obtained for some individual preference configurations instead of all of them. Within this framework of restricted requirements, it was however shown that the impossibility result was essentially preserved (see, for instance, Myles (1995)).

A second way consists of providing more information on individual preferences for the process of determining collective preferences, leading us to reconsider the various interpersonal forms of comparison and to show as well how both mentioned problems are closely dependent. Accepting interpersonal comparisons of utility increases the possibilities for making coherent social choices with the agents' individual preferences and thus gives rise to some possibility theorems.

Again, let us take the six distinction points' typology of ordinality–cardinality and comparability–non-comparability. The absence of comparability brings us back to Arrow's impossibility theorem and thus leaves no place for a social utility function with good properties. The four other options allow a certain form of comparability and leave open some possibilities for the existence of social welfare functions.

The social utility function concept constitutes the keystone of public economics. It is referred to in taxation problems as well as in the study of external effects. A field of application which concerns us particularly is cost–benefit calculus, a basic approach to a great number of public decisions.

3.1.3 Cost–Benefit Analysis

Cost–benefit analysis aims to measure economic agents' disadvantages and/or benefits from a public project in terms of social welfare. Thus it offers a rational basis for public action.[3] Insofar as it rests on a collective welfare measure, it requires a social utility function of Bergson–Samuelson type as a starting point:

$$U(s) = U[u_1(x_1(s)), \ldots, u_I(x_I(s))]$$

where $x_i(s)$ is agent i's consumption in state s and $u_i(\cdot)$ its utility function.

Each agent i maximises its utility under its budget constraint: $p(s)x_i(s) = R_i(s)$, where $p(s)$ and $R_i(s)$ represent price and i's income in state s, respectively. The program's solution represents optimal demands: $x_i(s) = d_i[p(s), R_i(s)]$. An indirect utility function can then be expressed as: $v_i[p(s), R_i(s)] = u_i\{d_i[p(s), R_i(s)]\}$. The social utility function can then be written in terms of agents' indirect utility functions, where variables are prices and incomes instead of consumption quantities: $U(s) = U[v_1(p(s), R_1(s)), \ldots, v_I(p(s), R_I(s))]$.

The planner decides on a production plan, z, which defines an equilibrium. A signal ω indicates that the equilibrium is obtained: $\omega = f(z)$. If we write w_i for an agent's indirect utility as a function of equilibrium, we get:

$$U(\omega) = U[w_1(\omega), \ldots, w_I(\omega)]$$

[3] See Drèze and Stern (1987).

If one considers an initial situation z and a small size project dz, which induces an equilibrium variation $\mathrm{d}\omega = \frac{\partial f}{\partial z}\mathrm{d}z$, we obtain:

$$\mathrm{d}U = \sum_{i=1}^{I} \frac{\partial U}{\partial w_i}\frac{\partial w_i}{\partial \omega}\mathrm{d}\omega = \sum_{i=1}^{I}\left(\frac{\partial U}{\partial w_i}\frac{\partial w_i}{\partial R_i}\right)\left(\frac{\dfrac{\partial w_i}{\partial \omega}}{\dfrac{\partial w_i}{\partial R_i}}\frac{\partial f}{\partial z}\right)\mathrm{d}z$$

Let:

$$b_i = \frac{\partial U}{\partial w_i}\frac{\partial w_i}{\partial R_i} \quad \text{and} \quad MWTP_i = \left(\frac{\dfrac{\partial w_i}{\partial \omega}}{\dfrac{\partial w_i}{\partial R_i}}\frac{\partial f}{\partial z}\right)\mathrm{d}z$$

where $MWTP_i$ stands for "marginal willingness to pay" in monetary units. We get:

$$\mathrm{d}U = \sum_{i=1}^{I} b_i MWTP_i$$

This expression is the subject of an economic interpretation that conditions its use: $b_i = \partial U/\partial R_i$, represents the social marginal utility of agent i's income and is the weight it is assigned in the social utility function, expressed in terms of a monetary variation.

Given that:

$$\mathrm{d}w_i = \frac{\partial w_i}{\partial \omega}\mathrm{d}\omega = \left(\frac{\dfrac{\partial w_i}{\partial \omega}}{\dfrac{\partial w_i}{\partial R_i}}\frac{\partial f}{\partial z}\mathrm{d}z\right)\frac{\partial w_i}{\partial R_i} = MWTP_i\frac{\partial w_i}{\partial R_i}$$

$MWTP_i$ is the dz project's net value for agent i, with dz in monetary units, i.e. its marginal willingness to pay for the project.

When the planner decides on a small size variation around the equilibrium of his production, the resultant social utility variation can be written as the sum of individual net values obtained from the project, weighted by the social marginal utilities of individual incomes. Then a project's net profit in welfare is the weighted sum of individuals' willingness to pay for the project, with weights representing those used in the social welfare function. This formula is the basis of cost–benefit calculus: starting from an initial situation z, a dz project will be accepted if and only if the social welfare marginal value is strictly positive ($\mathrm{d}U > 0$).

Consumer theory makes it possible to characterise consumer willingness to pay. A consumer is confronted with a space of commodities among which L are commercial commodities and N are non-commercial commodities. If the non-commercial commodities are of public commodity type, the quantities available of these commodities are identical for all individuals. The method to obtain willingness to pay consists of identifying an individual's commodity value with the satisfaction level (utility) it provides and in extending this principle to non-commercial commodities.

Initially, one represents individual i's preferences by a utility function: $u_i(x_i, z)$, where x_i is the consumption vector of agent i of commercial commodities and z is the consumption vector of non-commercial commodities. According to classical consumer theory, it is supposed that individuals maximise their utility by choosing among the commercial commodities (they do not control the stock level of non-commercial commodities): max $u_i(x_i, z)$ under budget constraint $px_i = R_i$, where p is the commercial commodity price vector and R_i is agent i's income.

The solution of this optimisation problem gives the traditional (Hicksian) demand functions:

$$x_i^l = d_i^l(p, z, R_i), \quad l = 1, \ldots, L$$

as well as indirect utility functions defined by:

$$v_i(p, z, R_i) = u_i[d_i(p, z, R_i), z]$$

It is assumed that an increase in a public good quantity, when all others do not decrease, leaves prices and income unchanged. Then consider a vector z^1 in which at least one component increased ($z^1 \geq z^0, z^1 \neq z^0$). We have:

$$u_i^1 = v_i(p, z^1, R_i) > u_i^0 = v_i(p, z^0, R_i)$$

The compensating variation measure is thus defined as the amount of WTP_i which leaves individual i's welfare unchanged by the non-commercial commodity modification from state z^0 to state $z^1 : v_I(p, z^1, R_i - WTP_i) = v_I(p, z^0, R_i) = u_i^0$. This compensating variation can be regarded as the "willingness to pay" for an increase in quantity of a beneficial non-commercial commodity.

The difficulty of determining individual prices and the obtained results' strong sensitivity to the method chosen, often encouraged economists to use market prices, whenever possible. This methodology poses two types of problems which will be analysed in the following sections. The first problem is the normative validity of valuation by market prices under uncertainty. The second concerns the adequacy of the approach for public investment choices.

3.2 MARKET AGGREGATION OF INDIVIDUAL PREFERENCES

Markets are places where commodities are traded; here they will be a formal representation of effective exchanges between individuals. They also provide price discovery. Prices play a double role. They give the exchange ratios between traded commodity quantities, whether directly (barter economy) or via currency units (monetary economy). They also provide signals used by agents to make decisions. In a decentralised economy, prices depend on individual supply and demand or, not fundamentally differently, on bid and ask prices, according to the mechanism used for their formation. In all cases, individual valuations (the maximum price at which traders are ready to buy and the minimum price at which they are ready to sell) condition the emergence of effective market prices, i.e. prices that will actually be paid. In this sense, one can assert that the market (as a setting for price emergence) aggregates individual behaviours. Looking at motivations of individual behaviours, prices aggregate agents' characteristics: tastes, budget constraints, attitudes towards risk, expectations, production possibilities and exchange constraints.

Obviously, a market is a particular aggregation mechanism for agent behaviours, while another one is the voting process. The pricing process in perfect markets is the main point of reference in this field of economic analysis for two reasons. The first one is not especially glorifying for economics: the equilibrium representation in perfect markets is the only available general model with strong internal coherence! The second is definitely a more avowable motivation: the equilibrium properties justify its use as a normative reference.

3.2.1 The General Equilibrium Model Under Uncertainty

One of the major characteristics of general equilibrium is that it takes the markets' interdependence into account: prices in a commodity market depend on prices in the other markets. This makes it possible to integrate the effects of substitutability and complementarity between commodities and between production factors, as well as joint production and economies of scope.

The general equilibrium model seeks to represent a perfectly competitive economy with two major characteristics: agents are price-takers (they consider that their behaviours have a negligible influence on prices) and the agents' preferences only depend on their commodity endowments, in the sense that they are independent of the other agents' behaviours (no rivalry and no altruism). The assumption that agents are price-takers can be understood as a consequence of atomicity, perfect information and the absence of frictions in markets. The "modern" version of general equilibrium (Arrow–Debreu model) gave birth to many alternatives and extensions. For instance, some kinds of imperfect competition have been introduced, or constraints on traded quantities (equilibriums with rationings), but the versions which are of interest to us here are relative to the introduction of uncertainty. The basic model which is used as a reference in this subsection originated in Arrow (1953) and was included in Debreu (1959, chapter 7). We shall present its general features, and shall insist on its properties later on.

The economic commodities (rare commodities) are defined by four data:

- Their physical characteristics (as they are perceived by agents).
- The place where the commodity is available.
- The date at which the commodity is available.
- The state of the world (or nature) in which the commodity is available.

The first two elements are often gathered under the denomination of "physical characteristics", for the sake of simplicity. The last two are known as the "pair dates–states", and even "states", to make it shorter. As one will be able to notice, the model is a static one, so the explicit loss of the time reference is not detrimental.

In this model, uncertainty derives from the fact that agents do not know which state will occur at each future date. These states gather all events relevant to the whole set of individuals; in fact, as we shall see below, all events which have an influence on the agents' welfare. What is necessary to notice from now is that all individuals agree on the same definition of uncertainty, even though they may have completely different opinions on the probabilities for the states' occurrences.

To simplify, we will suppose here that the number of commodities is finite, which is not very satisfactory considering the way in which they are defined. However, this assumption can be relaxed with no consequence for the results, at the cost of technical complexity. We will denote the commodities by $l = 1, \ldots, L$.

There are two types of agents: consumers and firms. Consumers (assumed to be in finite number) are denoted $i = 1, \ldots, I$. Consumption of agent i of commodity l is denoted x_i^l, and its consumption plan by $x_i = (x_i^1, \ldots, x_i^L)$. Agents carry out consumption choices as a whole consumption plan, and let X_i be the set of such plans on which they have preferences. Preferences satisfy the general properties presented in Chapter 2. For instance, we shall suppose that each agent's preferences are given by a monotonic and strictly convex, continuous, complete pre-order. It is then possible to characterise an agent's tastes by a continuous monotonic, strictly concave (ordinal) utility function u_i.

Consumers lay out an income R_i at the start of trades. Commodity prices are denoted $p = (p_1, \ldots, p_L)$ and a consumption bundle value is $px = \sum_{l=1}^{L} p_l x_l$.

A consumer has to constrain his choices to a value less than or equal to his budget,[4] the set of available bundles at price p, where the budget set is:

$$B_i(p) = \{x \in X_i / px \leq R_i\}$$

Agent i's choice will be a bundle in his budget set; the one preferred to all others in this set. This behaviour defines the demand function:

$$d_i(p) = \{x \in B_i(p) / u_i(x) \geq u_i(x'), \forall x' \in B_i(p)\}$$

Under the above assumptions, consumer demand is a continuous function of prices. Furthermore, it can be shown that the budget constraint is saturated (the whole income will be spent: $px = R_i$).

Firms are denoted $j = 1, \ldots, J$ (finite in number) and they are characterised by their production technique, a quantity vector for each commodity, where outputs are positive and inputs are negative, according to the usual convention: $y_j = (y_j^1, \ldots, y_j^L)$.

The set of feasible production techniques is called the "production set" and is denoted by Y_j. Fundamental assumptions on these sets are:

- Free disposal of the surpluses, which supposes that it is always possible "to waste" quantities of inputs or outputs without cost: $\forall y \in Y_j, y' \leq y \Rightarrow y' \in Y_j$.
- Convexity: $\forall y, y' \in Y_j, \forall \alpha \in [0, 1], \alpha y + (1 - \alpha) y' \in Y_j$.

The last assumption excludes the increasing returns to scale phenomenon. It is quite obvious that this type of restriction is not in agreement with reality, where many industrial situations make it possible to increase the productive efficiency along with the size of production plans. It is however coherent with the perfect competitive markets assumption formalised in the model. Indeed, the presence of increasing returns to scale implies the existence of important sized firms, who thus do not respect atomicity and make the assumption that producer decisions do not have a significant influence on the prices very unlikely.

The value of a production plan (profit of the company) is given by: $py = \sum_{l=1}^{L} p_l y_l$. Firm j's choice is a production plan belonging to its production set, yielding a profit superior to any other technique in this set. This behaviour defines the producer's supply function:

$$s_j(p) = \{y \in Y_j / py \geq py', \forall y' \in Y_j\}$$

Adding some assumptions (returns converge towards zero when production increases to infinity, and no constant returns to scale), the producer's supply is a continuous function of prices.

In order to complete the model, it is necessary to specify the assignments of firm profits and to specify consumer incomes at the same time. In a private property economy, the companies' results are paid to their shareholders in proportion to their ownership rights: η_j^i

4 Given that commodities are time-contingent, this does not exclude borrowing and lending, it is merely a solvency constraint.

is agent i's share of firm j. Furthermore, consumers hold initial commodities endowments: $w_i = (w_i^1, \ldots, w_i^L)$. Their incomes are given by the sum of their shares in firm profits and their endowments values:

$$R_i = pw_i + \sum_{j=1}^{J} \eta_j^i py_j$$

From the agents' demand and supply, it is possible to define individual excess demand and then global excess demand:

$$z_i(p) = d_i(p) - w_i, \quad z_j(p) = -s_j(p)$$

$$Z(p) = \sum_{i=1}^{I} z_i(p) + \sum_{j=1}^{J} z_j(p)$$

Allowing us to easily define equilibrium.

Definition *An economy is represented by consumption sets, preferences (i.e. utility function) and initial endowments for each consumer* $i = 1, \ldots, I{:}(X_i, u_i, w_i)$*, production sets of firm* $j = 1, \ldots, J{:}Y_j$ *and, finally, by consumers' shares in firms:* $\left(\eta_j^i\right)_{i=1,\ldots,I;j=1,\ldots,J}$. *A general equilibrium is a price* p^* *such that:* $Z(p^*) = 0$.

The equilibrium concept means/implies market clearing: at the equilibrium price, the demand sum is equal to the supply sum, for each commodity. Then, each commodity demander can buy the desired quantity at the equilibrium price, while each supplier can sell the quantity he wishes to buyers. Indeed, perfect information is assumed so that demanders and suppliers can meet each other. In such a situation, no agent will find incentives to change behaviour (no regret), a dominant equilibrium notion in economics. Only at equilibrium do agents' demand and supply functions make sense. Indeed, it is only then that *ex ante* demands and supplies are equal to *ex post* purchases and sales. Under the assumptions made on agents, total excess demand functions have particular properties:

1. They satisfy Walras' law – net total trades' values are nil. This comes from value being defined by prices, hence what is sold and what is bought have the same value.
2. They are continuous.
3. They are strictly positive when prices converge towards zero (commodities are desirable by consumers).
4. They are homogeneous of degree zero in prices (no monetary illusion).

Thanks to these properties and with the help of a fixed point theorem, the existence of a general equilibrium can be proved.

An important assumption for our present topic is that of the complete market. In the current context, it means that a market is open for each commodity, i.e. each physical characteristic, but also each state (time and uncertainty) on which commodities are contingent. Every agent may trade a contract delivering any commodity, at any future time and under any uncertain event considered as a possible state of the world: agents can hedge any risk. This does not imply that they actually do so, nor that their budget constraints allow them to buy complete insurance, but it is possible to express such a demand. Market completeness prevents any incentive to trade again (reopen a market) when some state occurs in the future.

Indeed, contingent commodities (contracts) have made it possible to allocate income at best to each possible future state. The classical example of such a contingent contract is one that yields the opportunity to buy an umbrella tomorrow if it rains. With such a contract, there is no reason to make another trade tomorrow if rain falls, given that demand was adapted to the two possible states. It should be clear that a contract for an umbrella contingent on it raining or a contract for an umbrella if it doesn't are bound not to have the same prices. One could expect the former to be higher than the latter because demand is likely to be higher for the first event-contingent contract. Contingent commodity prices reflect agents' preferences about commodities themselves (consumption) and also reflect their attitudes towards risks. Hence, market completeness justifies the static characteristic of the model: it is possible to write contracts contingent on any future date and uncertain event at the present time.

Let's turn to the general equilibrium properties in terms of collective choice. This question requires that we know how to decide which allocation is collectively desirable and which one is not for the agent population. We are looking for a purely economic criterion, meaning: an efficiency criterion such that no ethical value can contest it. Hence, economic efficiency should be unanimously recognised. An efficiency concept such that no rational[5] agent will reject its allocation is known as Pareto efficiency (or Pareto optimality). In order to define Pareto efficiency precisely, let us introduce the feasible allocations.

Let $w = (w^1, \ldots, w^L)$ be the vector of initial endowments in the economy. Let Y be the set of all production techniques available in the economy (and not for one firm only). A commodity vector $x = (x^1, \ldots, x^L)$ can be obtained for consumption if and only if: $x - w \in Y$.

Definition *An allocation is a commodity assignment among consumers:* $(x_i)_{i=1,\ldots,I,}$ *such that:* $\sum_{i=1}^{I} x_i = x$.

Definition *An allocation is feasible for initial endowment w and production set Y, if it can be obtained from the techniques in Y using inputs from w, i.e. if it belongs to the set:*

$$R = \{(x_i)_{i=1,\ldots,I} / \sum_{i=1}^{I} x_i - w \in Y\}$$

Definition *A Pareto optimum is a feasible allocation, x, such that there are no other feasible allocations that would be preferred by every agent and strictly preferred by at least one of them:*

$$[\forall i = 1, \ldots, I,\ u_i(x') \geq u_i(x) \text{ and } \exists k,\ u_k(x') > u_k(x)] \Rightarrow x' \notin R$$

No individual agent has any incentive to oppose a move from a non-optimal situation to an optimal (or efficient) allocation. Indeed, such a change is possible without any agent being made worse off (some agents' utilities may be increased without decreasing the others'). Conversely, in an efficient situation, it is not possible to modify resource assignments in the economy without diminishing some agent's welfare. The Pareto efficiency concept marks the limit between what it is possible to do under purely economical considerations and

[5] Rational is not an innocuous assumption, we shall come back to its meaning later on.

what would be required to make ethical judgements. Obviously, this limit is related to the assumptions made in the model, notably the one about preferences bearing upon an agent's own consumption only. It would not be possible to eliminate ethical considerations, if one introduced other agents' welfare components as arguments of an agent's utility function (such as altruism or rivalry). The link between general equilibrium and Pareto efficiency is given by the two welfare theorems.

Theorem (1) *Any general equilibrium yields a Pareto efficient allocation.*

Theorem (2) *Any Pareto efficient allocation can be obtained through a general equilibrium: there exists an assignment of initial endowments among agents and a price vector such that the efficient allocation is a general equilibrium of the economy for this price vector and with this endowment assignment.*

The first welfare theorem formalises Adam Smith's invisible hand descriptive fiction: it says that a collectively satisfying situation will be obtained in a competitive economy through trades between individual agents who act according to their own satisfaction only (consumer utility and producer profit maximisation). Collective satisfaction means Pareto efficiency in our setting. This result is invoked to justify market prices as a commodity social value.

If commodities and services are allocated efficiently, their relative individual values[6] are the same for every agent. This property is satisfied at equilibrium because all relative individual values are equal to relative equilibrium prices.[7] Using market prices, individual valuations are taken into account, which are collectively compatible at equilibrium and consequently the equilibrium is Pareto efficient.

Let us underline that efficiency has nothing to do with social justice or equality between agents, whatever their meanings. Indeed, an allocation where all available resources are allocated to a unique (and non-saturated) agent is Pareto efficient! It would be hard to maintain such an allocation in most human societies, though.

The second welfare theorem provides a way to ward off the difficulty: it is always possible to obtain a socially desirable allocation (i.e. one of the Pareto efficient ones) as the result of competitive trading in equilibrium. What is the advantage derived from this method with respect to a more direct one (e.g. regulation)? In order to plan a Pareto optimum, one would need to know agents' preferences, but this is hardly feasible in practice. Conversely, this method requires that initial endowments are modified, then competitive trading will achieve the desired Pareto optimum at equilibrium. Note that competitive markets do not allocate commodities more or less equally among agents: they are neutral in that respect. Markets only deal with achieving efficiency: if initial endowments are equally (unequally) assessed, then equilibrium allocations will be equal (unequal). Market mechanisms are not able to amplify nor to reduce social inequalities in a competitive economy. There is no critique nor argument in favour of public action in the previous remarks, they merely aim at clarifying the task allocation between economics and politics: the former can deal with efficiency, while the latter is able to look for socially acceptable resource allocations, with no contradictory involvements.

[6] Individual marginal rates of substitution.

[7] This result is easy to understand: assume that there is one relative individual value between two commodities that differs from the other agents'. Then the agent has an incentive to substitute one of the two commodities for the other and this is in contradiction with equilibrium where no agents would gain from modifying their net demands.

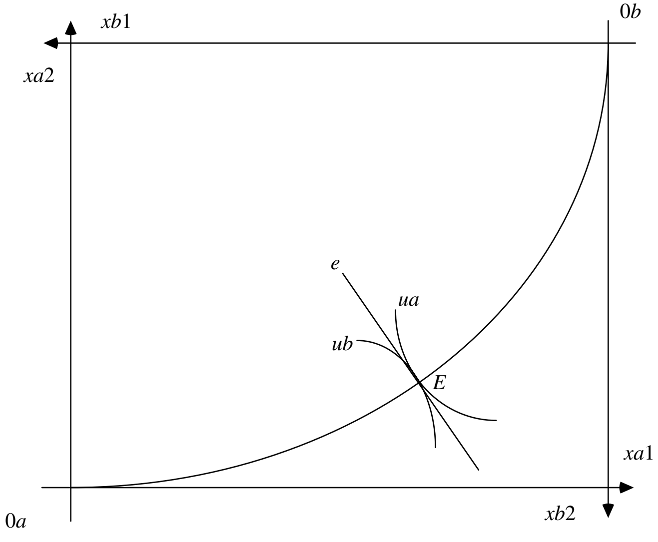


Figure 3.1 Edgeworth box

We illustrate this important result in Figure 3.1: an Edgeworth box, which corresponds to a very simple version of our problem. The economy is provided with two commodities, 1 and 2, and two consumers, *a* and *b*. In the absence of production, the problem is simply that of distributing an initial endowment of the two commodities, $W = (w_1, w_2)$, between the two agents. The Edgeworth box length is w_1 and its height w_2: it defines the locus of all possible initial endowment assignments and hence the locus of all feasible allocations. Axes representing agent *a*'s consumption are as per usual, while those for agent *b*'s are rotated by 180°.

Let us assume that the state wants to obtain an allocation of resources based on equity, in the sense that no agent would envy another one. If consumer preferences are unknown, the only way to achieve the state's objective is to assign the same amount of resources to each individual agent. This is point *e* in the diagram,[8] with $x_a^1 = x_b^1 = w^1/2$ and $x_a^2 = x_b^2 = w^2/2$. Such an allocation respects equity, but is not efficient in general. If agent *a* is relatively more keen about commodity 1 and agent *b* is more greedy for commodity 2, for instance, both their welfares can be improved by a trade. A solution is to organise competitive markets between the two agents from initial endowment *e*. Equilibrium is obtained at point *E*: this point is on the line 0_a–0_b (Pareto efficient allocations), where two indifference curves and the budget constraint lines are tangent. Budget constraints are those coming through initial endowment point *e*, with slope equal to minus the ratio of equilibrium prices. It can be proved that *E* satisfies equity and efficiency. Hence, competitive equilibrium is a means of reconciling efficiency and equity, or any other social justice requirement, without the state having to know the agents' preferences.

At this stage, we can compare the method we just saw with the social welfare functions we evoked in section 3.1. Bergson–Samuelson social utility functions are directly linked with the choice of an optimum point in Edgeworth box, *E*, for instance. Social welfare

[8] Geometrically, *e* is the centre of the Edgeworth box.

functions depend on agents' utility functions, so the latter must be known by a social planner. Furthermore, individual utility functions must be comparable, as we have seen before, so that the existence of a social welfare function is not contradicted by Arrow's theorem.

Reference to prices in a market economy is then a major teaching of the general equilibrium model: prices are the only information we can get about agents' preferences by simple observation. This is the reason why economic calculus is founded on market prices, whether directly or indirectly. We shall investigate precisely how these prices enter into cost–benefit analysis in what follows.

3.2.2 Reference to Market Prices in Public Capital Budgeting

We have presented the foundations of cost–benefit analysis in a preceding section; they are rooted in the general equilibrium model. In this context, market prices are a natural measure for commodity values, because equilibrium prices are equal to individual valuations.

However, market imperfections may distort prices. Take labour as an example: are hourly wage rates a commodity measure for this factor value? In equilibrium, the answer would be yes! Market clearing means that there is no excess demand or supply for labour and price flexibility should make any discrepancy disappear. However, unemployment prevails in most economies, whatever may cause it, introducing a distortion in commodity relative values. Labour value may be considered as lower than the observed wage rate under unemployment,[9] but there is no way of determining what this wage rate should be in whatever equilibrium occurs.

Market prices are always present in economic calculus, but their most difficult use is encountered in one application: risk valuation. The problem arises when costs or damages and benefits must be discounted in order to value a public investment. Obviously the problem is particularly important in a dynamic setting (see Part III), but it already has to be solved in a static approach. It has been a major issue in economic literature for about 20 years, and the answers have become clearer since the mid 1980s.

The debate was on whether to use the market riskless rate (e.g. Treasury bills rate) for public projects or a rate including a risk premium equal to the one paid for an equivalent private project.[10]

Let us present the arguments in favour of a risky rate first. Hirschleifer (1966) and Sandmo (1972) are two major contributions to this position, and both are founded on the general equilibrium model. This trend of thought recommends the use of the same rate that would be used in the private sector if a firm were in charge of realising the project, for discounting a public project's impacts. We have seen in the example of Millau's viaduct in Chapter 1, that discount rates invoked in one and the other valuations were not the same. More generally, the comparison is made between a riskless reference rate and a discount rate that would be used by a firm producing the same commodity allocations in every future state of the world. The main argument is that this rate yields an equilibrium between consumers' and firms' relative values and that any other rate would induce inefficiencies. If this method is turned down, a project which produces income in wealthy states of the world may be favoured at the prejudice of projects bringing wealth in states where it is needed. Let us take a simple

[9] There is no loss in productivity incurred by employing an unemployed agent, but negative utility of labour is difficult to measure.

[10] Investors require a higher rate for a risky investment, the difference between risky and riskless rates is called the "risk premium", details will be seen in Chapter 7.

example to illustrate this: consider a general equilibrium without production (pure exchange economy), with only one physical commodity, two dates 0 and 1, two future possible states at date 1: s_1 and s_2, and two consumers: a and b. Agents' utility functions are:

$$u_i[x_i(0), x_i(s_1), x_i(s_2)] = \ln[x_i(0)x_i(s_1)x_i(s_2)],\ i = a, b$$

Their initial endowments:

$$w_a(0) = 5.5, \quad w_a(s_1) = 7, \quad w_a(s_2) = 0$$
$$w_b(0) = 0, \quad w_b(s_1) = 7, \quad w_b(s_2) = 10$$

Budget constraints:

$$5.5p(0) + 7p(s_1) = x_a(0)p(0) + x_a(s_1)p(s_1) + x_a(s_2)p(s_2)$$
$$7p(s_1) + 10p(s_2) = x_b(0)p(0) + x_b(s_1)p(s_1) + x_b(s_2)p(s_2)$$

Utility maximisation under the constraints yields the optimal demands of the consumers:

$$x_a(0) = \frac{1}{3}\left[5.5 + 7\frac{p(s_1)}{p(0)}\right]$$
$$x_a(s_1) = \frac{1}{3}\left[5.5\frac{p(0)}{p(s_1)} + 7\right], \quad x_a(s_2) = \frac{1}{3}\left[5.5\frac{p(0)}{p(s_2)} + 7\frac{p(s_1)}{p(s_2)}\right]$$
$$x_b(0) = \frac{1}{3}\left[7\frac{p(s_1)}{p(0)} + 10\frac{p(s_2)}{p(0)}\right]$$
$$x_b(s_1) = \frac{1}{3}\left[7 + 10\frac{p(s_2)}{p(s_1)}\right], \quad x_b(s_2) = \frac{1}{3}\left[7\frac{p(s_1)}{p(s_2)} + 10\right]$$

Market clearing:

$$x_a(0) + x_b(0) = 5.5, \quad x_a(s_1) + x_b(s_1) = 14, \quad x_a(s_2) + x_b(s_2) = 10$$

Without loss of generality, the commodity unit price may be normalised at 0 and equilibrium prices are then:

$$p(0) = 1, \quad p(s_1) = \frac{5.5}{14}, \quad p(s_2) = \frac{5.5}{10}$$

Consider a consumption plan yielding a unit of the physical commodity in each state at date 1 (a riskless asset in this model). Its price must be: $p(s_1) + p(s_2)$ and the riskless rate is:

$$R_F = \frac{1 - [p(s_1) + p(s_2)]}{p(s_1) + p(s_2)} \cong 0.0606$$

Similarly, one can calculate interest rates for risky consumption plans. A plan paying one unit of the commodity in state s_1 only, has a rate:

$$R(s_1) = \frac{1 - p(s_1)}{p(s_1)} \cong 1.5455$$

A plan paying one unit of the commodity in s_2 only, has a rate:

$$R(s_2) = \frac{1 - p(s_2)}{p(s_2)} \cong 0.8182$$

Now, take a public project yielding $y(s_1)$ units of the commodity in state s_1 and $y(s_2)$ in state s_2. What is its value (price) at date 0?

If we apply risk prices to evaluate it, we get:

$$V(y) = \frac{y(s_1)}{1+R(s_1)} + \frac{y(s_2)}{1+R(s_2)} = p(s_1)y(s_1) + p(s_2)y(s_2)$$

If we value the project as its expected gain discounted at the riskless rate (with equal probability for each state in order to keep it simple), we get:

$$V_F(y) = \frac{y(s_1)+y(s_2)}{2(1+R_F)} = \frac{1}{2}[p(s_1)+p(s_2)][y(s_1)+y(s_2)]$$

Consider two projects:

$$y(s_1) = 1, y(s_2) = 5 \quad \text{and} \quad z(s_1) = 5, z(s_2) = 1$$

With respective costs at date 0:

$$C(y) = 2.9 \quad \text{and} \quad C(z) = 2.8$$

Using the Net Present Value (NPV) method at the riskless rate, project z will be chosen over y:

$$NPV_F(y) \cong -0.0714 < 0 \quad \text{while} \quad NPV_F(z) \cong 0.0286 > 0$$

If we discount at the risky rates, the choices are reversed:

$$NPV(y) \cong 0.2429 > 0 \quad \text{and} \quad NPV(z) \cong -0.2857 < 0$$

However, project y yields more wealth in state s_2 where the total endowment of agents is lower (10 against 14 in state s_1), while project z does the converse. We could then consider that y is socially worthier than z in terms of risk reduction. Then the choice induced by using the riskless rate is at fault.

Undervaluation of risks in public projects leads to dominated resource allocations. In all cases, a public (or private) project that creates value can be evicted in favour of projects that destroy it. Taking risks into account does not necessarily imply more or less public investment, but it leads to the choice of more efficient projects.

There are different arguments in favour of using a riskless rate to evaluate public projects, but they are all related to diversification and risk sharing. Risk sharing is based on the following principle: if N independent (and identical[11]) risks are put together (are summed), then an Nth of the total risk is less risky than any original risk. This results from the risk level being measured by variance and the variance of an Nth is a function of $1/N^2$. Furthermore, according to one of the laws of large numbers, if N converges to infinity, each Nth of the total risks converges towards a non-random value. We shall refer again to this principle in the economic analysis of insurance in Chapter 6. In our problem here, the state may act as a huge insurer, given its size, and it can organise a risk sharing between citizens through taxes and subsidies in such a way that they would be perfectly insured. Then the state is justified to use a riskless rate to value its investments in some new risky project because it would not change (much) the overall risk.

[11] Same means and variances, or, more generally, same probability distribution.

Diversification is founded on another property of the sum of risks: if some risks are negatively correlated, the variance of their sum is lower than the sum of their variances. This has nothing to do with large numbers, two risks can perfectly hedge each other as we have seen in the example above. Obviously, in a large economy, activities are diverse and the state is more able to diversify, at least some risks, than any private company (although larger ones do diversify between non-correlated activities). Then, if the state's total risk is considered as sufficiently diversified to be considered as riskless, the state is justified to use a riskless rate to value its investments.

Using these two arguments to value a new project introduces a confusion between a global valuation for the overall state risk and the valuation of some particular new project. Even if the project is considered as tiny enough to have its risk shared and/or diversified away, its valuation should consider it for itself. Otherwise, a project with a high risk may be valued more highly than a lower risk project with the same impacts. Let's go back to the previous example to illustrate this.

Projects y and z should not be valued together if they are not physically joint projects. Take their NPVs (including risk valuation):

$$NPV(y) \cong 0.2429 > 0, \text{ while } NPV(z) \cong -0.2857 < 0, NPV(y+z) \cong -0.0429$$

Project $y+z$ does not create an added value as y does, but it is riskless. If the two projects can be realised together, $y+z$ will be rejected, although y is desirable. In other cases, a project with a negative value may be chosen due to the added value of another one, artificially linked to it.

On the other hand, it is true that public projects, as big as they can be, are tiny compared to the whole economy, at least in large states. If the state was managing the national economy as a whole,[12] it could efficiently distribute risks among citizens through wages and taxation (risk sharing) and use the diversification effect. This would justify that all projects, and not only public ones, would be efficiently valued by a riskless discount rate. Some recent events have shown that this could lead to catastrophic investments (Tchernobyl, for instance). Some errors are due to the fact that, even in closed national economies, discount rates are linked to international monetary markets. Others come from a tendency to undervalue future possible damages, because of applying too low a discount rate.

Nevertheless, the above arguments in favour of using a riskless rate for public projects constituted the foundations of the first attempts to debate on the efficient way to value them (Arrow and Lind, 1970). If a particular public investment's returns are independent of other components in the national income, and if the risk is shared by a great number of individual agents, it must be valued by its expected returns, discounted at a riskless rate, in order to decide whether to implement it or not.

This way of reasoning has been criticised at two levels. Firstly, agents do not choose to participate in the state's risk sharing. They are forced to go along with it, whether they would take that risk or not if they had the choice. Because there are no markets for public project risks which would provide efficient risk allocations, the allocation decided by the state would be dominated. Furthermore, public firms are not owned by individual agents who would efficiently exchange risks, through a financial market for instance, and this is another source of inefficient risk allocation. Such state-based risk allocation may lead to

[12] It has been the case to a large extent in countries ruled under Marxist ideology.

the valuing of a public project at a higher rate than the one referred to for private similar projects (Stapleton and Subrahmanyam, 1978).

Hirschleifer (1966) proposes an analysis based on assumptions on the difference between public and private investments that are the exact opposite of the one referred to by Arrow and Lind (1970). The former suggests that any public investment can be related to a private one with returns highly positively correlated with the public project returns. Conversely, the latter considers that private and public investment returns are not correlated at all and/or are efficiently shared by citizens. A more general approach was presented by Grinols (1985). The context is a general equilibrium version with imperfect risk sharing because of restrictions on asset markets. The government maximises a Bergson–Samuelson social welfare function. In the general case, the difference between public ($G(y)$) and private ($V(y)$) project valuations can be decomposed into two terms:

$$G(y) - V(y) = A(y^*) + B(\bar{y})$$

In order to do that, a public project is split into two components: $y = y^* + \bar{y}$. The first one is called private sector ($y^* \in P$), and the allocations it provides may be replicated by a portfolio of private securities. The second component represents the non-private sector elements, which form a vector ($\bar{y} \in \bar{P}$) orthogonal to the first one: $\bar{P} \perp P$. If the project is in the private sector (or replicated by it) $y \in P$ ($\bar{y} = 0$), the second term of the difference, $B(\bar{y})$, is zero. Furthermore, if agents are treated equally in the social welfare function, the first term of the difference, $A(y^*)$, is zero as well. Hirschleifer's result is then obtained: the public project is valued as if it was a private one with the same risk. The first term may be interpreted as a risk assignment effect: in contrast with a private firm, the government may take into account inequalities introduced by a project (some agents get benefits and others are disadvantaged) as well as wealth production. Because the assignment term may be positive or negative, it gives no clues as to superiority or inferiority of the public discount rate with respect to the private one. If the assignment problem can be solved independently (by an optimal tax/subsidy scheme), the first term is zero in all cases.

If the public project is not completely replicated by private ones, the second term in the difference can be interpreted as a risk-sharing representation: it is the social value of risk-sharing variations introduced by the public project. If the project is completely out of the private sector risk domain ($y \in \bar{P}$) and the risk premium converges towards zero because it is shared by a great number of tax payers and/or diversified away, we are back to Arrow and Lind's case: valuation is done by the expected returns discounted at a riskless rate. Out of this special case, the second term in the difference may be positive or negative. A project that cannot be realised by the private sector alone does not indicate whether the public rate must be more or less than the private one. Arguments in favour of an inferior one, because it shares risks better or because public risks are not marketable, are not supported by the general model.

If we leave aside risk assignment problems because it is assumed they are dealt with otherwise,[13] a conclusion from Grinols' model is that the crucial point is the possibility of replicating public risks by private ones. Private risks are often financial ones: financial securities, insurance contracts, etc. In order to use these risks, it is necessary to refer to financial asset valuation, and we shall see to this in Chapter 7.

[13] Even though it is rarely optimal to solve problems separately, it is often a good way not to be able to decide at all if they are treated altogether!

In the applications that are presented in the next chapter, it is important to keep in mind the validity domain of the above arguments. They are founded on the normative power of market prices to reflect social value. It assumes that market imperfections are not so great as to deprive us of information about agents' attitudes towards risk. It assumes furthermore that the state decides on a way to take individual agents' interests into account, and that no superior "national" (or ethical) interest prevails. Such an interest above the individual ones (supposedly shared by every citizen) is difficult to define in economic terms and may contradict democratic concepts announced by a government.

4 Individual and Collective Risk Management Instruments

4.1 DECISION TREES

A decision tree is a graphical description meant to summarise and present a decision problem's most relevant elements. They are displayed on an expanding lattice similar to a tree. See Figure 4.1.

This rather intuitive method of gathering information and illustrating an ongoing analysis was developed during the early 1950s among operations resarch tools and as a device to represent a game (in extensive form). Luce and Raiffa's (1957) famous book made economic applications of decision and game theories clearer. The method's purpose is the analysis of a decision problem in order to eventually find a rational solution.

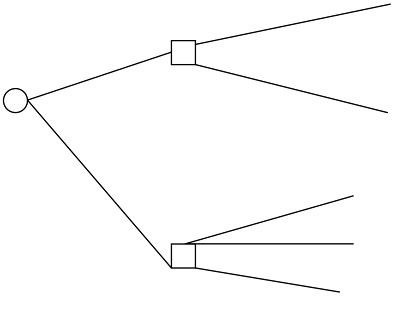


Figure 4.1 Decision tree

4.1.1 Decision Tree Structure

A tree, or part of a tree (subtree), can be decomposed (analysed) bit by bit so as to better spot feasible actions, information arrival, relevant time periods, etc. Obviously, not everything can be put on a tree, too many details would inhibit intuition, however, many things are added after a first analysis has been made: subtrees may grow. Tree construction necessitates the consideration of the different elements of a decision problem, which is a more difficult task for a collective choice than for an individual one. Let us stay in a sylvan domain to illustrate decision trees. We shall refer to a case study we made:[1] the decision problem was

[1] Case study (2002): "Methodological approach to decision assistance in forest protection against fires" (*Approche méthodologique pour l'aide à la décision pour la protection contre les incendies de forêt*), Financed by Ministère de l'Intérieur and Ministère de l'Agriculture, convention 61 45 27/01, MAAPAR-DGFAR.

about which means firemen and territorial authorities may use in the struggle against the starting of forest fires.

At the tip of branches, we find the youngest shoots, which represent ultimate consequences of decisions indicated in higher branches. Even when the time horizon is very short, the number of possible consequences may be very great. Shoots are linked by boughs to branch junctions, forming nodes, which are represented by circles or squares.

A square stands for a decision point, i.e. a stage where one decision has to be made: on the tree, one **or** another of the joint branches will be chosen. The tip of a branch is the consequence of the decision taken, and corresponds to a subtree in general and to an ultimate consequence when the time horizon is reached. See Figure 4.2.

Circles represent phenomena that are out of control for the decision-maker: all branches coming out of such a node must be taken into account and hence a whole set of consequences is considered. Again, each consequence corresponds to a subtree, in most cases. See Figure 4.3.

Time is visualised as the stages following circle nodes: they represent information states, i.e. different states which will be known at that time and according to which a partial decision can be made. The foot of the tree corresponds to initial information, the ultimate top branches correspond to final information. Final information becomes available after other information arrived and decisions were made. Following the uncertain outcomes branches (those sprouting from circle nodes) we obtain a "state of the world" as defined in decision theory. If we add to the uncertain outcomes branches, the one corresponding to decisions made, we have a trajectory from the initial time to the time horizon.

Tree construction (or growth) starts at the foot. It may be a circle or a square, depending on whether the first decision is made at the start or after a first piece of information has arrived. It is often the case that one has to go back to the tree and change things as the description develops. The decision tree grows according to the nodes that are added. Some are circles, when an information source may be expected at some future date, while others are squares, when it seems judicious to make a partial decision with the information available at that

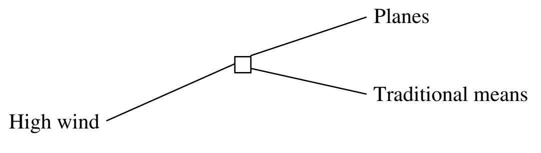


Figure 4.2 Decision point

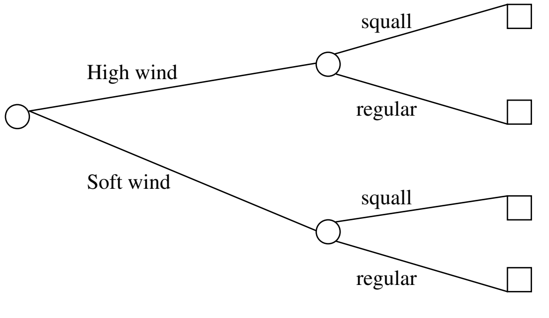


Figure 4.3 Chance nodes

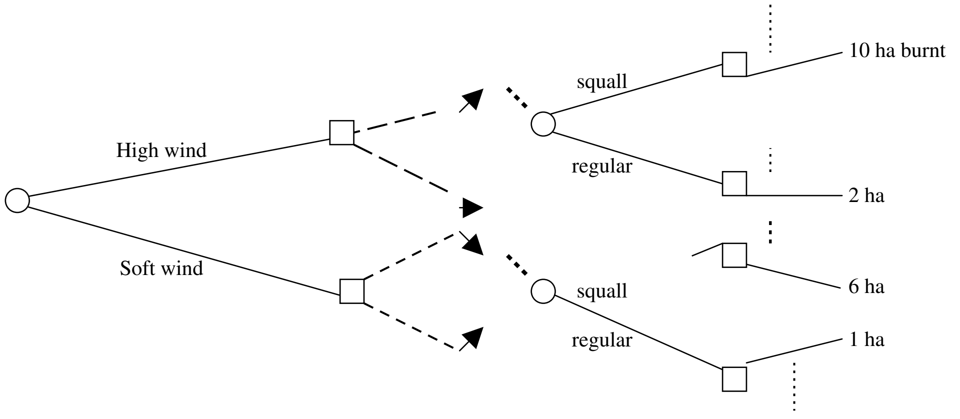


Figure 4.4 Forward induction

time. After each node, branches must be listed, whether from uncertain states or feasible decisions. Ultimately, one arrives at the time horizon and the top of the tree. There, partial decision trees collapse into pure consequences, written in a square or a rectangle, indicating the monetary value of the outcome or a numerical measure, when possible. Growing the tree as time passes, and developing subtrees to detail outcomes, corresponds to a forward induction analysis. See Figure 4.4.

In order to solve the decision problem, once it is conveniently described by a tree, an inverse process will be followed: starting from the tree top and folding back node after node until the foot of the tree and the first node is reached. This is the backward induction process, a common practice in dynamic programming (see Section 4.2 below). Folding back is achieved via two different processes, depending on the node shapes. All branches originating from a node are summarised.

For a square node, the optimal decision at that stage remains and the other, dominated ones are forgotten. This procedure is justified by Bellman's principle, which we shall recall below.

At a circular node, the emerging branches are aggregated into a summary. When consequences are numerical, the summary is a weighted average. Typically, when branch likelihoods are measured by probabilities, their summary will be their consequences' expected value (or consequence expected utilities). When consequences are not yet numerically measured, a summary may be any subjectively defined "middle" one, meaning that to further analyse the problem, the summarised branches should be decomposed again into a subtree.

In any case, such summaries will be used as consequences to choose a decision at the previous (square) node. This continues all the way until the foot of the tree is reached. See Figure 4.5.

Time, i.e. arrival of information, is an essential structural factor. Information is obtained after some uncontrolled, although not necessarily random, event has occurred. For instance, information which is available somewhere but takes time to gather may be taken into account in the decision tree. This is because, what matters about information is that it allows decisions to change: in a dynamic setting, a decision is a strategy of actions, each of them decided on the bases of available information (Part III). Such information arrivals are measured by time, whether it is continuous or discrete time, but in a tree a discrete version is necessarily adopted. It gives the decision problem representation its rhythm.

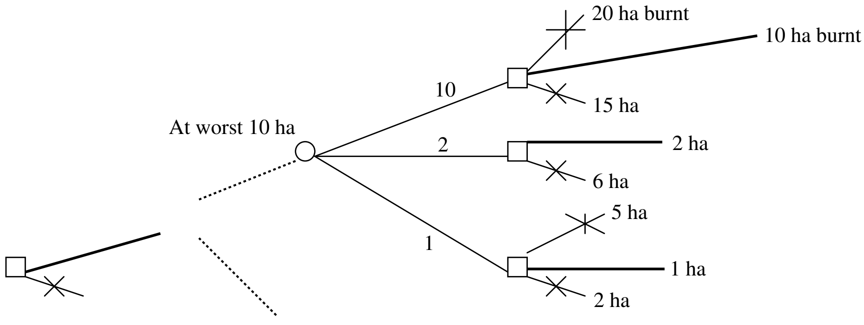


Figure 4.5 Backward induction

Another structuring factor is the nodes' analysis: circles or squares depending on if they represent controlled variable choices (action node, square) or uncontrolled event realisations (chance node, circle). For example, is truck or plane availability in a given spot the result of a decision or should it be considered as random *ex ante*?

After information arrives, an action is chosen, with the list of feasible actions represented by branches sprouting out of the control node. The set of feasible actions is defined by the decision-maker by a list (it may be a continuous set, e.g. quantity measures) and by constraints that exclude some of the set elements. It is often the case that constraints must be scrutinised because they may conceal hidden prior decisions. Indeed, expert advice and experience often induce decisions to be limited or repeated due to habit formation rather than to actual relevance. Most of the time, such habits are founded and may be cautiously respected, but this does not mean that they must be considered as absolute constraints for new decision problems. Another hidden decision occurs when some means are systematically employed simply because they are available, even though they are not the most efficient ones in a given situation (planes in place of fire trucks or vice versa). An easy way to understand that an assumed constraint is in fact a hidden decision, is to consider a monetary shortage (a new constraint) that forces you to reduce the means usually employed (for example give up planes, which are too costly and use more trucks). Other constraints are more difficult to spot as a prior decision, such as infrastructures: access roads, reservoirs, etc.

Conversely, a regulation may exist that enforces a way of doing things, priorities, or the means to be employed. Then it is a real constraint and it does limit the set of possible actions. For instance, in fire struggles a priority is given to threats to people and to their belongings, if this priority were not satisfied, the decision would be at fault and those responsible for it would be answerable to justice.

The other kind of node presents as many difficulties. Circular nodes are often called chance nodes, or random nodes, because this is the most common (and the easiest to deal with) kind of uncontrolled outcome. As we have seen in Chapter 2, the terms random and chance refer to some mechanism producing outcomes with known probabilities. Many variables that are not under the control of the decision-maker are not of that type (reactions of other actors, data that could be known but are too hard to find, etc.), and very few have known probabilities of outcome (we shall come back to the assignment of probabilities later, with other quantification problems). In any case, these nodes are present in a decision tree when the decision-makers want to view alternative scenarios. Figuring out several scenarios is a way to take different experts' views into account. More generally, controversies about future developments often occur in collective choice and will be represented by circle nodes. Branches coming out of a circular node are representations of events that have been observed, or are consistent with past observations of the phenomenon that may generate them.

During the analysis, some square nodes may appear to be better described by circle nodes. A way to deal with constraints that are too numerous or ill-defined is to consider them as unknown altogether. Reciprocally, some circular nodes may be better analysed as square ones: some alternative reactions of other actors or of natural phenomena, which could be represented by a circular node, may be decomposed into a succession of square nodes if such reactions may be driven by intermediary decisions.

The tree construction will alternate phases where complexity is increased with phases where simplifications are done in order to make the instrument handy for decision assistance. This is why the important task is to measure the decisions' sensitivities to possible modifications (added or reduced complexities), once they have been listed. To achieve this task and proceed to backward induction, it is necessary to value each of the nodes, which requires a thorough quantification of consequences and of the branches' relative weights.

4.1.2 Quantification of the Decision Tree

A qualitative approach can be conducted while still using a decision tree as an analytic tool. Then, the tree will be useful to obtain a clear view of feasible choices and a list of relevant consequences. In many cases this will be enough not to be oblivious of some alternatives and to facilitate the rational analysis.

Otherwise, a tree may come in handy as an instrument for making optimal choices but quantification is necessary to use it this way. Assigning meaningful numbers to the different tree (decision problem) elements is a three-level task:

1. Consequences have to be given monetary values (and/or cardinal utility levels, but this makes sense for individual decision-making mainly, and we shall only mention it here when it could make a difference).
2. Alternative circle node branches have to be measured by relative weights (frequencies in the case of random events).
3. Time intervals must be given a relative importance measure (discounting rates). It can be an individual one (preference for early consumption), a collective reference one, or a market one (monetary discount rates).We shall not analyse the time measure quantification in this chapter, given that it is one of the topics of Part III.

As regards consequences, quantification follows two stages. First, quantities have to be known. Then, the quantities will be given monetary values (and/or a utility value). Why monetary values when, in some cases, quantities could be sufficient to make some of the choices? This is because all elements must be measured in the same units so as to be comparable: a wood surface would be hard to compare to human life otherwise, while a statistical value of life is measured in the same unit as the market value of wood (plus eventually some environmental value). As we have seen in the foundations of cost–benefit analysis, money is the unit of measurement for value, as it is defined in reference to an underlying general equilibrium. Furthermore, it can be discounted through time, which would make little sense for quantities.

The first stage (quantities) is relatively straightforward, although it can be lengthy, costly and time-consuming: data-gathering, consulting experts and agents directly concerned by consequences, and estimation of likely quantities do not require economic theory references. Valuation of quantitative consequences at the second stage may be more or less difficult depending on the nature of the commodities. We can distinguish between three categories:

1. Traded commodities (on a market). Market prices reflect their value. This must be corrected, though, depending on the relative market imperfections.
2. Non-traded produced commodities. A common practice values them at their production cost. When produced by the public sector, this method is justified if public monopolies price at their mean production cost.
3. Non-traded, non-produced commodities (e.g. environment, social welfare, health, etc.). We shall evoke the different methods used to approximate these commodity values by market values in Section 4.3.

For the first two categories, a simple economic solution to the valuation problem exists, even though it may be disputed. However, difficulties are encountered in practice. As an example, consider damages to human belongings. One way to value them is to follow insurance companies who evaluate according to reconstruction costs or selling prices. However, damages are not always easy to estimate when insurance experts are not called for (e.g. uninsured natural catastrophes or when a public project's consequences are not covered by insurance companies).

Cost quantification presents several specific problems. One is the differentiation to be made between investments and functioning expenditures. Total cost of a working stock integrates its buying value and the expenditures of the staff. The cost allocation per year requires knowledge of the stock's lifetime, use periods, and choice of a discount factor.

The second problem is much harder to solve: what is the extent of expenditures, and how should they be allocated? These are well-known questions in management, addressing different structural activity cost allocation. No general theory reference exists, so each case requires an adapted allocation rule. In many situations, though, the problem can be avoided by cost–benefit analysis. The method values projects, i.e. modifications of existing production factors. As a consequence, knowledge of total expenditure is not necessary as it is only the additional ones that are required. Obviously, such an approach neglects synergies, i.e. different uses of the same production factor, and thus may induce an undervaluation.

Valuation of the third category of commodities is less direct and depends on the type of commodities in question. Approximation methods are of three types: contingent valuation, hedonist commodities and avoided costs. The contingent valuation method requires elaborate enquiries into population samples to estimate individual values and aggregate them. The hedonist commodities method is founded on the possibility of decomposing some commodities into market valued attributes (e.g. vicinity and view for housing). The avoided cost (or transportation costs) method values a commodity by agents' expenditures or what they say they would accept to spend to use or to enjoy a similar commodity.

Probability assignment to circular node branches is linked to the nature of their consequences and to the quantification problems that have been solved previously. Depending on the kind of available data, we can distinguish four cases.

Some events are caused by a natural (physical or biological) phenomenon that the relevant science describes by a deterministic law, in which case probability is scientifically based on error terms and is known. Unhappily, such situations are rather rare.

In some cases, a database authorises frequency calculations. This is the simplest case as it is the usual way to estimate probabilities of random events. However, frequency is an estimation of probability only if the sample is formed of independent random variables with the same probability distribution. In fact, the sample is generally obtained from time series and it is not easy to verify that the same probability distribution prevails as time passes, nor that variables are independent.

Frequencies are not often available, for instance, because data are lacking or because there are structural changes that cause observed variables to be in obvious violation of the above assumptions. Then specific methods can be employed to calculate probabilities. The most common one is to use relative measures, e.g. ratio between the damage caused by one event and total damages.

When probability calculation is not possible, the last resource is to turn to expert estimation, or to subjective probabilities assessed from observed behaviour (of the decision-maker) in the face of similar risks. In both cases, such estimation is generally controversial and/or, to say the least, not very reliable. Nobody would accept an expert assertion without questions today.

To take controversies into account, one can turn to measures that are less precise than probabilities: belief functions and capacities represent a likelihood order but do not satisfy additivity. They still allow one to calculate means, as we have seen in Chapter 2.

4.1.3 Decision Trees in Practice

Formalisation of a decision problem by a decision tree is merely a way to help communication between actors, because this instrument is meant to confront analysis and experiences. Obviously, the task is simpler when an individual decision problem is at stake than when a collective decision must be reached. However, even in the former case, consulting experts and experienced decision-makers is helpful if not necessary to get a relevant picture of the situation and to achieve quantification. So, let us concentrate on how to organise the construction and the analysis of a decision tree for a collective choice problem: a public building project such as the Millau viaduct for example.

A first schematic approach may be prepared by someone responsible for the case study. The scheme is presented to a group of public decision-makers: Ministry directors, local public decision-makers, engineers, architects, population representatives, etc. Depending on the capacity to organise such a meeting in order to be constructive without having it degenerate into lengthy discussions over details, this group may be limited to a restricted number and then a more general one gathered to verify that the aggregation result is still relevant. At this stage, other representatives such as local actors: entrepreneurs, town mayors, social workers, etc. may be consulted and their remarks and information integrated. At each stage, the prospective/retrospective process is conducted so as to end up with a relevant but not too complex presentation.

In order to go on with quantification, other groups (which may be reduced to individuals) are consulted to obtain numbers, learn which data are available and where to find them. After some homework, this yields a first quantified tree presentation. Some summarising mean values and some obviously dominated decisions have been removed from the tree. This simplified tree is presented again to the same groups as before, one after the other. Some simulations and calculations may be presented as well, and some decision and chance nodes may be removed at each stage. Generally, during the second round process, some alternatives appear, some biased decisions are spotted by opponents, and such controversies are written down and kept for special studies. At the quantification level, notably, some simplifications may be questioned by experts and analysed again. Once the decision tree is thoroughly constructed and reduced, a simple enough decision problem is obtained with which valuation and decision-making can be discussed. To each simplified decision, an aggregated consequence can be assigned and the best consequence indicates the best decision. In order to achieve that practical goal, a third round can be attempted, where decisions (at least most partial ones) are taken. They are not always the same from one group to another, and these differences are informative. A final presentation is ready. It may be presented again to the same groups, but, generally, the first tour is sufficient to reach a decision, or at least to sufficiently inform the group which will then take it on the basis of political considerations. The final decision tree plays two roles: as the technical decision-making instrument and as a useful communication tool for responsible public decision-makers and population representatives.

As far as the technical decision helping instrument is concerned, decisions that give the choice criterion the highest values are chosen. It is an indication obtained from the analysis of the problem and should not be confused with the political and social decision process. The criterion is optimised according to mathematical programming techniques.

4.2 OPTIMISATION UNDER CONSTRAINTS: MATHEMATICAL PROGRAMMING

Here, decisions are called “control” variables or control. The decision criterion is a function assigning a number to each control variable and is assumed to represent the decision-maker’s behaviour. The criterion is obtained from a utility function on consequences and an aggregating rule for these consequences (e.g. mathematical expectation). Utility functions are usually defined through some simple decision problems solved by the decision-maker from which axioms are checked and parameters are estimated. In practice, however, utility functions are often defined in a pragmatic way (making calculations easy), or by opportunistic considerations. In some cases, they are justified by habits, or by regulations and official recommendations.[2] In any case, decision-making means to find decisions (controls) in the set of feasible decisions for which the criterion takes the highest value. The set of feasible decisions is defined by bounds inside a larger space with “good” mathematical properties. In general, decisions are represented by control variables belonging to a topological vector space: the components correspond to decision characteristics. Bounds are called constraints, e.g. budget constraint, and looking for the control which yields the highest value of the criterion is mathematical optimisation under constraints, or mathematical programming.

When the function (criterion) has some “good” properties, such as being differentiable, the optimisation problem can easily be solved by differential calculus and elimination of solutions outside the constraints. Constraints may be integrated into the optimisation problem in some cases. In order to do so, artificial variables are introduced, called Lagrange multipliers. As an example, take the consumer optimisation problem:

$$\max U(x) \text{ s.t. } R \geq px$$

[2] For example, we are aware that a French road state department recommends to limit alternative scenarios to a high assumption and a low one and to take the mean value as the criterion.

It becomes:

$$\max\ [U(x) + \lambda(R - px)]$$

where λ is the Lagrange multiplier corresponding to the budget constraint.

Lagrange multipliers have an economic significance: in the example, λ corresponds to the marginal utility of income. They also have a mathematical meaning: they are control variables of a dual problem, i.e. an optimisation problem where the roles of constraints and controls are reversed (for example, minimise the budget $px - R$ under the constraint $U(x) \geq U^*$ where U^* is the optimal utility level).

In many cases, criteria do not possess properties such that they could be optimised by differential calculus and the problem has no explicit solution. However, some approximation methods and algorithms may be found to obtain "best" solutions. Mathematical programming is a set of techniques, meant to obtain approximate optimal solutions to the initial optimisation problem. Techniques have been invented in parallel with computer evolutions. They are founded on mathematical results and procedures evolve continuously. It is not the aim of this section to present mathematical programming techniques, we only indicate how to present a decision problem in terms of a program.

In a static version of a decision problem, the tree can be reduced to the first square node. Then the corresponding program is of the form, if F is the set of control variables:

$$\max\ V(x)$$
$$\{x \in F / C_1(x) < a_1;\ C_2(x) > a_2;\ C_3(x) = a_3;\ C_4(x) \geq 0 \ldots\}$$

When no explicit solution can be found (through, for example, differential calculus), mathematical programming offers different techniques such that a solution can be approximated by iterative procedures. Baumol (1961) was one of the mathematicians to promote mathematical programming, expressing with humour its method's status with respect to clean mathematics: "... just as the term *ragoût* disguises the fact that it is only stew ...". Nevertheless, he showed how useful such procedures can be. He proposed a criterion to judge whether a method was efficient on the basis of finding a **sufficiently** close to optimal solution, in an **acceptable** calculus time. At the time he wrote this book, a computer occupied a whole room and it was not rare that a program would take a week to be written and the computer would spend a fortnight to achieve the calculations. Today, programs are common software and can be used on personal computers, examples are: Mathematica®, Mathlab® and Mapple®, among others.

A dynamic version of a decision problem requires the solving of several programs, one for each time period where arriving information can change the possible consequences. We mentioned that the method is based on backward induction, i.e. starting from ultimate consequences and climbing down towards the initial decision node. The method is founded on optimal control and Bellman's[3] principle. The principle can be expressed in this way: "any subpath of an optimal path is optimal itself". Following the author more precisely, we find in Bellman and Dreyfus (1963) the following: "An optimal policy has the property that, whatever the initial state and decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision."

[3] Bellman (1957).

Another enlightening definition was given by Aris (1964): "If you don't do the best with what you happen to have got, you'll never do the best you might have done with what you should have had."

In practice, the best solution is looked for at an information node, assuming that the next best one got us there. Within a discrete time model (as is always the case in a tree, although the problem may be solved in a continuous time setting, using differentials with respect to time), the program at an information node at time t is:

$$\max_{\{x_t \in C_t\}} V_t(x_t)$$

where C_t are constraints at time t and:

$$V_t(x_t) = \max_{\{x_{t+1} \in C_{t+1}\}} [V_{t+1}(x_t,\ x_{t+1})] + u_t(x_t)$$

where $u_t(x_t)$ is the (expected utility of) consequence(s) at time t.

When time is economically important, the same formula has to be adapted with discount factors. At any square node of the decision tree, the best decision (i.e. optimal or approximately the closest to optimum given the programming method) is kept. Hence, the decision that is kept has for its consequence (summarised consequences) the one that depends on making the (future) best decisions. It must be clear that all this assumes that the decision-maker knows how to summarise future consequences, even in a remote future. This implies that there is a rule to condition valuation of future random outcomes by past information (past, in the future, but future at the time conditioning is done!). We shall address this difficult problem in Part III, let us only note now that there are no problems as long as probability measures are concerned. Bayesian conditioning is dynamically consistent. Classical dynamic programming has been developed in that setting.

If future consequence summaries are available, then each square node can be assigned a unique consequence and backward induction may proceed to fold back the tree, until a static decision program remains to be solved to yield the best solution to the initial decision problem.

4.3 RISK AND COST–BENEFIT ANALYSIS

A method can always be improved, however it is not possible to propose valuation techniques that are perfectly founded theoretically and hence safe from contestation. As a consequence, cost–benefit analysis cannot be thought of as a substitute to political decision-making. Acceptability[4] is a constraint and cannot be ignored with impunity. For instance, it can drive one to choose a poorly beneficial action over a better-valued one, because the first is more easily sustained by communication means. The role of cost–benefit analysis must be clearly limited to enlighten and justify public decisions.

In order to make this method credible, it must carefully value two main points: non-traded commodities and risk. Taking risks into account depends on the way non-tradable commodities have been valued.

[4] See Kast and Luchini (2002) for a way to integrate a project's political and economic issues into an experimental market game.

4.3.1 Valuation of Non-traded Commodities

There are several methods that may contribute to valuing commodities that are not directly traded. The first two use indirect market values.

The transportation cost method is naturally easy to apply to road construction. How much is the Millau viaduct worth to an individual? An indirect way to calculate this is to consider the time saved by using the viaduct instead of going through downtown Millau. Time is valued by an average salary rate, i.e. a market value. Similarly, the time spent by a student at university can be valued by the additional future expected salary obtained thanks to the diploma. The sum of added values is a way to valuate how much the university is worth, whether for the state to invest in it, or for a student to pay for it.

The second method is founded on Lancaster's theory: commodities are not bought for themselves but for the flow of services they yield. For instance, a statistical study on a set of houses for sale may show a relationship between house attributes and their price differentials and estimate a function (linear, if attributes are distinct) between the attribute "value-to-calculate" and the house prices. Decomposing selling prices according to the different attributes is a way to value them indirectly: view, quietness, town proximity, etc. For example, the value obtained for quietness may be used to valuate the noise impact on the vicinity of a public project such as a freeway or an airport. This is the so-called "hedonic price method".

Such methods are relevant for use values. For instance, road traffic has an influence on the neighbourhood house prices because it changes the use value for their inhabitants. Taking non-use-related values into account is a more difficult problem. Some commodities may have a value that is not related to their use or consumption. Such a value is linked to the commodity's existence: I'm willing to pay for some equatorial butterfly species' existence to be preserved, even though I've never seen a specimen and most likely never will. Obviously, neither the hedonic price nor the transportation costs methods can be used for valuing such commodities.

Conversely, the Contingent Valuation Method (CVM) is meant to value such tricky cases by a direct determination of commodity values through enquiries on a sample population. The first enquiry into studying agent preferences and values directly is due to Davis (1963). However, the CVM only developed during the 1980s. When no direct behavioural observation of existing markets is possible, CVM asks individuals directly what they would do if there was a market for some commodity. In practice, demands are made to try and learn how much agents are willing to pay for a modification of a commodity quantity (or how much they would require to accept it, if the modification is harmful). Given that the modification has not been observed, agents are placed in a hypothetical framework (a scenario), and their answers are intentional only (nothing proves that agents will trade the commodity at the value they said if a market was organised).

A scenario presents a hypothetical market for a commodity to be valued. The hypothetical market must be as credible as possible. The agent must face a situation similar to the ones encountered in real trades in order to deal the commodity effectively. This requires that the commodity is precisely described: quantity, quality level, production conditions, proposed measures to improve its quality or increase the quantity, and how the process is financed. The population sample is chosen in accordance with the type of value concerned (for a use value, take consumers of the commodity). Agents must be reminded that they have a budget constraint, otherwise they may reveal values that are inconsistent with what they could pay, and they must be aware that substitute commodities are available, as well.

Tversky and Kahneman (1981) have taught us that individual answers may be modified by the scenario's structure, or more generally because of the questionnaire's formulation itself. This problem is known under the name of "structural effect". Implementation is a delicate matter as revealed amounts depend on the scenario formulation. In order to write a scenario, an interdisciplinary collaboration is necessary: psychology, sociology and anthropology are required to help economists formulate the questionnaire. The scenario's four most important qualities are:

- Correspondence with the proposed population.
- Understandable by the interviewees (and also by the interviewers!).
- Respect of neutrality so as not to influence answers.
- Coherence with the interviewed agents' customs, ethical values, criterions, etc.

In order to satisfy these qualities, a first sequence of semi-direct interviews tests a sketch of a questionnaire. For instance, the way agents are used to paying for a commodity or to enjoying it without paying must be respected. Most of the time, people are not ready to pay for something they used to consider as free. Air, for example, is a public commodity and is not paid for directly. Indirectly it is, as pollution is costly to prevent in town or in the vicinity[5] of a factory.

After the scenario has been presented by interviewers, valuation is undertaken. It can be approached by different processes:

- Repeated increasing or decreasing auctions.
- Open or closed questions.
- Payment vehicle.
- Two successive offer mechanisms.

Answers about valuation are completed by other questions on individual characteristics: socio-economic, life customs, commodities' uses, social perceptions and representations, etc. They will serve as explanatory variables for the willingness to pay in the econometric model. Other questions are meant to control that the scenario was well understood and the payment means were readily accepted. Then, statistical methods (econometrics) are called for to integrate answers into global (average of some kind) amounts corresponding to agent characteristics.

In a risky universe, it is necessary to present several scenarios so as to simulate possible future states of the economy. Many enquiries are then necessary and, furthermore, global valuation requires the ability to weight the relative importance given to the different scenarios.

4.3.2 Risk Weighting

Several methods can be followed to obtain risk weights. In the following chapters, we shall study how to approach the risks to be valued by comparing them to risks that are priced on financial markets.[6] Here, let us concentrate on an alternative approach based on experimental

[5] Chanel *et al.* (2004) propose that interviewees move from a town identical to the one they live in but with a much better pollution level. Their willingness to pay is measured by the moving costs they are proposed to choose between.

[6] An objective measure can be calculated for the states corresponding to marketed risky assets under some assumption, as we shall see in Chapter 7. An example is given in an experimental procedure, below.

market games, which is consistent with the CVM because individual agents are directly implicated in the process.

We are looking for a way to weight the different scenarios chosen to represent future uncertainty about a project's impacts, so as to aggregate the willingness to pay obtained from CVM (or from other methods if more suitable). Financial market theory teaches us that asset prices give a relative weight to risks (represented by assets). These weights define the risk-adjusted measure: a social aggregation of agents' preferences revealed by demand and supply functions. Market games try and reproduce financial market theoretical characteristics: prices inform about information asymmetries between agents that can buy or sell assets for money (no producers–suppliers nor consumers–buyers). This branch of experimental economics has been developed along three axes:[7]

- Information transmission through prices from informed to uninformed agents.
- Different information aggregation.
- Endogenous treatment of information in simultaneous equilibriums in markets (e.g. bubbles formation).

Price formation processes are diverse and depend on financial market organisations. Three types of procedures can be considered, at least:

- Market-makers' mechanisms.
- Order book procedures.
- Auction processes.

Market makers (dealers) propose bid and ask prices to potential buyers and sellers, they act as intermediaries between the two parties to a trade (they provide clearing facilities). They face an inventory risk: if prices vary in contradiction to their forecasts, they may face a difference when trades are cleared at the bid and ask prices they proposed. That is, market makers have to hold an open position while they find a party to the demand/supply to deal with. The risk is compensated by the spread between bid and ask prices, a transaction cost that customers pay. Such a procedure is not appropriate for our purpose given that market makers are typically exogenous to the trading organisational mechanism.

Order books gather agent orders: supply or demand, proposed prices (maximum buying price or minimum selling price) and maximum quantity to trade. The market price will be the one that maximises the number of transactions. It maximises the minimum between cumulated selling positions at an accepted price lower or equal to the market price and cumulated buying positions at an accepted price higher or equal to the market price. Because the market price is figured out in one stage only, it cannot inform traders.

There are four main types of auctions.

1. English auctions: bidders (to simplify we talk of a unique commodity) call successively higher and higher prices until offers are fulfilled. An auction cannot be withdrawn. The agent with the best offer buys the commodity at his last price, if it is more than the reservation price of the seller. Dutch auctions propose prices decreasing from an initial price (too high on purpose) until an offer is called.

[7] See Sunder (1995).

2. First price auctions belong to the class of sealed auctions. At the opening, (sealed) price proposals for the commodity are consulted and the commodity is sold at the highest one. Second price auctions are the same, but the commodity is sold at the last before highest proposed price to the agent that did the higher bid.
3. Walrasian tatonnement needs buyers and sellers to meet together in front of a (Walrasian) auctioneer. The auctioneer proposes a first price and agents write down their demands or their supplies at that price on bills. The auctioneer collects the bills and figures out the net global demand. If the result is zero, trades are compatible and the market is cleared. If it is positive, bills are destroyed and the auction starts again at a higher price. If negative, the same happens with an inferior starting price.
4. In a double oral auction traders are in the same room and yell their supplies and demands (or prices if trades are done by units of commodity). Agents can accept the best (highest) bid price or best (lowest) ask price, or make an ask or a bid themselves. A new bid must be higher than the last one but not above previously accepted ones and a new ask price must be lower than the last one but not under previously accepted ones. The procedure goes on until there are no more orders left.

The double oral auction has interesting properties which make it the ideal candidate for experimental economics. It preserves symmetry on both sides of the market, contrary to the first two bidding procedures. Training plays an important role, but the principal advantage of the procedure is that it is very efficient at reaching equilibrium and doing it in a reduced number of iterations. Thanks to this mechanism and if a strict procedure is followed, it is possible to value public commodities with satisfactory information diffusion and aggregation (Plott and Sunder, 1988).

Let us return to the Millau viaduct case and suppose that we wish to measure an impact of its construction. Starting from previous investigations, in particular those made in the public utility procedure,[8] various scenarios can be built. They formalise sometimes opposite visions of the project consequences. Let us suppose that we retained N various scenarios, which correspond to states of the world $s_1, \dots, s_N$, modelling uncertainty. Economic surveys can make it possible to directly quantify monetary impacts and non-traded commodities are likely to be valued by the CVM, and this must be done for each scenario. This first phase of work yields a total monetary valuation for each scenario, i.e. a net profit or loss (benefits minus costs), signs may differ according to expectations. Let us denote such values as $Y_1, \dots, Y_N$. Now, let us define assets: $y_1, \dots, y_N$, contingent on states $s_1, \dots, s_N$, so that y_i, for $i = 1, \dots, N$, pays €1000 if s_i occurs and 0 if not. The assets are proposed to experimental subjects who are informed that they can trade them in a limited time period and that information on each asset value (the price at which it is traded) will be revealed after each trading round. Each scenario (state) is described in such a way that agents are conscious that it is indeed its realisation which will command the payoff of the asset contingent on it. This requires that the whole context of the scenarios is sufficiently familiar to the subjects. Furthermore, to simplify with respect to real financial markets, payoffs are immediate so that no discount rates will complicate individual valuations. Each agent is endowed with an asset portfolio, and with some amount of money with which to commence trading.

[8] In principle, such an investigation by open enquiry is required by law in France for any major public project. One has been done for the Millau viaduct project, where people could come to town council houses in the area and write down their observations in a book. These books have been summarised to inform the Ministry of Transport.

An experimental procedure of a double oral auction is set up. After each trading phase, the price is revealed. Subjects can figure out the balance between portfolio and money, and their gains and losses at market prices. At the end of the experiment, accounts are closed. Asset prices are known: $q_1, \ldots, q_N$.

If the viaduct project is to be valued, and scenarios are those represented by the states $s_i, i = 1, \ldots, N$, it may be represented as an asset (non-traded): y. Its payoff in state s_i is Y_i, obtained through the CVM, for instance. Asset y can be replicated by a portfolio, θ, of traded assets with weights: $\theta_i = \frac{Y_i}{1000}$ for asset $y_i, i = 1, \ldots, N$.

As we shall see in Chapter 7, at the competitive financial market equilibrium, the value of non-traded asset y, call it q, must be equal to the formation cost of the replicating portfolio, θ (otherwise arbitrage would be possible if asset y was marketed). Here, traded asset prices yield the asset and then the project valuation: $q = \sum_{i=1}^{N} \theta_i q_i$.

Note that this valuation did use a weighting for risks (payoffs contingent on scenarios) as we can see if we write the project valuation in the form: $q = \sum_{i=1}^{N} \frac{Y_i}{1000} q_i = \sum_{1}^{N} Y_i \frac{q_i}{1000}$, where payoff Y_i is weighted by $q_i/1000$. We shall see in Chapter 7 that such a weighting is a measure of uncertain states that has nothing to do with an objective probability measure nor with subjective ones. It is merely a collective measure of uncertainty aggregating information determined by equilibrium prices from the agent behaviours.

Thanks to this method, a value can be assigned to a non-traded risk. Weightings obtained from prices reflect agent beliefs, expectations and risk attitudes about uncertain states, on which traded assets are contingent. They result from an experimental aggregating procedure mimicking market pricing. In contrast with existing financial markets, prices (and then, weights) do not depend on agents' initial wealth, because, in the game, all endowments are the same. This is closer to a democratic philosophy: one citizen, one vote. As we shall see in Chapter 7, another advantage of this method is that the measure of uncertainty is obtained in a non-probabilistic context, thus avoiding controversies.

CONCLUDING COMMENTS ON PART I

In a public project study such as the Millau viaduct, many impacts are not valued and some are utterly ignored. As a result, many risks that will have to be managed collectively are not foreseen, not insured and not hedged by some financial strategy. The private contractor who constructed the viaduct did not have as many impacts to take into account as at the collective project phase. However, there were still many and they were studied, valued and taken into account in the construction process and within the financial strategy of the company. The difference of attitudes in front of risks between the individual contractor and the public ministry may be due to the state being its own insurer and risks being managed *ex post*, after unfortunate events occurred, while an individual is responsible to be insured for, and/or finance, casualties. But most often, impacts are not valued and risks are not formalised because valuation techniques would involve too much work or are merely ignored: this induces a risk! Another risk is introduced into a study when a method is employed outside its validity domain.

We have presented in this part the theories underlying valuation of impacts and of risks and some of the instruments they sustain. We want to insist here on the limitations that outline their range of applications.

Concerning individual valuation, if the theoretical criterion obtains the same decisions as the decision-maker's in observed (or observable) situations, it covers three dimensions of a decision-making process:

- The first dimension refers to the assignment of relative importance to consequences. The utility function does this through its cardinal property.
- The second dimension is about the relative importance of the "states of the world", i.e. elements describing uncertainty. Formally, this is achieved by the capacity measure (or subjective probability as a special case).
- The third dimension is relevant in the special case where uncertainty is generated by a random device (or phenomenon) with a known probability distribution. Again, it defines a relative importance of outcomes by a measure that may be a deformation of the known probability distribution. If the deformation function satisfies some properties (obtained by more conditions on the decision-maker's preferences), the measure is a capacity again. Notice that both capacities, the first one on the states of the world and the second one on lotteries outcomes, are rarely present in the same decision problem, however both represent the decision-maker's subjective perception of the uncertainty levels relevant to the decision problem.

The three dimensions are treated jointly by the criterion: expected utility, for instance. It would be totally out of place to transpose a utility function on consequences from one decision problem to another. A criterion is not a characterisation of human behaviour, but a definition of an individual decision-maker's attitude when facing a particular uncertain situation. Similarly, the subjective capacity measure, or the deformation of an objective probability, reflects the decision-maker's judgement of the relative importance of uncertain events in a given decision problem. If consequences are modified, the measure of events will be modified as well. These limitations are often forgotten in applied decision theory. For instance, it is commonly done to choose a particular form of utility function for practical reasons and then to assign parameters to characterise the decision-maker's attitude (e.g. risk aversion). Doing so may have some normative appeal and help to analyse a problem, but it cannot be justified by individual decision theory. The last remark excludes any possibility to justify the assignment of a "representative" criterion to a given category of decision-makers. In any case, such a representation would open more problems than it can solve, as we have shown in Chapter 3 with Arrow's impossibility theorem: the main limitation of individual decision theory is that it cannot be applied directly to collective choice!

On the other hand, Bergson–Samuelson's social utility functions are not subjected to Arrow's impossibility result, provided that one is willing to accept a certain form of comparability of preferences. It should however be recalled that the addition of comparability conditions is not expressed in terms of axioms on preferences, but affects functions in an ad hoc way. More deeply, the very concept of utility differs from the standard individual approach. In the standard theory of public choice, utility is an intermediate tool that need not be significant at the individual level. In order to assert the absolute meaning of the concept, conditions bearing directly upon utility and satisfied by every agent must be determined. There are many criticisms of such a concept, be it from a theoretical point of view, an experimental one or by simple introspection. One can also fear that it may not be possible, in practice, to compare characteristics that are unobservable by definition and are doubtfully perceived by individuals. Only a pirouette makes it possible to leave this dead end. It consists of seeing the formalisation of social welfare functions as a simple representation of the observed states' choices. Normative aspects require that we take precisely into account the conditions under which they can be enforced.

Instead of mimicking individual decision theory, it may be advisable to value consequences and risks by market prices. Existence of perfect markets is a major question for the implementation of decentralisation of efficient allocations by competitive trades. Shall we look for perfect competition in an economy where increasing returns to scale prevail? On the one hand, we can have consumers buying bigger quantities at lower prices if competitiveness is stimulated. On the other hand, small sized firms will be less efficient with higher production costs and will ask for higher prices. Arbitrage between these two positions has not been done, either theoretically or empirically.

Another remark is a more pragmatic one. Current economic evolution does not seem to be directed towards increased competitiveness. The worldwide market expansion phenomenon increases competitiveness in general, and the number of competing firms in each country. However, mergers and acquisitions of companies go in the other direction. Political and international competition regulators seem to be very lenient towards the regroupings of large companies.

The general equilibrium model under uncertainty has one more assumption: the complete market hypothesis. Any risk can be perfectly hedged (or insured). Obviously, a complete system of contingent commodities cannot be implemented in the real world for many reasons. Among the most serious ones, note that organisation of contract trading places would introduce many costs, and in turn transaction costs contradict the perfect market assumption. Another reason is that risks are defined by the set of states on which agents must agree if they want to write contracts contingent on them. This is unrealistic, even in an approximate way. However, these objections can be relatively well answered by another more simple, but equivalent, model that we shall see in Chapter 7.[1] Another objection is that, if all risks were insured, it would induce agents to generate new risks by investing less in prevention

[1] Arrow's model of commodity spot markets and the financial asset market: Perfect Foresight Equilibrium.

(moral hazard, see Chapter 6). From the normative point of view, difficulties are related to the distance between the real world and the very general features of the model. For instance, what meaning shall be given to "initial" endowments? That is, when does a government modify them to obey the second welfare theorem? A suggestion could be that government taxes bequests only, so as to modify initial endowments of younger agents, for example. In fact, as the model is static, time plays only a small role and so does anything related to time, such as bequests. For practical purposes a dynamic model is necessary,[2] at least as a reference, to analyse most normative applications. Then, in order to implement the second welfare theorem, we are forced to look for a way to tax and subsidise endowments (or income) so that resources are allocated properly, but prices remain the same. But this is difficult because, for instance, bequests are taxed during the heirs' lifetime, thus modifying the arbitrage they do between saving and consumption, and consequently changing prices. The general rule is that modifications of endowments should avoid distorting prices, and many works have focused on tax/subsidy systems which achieve this aim.

To the theoretical limitations we have mentioned, some more must be added at the instrumental level. A decision tree is a model and as such it cannot be but a simplified version of real situations. This method of analysis belongs to the cause–effect paradigm and, as such, it looks for causes of the phenomenon involved in the decision problem. The tree indicates its relation to its effect on consequences. Two main dangers may be encountered during this process. The most common mistake is to forget some essential causes (e.g. neglecting to describe available fuel for a fire). Such a mistake flaws the whole process, and is usually avoided by consulting experts on the matter. The second danger is more subtle, and as such more difficult to prevent: choosing a wrong cause (mistaking a malevolent incendiary fire for a natural fire setting). Mistakes of this kind are usually corrected during the tree's elaboration process.

Once all causes have been listed, a selection must be made among them, assuming nothing relevant has been omitted. Too many causes or explanatory variables would inefficiently increase the tree size and the analysis complexity. In any model construction, such a way of processing is necessary to clearly assess the limitations of the results derived from the model, so as to be able to enlarge them for application purposes.

Cost–benefit analysis has important limitations that we have underlined in Chapter 3. Furthermore, it must be used with caution because of the many difficulties encountered in practice. Let us consider three typical ones:

1. It is obviously impossible to take all impacts into consideration when doing calculations. In the Millau viaduct case, studies had been conducted only on tourism, on some regional economic activities and on the environment. In none of these domains was it possible to consider all impacts of the viaduct, or even think of doing so.
2. Many impact measures are not precise enough. Non-tradable commodities have no directly observable price, by definition. Alternative methods, that we have presented, can only give approximate values, at best. Possible environmental damages due to traffic and the building of the viaduct have been taken into consideration, but none of them have really been valued. However, precautionary devices to protect nature have been constructed at a cost that could not be compared to expected benefits.

[2] It will be introduced in Chapter 11.

3. Controversies arise, in a public project in particular, but in many others as well. Notably, environmental issues are very sensitive because they are not well understood and are difficult to measure precisely. But economic consequences are thoroughly discussed as well, even though they can be precisely valued. The viaduct's impacts on regional tourism have given rise to many controversies. Expectations were based on traffic variations. Official projections are constructed from a non-random model. However, many *ex post* traffic observations have shown results to be very different from the model's *ex ante* predictions. Even though some uncertainty was introduced into the model's result to take its imprecision into account, this is still a source of controversy. But there was more: why is traffic a relevant variable to predict tourist flows? It is obviously not because people drive on a freeway that they stop and spend their money in the surroundings. Most of the traffic on this motorway is aiming at the Mediterranean coast, in a hurry to get there for the paid-in-advance vacation rental accommodation, and the very reason for the viaduct was that going through the valley and its towns was creating too much delay. Many of these types of controversies are observed for most project impacts, they could be lessened if valuations of impacts were done thoroughly, e.g. using the CVM that relies on inquiries.

The CVM has some bias (difference between theoretical and observed values), but this may be avoided if some precepts are followed. Precepts have been elaborated along with experiments and theoretical analyses to deal with observed biases or inconsistencies, during the last 20 years, however some biases are still a problem. These are:

- Strategic bias – agents think their answers will have an influence on public decision and try to manipulate it by sending misleading signals.
- Hypothetic bias – individuals are not able to imagine themselves in the situation described in the scenario. The difficulty may come from the agent being unfamiliar with the question, or because of a lack of information, sometimes because the scenario presents too much uncertainty, e.g. about the commodity level offered. Given the scenario is a hypothetical one, there may be an exaggerated interest in the commodity if it is perceived positively because there is no real spending to get it.
- Inclusion bias – a faulty perception of the commodity leads the agent to give a global value instead of the one it would pay for the commodity alone (e.g. values the environment instead of the forest described in the scenario).

The user of methods derived from the theories must be aware of the limitations we have put forward and not get confused between an individual decision problem and a public one. In many circumstances, the theoretical frameworks presented in this part will not be sufficient to distinguish all the situations, these are the scope of the next two parts of this book.

Part II

Risk vs Uncertainty

INTRODUCTION TO PART II

There would be no risks if there were no uncertainties about the future, but there may be uncertainties without any risk being taken. Hence, there is something more in the concept of uncertainty than in the concept of risk. This may be why Knight (1921) enforced the use of the word "risk" for some situations and the word "uncertainty" for others, in economics at least. The Knightian distinction is relevant for insurance and finance businesses, which separated because each of them concentrated on some kind of risks (in the original general sense). Examples of insurable and uninsurable risks relate respectively to situations of risk and to situations of uncertainty. However, insurance and finance industries that had split apart for many centuries after the concept of risk appeared, have merged again during the 20th century. We shall understand why through the examples in Chapter 5. Moreover, the economics of insurance is founded on one theory, and finance on another one; they are the topics of Chapter 6 and Chapter 7, respectively. The instruments presented in Chapter 8 will show how new management methods rely both on the insurance industry methods and on the financial markets' ones.

The distinction on which this part is based can be seen in most risk situations, depending on the point of view of the decision-maker and/or the risk to be managed. Let's get the intuition through two different types of risks: financial (risky) assets and health risks.

- Take a newly issued security. Obviously, future returns are uncertain and one will hesitate to invest in it. Before doing that, an investor would take information from experts, or make its own enquiry, etc. Nevertheless, assigning a probability distribution to possible returns and deciding on the amount to invest based on it would be hazardous. In practice, even if one could describe the investor's behaviour by an expected utility maximisation, say, it would be a subjective and not very reliable one, at best. This clearly places security returns in situations of uncertainty.
- However assume that, some years later, the company did so well that its securities are commonly traded on a financial market. The security's returns are listed since it is traded, this forms a well-defined statistical sample. From this sample, estimation can be done and the probability distribution of *past* returns be well assigned. Obviously, this does not say much about future returns, even when all statistical assumptions are satisfied (e.g. i.i.d. returns, robust estimation, stationary process, etc.). Future returns are still uncertain, they depend on too many factors such as the internal policy of the firm, a huge workers' strike, etc. that cannot be probabilised. However, it is possible to formalise the risk on returns by the estimated probability distribution and therefore to study security market prices under a probabilised formalisation: a situation of risk.
- Heath is a collective as well as an individual problem. At the collective level, biology and medical science can give a description of diseases, most of the time. When a disease has been observed over long periods of time, data are collected and form a statistic that yields estimates for different possible states of sickness. We then have probabilities that are valid within a given population, and we can say that someone with given characteristics will die from the disease with a 0.1% probability, say.[1] Based on this, collective decisions can be taken that are founded on cost–benefit analysis (integrating priorities defined by

[1] Note that characteristics are descriptive explanatory variables statistically based and are not "causes" by any means! It may happen, though, that scientific knowledge suggests a characteristic, but it must prove to be statistically reliable all the same.

law, political choices and ethics): for example, vaccination and other preventive actions. National as well as private health insurance rely on such probabilities. They do make sense and yield risk premiums or cost–benefit manageable systems. Hence, health risk belongs to situations of risk.

- Now take the individual point of view. Most people do not know enough biology to understand the meanings of scientific models. But even if they do, or if they trust medical doctors, this only tells them to what risk class (as defined at the collective level) they belong. This does not say much about individual risk. Take vaccination, a usual preventive measure for epidemics. It is not a risk-safe measure, a portion of the population will be much sicker than it would have been without it and some people may die with a known probability. As an individual, one may fear this risk more than the disease. The population-based probabilities and risk profiles are not always relevant for an individual decision problem. For many individuals, health risk refers to situations of uncertainty.

Nevertheless, there are means to take into account subjective uncertainty perceptions about health. Insurance economics has been focusing recently on new decision models (RDEU, Chapter 2) that refer to subjective deformations of probabilities, a simple way to correct forecasts that turn out to be grossly wrong when it is assumed that individuals solve their individual health problems in the same way as a collective decision-maker does. The same arguments as the one for health risk can be opposed to individual decisions about buying insurance against a given hazard, or not. Let us look at other difficulties which may get us confused as to which one of risk or uncertainty situations the insurance business belongs. As in the examples we shall see in Chapter 5, the insurer refers to statistics to define risk characteristics and then calculate risk premiums. As a consequence, the insurance business developed at the same rate as statistical methods improved: it classifies the insurance business within situations of risk.

Insurance demand also concerns natural catastrophe risk, for example. In most cases, observations do not satisfy the statistical requirements to estimate probabilities in a reliable way, and even less the law of large numbers assumptions. On the one hand, catastrophe risk perceptions are very subjective and ill-informed. As a consequence, the insurance portfolio is risky, even in the case where catastrophe risk is probabilised. It must be shared with other companies (reinsurance) and/or diversified away by financial assets (financial hedging). This puts an important part of the insurance business in situations of uncertainty.

Chapter 6 concentrates on the foundations of the economics of insurance. Traditional insurance deals with risks having a known probability distribution. In this case, a risk is represented by a random variable and probability theory applies. The basic principles of insurance and the theorems they rely on are presented. The demand for insurance has been studied at length under special assumptions that are revisited. Expected utility and the classical risk-aversion concepts are confronted with situations where such assumptions may be violated. Recent results that complete and extend the traditional ones are presented. All the models in this chapter are static ones, consistent with the way insurees manage their risks: something is bound to happen within a given time period (a year, in general). An insurance contract covers this period, there is nothing to manage any more and then the insuree's situation is static. This is not true of the insurer's situation: The premiums are collected and managed in a dynamic way over the year, and, more generally from year to year. This distinction between static and dynamic models is the topic of the third part of the book.

Chapter 7 presents a theory of financial markets on which the management and valuation of uninsurable risks can be based. In this chapter, risks have their original meaning of assets with unknown payoffs. Mathematically, they are represented by measurable functions and no probability distributions need to be known on the state space. It is shown that markets provide a way to measure such risks. The measure induced by prices happens to have the same mathematical properties as a probability, however it has little to do with observed frequencies or with subjective probabilities. This “risk-adjusted” measure is defined by prices under a market equilibrium model that is detailed. Other models are referred to as well, they rely on more restrictive assumptions. All the models in this chapter are static, their dynamic extensions will be presented in Chapter 11.

Chapter 8 is split into two sections. The first one presents instruments related to the economics of insurable risks, they are based on the results presented in Chapter 6. The second section addresses the problems of valuing and/or hedging uninsurable risks through financial markets instruments. Management methods are proposed for catastrophes, social risks and controversial risks.

5
Insurable and Uninsurable Risks

Insurance is one of the oldest systems of risk management. It is also one that has known the most significant developments since it started in the middle ages, flourishing as from the 18th century and still increasing its range today. In modern societies, the extension of the insurance system goes along three paths. At a purely quantitative level, we observe an increase in the number of agents and commodities that are insured, this number is multiplied by an increase in the value amounts in the contracts. A second path involves a multiplication of the risk types covered by the insurance industry. The third path has been more recently traced out. It challenges insurers to go over the other paths' limits thanks to new hedging financial instruments for insurance portfolios. It also suggests ways to address risks that are still beyond the available insurance methods to manage, to cover, and/or to value.

The present chapter aims to stress the distinctions to be made between insurable risks and uninsurable risks (i.e. risks that cannot be insured using traditional methods). The first category corresponds to situations of risk, the second one to situations of uncertainty in the sense of Knight. Among the uninsurable risks, we distinguish those that require financial market instruments and new risks problems for which methods to apply are still questioned.

Insurance is defined by a contract between two parties: the insurer and the insuree (insured agent). The insuree pays a fixed "premium" to the insurer to get rid of (a part of) its risk. The insurer gets the cash payment and writes a contract, contingent on an expiration date and some random events, asserting that it will pay off the value of casualties defined by the contract, if the event occurs.

Insurance contracts differ, depending on what is being insured. Consumption goods can be insured: automobiles, houses, professional tools, industrial plants, etc. People can insure themselves: life insurance,[1] health insurance, insurance against injury, etc. Insurers can be private or public companies. States often act as an insurer (insurer of last resort), and corporate unions and mutual societies can manage insurance contracts for their participants, the latter taking different legal forms depending on the country.

Mutual insurance was first developed in the middle ages among craftsmen who organised themselves to mutually hedge their professional risks, mainly in Great Britain (Friendly Societies) but also in other countries (corporative mutual aid systems). This system is still very active in the European Union and benefits from particular tax and other legal advantages, but it is constrained not to generate profits, in general.

Social insurance is a particular case, concerning mainly health, work hazards and unemployment. Any institution may manage contracts, but they are generally linked to a state

[1] Insurance on one's life is different from "life insurance". The latter is not a proper insurance contract but a saving one, i.e. a financial asset, which can be linked (not necessarily) to an insurance on life. Insurance companies traditionally manage life insurance (we'll come back to it later).

organisation that sets general rules and acts as a super-insurer. All of them are based on solidarity between the insured agents and premiums do not depend on individual risks.[2] They have a distributive effect (equalisation) and reimbursements are often contractual.

This was just to give some general ideas about insurance; otherwise the domain is so vast that one book would not be enough. Let us now turn to the object of this part of the book: the distinction between randomised (probabilised) uncertainty and unmeasured uncertainty. The difference between the two kinds of situation is apparent within insurance contracts and insurance companies' attitudes in the face of risks.

Risk, as we mentioned in the Introduction, is a concept forged to grasp some particular situations. However, its use expanded[3] because the type of problem it addresses is present in any individual's, firm's or collective decision-making. The sense of the word extended as well to: hazards, damages, prudence standards, returns variability, etc. as well as to uncertainty measurements (variance, standard error, Value at Risk, range, etc.). The economic meaning of risk: *variability of an investment's future returns (losses and/or gains)*, goes well beyond the domain of insurance against losses. Indeed, in most cases, one is willing to take a risk because there is something to gain, at least expectedly. However, insurance contracts only address one-sided risks: a potential loss (damage) that the insurer may partially compensate. But why and for which kind of risks may an insurer be willing to take a risk that other individual agents don't want, or can't afford, to take?

5.1 INSURANCE OF RISKS WITH A KNOWN PROBABILITY DISTRIBUTION

The historical approach to risks, before the insurance industry developed, was founded on risk sharing through mutual aid. This form of insurance still exists. It addresses all risks, "probabilised" ones as well as some others for which no probability distribution is well known. It relies, in principle, on the possible diversification and equal repartition of the pooled risks among individuals. Let us call this the "mutuality principle". It has an economic foundation: the mutuality theorem that we shall see in Chapter 7.

In the case of risks with an ill-known probability distribution, diversification makes no sense and mutual insurance companies manage the risks by pooling and repartition (*ex post*) of extra costs. In the case of risks with a known probability distribution, diversification results from the total variance being less than the sum of the individual variances, if all risks are not perfectly correlated.

The diversification property is often confused with what drives another principle of insurance, on which the traditional insurance industry relies. The basic principle of the insurance industry is also founded on risk pooling and risk sharing, but it works differently and only makes sense with "probabilised" uncertainty. Indeed, it is founded on a theorem from probability theory: the law of large numbers. The theorem[4] requires independent and

---

[2] A worker with a high risk pays the same premium as a worker with a lower risk, for instance. The same applies for basic heath insurance, or unemployment.

[3] And still does!

[4] The first one is due to Poisson, in the 18th century. There are other ones that we will evoke in the next chapter, they slightly relax assumptions on identical and independent random variables.

identically distributed (i.i.d.) random variables. Let us refer to this as: the (independent risks) "large numbers principle".[5]

In most of this section, a risk can be represented by a random variable with known probability distribution. This assumes that such risks have already been observed in the past and that observations satisfy the properties required to form a sample of i.i.d. random variables. Then, the distribution parameters can be estimated: mean and variance (when they exist) are the main ones. The independent large number principle is as follows:

Given a population of N such risks (to be insured), if the risks are pooled and then shared equally among the N insured agents, each allocated risk has the same mean, but a variance divided by N.

We already evoked this principle in Chapter 2 through an example: here is another one. Assume two individual agents, i and j, have a 1 in 10 chance of having an accident but that an accident for i is independent of an accident's occurrence for j. Each of them faces a €1000 loss. The two risks can be characterised by Table 5.1. Assume the individuals decide to share their risks. This means they sign a contract that stipulates they will share losses equally if an accident occurs. All the different situations they may face are listed in Table 5.2. The shared expected loss is still the same as the one they faced privately, but the variance is halved. As a consequence, the contract yields each of them a "lower risk". We question the notion of "lower" risk in Chapter 6, here it can be justified by two remarks: the variance is half what it was for each individual and the maximum loss (€1000) probability has been divided by 10. If more identical and independent risks are added to the contract, variance (and maximum loss probability) become less and less, converging towards zero ("no risk"!) when the number increases without limit (law of large numbers, see next chapter). Then, what is the need for an insurer? Well, there is still a loss to face: the sum of losses, even if

Table 5.1 Independent risks

Situations	Losses (€)	Probabilities
i has no accident	0	0.9
i has an accident	1 000	0.1
Expected loss	100	
Loss variance	90 000	

Table 5.2 Risk sharing

Situations	Losses (€) for each agent	Probabilities
i and j have no accidents	0	0.81
i has an accident but not j	500	0.09
j has an accident but not i	500	0.09
i and j have an accident	1 000	0.01
Expected loss	100	
Loss variance	45 000	

[5] Both principle appellations are ours; they are often mixed up in the economic literature. Our attention was brought to this distinction by several remarks from our colleague, Professor Marie-Cécile Fagart. We thank her for these.

it becomes certain (the expected loss) at the limit (but then it is not a risk any more). This is why an insurer comes into the picture.

The insurer sells contracts promising to reimburse losses at a price (a premium, because it is paid before the loss has occurred). The sum of the premiums must be such that it covers the (expected) sum of losses. In the previous example, a fair premium is calculated as the expected loss (€100). Then, all the insured agents face no more risks (any losses will be reimbursed) at the cost of €100 (they sold their negative asset, or bought a hedging asset), and the insurer faces a risk (it bought the agents' risks, or sold the hedging contracts). Indeed, even under the law of large numbers, there is a risk (a non-zero probability) that the total reimbursement be much higher than the expected one.[6] But the insurer's risk is low (low probability), manageable (the sum of the premiums can be invested) and converges towards a riskless position as the number of independent identical insured agents increases. Besides, a slight increment in the premium will still be acceptable with respect to the insured risks, and enough to compensate the insurer's risk[7] and cover its management costs.

The previous principle helps us to understand what is meant by "the probability of accident is known". How do we know the probability of some event that will occur in the future? Because there are statistics on accidents and these form a sample from which the probability is estimated, if and (quasi) only if accidents are independent. If they are not, or if there is too much discrepancy between subsamples so that the estimated probability is not relevant for the whole population, accident risks are usually classified through statistical analysis by identified characteristics. Then, characteristics are the new definition of i.i.d. risks, and are often called "risk profiles", or risk types.

For example, age is an important characteristic (not a cause!) of the probability of accidents. Sex also, but this is a typical characteristic that cannot be employed (by law) to select among insured individuals. The fact that more men have car accidents than women may be explained by a job characteristic, for example: men are more likely than women to accept jobs involving lots of driving. Insurance contracts against robbery distinguish social environments by residential areas. Earthquake risk or flood risk insurances are sensitive to geography, etc. Obviously, characteristics must be observable (contracts are contingent on them) and besides, as we mentioned, there are legal limitations to the use of some of them (sex, religion, race, etc.). Some are clearly statistically significant but are difficult to observe. For instance, driving skill is certainly an important accident factor, however it is not observable and cannot be used.[8] In order for the law of large numbers to apply, independence is not enough, the number of insured agents counts. Then, the number of characteristics must not be too high, or risk classes will not be numerous enough to reduce significantly the insurer's risk, but it must not be too low either, or insured risks will not be i.i.d.

Traditional private insurance is therefore typically adapted to "probabilised" risks, i.e. to risks with uncertain returns that can be represented by independent random variables with the same known probability distribution. We have seen what is meant by "known" probability; this establishes a parallel between insurable risks and random devices such as lotteries or roulette wheels. We have seen in Chapter 2 that such devices correspond to a particular case

---

[6] Another "law of large numbers", the Central Limit Theorem (see next chapter) helps to manage this risk.

[7] Actually, increments are not slight, even in non-profit benefit insurance organisations, in order to cover expenses and be attractive to investors, and also because insured risks do not satisfy all the conditions necessary for the principle to apply (e.g. many are dependent random variables).

[8] Other characteristics appear from statistics, but they seem so strange that people would find themselves prejudiced if used: e.g. in France at least, red cars have more accidents than others!

of uncertainty (seldom encountered outside casinos and national lotteries!). It so happens that, thanks to sampling and estimation methods, insurable risks can be described by such a device. Take a roulette wheel with two sectors: one representing "accident", the other "no accident". If the first section is p, then the second one is $1 - p$ (normalising the surface to 1). Each spin of the roulette wheel is done independently and statistics verify that the proportion of "accident" is p. Probability calculus, and the mathematical theory that followed, referred to games of chance, and this is why insurable risks are often compared to such random devices in the economic literature. And, considering the previous principle, they should do!

This is the first reason why risk characteristics (risk profiles) are important to classify and observe. Another reason is that they will influence premiums and reimbursements. If an insurer were to propose the same premium to individuals with different risk probabilities, that would generate an adverse selection problem in the long run (we shall come back to this in Chapter 6). Indeed, insurance will be more advantageous, and therefore more attractive, for the risks with high probabilities than for those with lower ones. If too many high-probability risks enter the insurer's portfolio, its global risk will be modified and the premium should be increased. The traditional solution to this problem is to propose different premiums according to risk profiles. This solution cannot be used for social insurance, by definition: risks are shared equally, sick individuals are reimbursed by healthy ones … who may become sick and be grateful that ex-sick ones are healed in order to pay for them. Hence, social insurance does not refer much to this principle for its existence and its management, but instead to the first mutuality principle.

For the same reason, social insurance does not react to another problem faced by insurers, in the same way that traditional insurance does. The second problem results from insured individuals' behaviours in the face of their risk. No qualification in psychology is required to understand that someone who is insured may be less prudent than someone who is not. This phenomenon is called "moral hazard"; we shall come back to it in Chapter 6. The hazard is generated by insurance, not by the risk itself. To answer this problem, insurance companies use partial reimbursement contracts. The part reimbursed can be proportional to the damage amounts, or refer to a fixed sum called a "deductible". This way, the insured individual is still conscious that prudent behaviour may reduce the non-insured part of the risk. In the case of health insurance, this solution is partially employed, even at the social security level: a part of health expenses is left to individuals. However, that fraction is so small compared to total spending, e.g. to surgery costs, that the solution is not very efficient. This may be compensated for by the mutuality principle, if the insured individuals integrate it and understand that a lack of individual prevention will lead to higher expenses. In practice, laws enforce integration: e.g. vaccination is often obligatory.

The traditional insurance industry is adapted to well-known independent risks. It is founded on what we called the independent large numbers principle. Furthermore, the mutuality principle helps to reduce inequalities amongst the insured. All this is all right as long as probability distributions are well known, and this means in practice that risks are represented by i.i.d. random variables. However, many situations are not adapted to such a representation: there are uncertainties with respect to damage occurrences.

5.2 INSURANCE OF RISKS WITH UNCERTAINTIES

Let us still assume risks have known probabilities for the moment, but some uncertainties must be taken into account. For instance, there may be uncertainties about the independence of the insured risks. Then, the independent large numbers principle cannot be referred to: the insurer's risk portfolio is not riskless, even in approximation. If risks are not independent, some may hedge each other: if (and only if) they have a negative covariance. This reduces the insurer's portfolio global risk but does not eliminate it. Furthermore, diversification results from the sharing of an investment among imperfectly correlated risks. In any case, the global variance calculus requires covariance estimation, which is difficult if the number of risks is great, so that there are many uncertainties left with regard to the insurer's total risk. In many situations, correlation coefficients are neither 0 nor negative, and the difficulty for insuring is often due to strictly positive correlations. Natural catastrophe risks are a common example of the extreme case with perfect correlation. The mutuality principle can be applied to partition risks among insured agents so that they face less risk than they had to without pooling them; this still leaves some risk to be managed by the insurer.

How does an insurance company manage highly positively correlated insured risks? The first task is to accept risks in proportion to the insurer's capital so that it can fulfil contractual reimbursements. Then, it must diversify the insured risk portfolio by mixing types of risky activities: fire, automobiles, houses, industrial risks, natural risks, etc. Diversification can also extend to localisations: regions, countries and parts of the world.

Among highly correlated risks, natural catastrophes play a special role: earthquakes, floods, storms, etc. Obviously, damages cannot be considered as independent and identically distributed random variables. Furthermore, past catastrophe frequencies yield an unreliable estimate of future occurrence probabilities. But even when they do, catastrophes are not regular in the timing of occurrences nor in intensity. There is no way to estimate the precise damage amount per year.

Then, we are not in a situation where probabilities are well known (and/or reliable): the future is uncertain, i.e. a set of consequences (casualties) may be defined. The concept of risk which underlies such situations differs from the one that the traditional insurance industry knows how to manage, we are back to the origin of risk: an certain capital (premiums) is transformed into an uncertain one (reimbursements).

This is the main remaining risk that the insurer must get rid of. One way is to reinsure. Like any other company or individual agent, insurance companies need to insure themselves against the risk that reimbursements exceed their financial capacities. Insurers are still responsible for the contracts they signed with individual agents, but they pay a premium to another insurer to insure a part of the total damages they must reimburse.

There are different types of reinsurance contracts:

- Reinsurance treaties cover the first insurer's total portfolio in a given activity branch. A treatise is a contract stipulating conditions to be satisfied (contingent contracts). Some are proportional: "pro rata risk-sharing" contracts. In this case, the reinsurance company takes responsibility for a given share of the damages. Non-proportional contracts include, for example, cases where the reinsurance company only insures damages that are beyond a given ratio: the loss ratio. It is defined as the ratio between the sum of insured damages and the sum of premiums.
- The so-called "facultative reinsurance" contracts are specific to an individual's risk. They concern big industrial risks[9] and huge equipment works.

[9] The European space rocket Ariane is an example.

Co-insurance is another way to share risks among insurance companies. A contract binds several insurance companies that jointly warrant a risk or a risk portfolio. Each insurance company covers a percentage of the risk in exchange for the same percentage of the received premiums. The insured individual only deals with its insurance company, which is in charge of the contract. Co-insurers are responsible only up to their engagement with the other insurers.

Lloyd's unions form a particular organisation that offers a place and a set of rules for insurers to sign risk-sharing contracts. Insurers in Lloyd's are called "names", they can be individuals or companies; they have a large capital base and sell insurance contracts. Unions are groups of names that specialise in a particular branch of risk, but a name can belong to several unions. Lloyd's proposes insurance contracts for exceptional risks, but it also offers other types of contracts. One of Lloyd's particularities is that individual names are personally responsible on their own belongings.

However, all of the above management techniques have limitations, in particular in what concerns natural catastrophes:

- The increasing concentration of insurance companies, and even more so of reinsurance ones, limits the number of agents to share risks, be it by reinsurance or by coinsurance: the four major world reinsurance groups have increased their market shares from 22% in 1990 up to 34% in 1999.
- The number of damage claims due to natural catastrophes have increased strongly during the last three decades: from an average of 35 in 1970, up to 125 per year between 1994 and 1998. The ratio between economic damages and insured damages has grown at the same time. The risk increase faced by insurers is explained by: population growth, capital accumulation in risky areas, economic growth, social behaviour under insurance (moral hazard) and decreasing public subsidies. Forecasts announce that several catastrophes in the United States (Californian earthquakes and typhoons in Florida) could make damages climb to over $100 billion. Such amounts would reduce insurance company financial capacities dramatically. Indeed, $100 billion of damages represents about 30% of US global insurance capital.

In order to confront such difficulties, another way to share risks developed: securitisation. Risks are grouped into homogeneous classes, which are divided into security shares and sold on specialised markets, or directly to institutional investors. The securities' returns have higher interest rates than the riskless rate (government bonds interest rate) for the same expiration date: the risk premium (the difference) corresponds to the risk incurred on interests and/or on the capital.

Securitisation does not concern insurance only: mortgages or consumption credits, for example,[10] have been securitised. Securitisation is not limited to catastrophic risks either (some exist on automobile insurance), but this is where it finds its best justification. For instance, the $100 billion worth of damages we evoked could easily be absorbed by financial markets: their daily variation is more than ten times this amount, which is less than 0.5% of the total value of US security and bond markets. Securitisation mainly developed two instruments: "catastrophe bonds" (or cat-bonds) and "weather derivatives". We shall come back to how to use them in Chapter 8.

[10] See Chapter 12.

5.3 NON-INSURABLE RISKS

We shall advocate that "new risks problems", i.e. risks that are not insurable using the existing techniques mentioned above, can still refer to the theory of financial markets and to its recent extensions, at least for a gross approximation of their values. We have mentioned some of these risks in the Introduction and in Chapter 1. Most are linked to worldwide hazards, others to huge projects with uncertain impacts. Environmental issues, human health and huge natural catastrophes as well as industrial catastrophes do not respect state borders. As a result, there is no authority (or no very powerful ones) to set rules and enforce regulations on which contracts may be defined and signed by two parties. Furthermore, a contract has to be contingent on some observable events. In the issues we mentioned, possible events are highly controversial. This is due to two factors: the first one is scientific uncertainty; the second one is that ethical values and political priorities diverge among national cultures. Evidence of such divergence is provided by the US and EU opposite interpretations of the precautionary principle, for instance. Others come from economic priorities: underdeveloped countries would accept some reduction of deforestation to save oxygen, for example, if richer countries were willing to compensate their losses on expected missed profits by accepting higher wood prices. Aids is another example where poor countries cannot afford medications developed by partially government-funded research in richer countries and sold at a high price. But even within one country, controversies often cause tensions. For example, genetically modified organisms are perceived differently by consumers and producers, generating controversies about regulations.

However, all of these problems have high stakes: damages may be important, prevention and regulation measures are costly to enforce and manage. Risks cannot be well defined and represented by formal contracts, so their valuation is beyond the usual insurance industry methods. Basic cost–benefit analysis also lacks methods to include such risks; at least its traditional instruments fail to meet the requirements needed to apply them. This extends to huge national or transnational projects as well, for the same reasons: impacts are difficult to list without controversies, even more difficult to value, and, in the absence of scientific certainty (i.e. non-controversial probability distributions), summaries cannot be done by the usual averaging methods.

All of these problems are linked to one: unmeasured uncertainty about future event occurrences prevails. In some cases, uncertainty comes from scientific uncertainty. When science proposes a controversial probability distribution assigned to outcomes, there is a way out: capacities, instead of a probability distribution, may be constructed that take the different distributions into account and still allow averages to be performed (Choquet integrals, seen in Chapter 2). Other general measures and integrals may be referred to as well. When science cannot assert any probability distributions, or if asserting one doesn't make any sense (e.g. occurrence of a phenomenon never observed and without any theory to simulate it), measuring uncertainties is out of reach.

Are we completely at a loss then, to measure a risk and compare its valuation to preventive or precautionary costs? Our answer is: not completely! Similar problems have been solved by insurance and other organisations concerning natural catastrophes, for example. The solution they found, notably securitisation, is a reference that can be inspiring. We shall advocate that mimicking the risk by (a portfolio of) risky assets that are traded, may give a gross approximation of a collective value for it. Methods to achieve this are derived from securitisation methods and marketing.

6
Risk Economics

In the last chapter, we have seen examples of risks where a probability distribution of outcomes is known and others where it would be difficult to assign one without controversy. The distinction between these two kinds of situations has been established by Knight (1921).[1] Knight meant to characterise situations where risks are well evidenced by past experience and/or by scientific knowledge, he referred to the most common meanings of the word.[2] Situations of risk are well described by a known probability distribution on uncertain outcomes. All other situations are then labelled "uncertain" by Knight. In his work, these did not refer directly to scientific uncertainty, nor to the lack of sufficient data to estimate probabilities, but to the absence of sufficient evidence and knowledge, in general. By uncertainty, Knight really meant risks that are subjectively perceived without any objective evidence to convince of uncontroversial probability estimates. In such situations, it is not even clear what "probability" could mean, nor if one distribution exists at all. Nevertheless, it can be advocated that uncertain situations correspond to a description of outcomes and to some events that are expected to be realised. This is formalised in mathematics by a measurable space, i.e. a set of states of nature, S, and a set of events (an algebra or a σ-algebra), F:

$$(S, F) \quad (\textit{non-probabilised}) \text{ uncertainty}$$

Situations of risk are those in which knowledge is more precise, so that outcomes and events realisation are measured by probabilities. This is formalised in mathematics by a probability measure: Π, and situations of risks are represented by a probability space:

$$(S, F, \Pi) \quad (\textit{probabilised}) \text{ risk}$$

The distinction has been established in Chapter 5 between insurable and uninsurable risks, both present in the insurance industry.

This chapter focuses on situations of risk,[3] while the next one addresses the problem of valuing financial risky assets that formalise situations of uncertainty. This means that we assume in the following that a known probability distribution measures risks.

---

[1] And further developed by Ramsey (1926) and Keynes (1921).

[2] And not to the etymological sense we recalled in the general introduction.

[3] Given that "risk" has two meanings, we shall try and use "risk" for risk situations in the sense of Knight, and risk for risk taking, for example, risky asset, etc. which may correspond to Knightian uncertainty.

6.1 LAWS OF LARGE NUMBERS AND THE PRINCIPLES OF INSURANCE

Insurance and finance businesses developed because they offer ways to reduce risks. We shall come back to the meaning of "reducing", it assumes that risk is measurable and that its measure can be lessened. Indeed in situations of risk, the risk is measurable by a probability distribution. In the simple case of a Bernoullian risk, i.e. a random variable which only takes two values: 0 with probability p and 1 with probability $1-p$, reducing risk could mean to reduce p. But this does not take into account the relative importance between the consequence represented by 1 and the other one by 0. Another way is to reduce $p(1-p)$. It so happens that, in this case, $1-p$ is the mean value and $p(1-p)$ is the variance. The mean value does not give an idea of the range, while the variance does. This is why variance (or its square root, the standard error) is often chosen as a risk measure to be reduced... if it exists![4] We'll see an alternative risk measure criterion later in this chapter.

We have mentioned in the last chapter that a reason why insurance supply may be safe is that a large portfolio of independent risks, if shared among insured agents, is approximately riskless. We mentioned that this property came from a mathematical theorem, the Law of Large Numbers (LLN). In fact, there are many such laws corresponding to more or less relaxed assumptions, and there are two types of theorems: it may be confusing. All the theorems have been developed in order to relax assumptions in previous ones, they address two types of problems: the first one was meant to capture the idea of "errors" in physics measurements. That is achieved by the Central Limit Theorems, they are meaningful for physicists (or engineers) because both need to control risks (risk of errors for the latter), but we shall see that they are relevant for the insurance business as well. The second one was directly related to insurance problems and theorems are called laws of large numbers, indeed. Both types of theorems are about convergence of a sequence of random variables.

Convergence of random variables may take several meanings. Depending on the type of random variables, one can concentrate on three main ones. From the more to the less general, we have: convergence in distribution, convergence in probability and (Π-) almost sure convergence, where Π is the reference probability distribution over the states of nature. All random variables in the following are real and on the probability space: (S, F, Π) and let X_n be the generic term of a sequence, X its limit if it exists, Π_Y the distribution on R of a random variable Y, and F_Y its cumulative distribution function. The first convergence notion is practically the same as the pointwise convergence, which is applied on non-zero probability sets:

X_n converges almost surely towards X iff $\Pi\{s/X_n(s)$ does not converge towards $X(s)\} = 0$

This implies a weaker convergence notion:

X_n converges in probability towards X iff $\Pi\{s/X_n(s)$ does not converge towards $X(s)\}\underset{n\to\infty}{\longrightarrow} 0$

Which implies:

$$X_n \text{ converges in distribution towards } X \text{ iff for any } A \text{ in } F,\ \Pi_{X_n}(A)\underset{n\to\infty}{\longrightarrow}\Pi_X(A)$$

[4] Variability may be infinite. For instance, in the St Petersburg betting problem (Chapter 2) the mean, and hence the variance, are infinite.

Or, equivalently, if and only if for any real number x at which the cumulative distribution functions are continuous: $F_{X_n}(x) \underset{n\to\infty}{\longrightarrow} F_X(x)$.

For the insurance problem, the sequence represents a bigger and bigger portfolio of risks (insurance contracts) as n increases. Formally, let Y_k be the insurance contract (random) payoffs of agent k and let us assume from now on that all agents belong to the same risk profile so that they have the same casualties' means and variances: m and σ^2 (they do not need to have the same probability distribution). Let $S_n = \sum_{k=1}^{n} Y_k$ represent the insurer's risk portfolio. Its mean is nm and its variance is $n\sigma^2 + \sum_{k=1}^{n} \sum_{k'=1,k'\neq k}^{n} \mathrm{cov}(Y_k, Y_{k'})$. Obviously, if the Y_k's are not negatively correlated, there is no way that portfolio S_n is less risky than S_{n-1} (in the sense of reducing the variance).

However, there are two ways to deal with this, and both rely on an increasing number of agents. The first way is to have insured agents mutually share their risks (if they are similar enough) through insurance contracts with the company. This is the traditional basis of insurance, it is based on the LLN which says that the large insurance risk portfolio is almost riskless. However, "almost" may not be enough to induce investors to invest in the insurance company, especially because investors generally prefer to take risks that have some positive outcomes, while insurance portfolios have only a negative side. The second way consists of having investors take capital shares of the insurance risk portfolio, this is based on another convergence theorem: the Central Limit Theorem. It says that the insurance shareholders' risks will converge towards a symmetrical risk with a well-defined variance.

As we mentioned in the last chapter, an intuitive idea driving the LLN made it possible for financial managers to consider taking risks from different agents, before the theorem was formally proved during the 18th century. The intuition is that agents won't all face casualties at the same time, or more exactly contingent on the same event, during a given time period (a year, in general). As a consequence, the sum of premiums, if correctly calculated and reinvested, may cover the total damage to be reimbursed. The relative number of agents facing a damage in one year among the ones insured is close to the frequency of observed casualties time series observations. If the time series can be considered as a sample of independent random variables, then the number of accidents in one year is forecasted accurately. This is obtained by applying a LLN, a special case of the weak law, below. Furthermore, a LLN is applied to agents risks, once the risks have been mutually shared in a way to reduce their variances. This is obtained by letting the contract of each agent be written on: $X_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$, instead of its own risk: Y_k. Then, the variance of X_n is affected by $1/n^2$ and decreases as n increases, hence shifting from Y_k to X_n is beneficial for each agent because the former is "less risky" (its variance is lower) than its own risk. X_n is the risk insured by each agent k, $k = 1, \ldots, n$, and of course $nX_n = S_n$: the insurer risk portfolio. An insurance contract is paid by an insurance premium. Let us assume for the moment that the premium is "fair", i.e. the insurance company only charges the mean value of insured casualties: m. Then, $X_n - m$ is the risk "sold" by the insuree to the insurer and the insurance company's risky part of its portfolio is $n(X_n - m) = S_n - nm$. The convergence theorem bears on this risky part of the insurer's portfolio.

We can express two (main) laws of large numbers. They do not necessarily require independent identically distributed (i.i.d.) random variables, but that they are non-correlated (i.e. $\mathrm{cov}(Y_k, Y_{k'}) = 0$).[5] Furthermore, it is assumed that each Y_k has finite means and variances

[5] Independence implies no correlation but is not necessary except for the Gaussian distribution, in which case they are equivalent. However, independence is usually assumed as it is a requirement for samples.

m_k and σ_k^2 and then the centralised (zero mean) insurer portfolio is: $S_n = \frac{1}{n}\sum_{k=1}^{n}(Y_k - m_k)$. So, if the Y_k's are not correlated: $\sigma^2(S_n) = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2$.

Weak Law of Large Numbers *If the Y_k's with means and variances m_k and σ_k^2 are not correlated and if $\sigma^2(S_n) = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2$ converges towards 0, then $S_n = \frac{1}{n}\sum_{k=1}^{n}(Y_k - m_k)$ converges in probability towards 0.*

Otherwise stated, the probability that the insurer's risk portfolio is bounded away from zero gets closer and closer to zero when the number of insured agents increases: the bigger the insurance portfolio, the less risky it is.

The strong LLN is even more reassuring about the convergence of the insurer's portfolio towards a riskless position.

Strong Law of Large Numbers *If the Y_k's with means and variances m_k and σ_k^2 are independent and if $\sum_{k=1}^{n}\frac{\sigma_k^2}{k^2}$ converges (towards a constant), then $S_n = \frac{1}{n}\sum_{k=1}^{n}(Y_k - m_k)$ converges towards 0 almost surely.*

This theorem is stronger than the preceding one (even though the independence assumption is slightly stronger) because it says that the set on which the insurer's portfolio does not converge towards 0 has a zero probability (the limit portfolio is almost riskless).

As we mentioned at the beginning of this chapter, an investor may still be reluctant to invest in an "almost" riskless portfolio, especially because this portfolio only presents negative payoffs. The idea of changing the shape of the insurance portfolio's risk is formalised by the Central Limit Theorem: it says that, if risks are not correlated, the normalised insurance risk portfolio $S'_n = \frac{S_n - nm}{\sigma\sqrt{n}}$ has a distribution with mean 0 and variance 1 that converges towards the normalised Gaussian distribution $N(0, 1)$.

Central Limit Theorem (CLT) *If the Y_k's are not correlated and have the same means m and variances σ^2, then the sequence S'_n* ***converges in distribution*** *towards a $N(0, 1)$.*

Equivalently, the theorem could be written as:

$$\sigma S'_n = \frac{1}{\sqrt{n}}(S_n - nm) \text{ converges in distribution towards a } N(0, \sigma)$$

This theorem teaches us that a large normalised risk portfolio approximately follows a Gaussian distribution with zero mean and σ^2 variance. Notice that such a distribution is symmetric with respect to positive and negative payoffs. A way to obtain this symmetrical distribution is to equally share risks among the insured agents in the first step (this is the basis of insurance based on the LLN) and then among the insurance company's shareholders in a second step.

The basis of insurance consists of writing contracts that are the same for different agents with similar risks. Here these risks are the Y_k's and the contracts are based on an equal share of the total insurance portfolio S_n. Assuming the premium paid by each agent is equal to m, $S_n - nm = n\left[\frac{1}{n}\sum_{k=1}^{n}(Y_k - m)\right]$ is the portfolio of such *ex post* insurance risks. This is the first step.

The second step consists of changing the shape of this portfolio in order to make it more attractive to investors that may be willing to take shares of the insurance risk portfolio. There is still a need for a large number of investors, but not as large as the number of insured agents. Instead of n agents, the portfolio shares must only be $\sqrt{n}$. Let's take a simple example to see how it works.

Take the case of simple insurance for damage with a probability p for a given loss, L. If no losses occur, the insurer gets $0 + n\pi$ where π is the premium perceived. Obviously, π must be tiny as compared to L, otherwise no agent would get insurance. In practice, as we shall see, the premium is close to the mean value ($\pi = pL(1+\lambda) = m(1+\lambda)$, with λ a given "loading" factor). Then, the LLN assumptions and the CLT assumptions are satisfied if agents' risks are not correlated, and let us assume that an insurance company can gather a great number of insurance contracts.

Assume $n = 10\,000$, a reasonably large insurance portfolio for the LLN and for the CLT to apply, and $L = €10\,000$ with probability $p = 0.01$. Then $m = pL = 100$ and $\sigma^2 = p(1-p)L^2 = 990\,000$. Share this portfolio among $100 = \sqrt{10\,000}$ insurance company shareholders. Each shareholder would face potential loss $\frac{1}{\sqrt{n}}(S_n - nm)$, which is approximated by a symmetrical $N(0, \sigma^2) = N(0, 990\,000)$ distribution according to the CLT. If each shareholder wants to be warranted a given loss with less than a 5% probability to go over, this maximal loss for the $N(0, 99 \times 10^4)$ is €1636.4. In order to hedge this potential loss, the insurance premium may be loaded by €16 for each of the 100 insured agents.

In economic theory, most authors consider the distribution as being "objective". In fact, it is derived from times series frequency, i.e. it is an estimate of an objective probability, if and only if risks are i.i.d. When this assumption is not satisfied, insurance contracts and the economics of insurance have to be reconsidered. In many cases, however, samples over past series yield good estimates and such an assumption is satisfied, at least approximately. When there is no objective distribution, a subjective distribution is referred to. It is justified either pragmatically by an expert assigning probabilities to events, or theoretically as a representation of an agent's behaviour. Most of the results and analysis in insurance economics require a given probability distribution, the difference between an objective and a subjective one only matters if different behaviours are confronted with each other.

6.2 RISK AVERSION AND APPLICATIONS TO THE DEMAND FOR INSURANCE

Throughout this section, risks are represented by probability distributions over decision outcomes. Economic theory has developed concepts within this framework and mostly within the expected utility paradigm. With a given probability distribution over outcomes, decision-making under risk is similar to betting on a lottery number. A decision-maker who doesn't seek risk but wants to bet will choose a lottery different from the one chosen by a player who finds extra gratification from playing a riskier lottery. Characterisation of risk aversion has been a major contribution to the economics of risk. The basic idea of "risk aversion" is that a risk-averse agent will prefer a certain amount of money to a lottery ticket with a mean prize equal or superior to this amount. Comparison of risk aversions can be achieved by observing behaviours in the face of different lotteries.

Take the extreme case where one of two lotteries is degenerated: it yields an outcome of €10, say, with probability one. How can we compare it with another "normal lottery", a lottery where there is at least a slight probability $p = 0.01$, say, of getting €0 instead of €10 with probability $1 - p$? The question opens another one: how do we compare something certain with something uncertain? Otherwise stated, is there something certain equivalent to a lottery? A solution was proposed to the second question as soon as probabilities were defined by Pascal and Huygens in the 17th century: a lottery was proposed to be equivalent to the mean of its payoffs.

The mean is indeed a certain value which is central in the range of payoffs. But why and for whom, would it be equivalent to the lottery? A decision-maker must have preferences over lotteries, so we can define:

Certainty equivalent *A decision-maker's certainty equivalent of a lottery X is a degenerated lottery yielding a certain outcome CE*(X)*, such that the decision-maker is indifferent between the two lotteries.*

$CE(X)$ can be interpreted as the value for which the decision-maker is indifferent to bet or not to bet on the lottery. A decision-maker who is risk averse may not value a lottery by its mean, it will bet less than what it expects to win on average. This yields a first definition of risk aversion (Arrow, 1971).

(Weak) risk aversion *A decision-maker is weakly risk averse if and only if its certainty equivalent of any lottery is inferior to the lottery's mean value:* $CE(X) \leq E(X)$.

Reciprocally, we shall characterise a risk-seeking behaviour by a certainty equivalent being superior to the mean value: $CE(X) \geq E(X)$.

Both definitions reconcile for a particular type of behaviour, the one characterised by the Pascal and Huygens' criterion: the mean value. Such a criterion is often contradicted by observed behaviours, notably in front of the St Petersburg paradox example (the mean is not finite! See Chapter 2). A decision-maker who is rational in the sense of the Pascal–Huygens criterion is called risk neutral (indifferent to risk: neither risk averse nor risk seeking).

To take an example in insurance economics, the insurer is often assumed to behave as if it were risk neutral. The reason behind this assumption is not that the certainty equivalent of the insurer is equal to the insurance company's risk portfolio mean value. The insurer may be considered as using the mean as its certainty equivalent because, if the LLN applies, the portfolio of insured risks converges towards a riskless value: its mean value.

Obviously, a risk-neutral agent would not be ready to pay anything to get rid of a risk. A risk-averse agent would do: it should be ready to pay the difference between its certainty equivalent and the mean value. The "price" to pay for trading a risk against its mean value with certainty is the definition of a risk premium.

Risk premium *If* $CE(X)$ *is an agent's certainty equivalent of the random variable X with expected value* $E(X)$*, the agent's risk premium for X is:* $RP(X) = E(X) - CE(X)$.

According to the definition, a risk-seeking agent will pay a negative risk premium (accept to be paid a positive amount) to give up a risk, or pay a positive one to "buy" a risk. Buying a risk means that this risk is represented legally by a tradable contract. In order to be insured, a risk-averse agent buys an insurance contract with positive payoffs corresponding to the reimbursements or casualties if they occur, or nothing if they don't. By so doing, the agent "gets rid" of the risk of casualties (or a part of it). On the opposite side, the insurer sells a contract with positive payoffs, but by doing so it contracts the opposite sign payoffs to the risk-averse agent. This may be confusing because the insurer could be thought of as buying a contract with negative payoffs, which is formalised by selling a contract with the opposite sign payoffs. Indeed, the insurer is a risk taker, even though the LLN assures that the overall risk will converge towards a riskless position.

In the insurance business, the premium paid by the insuree is the "fair premium" (the risk's mean value) plus the risk premium.[6] The risk premium is usually designed as the risk's mean value multiplied by a fixed loading factor λ. The loading factor includes a price for risk, but it also includes transaction and management costs. A fair premium has a loading factor equal to 0. A risk-neutral agent will not buy insurance[7] if the risk premium is not zero and even in this case it will gain nothing by being insured. A risk-averse agent will get insured and pay the positive risk premium proposed by the insurer if the insurance premium does not exceed its expected loss plus its own risk premium (the difference between the mean value and its certainty equivalent).

In market finance, risks are represented by tradable assets, i.e. contracts with (usually) positive payoffs. Most of the time finance thinks in terms of rates of return, so that risk premium designates the difference between a risky rate and the corresponding riskless rate, e.g. in the CAPM (see Chapter 7 and Chapter 10). The CAPM formula for the expected rate of return of an investment i is expressed as: $E(r_i) = r_0 + [E(r_M) - r_0]\ \beta_i$, where r_0 is the market riskless rate (e.g. a Treasury bill rate), $E(R_M)$ is the market expected risky rate for the portfolio formed of all traded securities and β_i is the sensitivity of the investment i to this market risk. If we think in financial terms, the insurer is a borrower: indeed, it receives a positive amount, the insurance premium: $(1 + \lambda)m$, where m is the average loss, and later reimburses (part of) the casualties (L) if they occurred. The "borrowing rate" is $r = \frac{L-(1+\lambda)m}{(1+\lambda)m}$. The expected rate of return is negative: $E(r) = -\frac{\lambda}{(1+\lambda)}$, in contrast with a market investor or a lender. As a consequence, it's not $E(r) - r_0$ that has to be explained, but $r_0 - E(r) = r_0 + \frac{\lambda}{(1+\lambda)}$.

The insurer uses the insurance premiums to invest on the financial market, so that the excess rate of return $\lambda/(1 + \lambda)$ with respect to the market expected rate of return is justified if the insurance portfolio increases the risk of the financial portfolio of the insurer. This happens only in one case: if the sensitivity (β) of the insurance portfolio to the market risk is positive. In the opposite case, i.e. if the insurance portfolio is negatively correlated with the market risk r_M, then this excess return is not justified. In the special case where $\beta = 0$, then it should be the case that the insurer pays the riskless rate to its insurees in order for the insurance premium to be fair! Let us note that, in general, the insurance business is not positively correlated with the securities market risk, so the paradox applies.

Reasoning in terms of rates of return makes apparent an important element in the management of insurance companies that the traditional approach neglects because the reference model is a static one. An insurance company, as well as other investors, plays with the time passing between the date where premiums are paid and the dates where casualties are reimbursed. The relevant model for insurance management should be a dynamic one (the distinction between static and dynamic models is the topic of Part III). As a consequence, if profits from investing premiums are taken into account, insurance premiums could be much lower than in the traditional approach. However, the loading factors not only take risk premiums into account, they include management costs that increase the insurance premiums and are out of range of our approach as well as of the traditional one.

In individual decision theory, figuring out the risk premium that an agent is ready to pay for a given risk requires that its certainty equivalent be defined. This is why we need an axiomatic approach to define the agent's decision criterion. Standard insurance economics

[6] In practice, both are often confused and called indifferently "risk premium" or insurance premium.
[7] If regulation does not make insurance an obligation.

relies on the particular von Neuman–Morgenstern (1944) or the Savage (1954) axiomatic theory: expected utility. The criterion is then: $V(X) = \int u(X)d\Pi$, where risks are represented by random variables on a probability space $X:(S, F, \Pi) \to R$.

Within this individual decision theory, the only way to characterise behaviour in front of risk is the utility function: u. Three function forms are easy to distinguish: linear, concave (super-linear) and convex (sub-linear).[8] They are bound to characterise the three types of behaviours in front of risk, respectively: neutrality, risk aversion and risk seeking. Indeed, take a simple risk Y with two outcomes: y_1 and y_2 and take a non-random asset X (i.e. paying m with probability 1) such that X is weakly preferred to Y. Assume furthermore that $E(Y) = m = py_1 + (1-p)y_2$, so that the weak preference indicates (weak) risk aversion. The weak preference is expressed by: $E(u(X)) \geq E(u(Y))$. This is:

$$E(u(X)) = u(m) = u(py_1 + (1-p)y_2) \geq pu(y_1) + (1-p)u(y_2) = E(u(Y))$$

The inequality between the two middle terms characterises concavity of the function u. A risk-seeking agent will satisfy the opposite inequality (u is convex), a risk-neutral agent is indifferent between Y and its mean value m and the utility function is affine: $u(x) = ax + b$ for some non-negative a and an arbitrary number b. Note, however, that the utility function properties are local ones: an often "observed"[9] utility function is sub-linear for a mean value m close to zero, it becomes linear on a segment of higher values for m, then for values even higher, it becomes concave (Figure 6.1). Such a utility function characterises an agent who will buy lottery tickets for €1 with a hope to gain a high prize. For common expenditures, an object with a 50% chance of being worthless is rationally valued at half its price. When higher values are at stake, e.g. a house, a car, etc., then the same agent buys insurance and may accept a large risk premium.

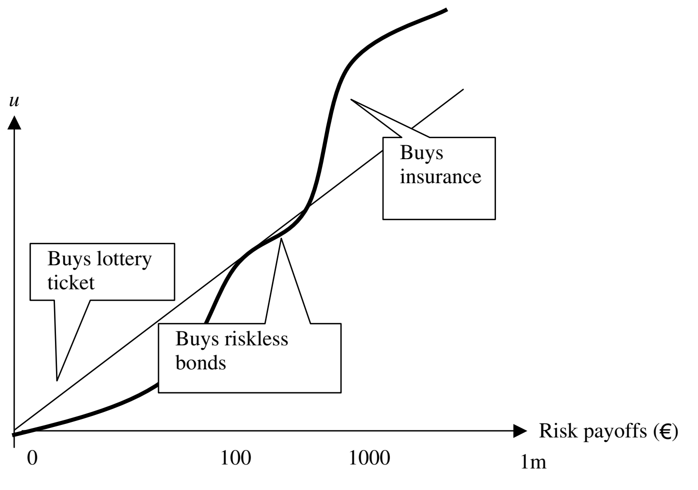


Figure 6.1 Behaviours under risk

[8] Given that von Neumann–Morgenstern utility functions are defined up to an affine transformation, linear really means affine.
[9] That is, consistent with observed behaviours.

Expected utility and the shape of the von Neumann–Morgenstern (vNM) utility function u are useful tools to proceed to marginal analysis, e.g. demand for insurance, etc. It has, however, a poor descriptive relevance for choices in front of large rates of return: e.g. in finance. Going further in the local characterisation of risk aversion within the EU model, Arrow (1971) and Pratt (1964) found indexes measuring different levels of risk aversion. Let us first define:

More risk averse *An agent with utility function u_1 is "more risk averse" than an agent with utility u_2 iff there exists a concave function f such that: $u_1 = f(u_2)$.*

This is consistent with u_1 being "more concave" than u_2. In the case where utility functions are twice differentiable at a point x, concavity is expressed for an increasing function ($u' \geq 0$) by: $u'' \leq 0$. Then "more risk averse" could be grasped by "more concave", i.e. by some concavity curvature index being higher. The Arrow–Pratt index is exactly this: at a given income level x, the absolute risk aversion (ARA) index of an agent with vNM utility function u (twice differentiable in x) is: $ARA(x) = -u''(x)/u'(x)$.

The Arrow–Pratt index is obtained by a simple mathematical development of the utility of a small income random perturbation around $x : x + \varepsilon$. The risk premium $RP(x+\varepsilon) = E(x+\varepsilon) - CE[(x+\varepsilon)]$, or $Eu[E(x+\varepsilon) - RP(x+\varepsilon)] = Eu(x+\varepsilon)$, is calculated using two local approximations with derivatives and an approximate value is obtained: $RP(x+\varepsilon) \approx -\frac{u''(x)}{u'(x)}\frac{\sigma^2}{2}$. Note that approximations assume a "small" perturbation, this means that: $E(\varepsilon) = 0$ and $E(\varepsilon^2) = V(\varepsilon) = \sigma^2$ is small (and the smaller, the better the approximation!).

Under this limitation, the Arrow–Pratt index[10] is an easy way to compare two risk-averse decision-makers at a given level of income, when they face the same risk. Here is an example. Assume agent 1 has preferences represented by: $u_1(x) = \ln(x)$ and agent 2 by: $u_2(x) = 1 - \exp(-x)$. Their respective risk-aversion indexes for an income level of €10 are:

$$ARA_1 = -\frac{u''(10)}{u'(10)} = -\frac{-10}{100} = 0.1 \quad \text{and} \quad ARA_2 = -\frac{-\exp(-10)}{\exp(-10)} = 1$$

Hence, agent 2 is more risk averse than agent 1 at the income level 10. Note, however, that at the income level 1, agent 2's Arrow–Pratt index would not have changed, while agent 1's becomes equal to 1: at that income level, both have the same risk aversion! As a consequence, there is no way to characterise the whole behaviour of an agent by such a parameter: it is an approximation and is only valid locally.

The demand for insurance has been analysed within the expected utility model and forms the basis of classical insurance economics, results can be found in the usual textbooks (Borch, 1990, for instance). Risk aversion, characterised by concave utility functions on income, is crucial in this setting. A famous theorem, due to Mossin (1966), summarises results obtained by himself and others. It characterises the maximum risk premium $RP_{\max}$ an agent is ready to pay to buy insurance for a risk Y with a riskless wealth W, of a given damage level D occurring with probability p.

[10] There is another Arrow–Pratt index: the relative risk aversion (RRA) index. It is obtained by a multiplicative random perturbation around $x : x\varepsilon$, and is $RRA(x) = -\frac{u''(x)}{u'(x)}x$.

Demand for insurance theorem (in EU)

1. *If u is concave, RP_{max} is superior to the risk to be insured mean value (pD).*
2. *RP_{max} is a non-decreasing function of p and of D.*
3. *If the absolute risk aversion for Y is non-increasing in wealth, then RP_{max} is a non-increasing function of W.*

The proofs of 1 and 2 are direct, the proof of 3, however, relies on more sophisticated mathematics, this shows how sensitive to particular assumptions of the model such a seemingly intuitive result is.[11] Let us insist on the fundamental ones:

- The result is obtained in an EU model.
- ARA is an approximate characterisation of risks with a "small" variance.
- The utility function properties are locally defined around the mean value and rely on it being twice differentiable.
- Risk aversion is characterised by a comparison between a risk and an amount of riskless wealth. In most situations, however, wealth levels are risky.

We shall soon come back on the last remark and characterise a more general concept of risk aversion: aversion to risk increases. Let us concentrate on a simple example to make things easier.

The owner of a house wants to insure it against fire damage. The value of the house is x_1 but it may burn, so the owner really has a risky asset X that takes two values: x_1 with probability p and $x_2 = x_1 - D$ with probability $1 - p$, where D is the potential damage if it burns. An insurance contract proposes to reimburse the amount D in case of fire damage. With this insurance contract, the value of the insured house is certain and equal to x_1. It is assumed, furthermore, that the house owner has a non-random wealth W and faces no other risks, so that the total initial wealth is $W_i = W + X$. A risk-averse house owner would prefer a riskless final wealth $W_f = W + K$, where K is a constant to be defined, to his initial risky one, if W_f yields him a higher satisfaction. In terms of expected utility, this means that $Eu(W_i) \leq u(W_f)$, given that W_f is not random. Several opportunities are open to such a house owner to change his initial W_i to W_f:

- Sell the house at a price $P(X)$, then $W_f = W + P(X)$. He then must find a buyer accepting a price $P(X)$ high enough to obtain: $Eu(W_i) \leq u[W + P(X)]$.
- Insure damages, i.e. find an insurance contract $C(X)$ such that $C(x_1) = 0$ if the house does not burn, and $C(x_2) = D$ if it does, so that $X + C(X) = x_1$ is riskless. Such a contract is traded at a price, the insurance premium: $P(C)$, and the final wealth is $W_f = W + X - P(C)$. The insurance contract will only be purchased if the insurance premium $P(C)$ is low enough, so that: $Eu(W_i) \leq u(W_f) = u[W + X - P(C)]$.
- Compensate possible losses by gains obtained from another asset, for example an investment with random future payoffs Y such that $W_f = W + X + Y - P(Y)$, where $P(Y)$ is the price of asset Y, and is low enough for $Eu(W_i) \leq Eu(W_f)$. Asset Y is a hedge against X, it is more common to use such means for a risky investment than for a casualty risk.

[11] Postgraduate students, when they are submitted to this lengthy and delicate proof and realise what it does prove, have a sane tendency to laugh among themselves . . . when they are nice!

Note that, formally, the two last cases are identical: an insurance contract is a hedging asset (in the previous example it was a perfect hedge, but most insurance contracts only propose partial insurance and hence the hedge is not perfect). The first case consists of selling a risky asset at price $P(X)$. An insurance contract is usually interpreted as being bought, as $P(C) > 0$, but it can also be considered as selling the risk casualty of the asset $C(X)$ at price $-P(C)$. Arbitrage between the two last solutions is done by comparing the two prices, this usually goes in favour of the second one because the risk premium is relatively low. According to both interpretations, the house owner wants to get rid of its risk (or part of it). Conversely, in the third case, the house owner prefers to keep his risky asset and uses part of his riskless wealth W to invest in another risky one, such that his total portfolio has a higher expected satisfaction: $Eu(W_i) \leq Eu(W_f)$.

We shall come back in Chapter 8 to other particular results on insurance demand in the EU setting.

Everything we have presented up to now in this section was obtained within the expected utility model, a model in which the whole agent's behaviour is represented by its utility of income. If we turn to the dual model of Yaari (1987): $V(X) = \oint X \mathrm{d}\phi(\Pi)$, the whole agent's behaviour is characterised by its deformation function ϕ. We obtain similar results, but they bear upon the agent's attitude in front of lotteries, whatever its behaviour on incomes (Doherty and Eeckhoudt, 1995). For instance, risk aversion, risk seeking and risk neutrality are respectively characterised, up to an affine transformation, by convex functions (under the identity function), concave functions (above the identity function) and the identity function. As a consequence, the premium to pay, or the price at which the house could be sold, in order to reduce the risk in the previous example will be different and the decision changed depending on the transformation function instead of the marginal utility.

To be more complete (and add a degree of freedom in the descriptive power of the criterion), we can refer to the RDEU model, where both the utility and the probability transformation functions are present. Then, risk aversion plays on two sides (income and probabilities):

- If the utility function is concave and the deformation function is convex, then the decision-maker is risk averse.
- If the utility function is convex, the decision-maker can still be risk averse if its deformation function is sufficiently convex.
- Similarly, with a concave utility function but a very concave deformation function, the decision-maker can behave in a risk-seeking manner.

The previous results were obtained by Chateauneuf and Cohen (1994) in a seminal paper which opened the way to many more relevant characterisations of risk aversion.

The more general behaviour representations, however, are better armed to front a problem that has shaken insurance economics: most of the previous results generally don't hold any more if an agent's risk is defined in reference to an already risky position instead of a riskless one, e.g. if removing the undesirable risk by an insurance contract does not make the agent's situation riskless. This is known as the "background risk" problem.

6.3 BACKGROUND RISK

The weak point of the previous characterisation of risk aversion is that it refers to a "riskless" wealth that could be an alternative to a risky one. There are two situations in which such a choice is present: chance games and "simple" insurance contracts, i.e. contracts concerning a great number of little i.i.d. risks. In a chance game the alternative to gamble is not to bet: a riskless situation. A simple insurance contract removes the risk that it insures, hence the result can be considered as riskless. Furthermore, in both cases, the mean value of gains is comparable to the amount to pay in order to participate. In most other risk situations, however, there is practically never a riskless situation to be referred to. Consider an investment opportunity instead of a chance game: the no-investment alternative is not riskless, because the amount to be invested may depreciate. As far as insurance is concerned, it is obvious that not all the risks faced by an individual can be removed by insurance contracts so that the alternative to insure a given risk (say, an automobile) is not a riskless situation but only a "less" risky one.

In the presence of what has been called a "background risk" in insurance economics, this risk interacts with the risk to be insured. Given that the previous definition of risk aversion is crucially dependent on the riskless wealth level, it is not a surprise that none of the classical insurance results still hold in the presence of a background risk when particular assumptions are not assumed. For instance, Kihlstrom *et al.* (1981) have proved that an increase in risk aversion (in the EU model) does not imply that an investor will invest less in a risky asset, if the initial capital is already risky. Similarly, Doherty and Schlessinger (1983) have shown that the demand for insurance does not need to be a non-decreasing function of the risk-aversion index, if the non-insurable wealth is risky. Since these two striking results are in contradiction with the above theorem in classical insurance economics, other in the same vein have flourished: they all show that the attitude in the face of risk depends on the whole risk situation and attention cannot be restricted to the particular risk that can be dealt with. Consider an extreme example to make things clear: someone who knows she is going to die of cancer within three months is likely not to spend money on insurance contracts, while someone who does not face such a risk (or doesn't perceive it) may well do it in order to lessen future losses.

When the background situation is risky, attitudes towards a particular risk will depend on the increase or decrease in the overall risk and not from a comparison between a risky and a riskless situation. Let us consider two examples to illustrate this point. The first one is relative to insurance: someone who insures his house against fire hazards faces a monetary risk (losing or not the house's value). Buying an insurance contract induces a reduction of this risk, but this is because the insurance contract itself is risky. Indeed, if the house burns, the insuree gains the damages' reimbursement minus the premium, and zero minus the premium otherwise. To insure is to take a risk. This risk is taken because it hedges the insured risk and hence reduces the overall risk faced by the risk-averse agent.

The second example is about investment and addresses the same phenomenon. Assume you inherited an important amount of money and that you consider altruistically to invest it in a poor country's industry in order to help its development. You take a risk: whether the country's difficulties increase (political crises, for example) and you may lose your investment, or the situation improves and there are chances that returns are high. Generosity is not blindness and you are not ready to lose all your investment opportunity. Fortunately, financial markets are organised for investors who need to get rid of a risk they have taken, or to hedge it, or at least to share it. Here, sharing means that in order to reduce the risk of losing everything, you can diversify your investment between several similar countries. If you want to hedge a country risk you can look for assets with values that are bound to vary in opposition with the country's difficulties (negatively correlated). For instance, if this country's wealth is essentially based on cocoa production, it may be wise to sell futures on cocoa. Indeed, if the cocoa's value decreases in the future, the country's difficulties will increase and your investment will lose value, but you will be able to get some money back from your futures contracts to compensate. Taking the futures' risk hedges the risky investment.

These two elementary examples have pointed out the importance of hedging when two risks interact. This property had been referred to in the generalisation of expected utility theory (Chapter 2). It had been remarked that the fundamental independence axiom can be contradicted if adding a risk to another one reduces or increases the total risk. In the generalisations, this situation only happens when the additional risk cannot hedge the original one: then the independence must be satisfied (comonotonic independence axiom). In order to deal with the so-called background risk "problem", economics has reconsidered the risk-aversion concept in a better adapted context: aversion for risk increases. Let us now give a clear definition to an "increase in risk", and question how to measure risk in the first place.

6.4 RISK MEASURES: VARIANCE AND VALUE AT RISK

There are several "natural" measures of risk. Risk is linked to the consequences' variability with respect to the mean value (or the median, if the mean is not finite). Naturally, then, variance is the most common risk measure, when it exists. Indeed, variance is a measure of variability because it is defined as the mean of quadratic spreads between consequences and their mean value. As such, it is a measure of risk that is independent of the decision-maker's perception of risk and its attitude in the face of it. However, variance is not sufficient to characterise risk, as the following example shows.

Flip a coin and bet €100. If heads, you lose your bet, if tails you win €200, meaning the net risk is a ½ chance to gain or lose €100. The mean value is 0 and the variance is $10\,000 = 1/2 \times (100-0)^2 + 1/2 \times (0-100)^2$. If €10 was at stake instead of €100, the variance would be 100, which is consistent with intuition that the second game is less risky than the first one. Now assume that you win €10 if tails and 0 if heads. The variance would be $1/2 \times (10-5)^2 + 1/2 \times (5-10)^2 = 25$, which is a quarter of the preceding one, according with the intuition that the risk has been "divided". But another way of dividing the risk is that you *lose* €10 if tails and 0 if heads. The variance is still 25, but few gamblers would consider the risk is the same in the two situations. There are two reasons for that: the first one is that the second gamble's mean value is -5 instead of 5. Variance only allows us to compare risks with the same mean. The second one is that risk perception is induced more by possible losses than by expected gains.

The first reason is behind the intuition that risk can be measured by two parameters: the mean and the variance. This is the double criterion proposed by Markowitz (1952) to formalise investors' behaviour when they select portfolios of risky assets. This double criterion reflects the observed investors' behaviour who prefer a higher mean return if the returns' variability is not greater. However, there is a weakness in such a criterion: it does not define a complete order on return and we only get non-dominance. We shall come back to this criterion in Chapter 10, where we shall prove that it is sufficient to define a market equilibrium, if all agents behave accordingly.

How can we take into account the difference between attitudes towards gains and losses, then? A natural measure is to consider the probability that a given maximum amount may be lost or a minimum amount may be gained (percentiles). Take 5% as the fixed probability: the "Value at Risk" (VaR) is defined as the maximum amount that can be lost with this acceptable probability. It corresponds to the left side of the cumulative distribution function (also called the "tail"). Between two assets, a risk-averse investor will certainly prefer the risk with a lower VaR.

The use of VaR as a prudential recommendation is widely referred to by insurance portfolios managers and banks since it was put forward by the Basle committee[12] as a minimum safety requirement. Consider a risk formalised by a random variable X. The cumulative function F_X is defined by $F_X(x) = P[X < x]$. Conversely, for a given probability level p, the p-percentile of X, $Q_X(p)$ is the number such that X does not exceed $Q_X(p)$ with probability $p : P[X < Q_X(p)] = p$. VaR_p is the absolute value of this percentile, it is usually defined as $VaR_p(X) = |Q_X(p)|$. In general, the VaR is meant to be a loss, hence a negative value, in practice however the absolute value is directly given as being the VaR (a portfolio with a $VaR_{0.01}$ of €1m means that $-€10^6$ may be its payoff with probability 1%). Because we are interested in insurance where risks concern losses in this chapter, the percentiles can be considered as being always negative and in order to simplify notations in the following, we shall define $VaR_p(X) = -Q_X(p)$. Notice however that, for distributions with positive percentiles, this definition may be confusing.

Definition of the value at risk of a risk X at the confidence level p

$$P[X < -VaR_p(X)] = p$$

VaR, or percentiles, are easy-to-use parameters to compare two risks, probability theory and statistics achieves this by characterising stochastic dominance (i.e. an incomplete order on random variables), independently of agents' behaviour. We shall come back to VaR in the third part of the book as a static instrument to manage risks (Chapter 12). In this chapter, it will be used as a mathematical instrument to measure risk increases and define stochastic dominance.

6.5 STOCHASTIC DOMINANCE

First-order stochastic dominance (SD_1) is a rather natural way to compare two risks (random variables): the first one "first-order stochastic dominates" the second one if and only if the cumulative function of the second is higher or at least equal to the cumulative function of the first (Figure 6.2).

SD_1 *X first-order stochastic dominates Y, $X\ SD_1\ Y$, iff* $\forall x \in R,\ F_X(x) \leq F_Y(x)$

This definition can be expressed equivalently in terms of VaR, indeed, as can be seen in Figure 6.2, first-order stochastic dominance means that, for the same probability level, the value that risks to be lost with X is less than the value that can be lost with Y. Or, because VaR is a positive number in our notations, the VaR of X is less than the VaR of Y:

$$X SD_1 Y \quad \Leftrightarrow \quad \forall p \in [0,\ 1],\ VaR_p(X) \leq VaR_p(Y)$$

First-order stochastic dominance is a good characterisation of the risk to lose something. It doesn't take all the risk range into account, though, for symmetric distributions. Furthermore,

[12] Basle Committee on Banking Supervision, *Amendment to the Capital Accord to Integrate Market Risk*, Basle, January 1996.

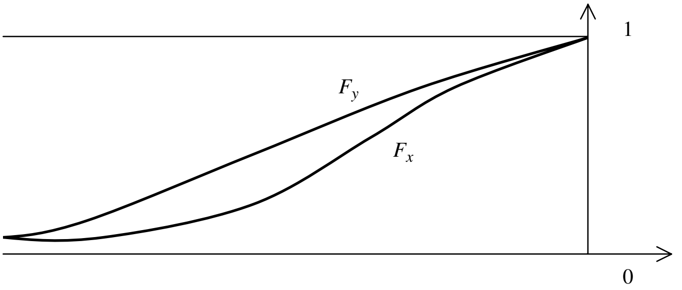


Figure 6.2 First-order stochastic dominance

another drawback of SD_1 is that there are few couples of random variables that satisfy it. In most cases, the inequality is satisfied up to a certain probability level (or up to a certain VaR level), but it is not for higher values any more. Indeed, most cumulative functions cross at some point, and may cross more than once.

If first-order stochastic dominance is not satisfied everywhere, then it may be the case that it is satisfied on average. This is the idea of second-order stochastic dominance (SD_2):

SD_2 *X second-order stochastic dominates Y, X SD_2 Y, iff* $\forall x \in R, \int_{-\infty}^{0} F_X(x)\mathrm{d}x \leq \int_{-\infty}^{0} F_Y(x)\mathrm{d}x$

Equivalently, in terms of VaR:

$$X\mathrm{SD}_2 Y \quad \Leftrightarrow \quad \forall p \in [0,\ 1], \int_0^1 VaR_p(X)\mathrm{d}p \leq \int_0^1 VaR_p(Y)\mathrm{d}p$$

However, even SD_2 is not a complete order on random variables, even though more risks can be compared, furthermore it has the same drawbacks as SD_1: it is symmetric, in the sense that if one risks to lose more with Y than with X on average, one can earn less on average. So stochastic dominance is not always a way to choose between two risks, unless one is concentrated on losses (or on gains). The same problem arises with further stochastic dominance orders, only they do allow us to compare more risks each time. SD_2 can be satisfied up to a certain probability (or value) level, but not after. In most cases it will be satisfied if the average is taken over an interval. Third-order stochastic dominance, for instance, says that X SD_3 Y iff, on average over intervals, the mean of $VaR(X)$ is less than the mean of $VaR(Y)$.

Still, at whatever order, stochastic dominance is not a criterion to choose between two risks, it is a characterisation of an incomplete order on random variables. In order to choose, some characterisation of the possible gains must be put in balance with risks of losses. A way to avoid the problem is to concentrate on risks with the same mean value. This way implicitly introduces a behaviour in the face of risks, it is similar to Markowitz's double criterion: for the same mean value, the risk with the inferior variance is preferred. In stochastic dominance, variance is not the referred measure of risk, VaR is used instead.

Expected utility is a global criterion that has the following interesting property: for any measure of risk (variance, VaR, and for all stochastic dominance definitions), a concave utility function characterises a preference for the less risky random variable (with the same mean). There is a drawback to this general property: expected utility cannot characterise aversion to risk increases, as we shall see now.

6.6 AVERSION TO RISK INCREASES

In this section, VaR will be the parameter to measure risks (remember that in our notations: $VaR_p(X) = -Q_X(p)$, even if $Q_X(p)$ is positive). In order to get around the background risk problem where two risks may interact, we shall characterise risk increases (or decreases). We shall refer to stochastic dominances, as we defined them in terms of VaR, and to other less known dominance concepts. To each characterisation of a risk increase, according to the different dominance concepts, we shall define a behaviour that shows aversion to such a risk increase. We shall however limit ourselves to risk increases that preserve the mean, so that two random variables with the same mean can be compared by their VaR, which is a relevant risk measure for losses.

Let X and Y be two random variables representing two risks, and assume $E(X) = E(Y)$ in the following. We shall say that Y is a Mean Preserving Risk Increase (MPRI) of X, if Y is "more risky" than X according to some risk measure.

The first definition of MPRI corresponds to first-order stochastic dominance. We shall say that Y is a weak mean preserving risk increase of X, if and only if X SD_1 Y, or equivalently, if $VaR(X)$ is less than $VaR(Y)$ for any confidence level:

$$Y \text{ is a weak MPRI of } X \quad \Leftrightarrow \quad \forall p \in [0, 1], VaR_p(X) \leq VaR_p(Y)$$

In the special case where X is not random and $X = E(X) = E(Y)$, any random variable Y is a weak MPRI of X. In this case we say that an agent who prefers X to Y is weakly risk averse (Arrow, 1971). This notion of risk aversion corresponds to a special case of aversion to weak MPRI. As we have seen, in EU models, weak risk aversion is characterised by concave utility functions.

The second characterisation of MPRI corresponds to second-order stochastic dominance. We shall say that Y is a strong mean preserving risk increase of X, if and only if X SD_2 Y, or equivalently, iff on average $VaR(X)$ is less than $VaR(Y)$:

$$Y \text{ is a strong MPRI of } X \quad \Leftrightarrow \quad \forall p \in [0, 1], \int_0^1 VaR_p(X)\mathrm{d}p \leq \int_0^1 VaR_p(Y)\mathrm{d}p$$

The corresponding strong risk aversion characterises a decision-maker who prefers a risk X to Y if X second-order stochastically dominates Y. Rothschild and Stiglitz (1970) have shown that in EU models, any concave utility function will show a preference of X over Y if Y is a strong MPRI of X. In their paper, a strong MPRI was characterised by $Y = X + Z$, with $E(Z) = 0$. The authors show in a first step that such a risk increase is equivalent to second-order stochastic dominance (and hence, to strong MPRI in our terms). Otherwise stated, in EU models, weak and strong risk aversion are equivalent.

A third characterisation of MPRI is less strong than the preceding one. It takes into account the possibility that X and Z in the sum $Y = X + Z$ may hedge each other. We have seen in Chapter 2 that, indeed, this important property may contradict the independence axiom (and then the EU models) because a risk-averse agent may consider that $X + Z$ is less risky than X if Z and X are negatively correlated. This suggested to Quiggin (1991) to consider only monotonic risk increases as being "more risky". Indeed, if X and Z are comonotonic we have seen that their correlation must be positive for any additive probability measure (Chateauneuf *et al.*, 1994). We shall say that Y is a monotone MPRI of X if and only if there exists a zero mean Z, comonotonic with X and such that $Y = X + Z$.

Landsberger and Meilijson (1994) have shown that Quiggin's characterisation is equivalent to Bickel and Lehman's (1979) dispersion order, which is well known in statistics. This order is defined in terms of percentiles, or equivalently, in terms of VaR: Y is more dispersed than X iff $VaR_p(Y) - VaR_p(X)$ is a non-decreasing function of the confidence level p. As a conclusion, we can state:

$$Y \text{ is a monotone MPRI of } X \Leftrightarrow \quad VaR_p(Y) - VaR_p(X) \text{ is non-decreasing in } p \in [0, 1]$$

Aversion to monotone MPRI is typical of an agent who prefers a hedged position to a risk without hedge. This is indeed a requirement for buying insurance. However, again, EU models do not differentiate this risk aversion from the others, all characterised by concave utility functions.

As we have underlined before, what matters in insurance is not risk in general, which includes gains or losses, but the risk of losses. This is indeed the reason why we choose VaR over variance as a risk measure. Hence, a decision-maker may be interested in risk increases that add to losses, instead of a general risk increase. Jewitt (1989) introduced an increase in risks that models behaviours in a context of partial insurance, he called it location-independent risk. This MPRI is characterised by:

$$\forall p \in [0, 1], \quad \int_{-\infty}^{VaR_p(X)} F_X(x)\mathrm{d}x \leq \int_{-\infty}^{VaR_p(Y)} F_Y(x)\mathrm{d}x$$

Because increases of losses concern the left-hand side of the cumulative distribution function, we could say that they are "left" risk increases. We shall follow Chateauneuf *et al.* (1997), who have proposed "left monotone" instead of Jewitt's appellation. They show that the previous definition is equivalent to the following:

$$Y \text{ is a left monotone MPRI of } X \Leftrightarrow \int_0^1 [VaR_p(Y) - VaR_p(Y)]\mathrm{d}p \text{ is non-decreasing in } p \in [0, 1]$$

Aversion for such risk increases is typical of an agent who wants to avoid increases in losses on average. It may be called aversion to increases in losses or left monotone risk aversion. Such risk aversion is consistent with the Basle committee's recommendation, which insists on banks maintaining their portfolios' VaR above a given level. Obviously, in practice, this is to be understood as taking a higher VaR level as an objective and managing the portfolios so that this level is obtained on average and keeps above the recommended one.

However, here again and for the same reason as before, EU models cannot distinguish between this risk aversion and others because they are all characterised by concave utilities only. This is why in order to pursue finer risk aversion analysis, more general models are called for. In the RDEU models, the utility function u and the probability deformation function ϕ play a role. Thanks to these two components, the different risk aversions can be distinguished. However, for the first two no necessary and sufficient conditions are known,[13] so we give one in Yaari's model in which there is only the deformation function ϕ:

[13] See Chateauneuf and Cohen (1994), who show that no characterisations are available but give different sufficient conditions (making the above one precise), they do not necessarily imply u concave.

Weak risk aversion in Yaari's model, $\phi(p) \leq p$
Monotone risk aversion in Yaari's model, $\phi(p) \leq p$[14]
Left monotone risk aversion in RDEU, $\frac{1-\phi(p)}{1-p}$ is increasing and u is concave
Strong risk aversion in RDEU, ϕ is convex and u is concave

Chateauneuf *et al.* (1994) present all results on risk increases, their properties, and aversion to risk increases. They have shown the following implications:

$$\text{strong risk aversion} \Rightarrow \text{left monotone risk aversion}$$
$$\Rightarrow \text{monotone risk aversion} \Rightarrow \text{weak risk aversion}$$

6.7 ASYMMETRIC INFORMATION: MORAL HAZARD AND ADVERSE SELECTION

The main difficulties encountered when calculating a risk premium come from the probability distribution that is referred to. We have assumed up to now that this distribution was well known and unique. In fact, insurers face many different distributions according to their clients' risk classes; we have evoked the problem in Chapter 5. How can the insurer know if a potential client belongs to one class or to another? A more difficult question is: how can it be known whether the client will behave in the same way as when it had been classified, once insured, or change its behaviour?

We see that the insurer faces two risks related to information, besides the risk to be insured proper. Because hidden information may misguide the insurees' selection, the theory of insurance has called it "adverse selection". The other informational problem is about the hidden behaviour of the insurees in the face of the risk they want to insure, this is called "moral hazard".

A huge amount of economic literature has been devoted to these problems of asymmetric information (Rothschild and Stiglitz, 1970). It developed concepts to address problems such as: imperfect and incomplete information, incentive compatible contracts, the agency and principal–agent relationships, etc., which found applications in industrial economics and are applied to the insurance industry.

A general theory of incentives developed the links between the findings of different approaches to imperfect and incomplete information in a two-sided (principal and agent) relationship (Laffont and Martimort, 2002). The principal and the agent(s) cannot rely on a one-shot relationship to obtain efficient contracts, but they can almost achieve it through repeated relationships (Osborne and Rubinstein, 1994). As we shall see, using past results in repeated insurance contracts is indeed the way to achieve a better selection among risk classes, and to get closer to optimal risk premiums when hidden actions prevail.

In order to illustrate the two problems and give some hints about tentative solutions, let us concentrate on the non-obligatory part of automobile insurance: the casualties concerning the owner's vehicle. Assume there are only two types of drivers as a relevant simplification: prudent drivers (Bears) and imprudent ones (Bulls). Bulls have a higher probability of accident than Bears, as a consequence the insurance company would propose a higher premium to Bulls than to Bears if they could be distinguished. If the insurer proposes a unique

[14] Here again, because there is only one parameter to play on, Yaari's model is as poor as the EU model to compare the two types of risk aversion. Notice that $\phi(p) \leq p$ is implied by ϕ convex.

contract based on a premium calculated with the mean damage of the whole population, Bears will pay too much and Bulls not enough as compared to the fair premium they deserve. There are two consequences to such an undiscriminating premium. The first one is illustrative of the adverse selection problem, the second one is typical of moral hazard.

As regards the selection problem, Bears will be reluctant to insure at that premium level, while Bulls, on the contrary, will have a tendency to rush in. As a result, the insurance's portfolio will have a risk with a mean value higher than the one calculated over the initial population and the proposed risk premium will be too low to compensate that risk.

A solution to this problem is to write contracts that include a clause that incites Bears to indicate that they are prudent drivers (or make Bulls reveal that they drive hazardously). Such contracts will provide an autoselection incentive. How can a contract be more attractive to Bears than to Bulls so that its premium is fair to the client who buys it? Offering a list of contracts with partial casualties coverage is a way to deal with this selection problem. Stop-loss contracts, for example, propose to insure damages up a given amount of money. A Value at Risk for a given confidence level is a way to determine the stop-loss level. The less prudent the driving, the higher the potential damages, and the higher the remaining amount left at the charge of the insuree. Assume two stop-loss levels are proposed, based on the distribution on the whole population among which the insurance company cannot discriminate: high and low. The high stop-loss level (casualties over 1000, say) is designed to correspond to Bears with a probability of accident at most equal to 1%, and the low level (500, say) would correspond to Bulls with a probability of accident that can go up to 5%. If the probability of accident is estimated over the whole population, say 3%, the fair premium associated with the high level stop-loss will be lower than the one going with the low one if the contract includes a stop-loss clause. This should incite Bears to choose the high level contract and vice versa. In contrast, with the same premium for both contracts, the Bulls would save on the premium. In order to avoid this contrary effect and to incite drivers to reveal if they are prudent or not by the contract they choose, the insurance premiums must be adapted to the corresponding risk class revealed by the type of contracts chosen. In practice, such types of contracts are more common to insure industrial risks than for automobile insurance.

More common for automobile insurance are two other types of contracts that present similar incentives to reveal hidden information effects. Contracts with deductibles deduce a fixed-in-advance amount of money from the reimbursements. So, if the damage is high, the deductible is relatively less important. Assume again two deductible levels are proposed, high and low. This time, Bears may be inclined towards the high deductible and Bulls may prefer the low one. This is because high deductibles amount to an overall damage reimbursement lower for the insurer, and hence should define a lower risk premium for the high deductible contract. However, a Bull may feel that with his high probability level of accident, he gets a better deal with the low premium one. In that case, the high cost left at their charge with the high deductible is an incentive for Bulls to choose the low deductible, and thus reveal their class risk. The converse incentive works for the prudent drivers.

When contracts with full insurance are repeated, driving behaviours may change accordingly and the second problem may arise: the way the driving is modified as a consequence of being insured is a hidden action, so that moral hazard may prevail. Bears will feel an incentive to drive in a less prudent way than they used to (given that prudent driving may incur some costs, in terms of time for example) because they are not rewarded for their ("good") behaviour. Then, they may become bully so as to justify the premium they pay, *ex* *post*, given that it is unfair to them. The result on the insurance company's portfolio risk is the same as the previous one: it increases as all drivers become Bulls and the probability of accident increases.

In order to reduce this behaviour and incite drivers to stay within their risk class, retaliation may be a threat. The insurance company keeps records of private results over past years' contracts, so that each accident amounts to a penalty on next year's individual[15] risk premium and/or the offered choice of deductible levels.

Another way to deal with the asymmetric information problems is to promote co-insurance: the insurance company only takes in charge a proportional share of the damages, the complement is left as a risk to deal with by the insured contract. Here again, different proportions may be proposed to help and select among risk classes, while the proportion of the risk that is not insured is an incentive for the insuree to take proper preventive measures. We shall come back to the comparison between the co-insurance and contracts with deductibles in Chapter 8, which gives but a rough idea of the many ways to write contracts depending on the risks.

The theoretic elements presented in this section form the foundations of the insurance contracts' design. Such contracts are adapted to situations of risk, given they all refer to a given probability distribution. As far as situations of uncertainty are concerned, such instruments are not adapted, either to hedge the insurees' risks or to be managed in an insurance portfolio according to the large numbers and the mutuality principles. Other instruments exist to hedge risks in situations of uncertainty, they are based on the analysis and the theoretical approach presented in the following chapter.

[15] This individualisation of risk premiums is considered as a prejudice in some countries and then is forbidden by law.

7
Marketed Risks

At the origin of the concept of risk, there were no such things as "probability" measures, however insurance and finance did trade risky contracts. Trades yield prices and prices measure risks. A theory for such trades is presented in a simple model and explains how prices do measure risks. Then valuation is developed in order to address insurance and financial hedging contracts. In Chapter 11, we shall come back to the models presented here and we shall extend them to a dynamic setting. The static approach chosen here is simpler and is sufficient to introduce the basic concepts relevant for the valuation of uncertainties and for management of the insurance companies' risky portfolios.

7.1 A GENERAL THEORY OF RISK MEASUREMENT

The aim of this chapter is to study risks that cannot directly be hedged by classical insurance contracts. We introduce a valuation model that is consistent with the general competitive equilibrium model we have seen in Chapter 3. The risk measurement theory we shall present is quite general: it exhibits fundamental risk features in the insurance and finance domains and is flexible enough to be adapted to further specific assumptions. For instance, the expected utility-based valuations we have seen in the last chapter can be obtained as a very special case of the model presented here: assuming there is only one agent (i.e. all agents are described the same way), knowing the risk probability distribution and satisfying the expected utility axioms (Borch, 1962; Mossin, 1966). Other extensions may be technical (set of states of nature, number of agents, etc.) and/or conceptual: dynamics, in Part III, shall refer to other extensions. The main generality of the model here is that it embraces the two sides in this part's distinction: situations of risk vs situations of uncertainty. Each situation has its own requirements in terms of formalisation; we need to distinguish them clearly so as to be able to understand the relevance and the limitations of application domains in the next chapter.

In contrast with the general equilibrium model of contingent commodity trades, in this model risky contracts deliver money: they pay off. If models are reduced to only one commodity, there is not much difference between the two models,[1] obviously. However, this is a very special case and differences are important otherwise. Conceptually, risky monetary contracts are closer to most actual contracts. Risky assets as well as insurance contracts pay off in money and not in physical goods. Money is then used to trade commodities. Trading contingent monetary contracts synthesises two real transactions: trading risks and trading commodities. In the model below, at a first date risks are traded under the form of

[1] This explains why the two approaches are sometimes mixed up in the economic literature. For instance, the one we present below is often called the Arrow–Debreu model, when it is due to Arrow and is nowhere to be seen in Debreu's works.

contracts with payoffs in (future) money, and commodities are traded at a second date, once an uncertain state is realised. This corresponds to the way an insurance contract proceeds: it pays off in money if the damage is observed, and then the insured agent goes to market with the money in order to repair or to replace the damage. Such monetary contract trades are much easier to organise than contingent commodities ones; this is an advantage that may explain why the latter are much less developed than the former. We shall see that much less marketed assets are necessary to offer complete hedging against all (represented) states.

7.1.1 Risk Valuation

As we announced, this is still a static model: a financial asset y is traded at an initial date 0 and pays off at a final date 1. Payoffs are in money, or income unit,[2] with which desired commodities are bought. We shall be more precise on the agents' preferences later; for the moment, suffice it to say that they prefer more money to less. We assume a finite number of assets:[3]

$$Y = \{y_1, \ldots, y_M\}$$

They can be directly traded, let us call them the "marketed assets".

Asset payoffs are contingent to (future) states of the world that can obtain at date 1:

$$S = \{s_1, \ldots, s_N\}$$

they are obviously uncertain as seen from date 0. We assume S finite for the moment, and shall introduce the infinite case at the end of this section. Formally, a marketed asset is a real non-negative function $y : S \to R_+$, where $y(s)$ is the asset's payoff in state s. We assume assets to be non-negative as in real life; negative payoffs can be obtained by selling an asset short (i.e. writing a contract that ascertains that the buyer will be paid the asset's payoffs, even though the seller does not own it).

The asset price at date 0 is $q(y)$: this is the money amount that is paid or received in exchange for the asset y. An asset portfolio $\theta = \{\theta(y_1), \ldots, \theta(y_M)\}$ represents a list of asset quantities[4] that can be detained by an agent: $\theta(y) > 0$ if the agent bought y (long position) and $\theta(y) < 0$ if the agent sold y (short position). The portfolio's formation cost at date 0 is:

$$K(\theta) = \sum_{y \in Y} \theta(y) q(y)$$

The total payoff (positive or negative) of portfolio θ at date 1 in state s is:

$$\Theta(s) = \sum_{y \in Y} \theta(y) y(s)$$

[2] We shall make precise what is meant by "money" and by income unit in the following, it covers the usual sense, e.g. \$, £ etc., at a certain date.

[3] This assumption can be weakened, however notations and mathematics are more involved with an infinite discrete set.

[4] Assumed continuous as an approximation of very small shares.

Here, we have a real function $\Theta : S \to R$. Such payoffs define a financial asset $\Theta : S \to R$,[5] which is not directly marketed but may be obtained indirectly by trading on the assets market. Let us call such an asset: a "marketable" asset, i.e. an asset that can be obtained by a replication portfolio such as θ. A marketed asset y is obviously a special case of a marketable asset, i.e. a portfolio with only one $\theta(y) = 1$ and all others $\theta(y_i) = 0$. A marketable asset can be replicated by several portfolios, in general, and it can be the case that a marketed asset can be replicated by a portfolio formed of other marketed assets.

Let us call *SpanY* the set spanned by the marketed assets in Y, i.e. the set of finite linear combinations of assets in Y. *SpanY* is a subvector space of R^N. The subset of marketed assets with linearly independent payoff vectors form the vector basis of *SpanY*.

Let us now introduce some restrictions on possible trades on such asset markets. The first one excludes gains at period 0 which are not justified by risks at period 1.

No arbitrage 1 (unique price) *If two marketable assets Θ and Θ' have the same payoffs in each state (i.e. the two random variables are equal), then any two portfolios θ and θ' replicating the marketable assets have the same formation cost:*

$$\forall \Theta, \Theta' \in SpanY, \forall \theta, \theta' \left[\Theta = \sum_{y \in Y} \theta(y) y = \sum_{y \in Y} \theta'(y) y = \Theta' \right] \Rightarrow K(\theta) = K(\theta') \qquad \text{(NA1)}$$

Without this assumption, it would be possible to make money at no cost by buying and short-selling two portfolios replicating the same marketable asset. Assume the first one's formation cost is more than the second one: then sell short the first one and buy the second one with the money. You have made the cost difference as a profit and, at date 1 in any state, what you have to pay from your short position is exactly compensated by what you receive from your long one. If market prices did not satisfy NA1, you could repeat this operation infinitely and get an infinite profit! Furthermore, all agents will do the same and the net asset demand would not be zero. So NA1 is a necessary condition for equilibrium prices to exist.

Are there such arbitrage opportunities in real life markets? In real life, trades take time, so there are periods where prices are adjusting to new demands and supplies, bid and ask prices are not the same[6] and there are arbitrage opportunities. Arbitragers react immediately to grasp these short-period opportunities, they make their living by adding hundreds of tiny cost differences a day. In real life, there are always some remaining risks, because states of the world are not precisely defined and future prices can vary outside the forecasted range. Furthermore, the immediate profit itself is not risk-proof either because the arbitrage trading may take some time and prices may vary during that time. If we are not interested in the short-periods dynamic, we can mimic the discrete one-period (static) model features and then conclude that NA1 is satisfied in real-life markets.

Another condition excludes other arbitrage opportunities, those seizing possible gains (and no losses) in period 1 at no cost in period 0.

[5] Notice these assets can be negative while a marketed asset cannot.

[6] In some markets, different bid and ask prices prevail, while they are assumed to be equal in this model. But an equilibrium bid–ask spread is the risk premium taken by dealers to pay for the risk taken during the time period necessary to conclude the deals.

No arbitrage 2 (no free lunch) *If a marketable asset Θ has non-negative payoffs in all states of the world and a strictly positive one in at least one state, then any portfolio replicating Θ has a strictly positive formation cost:*

$$\forall \Theta \in SpanY,\ \forall \theta \,\text{s.t.}\ \sum_{y\in Y} \theta(y)y = \Theta,\ [\Theta \geq 0,\ \Theta \neq 0 \Rightarrow K(\theta) > 0] \qquad \text{(NA2)}$$

Take a portfolio with positive payoffs and at least one non-negative payoff in some state s. If it has a zero formation cost (short sales compensate long positions), it would cost nothing to make a possible infinite gain if state s obtains and there would be an infinite demand for assets forming this portfolio at zero cost. At equilibrium prices, such opportunities cannot happen.

Theorem (price extension) *Under NA1 and NA2, the pricing functional* $\bar{q} : SpanY \rightarrow R$ *is a non-negative linear form on SpanY and its restriction to Y is q: SpanY* $\rightarrow R$*. Then there exists a non-negative real function* $\gamma : S \rightarrow R_+$*, such that* $\forall \Theta \in SpanY$, $\bar{q}(\Theta) = \sum_{s\in S} \gamma(s)\Theta(s)$.

The theorem is a direct application of a first-year course in linear algebra, and extends to more general vector spaces.[7]

From the theorem we get a value function for risks, which takes the form of a weighted average of payoffs. This is a fundamental finance valuation formula (i.e. all finance valuation formulas can be written under this form as we shall see here and in Part III). Indeed, assume there exists a riskless marketed (or marketable) asset y_0 paying one monetary unit (without loss of generality) in each state at price $q(y_0)$. It defines a riskless interest rate: $R_0 = \frac{1-q(y_0)}{q(y_0)}$.

From the theorem we have: $\bar{q}(y_0) = \sum_{s\in S} \gamma(s)$. Then let, $\forall s \in S,\ \upsilon(s) = \frac{\gamma(s)}{\sum_{s'\in S} \gamma(s')}$, where υ defines a non-negative additive measure, i.e. a (mathematical) probability measure, on $(S, 2^S)$ as: $\forall s \in S, 0 \leq \upsilon(s) \leq 1$ and $\sum_{s\in S} \upsilon(s) = 1$.

We can write for any marketable asset Θ:

$$\forall \Theta \in SpanY,\ \bar{q}(\Theta) = \frac{1}{1+R_0} \sum_{s\in S} \upsilon(s)\Theta(s) = \frac{1}{1+R_0} E_\upsilon(\Theta)$$

where E_υ is the expectation with respect to probability measure υ.

$\bar{q}(\Theta) = \frac{1}{1+R_0} E_\upsilon(\Theta)$ is the Fundamental Finance Formula (FFF) for risk valuation. The FFF says: any marketable asset price is the expectation of its discounted payoffs. The expectation is calculated with respect to a probability measure obtained from market prices. It must be remarked that probabilities are not founded on a prior likelihood or frequency-based uncertainty measure. Indeed, as a deduction from market prices, it reflects agents' preferences and their initial endowments and subjective expectations, as we shall see below. In that sense, it is a collective risk measure. In a purely financial approach, it is obtained from observed market prices without reference to agents' behaviours. The only requirement is that they all prefer more money to less, this implies the no arbitrage opportunity conditions at equilibrium. This requirement excludes agents who would accept to take risks without any counterpart, just for the fun of betting. Such agents would accept Dutch books, i.e. bets in

---

[7] See Duffie (1988), for instance.

which they are bound to lose with certainty, a basic violation of rational behaviour in front of uncertainty.[8]

How can we interpret this probability measure? Two versions are usually referred to.[9]

1. The "risk-neutral" probability. An expected utility maximiser is risk neutral if it values assets by their discounted payoffs' mathematical expectations with respect to some prior objective probability distribution. If all agents behave this way, then market prices would be identical to individual common valuation. As a consequence the probability distribution in the FFF would be the common prior. Remark that, such an assumption on agents' behaviours and beliefs is inconsistent with the existence of a market for risks (a risk-neutral agent has no incentive to trade risks!). Furthermore, expected utility maximisation and existence of priors are by no means required by the general equilibrium model, nor by the one we shall see below. To this interpretation, one could prefer the less outstretched one following.
2. The risk-adjusted probability measure. Risky assets prices include a risk premium (Chapter 6), i.e. a difference between a "riskless price" and a price for specific risk. A common prior, μ, is assumed to be known objectively on the set of states. The riskless price does not take into account the risk of payoff Θ and is given by $\frac{E_\mu(\Theta)}{1+R_0}$. The risk premium is $\frac{1}{1+R_0}[E_\mu(\Theta) - E_v(\Theta)]$. If the market does not allow Dutch books, then the risk premium is positive for assets that increase the overall risk in the economy.

The FFF formula is important from two points of view:

- On the one hand, all assets in a financial market are valued by their payoffs' mathematical expectation with respect to the *same* probability measure.
- On the other hand, the FFF separates the price of time (here for a unique time period), represented by the riskless discount rate, or more precisely by $(1+R_0)^{-1}$, and the price of uncertainty, represented by the risk-adjusted measure, or more precisely by the expectation with respect to it. This avoids many confusions when valuing future risks (or projects).

For application purposes, the analysis extends to conditions under which the risk-adjusted measure is unique.

Definition (complete market[10]condition) *The set of marketed assets is a complete market if any risk contingent on states in S (any payoffs scheme) can be replicated by a portfolio of marketed assets:*

$$\forall \Theta :\to R, \exists \theta \text{ s.t. } \sum_{y \in Y} \theta(y) y = \Theta$$

Under the complete market condition, the risk-adjusted measure v is unique. Indeed, within the linear algebra context of this model, the complete market condition means that there

[8] De Finetti (1930), see also Diecidue and Wakker (2002) and Kast and Lapied (2003).
[9] Both interpretations assume that agents all refer to the same probability distribution on the payoffs' state, a prior. The empirical distribution estimated from past returns is usually invoked in textbooks.
[10] Here and in the following, the "market" means the asset structure that models an institution such as, say, the Chicago Board of Trades.

exists a subset of Y formed of N linearly independent vectors (this necessitates that $\#Y = M \geq N = \#S$). Otherwise stated, Y contains a basis of the vector space R^N representing risky payoffs contingent on states.

The complete market condition[11] is basic for arbitrage valuation in financial markets. If marketed asset prices are known and if the market is complete, then *the* risk-adjusted measure can be derived easily from market prices. Once obtained, the risk-adjusted measure can be used to calculate the price of any marketable asset by the FFF (e.g. the famous Black and Scholes (1973) derivative assets formula, which we present in Chapter 11).

In a complete market, any payoff scheme at date 0 (any risk) can be obtained by a linear combination (portfolio) of marketed asset payoffs: by replication. Then the risk is priced by the replicating portfolio's formation cost: by arbitrage.

Alternatively (and equivalently), the unique risk-adjusted probability measure is derived from marketed asset prices and then any risk can be priced by the FFF. Calculations are simpler with the FFF than with the replication–arbitrage method, however we shall see in Chapters 11 and 12 that in some dynamic cases, there is no simple FFF available and we must resort to replication.

With either of the two methods, we can assign a price to any risk. Is such a "price" the same thing as the one defined by microeconomic theory? Microeconomics defines prices as the result of trades, but among marketable assets some are traded and some are not (although they could be, indirectly). Take a marketed asset: it has a price at which it is traded. Some marketed assets can be replicated by a portfolio of other marketed assets ("redundant assets"), then they can be re-priced by the replication–arbitrage method or by the FFF: the price found should be the same as the trading price. Hence "price" is the same thing in arbitrage valuation and in microeconomics, for marketed assets. Now a marketable asset is not always traded, but assume there is a demand for it, e.g. because the marketable asset would come in handy to hedge some positions. A dealer can make up the replicating portfolio and sell shares of this portfolio under the marketable asset's name (e.g. a call option). The price at which it will be sold must be the arbitrage price. Hence, the arbitrage "price" is the same as the price at which the asset would be traded. There is an apparent paradox here: why sell an asset that can be replicated? It is due to the theoretical model neglecting transaction costs induced by replication. This is what happened in 1973 on the Chicago Board when call options were traded for the first time on a market. The opening prices were obtained by the famous Black and Scholes formula (which is a FFF). Prices didn't fluctuate much around the proposed prices. If they did, arbitragers, using some version of the FFF, were able to seize any arbitrage opportunities and drive back trading prices towards the no arbitrage value.

Up to now we stayed within a finite set of states of nature to avoid mathematical difficulties. Most financial models however rely on an infinite state space. The results still hold, let us look through one of the most famous models to see this. Assume that the set of states is a probability space (S, F, μ) and that a financial asset is represented by a random variable: $X : S \to R$. We shall restrict the set of assets to the set of random variables with a finite variance: $E(X) = \int_{s \in S} X(s) \mathrm{d}\mu(s)$ and $\sigma^2(X) = E(X^2) - [E(X)]^2$ are finite. This set is a well-known Hilbert space sharing many common properties with R^N; in particular, it is its own topological dual space.

[11] Or its adaptations to more general models, see below and Chapter 11.

The no arbitrage conditions are straightforward to reformulate in this setting and they yield the same properties as above, i.e. there exists a non-negative bounded function $\gamma : S \to R_+$ such that:

$$\forall \Theta \in SpanY,\ \bar{q}(\Theta) = \int_{s \in S} \gamma(s)\Theta(s)\mathrm{d}\mu(s)$$

Once a riskless interest rate has been defined as before, let a function g be defined by:

$$g(s) = \frac{\gamma(s)}{\int_{s' \in S} \gamma(s')\mathrm{d}\mu(s')}$$

Then:

$$\forall \Theta \in SpanY,\ \bar{q}(\Theta) = \frac{1}{1+R_0}\int_{s \in S} g(s)\Theta(s)\mathrm{d}\mu(s)$$

The Radon–Nikodym theorem defines an equivalent probability measure v, absolutely continuous with respect to μ and such that g is the Radon–Nikodym derivative of v with respect to $\mu : g = \mathrm{d}v/\mathrm{d}\mu$. Then the FFF obtains:

$$\forall \Theta \in SpanY,\ \bar{q}(\Theta) = \frac{1}{1+R_0}\int_{s \in S} \Theta(s)\mathrm{d}\ v(s) = \frac{1}{1+R_0}E_v(\Theta)$$

This formula is formally the same as in the finite case, v is the risk-adjusted (also called risk-neutral) probability. Here, it is more relevantly called the equivalent measure (meaning that μ and v assign probability zero to the same events). In contrast with the finite case, this distribution is defined with respect to the given probability distribution over the state space. The status of the probability distribution μ on the state space can be questioned. An empirical distribution is usually evoked as a candidate for μ, this assumes that all agents agree on the same estimates of future distribution parameters on the basis of observed past time series. Such an assumption is not always easy to accept, depending on the problem, as we shall see later on.

The complete market condition has the same interpretation as in the finite case: any payoff scheme can be replicated by the payoffs of a marketed asset portfolio. However, the condition is much more demanding and is a severe limitation to applications. The limitation will be turned around in Chapter 11 by the dynamic generalisation approach. In any case, the complete market condition yields uniqueness of the risk-adjusted measure.

In a normative approach, valuation has to be consistent with an equilibrium theory, let us show that arbitrage prices do.

7.1.2 Equilibrium Models of Risk Trades

We develop here a model proposed by Arrow (1953) as an alternative to the general equilibrium of contingent commodities model (now called Arrow–Debreu) that we have seen in Chapter 3. In this model, financial assets are traded at time 0 and pay at time 1 when commodities are traded after one state of nature is realised. There is no need to trade contingent contracts any more, the financial asset portfolio made up in period 0 by each agent is meant to relax its budget constraint in period 1 so as to hedge its needs. Then, instead of a unique budget constraint in period 0 in the Arrow–Debreu model, the Arrow model has as many budget constraints in period 1 as there are states of nature ($\#S = N$), for each agent.

Commodities are numbered $l = 1, \dots, L$ and consumers $i = 1, \dots, I$. Agent i's consumption in state s is $x_i(s) = (x_i^1(s), \dots, x_i^L(s))$ in set X_i and its consumption plan is $x_i = (x_i(s_1), \dots, x_i(s_N))$. Agent i's preferences are represented by a monotonous, continuous and strictly concave utility function u_i and its initial endowments in state s are: $w_i(s) = (w_i^1(s), \dots, w_i^L(s))$. Let $w_i = (w_i(s_1), \dots, w_i(s_N))$ be the vector of all initial endowments. Note that there are no endowments in period 0 in this model, even though it could accommodate some as well as a commodity market at the price of unworthy heavier notations.

Prices in state s are $p(s) = (p^1(s), \dots, p^L(s))$ and the price vector is $p = (p(s_1), \dots, p(s_N))$. Now, let $Y = \{y_1, \dots, y_M\}$ be M assets (i.e. payoffs contingent on states, $y : S \to R$). They are traded among agents at period 0 and let $q(y)$ be the price of asset y on the financial market. Agent i's portfolio is $\theta_i = \{\theta_i(y_1), \dots, \theta_i(y_M)\}$ and its budget set is:

$$B_i(p, q) = \{x \in X_i / \exists \theta \sum_{y \in Y} \theta(y) q(y) = 0,\ \forall s \in S : p(s)x(s) \leq p(s)w_i(s) + \sum_{y \in Y} \theta(y) y(s)\}$$

Agent i chooses a portfolio of financial assets and a commodity bundle in his budget set such that it maximises its utility:

$$\max u_i(x) \text{ s.t. } x \in B_i(p, q)$$

The solution defines agent i's demands for commodities and for assets: $d_i(p, q)$ and $\delta_i(p, q)$. We can define:

Perfect Foresight Equilibrium *In an economy defined by consumers' consumption sets, preferences and initial endowments:* $(X_i, u_i, w_i)_{i=1,\dots,I}$*; a Perfect Foresight Equilibrium (PFE) is defined by a price vector for commodities* $\hat{p}$ *and a price vector for assets* $\hat{q}$ *such that the financial asset market is cleared as well as every commodity market in all states:*

$$\sum_{i=1}^{I} d_i(\hat{p}, \hat{q}) = \sum_{i=1}^{I} w_i, \qquad \sum_{i=1}^{I} \delta_i(\hat{p}, \hat{q}) = 0$$

Remark that the two no arbitrage conditions are necessary for a PFE to exist in frictionless markets.

The complete market condition is satisfied as soon as there exist at least N linearly independent assets ($M \geq N$).

The model we just saw seems much simpler than the general equilibrium model of contingent commodities (less demanding in terms of market organisation: $L + N$ instead of $L \times N$). However, the following theorem asserts that they are equivalent in terms of allocations, if the contingent commodity markets are complete (i.e. there is a commodity contract contingent on each state of nature).

Theorem (equivalence) *In an economy:* $(X_i, u_i, w_i)_{i=1,\dots,I}$*, and under the no arbitrage and complete market conditions, a general equilibrium price vector* $p^* \in R^{L \times N}$ *exists if and only if there exists a perfect foresight equilibrium* $(\hat{p}, \hat{q}) \in R^{L \times N} \times R^M$*. Furthermore, they both yield the same commodity allocations to each agent and equilibrium prices are such that:*

$$\forall s \in S,\ \forall l \leq L : p^{l*}(s) = \gamma(s) \hat{p}^l(s)$$

where γ (the risk adjusted measure) is defined by:

$$\forall y \in Y, \hat{q}(y) = \sum_{s \in S} \gamma(s) y(s)$$

or by its normalisation by a riskless rate.

This result extends to models with an infinite set of states, as long as the commodity and the asset spaces have enough topological properties and under the limitations we underlined before. Equivalence concerns existence, uniqueness when it holds, and stability. It concerns the two welfare theorems seen in Chapter 3 because agents obtain the same allocations with one market organisation as with the other. Then, it is coherent to measure economic values by arbitrage prices in a Pareto efficient economy.

Risks formalised as financial assets[12] (instead of contingent commodities) are closer to the ones we know, let us see this further in detail.

7.2 APPLICATIONS TO RISK VALUATION

We are going to apply the general concepts presented above to two classes of financial risks that are traded: financial assets and insurance contracts. The dynamical aspects will be studied in Chapter 11, here we concentrate on the above results that have been obtained in static models. Both types of risks have prices that are observable and numerous past data exist, particularly for financial markets. Furthermore, these data satisfy the good properties to form samples, at least in financial markets, so they yield reliable estimates of probability distributions. This has enforced the reference to probability models by economists and financial theoreticians (situations of risk), even though the results apply to more general situations (of uncertainty), as we have seen. The objective distribution can be assumed to be known by all agents (at least theoretically), it is referred to as the empirical distribution.

7.2.1 Financial Assets

In many cases, it is relevant to distinguish between two sources of risk for marketed assets: the first one is called systematic risk, it corresponds to macroeconomic and politic risk factors that affect all traded assets. The second source of uncertainty is specific to each asset (specific risk, or idiosyncratic risk), e.g. risks inherent to a particular production firm. In order to formalise this distinction, let us introduce a particular asset: the market portfolio M. The market portfolio is formed of all traded assets (in practice, a relevant and sufficient portion of them, e.g. 500 or 1000 for the Standard & Poor's stock market indexes: SP_{500} and SP_{1000}). An asset weight in the portfolio is obtained by the ratio of the capital invested in the asset to the total capital invested in the market. It is reasonable to assume that specific risks are independent, at least for assets in different production sectors, then the market portfolio diversifies specific risks away (this results from applying the LLN). Thus, the market portfolio is assumed to represent the systematic risk.

[12] Which corresponds to the original concept.

Referring to the no arbitrage or PFE results, the risk-adjusted distribution v yields the price of an asset y as the mathematical expectation of its payoffs, discounted at the riskless interest rate R_0:

$$\forall y \in Y, q(y) = \frac{1}{1+R_0} E_v(y) \qquad \text{(FFF)}$$

Let us make the following change in variables:

$$v(s) = \frac{b - M(s)}{a} \mu(s)$$

where $M(s)$ is the market portfolio's payoff in state s, μ is the empirical distribution and

$$b = E_\mu(M) + a, \qquad a = \frac{Var_\mu(M)}{E_\mu(M) - (1+R_0)q(M)}$$

where $q(M)$ is the market portfolio's formation cost (price).

The FFF becomes:

$$\forall y \in Y, q(y) = \frac{1}{1+R_0} \left\{ E_\mu(y) - \frac{Cov_\mu(y, M)}{Var_\mu(M)} [E_\mu(M) - (1+R_0)q_M] \right\}$$

The last term in the formula, $[E_\mu(M) - (1+R_0)q_M]$, represents the market risk premium, it is the difference between the expected payoffs of the market portfolio and its capitalised (at period 1) price. If the covariance between the risk to be valued and the market portfolio is positive, the risk premium for this risk is positive, it is negative otherwise because the asset hedges the systematic risk.

If we rewrite the FFF in terms of rates of return:

$$R(y) = \frac{y - q(y)}{q(y)}$$

we find the CAPM formula:

$$E_\mu[R(y)] - R_0 = \beta_y \{E_\mu[R(M)] - R_0\}$$

where

$$\beta_y = \frac{Cov_\mu[R(y), R(M)]}{Var_\mu[R(M)]}$$

The Capital Asset Pricing Model (CAPM) is the first and most famous financial valuation model (Sharpe, 1964; Lintner, 1965). It was initially derived from Markowitz' mean–variance behaviour model and we shall present it according to the more traditional approach in Chapter 10. It measures the excess expected risky rate of return with respect to the riskless rate: The difference is the risk premium weighted by the asset's sensitivity to systematic risk (the β_y or "beta" coefficient). As could be expected, only the systematic risk is remunerated by the market risk premium and the asset's specific risk is taken into account by the sensitivity factor β_y.

7.2.2 Insurance

In a seminal paper, Borch (1962) proposed the following specifications for a reinsurance market:

- There exists only one commodity ($L = 1$).
- The marketed assets, $Y = \{y_1, \ldots, y_M\}$ are the so-called "Arrow's assets", there is one corresponding to each of the N states of nature:[13] $\forall s \in S,\ y_s(s) = 1,\ y_s(s') = 0,\ \forall s' \in S,\ s' \neq s$ (hence, the complete market condition is satisfied).
- Agents $i = 1, \ldots, I$, preferences are represented by twice-differentiable and strictly concave utility functions u_i and satisfy the expected utility axioms.[14]
- There exists an empirical probability distribution μ, on the set of states of nature, which is known by all agents.

Remark that the above assumptions have introduced some long-lasting confusions between several models in the literature:

- Because of only one commodity, a financial asset that pays in unit of value (money) by definition, can be confused with a contingent commodity (confusion between Arrow and Debreu's General Equilibrium model and Arrow's Perfect Foresight Equilibrium model).
- Because of the expected utility specification, the agents' preferences lose their general properties.
- Because of the reference to an empirical distribution, the model loses its general uncertainty feature to become relevant to risk situations only.

Within this simplified version of the PFE model, the valuation formula for financial assets is simply:

$$\forall s \in S,\ q(y_s) = \gamma(s)$$

The consumer budget set:

$$B_i(p, q) = \{x \in X_i \ / \ \exists \theta \sum_{s \in S} \theta(y_s)\gamma(s) = 0,\ \forall s \in S : \hat{p}(s)x(s) \leq \hat{p}(s)w_i(s) + \theta(y_s)\}$$

where the price of the unique commodity in state s is $\hat{p}(s)$.

Agent i maximises its expected utility:

$$\max \sum_{s \in S} \mu(s)u_i[x_i(s)] \quad \text{such that} \quad x_i \in B_i(p, q)$$

First-order necessary conditions yield:

$$\exists \lambda_i > 0, \forall s \in S : u_i'[x_i(s)] = \lambda_i \frac{\gamma(s)\hat{p}(s)}{\mu(s)}$$

Let us denote by $W(s)$ for the total amount of commodity in state s, then the clearing market conditions are:

$$\forall s \in S : \sum_{i=1}^{I} x_i^*(s) = \sum_{i=1}^{I} w_i(s) = W(s)$$

[13] This is easily generalised to the non-finite case.

[14] It was the only known behavioural model at that time, the results extend without difficulty to the RDEU model, see for instance Landsberger and Meilijson (1994).

Borch's paper is famous for having put forward the mutuality principle and expressing it as an equilibrium allocation property. We know that equilibrium is Pareto-optimal, the specific assumptions allow us to characterise this property in terms of risk sharing. The proof was rather intricate in Borch's paper, here is a simpler one.

If $p^*(s)$ is the General Equilibrium price of the commodity contingent on s, the equivalence theorem expresses the relationship between spot and future prices:

$$\forall s \in S : p^*(s) = \gamma(s)\hat{p}(s)$$

Utility functions properties and the first-order conditions above are easily shown to imply the existence of a non-increasing function f such that:

$$\forall s \in S : f[W(s)] = \frac{p^*(s)}{\mu(s)}$$

and

$$\forall s \in S : u_i'[x_i^*(s)] = \lambda_i f[W(s)]$$

Then, a consumption plan value can be written as:

$$\sum_{s \in S} p^*(s)x(s) = \sum_{s \in S} \mu(s) f[W(s)]x(s) = E_\mu[f(W)x]$$

We can state Borch's result: it says that the marginal rates of substitution for the insurer and the insured agents are equalised. It can be stated as a principle for insurance that we have evoked in Chapter 5, together with the LLN principle.

Mutuality theorem *If all agents in the economy equally share the total risk (the "would be" insurer's portfolio), then their risks in incomes are obtained by a (decreasing) function of the systematic risk only.*[15]

There are two sides to this result. On the one side, we see that the price functional (and hence the risk premium) only depends on systematic risk (i.e. the aggregate of the agents' risks). The second side is the mutuality property: at equilibrium, the only remaining risk is the aggregate W, so that the equilibrium allocation of any agent only depends on the aggregated risk (and of its own characteristics): individual risks have been diversified away under the complete market condition. In applications, the function f can be estimated from reinsurance contracts premiums, even though reinsurance markets are not complete.

Let us look at two special cases in order to get a clearer interpretation of the results.

Special case 1: no systematic risk

Consider a situation where

$$\forall s \in S : W(s) = W$$

This is the usual context for insurance economics. If individual risks are not correlated, the LLN implies that when the number of agents increases, the global loss for the insurance

[15] As was noted by Landsberger and Meilijson (1994) who generalised Borch's result to an RDEU model, all agent allocations are comonotonic at equilibrium. As a consequence, they cannot hedge the systematic risk.

company converges towards a constant. The company has a good approximation of the number of casualties, even though it ignores who is going to face a loss.

Assume furthermore that all agents face the same risk, e.g. consider agents to belong to the same risk profile, so that they all have the same expected wealth:

$$\forall i = 1, \ldots, I : \sum_{s \in S} \mu(s) w_i(s) = E_\mu(w_i) = \bar{w}$$

Then:

$$\forall s \in S : u_i'[x_i^*(s)] = \lambda_i f(W)$$

The equilibrium allocation of any agent is riskless (constant over states):

$$\forall s \in S : x_i^*(s) = x_i^*$$

Summing up budget constraints: $x_i = \sum_{s \in S} \mu(s) w_i(s) = \bar{w}$. Each agent obtains its expected wealth and perfect insurance is possible, thanks to risk diversification.

Example: Agent i's endowment is $w_i(s) = w$, with probability $1 - \pi$, and $w_i(s) = 0$, with probability π.[16] Global wealth is equally allocated among agents:

$$\forall i = 1, \ldots, I : x_i = (1 - \pi) w$$

Special case 2: equal systematic risk

In this case, if one among the agents is risk neutral (linear utility), it could be an insurer for the others as far as the systematic risk is concerned. The idiosyncratic risks can be mutually shared by the agents, as before. If all agents have the same characteristics, then the (common) specific risk is mutually shared among agents but the systematic risk cannot be insured.

The previous specifications clearly delineate two types of insurance profiles:

- When risks are the same for all agents and without common components (they are not correlated), then the independent large number principle plays fully and risks are diversified away. Incomes can be equally assessed among agents and each of them is allowed its expected return, i.e. risks disappear. Two phenomena can interfere with this result, however. We have mentioned in Chapter 6 that agents may modify their risk after being insured (moral hazard) and that different risks may induce adverse selection by the insurer if it lacks information about agents. Then, perfect diversification is seldom possible.
- When there is a systematic risk among the agents' population, agents must share it. If they all have identical characteristics as above, perfect insurance is not possible any more. This is the case with natural catastrophes, for example. In order to achieve better protection, different populations with independent (uncorrelated) systematic risks may be interested to cooperate: this is the principle of reinsurance. Indeed, the mutuality principle suggested by Borch's theorem above has been obtained in a reinsurance market model.

The perfect foresight model yields a theory from which it is possible to construct managing and valuation instruments for risks that escape the traditional insurance domain. Some are proposed in Chapter 8 next.

---

[16] π is the probability of a loss.

8

Management Instruments for Risk and for Uncertainty

In situations of risk, most risks can be hedged, reduced or written off by insurance contracts. The first section presents a discussion of the two main forms of such contracts and compares them. The second section concentrates on managing instruments built for situations of uncertainty or to manage insurers' portfolios. In the third section, we address the problem of valuing and hedging controversial risks, in a collective choice problem.

8.1 CHOOSING OPTIMAL INSURANCE

An insurance contract, when available, is the most usual instrument to manage risks. In any contract, terms and conditions are important and insurance companies propose a long list of such conditions "written in small type" of the non-insured risks, exclusions and avoidance clauses. We shall not enter here into details of the insurance contracting practice, but we shall enforce arguments that make it easier to distinguish two essential types of contracts: contracts with deductible and proportional, or co-insurance, contracts. We have evoked them in Chapter 6 as an answer to the asymmetric information problems.

In order to compare these two types of instruments, let us analyse them in the context of a simple model: let an individual's initial wealth be $W_i = W + X$, where W is assumed non-random for the moment and X represents a risky asset taking two possible values: $x_1 = L$ if there are no accidents and $x_2 = L - xL$ if one happens. The proportion x of the damaged wealth is a random variable in $[0, 1]$ with known probability distribution Π. Let $C(x)$ be the (random) reimbursement, it is fixed by the insurance contract's form, and let $PR(x) = (1 + \lambda)E[C(x)]$ be the insurance premium, where λ is the loading factor. The final wealth of the agent is $W_f = W + L - xL + C(x) - PR(x)$. We shall analyse and compare the agent's demand for the two following types of insurance contracts:

- Contracts with a given deductible $D < L$ under which no insurance is provided. In this case, $C(x) = 0$ if $xL \leq D$, with probability $1 - p$, and $C(x) = xL - D$ if $xL > D$ with probability p. Then $PR(x) = p[(1 + \lambda)E[xL - D/xL > D]$.
- Or contracts with proportional co-insurance where the insurer reimburses damages on the basis of a proportion α of the damage: $C(x) = \alpha xL$ and $(1 - \alpha)xL$ is left at the charge of the agent. The risk premium is then: $PR(x) = (1 + \lambda)\alpha E[\alpha xL]$.

The choice of the agent is an arbitrage between the risk left to its charge and the risk premium that it is ready to pay, both depend on (the control variable) D in the first case and α in the second one. If V is the agent's choice criterion, for each value of the control variable, the decision to insure or not will be based on the comparison between $V(W_i)$ and $V(W_f)$. The solution will be found by the usual mathematical programming under the budget constraints technique as in Chapter 4.

For instance, if V is the expected utility criterion, it can be shown that the existence of a solution is subject to the following condition in the case of the contract with deductible: $\lambda \leq (1-p)/p$. Let D^* be the optimal deductible solution, the result is: $D^* = 0$ if and only if $\lambda = 0$. Similarly, for co-insurance contracts, if α^* is the optimal proportion insured, full insurance ($\alpha^* = 1$) will be obtained, if and only if $\lambda = 0$.

More generally, the optimal insurance level is an increasing function of the degree of risk aversion of the agent. In the context of expected utility, the degree of risk aversion can be measured, for example, by the Arrow–Pratt index (for relatively low risk levels). This index is a function of the riskless wealth level (W), so that, if the risk aversion index is a non-increasing function of wealth, the optimal insured level decreases with wealth. The agents with the highest risk aversion level will be the ones to insure the most (the lower D^* or the higher α^*).

This result has been evoked by mutual insurance societies, which are not profit-seeking and have a loading factor close to 0 (0.2 on average, due to managing costs) to refuse to impose the principle of co-insurance or contracts with deductibles to the members of their friendly societies. However, they could not maintain this position because they had to reinsure with other private insurance societies and/or with reinsurance companies. These companies could not accept the moral hazard and adverse selection risks (Chapter 6) that the mutual societies would have faced in a competitive world, and they imposed that such instruments be used by their clients in order to (partially) avoid them.

Obviously, these results[1] are not robust to a change of the model. For instance, if the agent's criterion is a rank dependent expected utility, we have seen in Chapter 6 that risk aversion is described at the same time by the shape of the utility function (behaviour with respect to riskless wealth) and the shape of the probability transformation function (behaviour with respect to probabilities). The above results still hold for co-insurance contracts, however they have to be revised to a less precise condition in the case of contracts with deductibles: full insurance will be chosen as soon as the loading factor is not too high ($\lambda \leq \lambda^*$ depending on both the utility function and the deformation function). Then if the agent's behaviour is formalised by such more descriptively relevant criterions, only very risk averse agents would choose full[2] insurance.

The question that remains open is then: which of the two types of contracts is preferable (assuming the choice is offered!). One result (Borch, 1962 and Arrow, 1971 theorem for EU and generalised to RDEU by Karni, 1992) explains why contracts with deductibles are more common than co-insurance contracts: contracts with deductibles offer a better risk transfer from the agent to the insurer, as soon as the loading factor is not 0.

These results have helped insurance companies to design the contracts they offer. How much relevance do they have for an individual to choose among them?

The RDEU model is more descriptively relevant than its restriction to the EU one, but it still has limitations. Even though, in practice, it is possible to estimate parameters for the utility function and for the probability deformation function, few individuals can support the cost of doing it and insurance companies would have difficulties in aggregating such

[1] See Gollier (2001) for all these results (and much more) within the expected utility model.

[2] Unhappily for them, the full insurance choice is usually not offered by the insurance companies, so the poorer and/or the more risk-averse agents will choose among the lower deductible levels.

estimates for each risk class in order to adapt contracts to the demand of their potential clients. The Yaari (1987) criterion, which is based on the probability deformation function only, is more tractable, however it assumes that the wealth level does not influence the agent's decision. In fact, not only does it influence the decision, but, in most cases, it is not riskless, contrary to the assumption we made in the example above. As we have seen in the discussion in Chapter 6, background risk invalidates, strongly in some cases, most of the classical results obtained in the economics of insurance. Aversion to risk increases, instead of risk aversion with respect to a given riskless wealth reference level being taken into account. In the context of insurance, aversion to the risk of losses (left monotonic mean preserving spreads aversion) is the adapted reference. As we have seen, the relevant risk measure for such cases is the Value at Risk. Furthermore, VaR is an easy-to-compute index of the level that is at risk (even grossly, for an individual agent with no statistical expertise nor available calculus means).

The demand for insurance can be defined from the level of VaR an agent wants to preserve. Let us keep the same simple model as above: the agent's initial and final wealth are: $W_i = W + X$ and $W_f = W + L - xL + C(x) - PR(x)$. However, we assume now that W is random as well as X so that the relevant probability distribution Π is the joint distribution of both random variables. Insuring as a function of the VaR at the confidence level $q\%$, consists of deciding on the contract's reimbursement $C(x)$ and risk premium $\mathrm{PR}(x)$ so that the probability of the VaR final wealth must be inferior to $q\%$.[3] Obviously, this is only possible if W and x are not independent and the potential losses on W are not high enough. In this case, a co-insurance contract will be taken only if:

$$\Pi[W + L - xL + C(x) - PR(x) < VaR] < q\%$$

Luciano and Kast (2001) have shown that for a relatively low level of loading factor ($\lambda \leq \lambda^*$ with λ^* a function of the distribution parameters), the demand for insurance can be calculated[4] (and is not null). Furthermore, the authors show that the contracts with deductible still dominate the co-insurance contracts as in the case of EU and RDEU models.

The advantages of VaR over other risk measurements when background risks prevail are the following: the first one is that it is present in all the definitions of aversion to risk increases. The second reason is that VaR takes the possible hedging opportunities offered by background risks, given that it is the joint probability that is referred to, instead of the probability of the risk to be insured in the criterions defined by individual decision behaviours. Another argument in favour of VaR as a managing instrument to measure and to decide both on insurance and on other hedging means, is that it is still valid in the case of public project choices where it is not always the case that insurance is available in order to reduce risks. In particular, this remark is relevant for health insurance and public health policies (e.g. prevention means such as early diagnosis and vaccinations). In public choice, indeed, statistics and/or experts can be referred to for estimating or assigning probability distribution parameters. We shall come back to the role of VaR as a general risk-managing instrument in Chapter 12.

[3] In practice, stop-loss contracts are defined in such a way. We have seen their similitude with contracts with deductibles in Chapter 6.

[4] In Luciano and Kast (2001) the calculations are made under the assumption of Gaussian distributions for which explicit solutions are obtained. For more general probability distributions, results can be approximated with the usual methods.

Obviously, all these instruments are only valid in the case where the probability distributions are known. Let us insist on the difference between the probability known by the insurer (which is based on relevant statistical data) and the probability referred to by the agents (which may be subjective, at least in the way they perceive an objectively defined one). In both cases the analyses are the same, even if the calculations may differ from one agent to another, and from potential clients to insurers, but the results may differ greatly.

In situations of uncertainty, where no reliable and non-controversial probability distribution is available, financial assets markets can be referred to. Both for the valuation of risks, and for the diversifications and hedging instruments available. This is the topic of the following sections.

8.2 INSURANCE CLAIMS SECURITISATION

In the previous section of this chapter, we have underlined the limits of insurance possibilities. Competitive markets for risky claims, i.e. financial markets, are an alternative to the classical approach to the economics of insurance. As we know, financial innovations on these highly regulated (so as to be) competitive and easily cleared exchange markets, have flourished after the seminal papers of Black and Scholes (1973) and Merton (1973) were published. The call options' market on the Chicago Board Of Trades (CBOT) during the same year, opened the door to a voluminous exchange amount of "derivative" assets markets. Many have been invented year after year to complete the market (in the sense of offering more and more hedging instruments). Later, new hedging instruments were specially designed for the insurance business and proposed to reinsurance companies by specialised agents, and some of them were inspiring for the creation of institutional trade markets.

8.2.1 Hedging Catastrophe Risks

Catastrophe-bonds, or cat-bonds for short, are the most famous among these instruments. As we shall see, they have some features in common with bonds, justifying their appellation, but these assets can be contingent on three types of random variables: whether a damage index (insured casualties), or indexes that are specific to particular insurance companies, or/and a parameter index based on (statistically measured) catastrophe characteristics. The choice between such indexes as a reference to define a cat-bond, requires us to take two risks into account: moral hazard, due to insured agents and/or insurance societies behaviours on the one hand. And, on the other hand, basis risks, due to the imperfect correlation between the insured risks and the insurance claims. Securities defined on an insurer's specific risk have no basis risk but make the investors face likely moral hazards.[5] Conversely, securities founded on industrial damage indexes, sometimes completely, eliminate moral hazards, but the hedger is running basis risks.

Cat-bonds are issued either to address specific risks, or for well-delimited geographical zones, for a fixed time horizon in both cases. Cat-bonds are called this way because, indeed, they are bonds: i.e. tradable debts based on market exchange indexes such as the Euribor or the Libor, which is calculated in the London market in US$ or in UK£ or other currencies.

[5] When perfectly hedged, an insurance company may be less meticulous with respect to the damage claims it insures.

Expiration dates are often between 3 and 10 years and these securities are contingent on conditions: if no catastrophe occurs, or more exactly if no damage claims are above the determined in advance level, investors perceive the due payments integrally. The returns on investments are above the Treasury bond ones (riskless), often by more than 300 points. On the contrary, if the claims exceed the fixed level, then coupons and/or the principals are reduced so as to reimburse the concerned insurance companies.

Because of regulation and tax concerns, a specific offshore structure called a "Special Purpose Vehicle" (SPV) proceeds to the bonds' emission. The SPV offers a reinsurance contract to the insurance companies seeking for one, at a cost. The total costs' amount is invested in Treasury bills at one riskless rate, for one part, and in a short-term securities' portfolio, for the remaining part. The riskless rate investments are meant to warrant the pledge on the investors, the short-term high risky rate portfolio aims at hedging potential reinsured damage claims.

Catastrophe risks can also be hedged by so-called "derivatives". The futures market is open on the CBOT since 1992. From 1995 on, claims are based on a group of indexes given by an official organism: the Property Claims Services (PCS). There are nine such index-based insured damage claims, which are estimated from enquiries among insurance societies and other available information sources. Such indexes are revised day by day, they are relevant to determined geographical zones and for fixed expiration dates (trimesters or years). One index concerns the US territory as a whole and five indexes are specialised in states that are running particular risks (California, Florida and Texas, notably).

Call options on catastrophe exist as well, they are called cat-options: in exchange for a subscription that is paid in advance (premium), options give the right, but not the obligation, to buy the PCS index at a fixed in advance price (exercise price) at a fixed expiration date. Only call options are traded on the CBOT.

Cat-spreads can be found. A spread is a combination of buying a call for a given exercise price and selling another one for a different exercise price, both calls with the same expiration date. Buying a cat-spread is a way to hedge an insurer's portfolio of claims, as an alternative to buying traditional insurance with deductible or stop-loss insurance contracts.

Comparisons between the different types of instruments can be based on: transaction costs, market liquidity and moral hazards.

- Cat-options can be traded at little cost (bid–ask spreads). Conversely, cat-bonds incur important transaction costs due to the organisational complexity and to the analysis of the underlying risk.
- Markets for cat-options are bound to be easily cleared because of the participants' anonymity and if claims have a standardised form. Up to now, however, the cat-bond market is not very active and trades can take time to be cleared due to a lack of agreements on some standard forms for contingencies.
- The main advantage of cat-bonds over cat-options is that the basis risk that an investor is facing is much lower.

Seen from the insurer's point of view, the basis risk value is a fundamental argument in favour of bonds. On the other side of the market, investors are mainly interested in claims contingent on catastrophes because they offer an alternative to traditional securities for diversifying their portfolios. Froot *et al.* (1995) have shown that the returns of securities contingent on catastrophes have zero, or close to zero, correlation with other major traditional assets such as stocks and bonds.

This new class of securities attractiveness can also be measured by the Sharpe ratio between excess return with respect to the riskless rate and the risk's standard error. Litzenberger *et al.* (1996) have shown how interesting cat-bonds may be with respect to this performance measure.

For the first time, some difficulties have been encountered in the use of claims contingent on natural catastrophes. Among them, the most important ones are related to the valuation problem, which comes from the representation of uncertainty. We have already seen why observed frequencies of natural catastrophes are not a sufficient and reliable measure of the future risks. Risk factors are modelled in four stages:

- Probabilities are obtained by simulation method from the observed catastrophes (earthquakes, typhoons, floods and the like).
- Vulnerability of the insured goods, which is a function of the events' intensity.
- The probability distribution of insured claims values (simulation based on extension past observations).
- The insurance contracts' conditions, such as the proportion of the damages that are insured.

A cat-bond's price depends on the damage claims' value above which the capital and/or the interest may be lost. The difficulty arises because the probability that damages reach this amount may be hard to estimate. First, reliable data are insufficient over a hundred years period of observations for rare events. Most of the data used for probability calculations are obtained by simulation methods from a sample of events that are assumed to be likely to occur. Then, the probability that the maximum loss outranges a given level is difficult to represent by a curve that is empirically constructed. Because of this, there is no market risk premium integrated in the opening prices of cat-bonds.

Climate-linked instruments (weather derivatives) have been invented in order to take into account the dependence of many economic sectors on weather conditions. About one-seventh of the whole US economic production is sensitive to climate. Some sectors are particularly dependent on meteorological conditions. For example, the energy sector is obviously sensitive to warmer winters or cooler summers because the demand for petrol, gas and/or electricity will go under the average consumption level. Conversely, industrial sectors that are dependent on energy may be affected negatively by an increase of energy prices provoked by unexpectedly high demand from households. Because of this discrepancy between sectors, compensation means have been sought that could be mutually beneficial.

Such are weather derivatives, they have been traded on a private contracts basis in a first stage. Then, futures contracts contingent on the temperatures in different US cities have been introduced on the Chicago Mercantile Exchange (CME). More recently, markets have been created on the Web, notably on the site *i-wex.com*, which is supported by the London International Financial Futures Exchange (LIFFE).

The energy sector has been the first to develop solutions to the problem of the fluctuations of its business volume due to weather. The deregulation of the energy market was a favourable factor to the increase of the weather derivatives' market, because it increased competition on prices, and limited possibilities to make consumers pay for the climate changes' effects.

Two main features differ between weather derivatives and insurance claims:

- Insurance contracts only offer a protection against exceptional events, while derivatives can hedge against small climate variations from the average temperature, which is calculated on the basis of historical databases.
- Insured damages are reimbursed after losses have been evidenced, when weather derivatives payoffs only depend on the observed climatic conditions.

Derivative assets (futures or options) are usually based on a underlying asset: commodity, such as copper or cocoa, or a security value, such as a stock or a market index. In the case of weather derivatives, the underlying "asset" is not a security (it could be an insurance claims' value) but a weather measure. The kind of measure referred to depends on the contract's specific characteristics. However, most weather derivatives are based on Heating Degree Days (HDD) or Cooling Degree Days (CDD). A HDD is the number of daily average temperatures that are under the reference level, during winters. Conversely, a CDD is the number of average daily temperatures that are above the reference level, for summers.

Daily HDD and CDD are calculated as follows:

- HDD = max (reference level of temperature – average daily temperatures, 0).
- CDD = max (average daily temperatures – reference level of temperature, 0).

HDD and CDD measures are usually computed over a given time period, the most common ones are cumulated HDD and CDD during a month or during a season.

Most of the weather derivatives traded up to now are swaps,[6] options and standardised assets such as straddles, strangles and collars.[7]

Payoffs are defined by a specified-in-advance value in USD: the tick size (e.g. $1000 for each degree day), it is multiplied by the difference between the HDD or CDD level during the exercise period and the HDD or CDD level specified in the contract (exercise level).

8.2.2 Social Risks

Among social risks, some can be considered as being independent. This is a valid assumption for health risks, for instance, except for epidemics. The observed increasing reimbursement amounts are due to moral hazard and to the costs of attendance growth, more than to unusual variations in health risks. Independence cannot be assumed, however, for unemployment risk because the fluctuation of economic activity has direct impacts on the job market as a whole. As a result, unemployment risks are highly correlated, although, at the individual level, unemployment probabilities vary according to well-known social characteristics: education, age, sex, occupation, activity sector, etc. At the aggregate level, the risk is high and the amount of allowances paid each year increases in states where unemployment risk insurance is provided.[8]

---

[6] Marketed claims on interest rates and/or foreign currencies for a given amount and a fixed period of time.

[7] Combinations of call options in such a way that the desired payoffs configuration is obtained.

[8] For instance, in the EU, all states provide an unemployment insurance plan, with obligatory subscriptions (insurance premiums) proportional to wages when employed and with allowances when unemployed. Depending on the states, allowances and subscriptions vary according to the amounts and to the length of time they are paid.

As it exists, the State organised unemployment insurance portfolio cannot be diversified. The reason is high correlation and the fact that the State is its own insurer. At this point, this consensually admitted fact must be questioned. It works on the principle that forecasted spending is provisioned by incomes and adjusted *ex post* to actual payments if they differ from forecasts (and they always do!). A simple example shows the logic of the system, here are two alternative versions:

- Assume a State faces two types of economic conjuncture: High (H) or Low (L) and that the government's experts are not able to precisely forecast which conjuncture the country will face during the next state budget. In conjuncture H, spending due to unemployment allowances amounts to 100, but it will rise to 200 in conjuncture L. In the face of this uncertainty, the state decides to gather incomes through taxes and subscriptions up to an amount of 150. If H occurs, it will have an excess budget of 50 and a deficit of 50 if L realises. Management is achieved through budget balance.
- Assume now that the government is able to forecast exactly the future conjuncture and that it decides to clear the budget. Then it will gather 100 for H and 200 for L. Management is done through distribution.

These examples evidence that diversification depends on the link between the risk's management and the economic policy. Let us see how by considering the distribution management: individuals are affected in different ways by conjuncture changes but the adjustment variable is the quantity of labour, given that wages do not vary much. As a result there are important differences in distribution effects for those who work and those who lose their job during an economic recession.[9] An increase in social subscriptions during a low conjuncture and a decrease during a high one, distribute the adjustment weight among workers. If taxes vary instead, the distribution effect concerns a larger population. Indeed, economic agents, as a whole, are negatively affected by a recession, whatever their activity, even though microeconomic effects are not comparable.

This approach's limitations are easy to see: first, it necessitates perfect forecasts; second, it does not use diversification through time. It does not use diversification through space, either, although different economic zones are not identically affected by conjunctures.

Now, let us consider management by budget adjustment. Such an economic policy aims at allowing a distribution of risks through time, this is achieved by means of financing the budget balance. According to a rather free interpretation of the Keynesian doctrine, it could be considered that an economic revival method by spending or by public investments consists in distributing incomes without any real counterparts (pay unemployed workers to dig holes and then to refill them). Its positive side is that it shows that if public action is efficient, it requires the agents' approval of the principles underlying it (conscious approval or unconscious compliance). Indeed, as no wealth is created, there are no justifications to an increase in agents' spending. However, assume that individuals who are allowed an income have a greater marginal propensity to spend with a lower income than the ones who contribute to the allowances. Then, a distribution effect would induce an excess consumption relative to savings. Other modalities of managing budget policies could be analysed the same way. For example, reducing taxes doesn't create wealth out of nothing, at least as far as the conjuncture is concerned. It has other properties such as a distributive effect

[9] There are also distributional effects between labour and capital.

between the public and the private sectors, and this may increase productivity and potential economic growth in the long and middle run. However, these effects do not act against the business cycle.

The creation of money must be added to the financing of excess spending, to go (a bit far) in the sense of the agents' illusion. A less extreme method consists in financing deficits by borrowings. The advantage of this way to deal with the problem is that it fully takes the business cycle into account: injection of income during low conjunctures (deficits compensated by borrowings) is drawn from high conjuncture phases (budget excess and reimbursement of loans). However, this is still founded on the principle that agents comply with the policy because, here again, the counterpart of spending is not real but monetary: borrowings do not create wealth.

This short review of discretionary policies questions the capacity of the budget balance policy to diversify risks through time (besides the fact that their practical efficiency and their possibly perverse effects have not been considered).

In the case of unemployment insurance, most economic agents (companies as well as workers) are negatively affected by a low conjuncture, diversification cannot be expected, without reference to other national situations. Conjunctures in different countries are far from being perfectly correlated, even though internationalisation of exchanges smoothes them. Even inside the Euro zone, conjunctures are not totally in the same phases, even though there is a strong interdependency due to the open market and the homogeneity introduced by the Maastricht agreement. Obviously, conjunctures in Europe, in North America and in South Asia are often very diverse.

Theoretically, international diversification could be achieved through a grouping of unemployment insurance schemes. However, what international institution could mutualise unemployment risks? Only financial markets appear as a credible candidate to enforce the mutuality principle. Besides its high ability to diversify risks away, it can distribute risks among a great number of agents. This offers the possibility to find capitals well above what States can do.

An efficient risk sharing would necessitate the creation of markets for assets contingent on unemployment rates in the different countries. We could call such assets "unemployment-bonds" (u-bonds) as a parallel to cat-bonds. The functioning of such hypothetical u-bonds would be founded on decentralised income transfers from high conjuncture countries towards low conjuncture ones. This would not require agents' compliance nor illusion, because it would induce an increase of real wealth to be spent in recession phases.

Let us enter now into the details of how our proposal for this method could be managed to share and diversify unemployment risks among countries. Let us call "u-bond" an asset contingent on unemployment rate. It is a bond with payoffs indexed by this unemployment rate. It could be imagined that the coupon rate is equal to a money market reference to which would be added the difference between the observed unemployment rate and the reference rate. Let a coupon at date t be defined by:

$$C_t = R_t \times N$$

where:

N is the nominal capital of the claim
$R_t = r_t + \alpha(u_t - u^*)$ is the rate paid by the claim at date t
r_t is the LIBOR or another market reference

u_t is the unemployment rate observed in t
u^* is the expected structural unemployment rate
α is a coefficient aimed at equilibrating the claim.

The α coefficient must make the claim attractive for agents looking for a hedge and avoid extreme return rates.

Whatever the formal definition of these instruments, they would require that several conditions be satisfied in order to put them into function:

- A standardised measure for unemployment rates, non-controversial and independent of organisations that insure the risk.
- A correct and standardised estimation of the expected unemployment rates, assumed to represent the structural unemployment rate in a country or a region's economy.

Existence of u-bonds and of derivative assets (futures and options) would allow us to apply the arbitrage valuation methods, even in the absence of probability distributions for individual risks, in the same way as is done for cat-bonds and weather derivatives.

A good functioning of a u-bonds market requires that demand and supply are expressed. Institutions that insure unemployment risks are obvious candidates to buy them. Counterparts must be found for such long positions. Suppliers must be external to the country that looks for a hedge to their unemployment risk. U-bonds would be attractive to the manager of an international portfolio of u-bonds from different economic areas (Europe, America, Asia, etc.) because they would provide a hedge against the specific risk of each area. Then, the buyer's price would be above the supplier's. Indeed, the buyer faces both the systematic and the specific risk, whereas the supplier only faces the systematic one because the specific risk has been diversified away. This satisfies a feasibility condition required on a financial market.

U-bonds will provide two diversification opportunities: both on the market for this category of claims and on the traditional financial asset market. The second opportunity is difficult to figure out until such instruments remain virtual. Conversely, the first one can be approximated by calculating the correlation coefficients between standardised unemployment rates among the OECD countries.[10] This preliminary investigation puts forward two important features:

- Japan has negative correlation coefficients with all other OECD countries.[11]
- Some countries belonging to the European Union have lower correlation coefficients with some countries inside the EU than with some countries outside it: for example, Belgium's correlation coefficient with Portugal is 0.254, while it is 0.707 with Canada.

It would be useful to study what is the optimal area for claims emission, given that it may not be restricted to state borders but extend to economic regions. In the same way, portfolio diversification would be better achieved if claims were defined according to sector activities.

Cat-bonds, weather derivatives and u-bonds have common characteristics as regards insured risks:

- Risks are highly correlated inside a zone (geographic in one case, economic in the other).
- Risks are relative to very high value amounts, however, their levels are of the same order of magnitude as that of the trades on financial markets.

[10] See Kast and Lapied (2001).
[11] Between November 1997 and May 2000, for instance.

- Risks increase with time (as far as unemployment is concerned, this is due to the demographic growth and to the development of social behaviours).

Claims contingent on unemployment rates, as well as claims contingent on natural catastrophes, are facing the same three main types of risks: default risk, basis risk and moral hazard.

As far as default risk is concerned, there is no problem as long as the claims are emitted by public institutions or if they are warranted by the state. In the case of a private insurer, a legal intermediary such as a SPV is necessary.

In order to eliminate the basis risk, u-bonds emission must be made directly by the unemployment's insurer (a state, usually).

Moral hazard is linked to the way unemployed individuals are counted in the different zones where u-bonds would be used as instruments. In order to limit this risk, claims payoffs must be linked to regulations imposed by an institution independent of the beneficiaries. On the cat-bond market, the futures are based on the PCS index, which is calculated by an official organisation in charge of figuring out the insured damage amounts among a sample of insurance companies. In the case of unemployment rates, the market could refer to an independent organisation that would calculate unemployment rates in each country according to the same rules. In a similar way, OECD calculates on a standardised base the unemployment rates among countries that belong to it. It is then necessary that an international and independent official organism warrants the reference rates.

8.3 VALUING CONTROVERSIAL RISKS

We have put forward difficulties found in valuing assets that are not directly linked to economic activities represented on financial markets, in the preceding section. More generally, the problem is to value a financial risk when there is no consensus about how close to other market-valued risks it is. For instance, we have seen that catastrophe risks, weather-dependent risks or unemployment risks were not straightforward to value because:

- Data on past observations of outcomes are not sufficient or do not satisfy the assumptions required by statistics, to assign a non-controversial probability distribution to its outcomes.
- Instruments designed to hedge such risks on financial markets are not clearly correlated to marketed assets.

The first point excludes the reference to individual valuations and its extensions in the traditional approach in insurance economics. In any case, individual valuation, although an indication of a value, does not yield a reference that can be put in balance with assets valued by trades and with financial instruments adapted to manage the risk.

If a clear correlation (positive or negative) is evidenced between an asset to be valued and securities traded on financial places, a good indication is obtained as to what its price would be if it were introduced on the market. This results from the theory of arbitrage pricing (Chapter 7), because a non-zero correlation between a risk's variations and market movements yields a way to construct a replicating portfolio of marketed assets that mimics the risk's variations. Then, no arbitrage implies that the portfolio and the risky asset should be priced the same. This principle is more difficult to apply when no clear correlation between a risky asset and the market variability can be put forward. We have seen to what extent and by which means, the difficulties can be overcome about valuing cat-bonds, weather derivatives and hypothetical u-bonds in the last section.

Efficient valuation consists in looking for some reference priced market assets in order to compare them with risks that are not priced by trades. We have already put forward three elements of a method to achieve this.

In Chapter 3, we have seen why and how public project future returns can be discounted to its present value, at a market discount rate. This rate is obtained by the one used by a private company, if it developed a similar project (even with a lower level of magnitude). A similar project means that it would yield the same outcomes in all future states of the world, in the sense of the general equilibrium model.

In Chapter 3, we have emphasised also that risks have impacts on consumption commodities, but these impacts have to be valued in monetary terms in order to define the risk as a risky asset. This is what insurance experts do concerning damages to properties, for instance. The method we have shown to be favoured on a theoretical basis is contingent valuation (CVM, Chapter 4), even though it is difficult and costly to put into practice.

The third element leading to valuation of risky financial assets is the theory of pricing by arbitrage, that we have presented in Chapter 7.

We shall concentrate here on how to try and relate controversial risks to monetary risks in order to valuate the first ones by the second ones, when they are marketed. We propose to extend the principle of valuation by arbitrage to risks that are already expressed in monetary terms and that will be compared to a portfolio of marketed assets built for that purpose. By extending the arbitrage principle, we mean that the interpretation of prices (value obtained) is different from the one proposed by the theory of financial markets. In the theory, markets are complete, in the sense that all risks can be perfectly hedged by a portfolio of marketed assets. Arbitrage means that the portfolio's formation cost must be equal to the risk price if it were marketed, because both have the same payoffs in all future states of the world. And, indeed, this theory has led us to introduce options on the CBOT and then all kinds of other futures and derivatives with prices that verified the arbitrage valuation that had been calculated before introduction. In our case, there is no way that the controversial risky asset be introduced on a financial market, at least in most cases. As a consequence, the price obtained by equalizing it with the formation cost of a mimicking portfolio is hypothetical: there will be no trades that would force that price to equal the portfolio's formation cost at equilibrium in the future.

As a conclusion, pricing non-tradable risky assets by arbitrage only yields an *a priori* valuation. In the same vein, an individual valuation doesn't tell us at what price an asset will be traded, when supply will equal demand. In both cases, valuation is an indication, the advantages of the arbitrage one over the individual one are several:

- The arbitrage valuation is expressed in terms of economic value, in the sense that this value is defined by the wealth and the trades of economic agents, while an individual valuation reflects an individual's preferences and may be completely foreign to the economy's value.
- It is valid for collective choice as well as for individual choices, contrary to individual valuation (except in the case of a dictatorial system, of course!).
- The principle of arbitrage necessitates constructing a replicating portfolio, thus this portfolio is a managing instrument for the risk, given that it perfectly hedges it. Individual valuation doesn't give any indication about how to manage the risk.
- Market valuation is a collective one, as such it takes controversies into account, when an individual valuation obviously doesn't.

After these general considerations that were meant to underline the advantages as well as the limitations of the methods we propose below, let us enter into the problem of controversies about a risk, already expressed in financial terms, but which is perceived, measured or directly valued differently by several economic agents.

Let X be this risk, a random variable with no known *a priori* probability distribution. We assume it can be decomposed into two parts:

1. A known deterministic function f of a marketed (or marketable) asset, Z. It will be called the natural underlying asset of X, when it exists. It corresponds to a systematic risk.
2. A random variable Y representing all the unknown factors that provoke the risk. This is the idea of a specific risk.

Then:

$$X = f(Z) + Y$$

In the "good" cases, Y is zero and we are back to the theory of arbitrage pricing. In some cases, Z is zero, or more exactly, no relevant Z was found, and it will be much harder to relate X to marketed assets.

Controversies can appear at different levels.

(1) Controversies about *Y*: the part of the risk X that is not explained by (or related to) a marketable asset. Some parties may contest the randomness of Y. Most of the time the converse is observed: the first proposed model assumes Y is not random (costs, or fixed-in-advance income). Non-random means that an average value is taken as an approximation, this is sufficient for some agents and for many purposes (e.g. to order materials for future use) but it is obviously not for managing the risk. More informed agents argue that this is a hazardous simplification indeed: past costs are usually known but with some (random) error terms and future incomes always contain a part of uncertainty. Thus, taking a mean value reduces the relevance of the model too much. Indeed, a mean value is assumed to be agreed on by everybody. When there is not agreement controversies arise depending on the point of view of the concerned agents. Moreover, taking a mean value assumes some kind of indifference in the face of uncertainty. Indeed, it can be argued that a certainty equivalent instead of a mean value should be taken. For instance, a certainty equivalent can be $u^{-1}\{E[u(Y)]\}$, where u is a concave utility function assumed to represent an aggregate collective behaviour under expected utility assumptions. We have seen in Part I the limitations of the EU model and of collective utility functions of the Bergson–Samuelson's type for applications. Such a certainty equivalent is relevant in the case of individual behaviour, but then it will not be controversial only if all agents have the same risk aversion (and their behaviours satisfy the limiting conditions of expected utility theory).

An alternative approach consists in looking for a market valuation of a risk similar to Y, i.e. to look for a portfolio that could replicate Y's payoffs.

(2) Controversies about *Z*: the natural underlying asset. Depending on information available to economic agents, the relevance of a given Z may be questioned. For example, if X is a project concerning an energy production plant, some ecologically oriented agents may consider that environmental issues are not represented by energy-related assets (Z), even though the management of energy production does take environment protection into account and the energy price reflects some of it. Depending on experts, also, the choice of Z may be different. In such cases, it is necessary to choose Z in an endogenous way so as to be able to verify on an objective (and then non-controversial) statistical basis the relevance of the underlying asset.

This takes us back to the same replication problem evoked for Y, applied to Z.

(3) Controversies about f: the deterministic function which expresses the relation between the natural underlying asset and the risk to be valued. Most of the time, f will have to be estimated through data samples of past realisations of both X and Z. Such estimation procedures can be controversial, mainly about the choice of the database. When X is not a directly observed variable, explicative variables are used, here again the choice of these variables is subject to controversies. Often, simulations are conducted in the absence of enough observations, then again, the simulation method and the results may be questioned. As we shall see in the following method we propose, the construction of portfolios of marketable assets replicating Z and/or Y includes the determination of a functional relation: this yields f.

(4) Controversies about Z and Y probability distributions. Assumptions on the distributions form (e.g. Gaussian, binomial, etc.) may be disagreed on, even among statisticians specialised in financial markets. Estimates as well may be considered as being not robust enough, or with too low a degree of reliability. Generally speaking, a statistical model is always subject to question (estimation risk, model risk, robustness, etc.). In the case where the values taken by X can reasonably be approximated by discrete data, there is no need to know a probability distribution, nor even to assume one exists to describe and to value the risk (as we have seen in Chapter 7). In this case the list of possible payoffs is sufficient and it can be extended after enquiries so as to take any outcomes considered as likely by any agent to avoid controversies.

Whatever the procedure and the economic theory underlying it, identification of two risks requires a data sample of past observations, or partially simulated data at least. The method we propose now relies on the construction of a portfolio of marketable assets that can be identified (i.e. that replicates, that perfectly hedges, etc.) to the risk to be valued. Several techniques to achieve this goal are presented, which depend on the assumptions made, all of them end up by solving a mathematical programming problem. In the program, the control variables are the coefficients to be assigned to the marketable assets in the constructed portfolio meant to replicate the risk to be valued.

All these techniques apply to risks defined by quantitative monetary values that have been measured and observed during a past period of time. Observed values are assumed to form a sample of i.i.d. random variables. During the same period, the financial markets offer samples of observations on asset prices (rates of return instead of prices are usually taken because of the i.i.d. assumption). A portfolio is a linear combination of marketed assets, the portfolio θ's "price" (its formation cost) is the same linear combination of the prices of assets composing it.

Let us now examine different techniques and methods that can be referred to in order to replicate the risk X. There are two cases to be distinguished: whether we have Z and we have to find an f such that we replicate $X - f(Z)$ with a portfolio θ. Or we don't have Z and we look for a portfolio θ that replicates X, which, in this case, depends on Y only.

8.3.1 Reference to CAPM[12]

The CAPM yields a simple formula which requires relatively few measurements. At equilibrium, the expected rate of return of any marketed asset is obtained as the riskless rate

[12] We have seen the formula in Chapter 7, but see Chapter 10 for the classical model.

to which is added a risky rate of return. The risky part of the expected rate of return is obtained by a sensitivity factor, β, multiplied by the market risk premium. We see that only the market risk premium enters into the valuation of the risky asset, hence, its specific risk is not taken into account. Obviously, this is a limitation when the CAPM applications are extended beyond the range of assets that are closely related to the market's ones. In our model, this means that the Y part of the risk X will not be valued.

8.3.2 Tracking Error Minimisation

This method is common practice among portfolio managers when they want to replicate a market portfolio with a restricted number of assets. This is the case if one needs to hold a portfolio that yields the same rates of return as the New York Stock Exchange market portfolio, under the constraint that more than 50 assets cannot be held and managed, instead of 500 (or even 1000) in the Standard & Poor's index.

The method can be directly transposed to our problem in order to construct Z, and/or an asset replicating Y: among a given set of marketed assets, choose the portfolio that minimises its standard error to the asset to be valued. The risk X and the optimal portfolio θ are considered as "identical", i.e. having the same returns, as long as the results are sufficiently stable and reliable. In this case, we can say the assumption of complete market is satisfied (approximately, given the precision level). Then, we apply the no arbitrage condition (NA1, Chapter 7) and assert that the value of X is the formation cost of portfolio θ.

8.3.3 Looking for a Functional Relationship

The reference tool for this method is the functional correlation coefficient (FCC).[13] The function is defined (approximated) by a linear combination of elementary functions (e.g. exponentials). The coefficients in the combination are such that the obtained function of the underlying asset Z minimises the difference between 1 (perfect correlation) and its FCC with X, the risk to be valued. Then X can be considered as a derivative asset of Z and known explicit valuation formulas can be applied if relevant. Otherwise, approximations by replication portfolio and numerical methods will yield the value of X (see Chapter 11).

8.3.4 Looking for a Comonotonic Relationship

When two random variables have increments always of the same sign, they are said to be comonotonic. We have a formal definition in Chapter 2 where it was argued that two comonotonic variables cannot hedge each other. An equivalent definition will explain why we may look for a comonotonic relationship between a portfolio of marketed assets and the risk to be valued.

Comonotonicity *Two random variables X_1 and X_2 are comonotonic, if and only if there exists a third random variable X_3 and two non-decreasing functions g_1 and g_2 such that: $X_1 = g_1(X_3)$ and $X_2 = g_2(X_3)$.*

Otherwise stated, if g_1^{-1} exists: $X_2 = g_2 o g_1^{-1}(X_1)$ and X_2 can be considered as a derivative asset of X_1.

This method is adapted to the case where the risk to be valued is well represented by a discrete variable and its rates of return can be represented by a binomial tree. No probability

[13] Generalisation to more general functions of the linear correlation coefficient.

distribution is needed at this stage, binomial tree representation only means that there is a finite number n of outcomes and that they can be obtained from a single one by multiplicative increments. If Z is the found portfolio of marketed assets rates of return that replicates X (expressed in terms of rates of returns) and Z is comonotonic with X, then the same binomial tree applies to Z and to X rates of returns. As we shall see in Chapter 11, the binomial formula yields a risk-adjusted probability and does not refer to a probability distribution for the risk to be valued (consistent with the model in Chapter 7). This is an important point in the context of scientific uncertainty about the outcomes.

Comonotonicity can be tested by the Kendall coefficient being equal to 1. The method consists of choosing a portfolio of marketed assets such that it minimises the difference between 1 and its Kendall coefficient with the risk. A difficulty arises because the Kendall coefficient uses the observed values directly and the portfolio's coefficients do not intervene in the formula. The programming method must then make a detour by approximation methods such as genetic algorithms.[14]

Let us come back to the four types of controversies we had evoked to introduce the tentative valuation methods presented above. Any of the four methods puts an end to controversies (2) and (3) about the existence of an *a priori* "natural" underlying asset, because they offer a way to construct one, which is adapted to the risk X. They could be applied in any situation where the choice of Z and/or f is not obvious. These statistical methods also answer the question founding controversy (1) by making endogenous the choice of Y when the underlying asset Z is already determined.

Controversies evoked in (4), about the probability distribution of Y, are responded to differently by the methods. The first three (CAPM, tracking error minimisation and functional relation) explicitly assume that uncertainty is probabilised. Conversely, a comonotonic relation between X and the constructed Y does not require an empirical probability distribution for X. Its application in a discrete (binomial) valuation model also avoids knowing an empirical distribution on outcomes, although the binomial approximation of the marketed asset returns process does use one (a binomial distribution approximating the estimated log-normal one).

In order to address this last aspect, i.e. valuation of derivative assets in a stochastic price process setting, we need to extend the reference equilibrium model to a dynamic version. This is the topic of the models presented in Chapter 11. Part III enforces the distinction between static and dynamic solutions.

[14] See Kast *et al.* (2001) and (2004).

CONCLUDING COMMENTS ON PART II

Situations of risk and situations of uncertainty are clearly distinguished: the former refer to a known probability distribution while there is no uncontroversial one available for the latter. The relevance of this distinction for insurance has been exploited in this part of the book: traditional insurance contracts require known probabilities. When there are none available, risks require other contracts to be hedged or insured. The application range of this distinction is not limited to insurance. Its relevance can be seen in public projects such as the one presented in Chapter 1 or for the management of investments (credit risk, venture capital, etc.) that will be considered in Chapter 9.

The teachings of the economics of risks have been motivated by the economics of insurance, more precisely: the demand for insurance. They extend to all risks that are defined by a known probability distribution. Most of the results that are referred to in the literature are founded on the dominant expected utility model, they are not always relevant for management problems in situations of risk. As was put forward by Allais (1953), the shape of the utility function on income has little to do with the agent's behaviour in front of risk: it is a local characterisation of its attitude for (riskless) money amounts. Indeed, concavity is a well-known way to express decreasing marginal utility of income: €1 added to zero makes a big difference in a poor man's welfare, the same amount added to €1m is hardly perceivable by rich people. The connexion with risk attitudes is not clear. The recent developments in individual decision theory have relaxed the assumptions of EU and allowed the enrichment of risk analysis.

An important measure of risk, the VaR, characterises aversion to risk increases. In real-life problems, there is seldom a riskless initial situation so that taking a risk entails an increase in risk, or a risk reduction in the case of a hedge. Similarly, in the management of an insurance portfolio, the main problem is to limit losses and manage risk increases. This requires adapted management instruments and economic concepts that are beyond the reach of EU models.

When the LLN cannot be invoked and more generally when no reliable probability distribution is available, then the classical results in the economics of risk are not relevant for managing risk portfolios.

The assumption of a known probability distribution that characterises situations of risk is clearly not satisfied when few observations and/or an evolution as times goes of the phenomenon provoking casualties, prevent any reliable statistical estimation. This is when the original Perfect Foresight Equilibrium model is really called for: it doesn't require a probability distribution over the set of states. A difficulty is encountered because of the discrete (finite) formalisation of uncertainty, which corresponds to a crucial modelling choice, given that controversies can arise about the relevant states to consider. However, the model yields a theory from which it is possible to construct managing and valuation instruments for risks that escape the traditional insurance domain. Even though this model is known since Arrow (1953), traditional textbooks in insurance (Borch, 1990; Gollier, 2001) have kept on a specific version based on the EU model that is handy but empties it of its generality. It is not until the theory of financial markets developed and made the link with this pioneering work that its relevance was enforced (Kreps, 1982 for example; Duffie, 1988; Duffie and Sonnenschein, 1989).

Contingent contracts are the natural way to formalise risks in general, while the reference to lotteries only makes sense in the case of known and non-deformed probability distributions, i.e. a "certain" perception of risks. It is obvious that lotteries are not relevant in the presence of scientific uncertainty, for example, while contingent contracts still are. The same remark applies for natural and industrial catastrophes as well as for country risk, for capital ventures and for most real investments in innovative production processes. We do not address production problems under uncertainty in this book, but the same distinction is relevant (see Chambers and Quiggin, 2000 for an approach to production favouring contingent contracts).

The complete market assumptions seem very restrictive, even in the setting of the PFE: it is well known that in real life not all risks can be perfectly hedged by financial assets. We have indicated how constructing an adapted replicating portfolio may be a relevant way to value and to hedge risks, however this requires a great number of assets in a static setting. The next part is devoted to the distinction between static and dynamic models, where we shall see that adjusting portfolio strategies to information arrivals allows markets to be complete with a restrictive number of assets (two, in many cases!).

Part III

Static vs Dynamic

INTRODUCTION TO PART III

There are risks as soon as we see some future ahead[1] and, each time we take a risk, we foresee some future. There are two ways to consider the future: whether we overview it as a whole and we represent it in a static framework; or we look ahead with more attention to information arrivals and we forecast our reactions to future events, in a dynamic approach. Dynamics is the science of forces and of how they act. It used to be a part of mechanics, but it is now extended to many situations where time and movements (or changes) are involved. In a decision-making process, information arrivals play the role of forces: they make decisions change.

This part concentrates on the difference between static and dynamic models. The same problem can be treated by one or the other model, the solutions will differ in general. The choice between the two methods relies on the nature of the question to be analysed. Dynamic aspects are generally more relevant, however, because static models can be considered as particular cases of dynamic ones. Consider a unique decision as in the case of the Millau viaduct: the question was to build the bridge or not, it can be answered in a static vision of the future if payoffs are well forecasted. However, the forecast relies on a decomposition of the future payoffs at each future date and these depend on intermediary decision and information arrivals. This is a dynamic vision of the future where decisions are adapted to the resolution of uncertainty and to information coming in the future.

Dynamics modifies fundamentally the problem's structure if decisions can be revised at times where information is obtained. Without knowing the content of a new piece of information, multiple decisions can be summarised by a unique choice, but increases in knowledge make the problem more complex. Indeed, when a decision is taken at some date, the fact that future choices will be made according to information that is yet unknown, must be taken into account. This is known as integrating option values in the present decision criterion.

The individual valuation approach, as well as the market one, has been enriched by the use of the option concept. It applies in situations where flexibilities are available, i.e. each time a decision is (partially) reversible or can be delayed in time. Because it modifies or delays choices, flexibility makes it possible to adapt them to new information. Constructing a prototype may be a way to investigate knowledge about the possible returns of an important or irreversible investment. This offers an option: the possibility to use information when it will become available in the future. An option cannot have a negative impact on the overall value, because it corresponds to a right, and not an obligation, to realise the project or not. The option will be exercised (i.e. the irreversible choice will be implemented) if the project overvalues the option (then, the option is killed by the exercise).

These considerations do not address the problem of representing the future, instead they assume that it can be decomposed into two components: time and uncertainty. We have mentioned this decomposition as from Chapter 2, in this part of the book we concentrate on the results obtained under this usual assumption. We shall question the more arbitrary assumption on the hierarchy between time and uncertainty: first, uncertainty is measured for a given date and then time is discounted to aggregate certainty equivalents into a present value. Is it equivalent to proceed the other way around? We shall see in Chapter 10 that it

[1] There may be uncertainties about the past (we don't remember well . . .) but there are no risks to take or to bear: they have been realised.

depends on the measures. Other questions about the future's representation are: Would it be the same if time and uncertainty interfered with preferences and their measures? Can we consider the future as a whole instead of this arbitrary decomposition into two components? We'll come back to these questions that have not yet been solved in the general conclusion of this book.

Chapter 9 presents several risk businesses for which the two ways to integrate time may be relevant, depending on the point of view and the major problem to solve. For instance, the gambling industry considers bets in a static way (identically repeated), while most gamblers play in a dynamic context. Similarly, credit risk can be managed in a dynamic way or in a static way if the debt is kept until the expiration date. We already made this remark about insurance, where insurees decide on the basis of a static model while the insurance portfolio has to be managed dynamically.

Chapter 10 is about static financial models and risk measures, it is short because the results have already been evoked before in Chapters 6 and 8.

Chapter 11 is long, because the dynamic models are rather mathematically involved, even though the concepts they use and their results are very close to the ones presented in Chapters 3 and 7. Their range of application, however, is very wide: valuing flexibilities and options is a problem in most new risk management methods.

Chapter 12 deals with static and dynamic instruments. They are applied to credit risk and to valuation of controversial risk.

9
Risk Businesses

In the examples we present in this chapter, two ways to look over the future can be followed: static and dynamic, they generally entail different decisions. Depending on the particular characteristics of each problem, the approach will propose a more relevant analysis and suggest managing instruments that are more adapted over the other.

9.1 LOTTERIES AND THE GAMBLING BUSINESS

Decision theory and probability calculus were inspired by the study of chance games. Drawing lots, dice and card games, lotteries, bingos, roulette wheels, etc., all have a common origin with other predicting devices such as interpreting birds flying (auspices), or sacrificed victims (augurs), soothsaying, etc.: magical ways to learn about one's lot. Even today, we can observe some of these behaviours as regards situations when some uncertainty prevails.

There is more than that in the attractiveness of modern chance games. It's a way to take the risk of changing a small amount of money into a possibly very large one (in general). As such it's a way to take risks without hazarding one's safety. Some people get hooked on gambling, though, because the excitement they get from waiting for the joy of winning acts like a drug. Some people like to give more interest to a card game necessitating concentration (e.g. bridge) by risking some money in it rather than playing for the sole fun of playing. Many people simply enjoy a lottery game where there is no mental effort to make to have some fun. Chance games were forbidden by the religions based on the Book until the Renaissance, and they are still suspected of some immorality because of social aggressiveness and inequalities they may generate. A more serious problem is that gambling addiction may push some people to ruin.

However, chance games are well accepted today and have been developed by a large industrial business that proposes (relatively) cheap entertainments. Obviously, gains in this case are not the result of effort but come directly from spending. This is the exact opposite of money earned through work, where gains (salaries, returns) result from effort: this contrast is part of the entertainment working people are eager to find in chance games.

One of the first gambling societies was established by French King François the First in the 16th century in the city of Paris[1]. Besides casinos (an industry that is rather secretive about its results), most gambling societies are state-owned, at least partly. In France, for example, state-owned "Française des jeux" is among the societies with the highest turnover for a population of 35 million players over 53 million inhabitants (in 1985), each spending €3.5 a week on average. Comparisons with other countries reveal how large is the set of

[1] Caillois (1967), in French.

Table 9.1 Number of players

UK 56.8	Italy 57.4	Spain 38.5	Japan 120	France 53.5
49.5	41.6	31.2	39.9	35.7

Source: "Française des jeux". Reproduced by permission of Française des Jeux, 2005.

Table 9.2 Turnovers for instantaneous games

Française des jeux	Massachusetts lottery	Texan lottery	Loterie nazionali (I)	UK national lottery
3.23	1.68	1.37	1.1	1.01

Source: "Française des jeux". Reproduced by permission of Française des Jeux, 2005.

potential customers for this industry (in millions of players, the total population in 1985 is indicated for each country in Table 9.1).

The turnovers for instantaneous games are comparatively even more interesting (in million €), the comparison is made for states that have similar population sizes (about 50 million) in Table 9.2

We have seen and understood why risk aversion justifies the existence of financial societies, insurance companies and markets for financial instruments. The above numbers show that risk eagerness may also give rise to profit-earning activities.

In Part II, we insisted on the distinction between situations of risk where a probability distribution is known and situations of uncertainty where there are no non-controversial probabilities available. The distinction between the two situations applies for the gambling industry as well. Anscombe and Aumann (1963) proposed calling roulette lotteries the risky ones and horse lotteries the uncertain ones, to make this distinction clear: roulette has a well-known probability distribution, while there is no way to have a horse run a race twice in the same conditions, so betting on a horse is a decision in a situation of uncertainty.

Several studies have shown that the population that bets on horse races is clearly distinct from the population that plays lotteries and other such games of chance. For example, in France where gambling on horse races is a state monopoly, a study[2] shows that horse race gamblers don't bet on horse numbers in a race as if they were random. The population of gamblers that cannot afford going to race stadiums is stable since the 1970s and is formed in the majority by low and middle class workers and shopkeepers, mostly men. They gather in special bars where there is a gambling bureau (and generally a special television channel showing the last races), where they form a kind of community with its rituals, where information is exchanged and discussions are animated, all smoke (at least passively!) and most drink wine or "pastis", except Muslims![3] The competition aspect is dominant, first because a race is indeed a competition and knowing the horses' past performances is important, but mainly because gamblers are conscious that they play against each other: there is competition among players, besides the competition they bet on. The random aspect

[2] Yonnet (1985), in French.

[3] If you are a non-smoker, and even if you are and don't want to choke on others' smoke, or if you want to have a quiet conversation, avoid bars advertising "PMU" (Pari Mutuel Urbain) when you visit France!

is not important and no gambler assigns a subjective probability to decide on which horse to bet.[4] Another observation is important: gains are rarely high, but they are regular each week.

The converse is true for gamblers in the National Lottery: they don't meet, don't discuss and there is no information to exchange, probabilities are known or could be (even though gamblers don't refer to them), gains are generally zero on average, but a rare gain is often very large. Furthermore, gamblers are not aware that they play against each other, even though they do indeed: if there are several winners, they share the prize.

It may be a surprise that forecasts based on average gains are completely faulty for gambling on a random mechanism such as a lottery. It may seem strange that someone with a low income that buys insurance, spends an important part of the family earnings to bet on a lottery or on a bingo number with objectively (quasi) zero mean prize. We have seen in Part II that expected utility theory can explain this contradiction by the utility function concavity depending on the income level at stake. The utility function would be concave (risk aversion according to this theory) for the insured value, and convex for the few Euro of bets.[5] The problem is that a few Euro each week may amount to more than the insurance premium over the year! A more convincing explanation is given by the extensions of expected utility: the certainty equivalent of a lottery depends on a subjective deformation of probabilities (which are not well perceived in general) or a subjective capacity. The criterion depends on gain increases and not on gains themselves. Even within this framework, in most cases, the certainty equivalent would be inferior to the minimum bet, so that a rational gambler should not enter the game.

Another behavioural aspect must be taken into account: people play a repeated game, they know they are not likely to win if they play only once. Objectively, it is false that the probability of winning at each drawing increases when one plays repeatedly, because lotteries are independent each time. However, it is true that the probability to win if a given number is played repeatedly increases, so many gamblers are confused. This confusion is helped because they have heard (and the National Lottery advertises[6] much on this) that other players have won by playing repeatedly. Furthermore, it is well known that most gamblers think that a number that has not been drawn for a long time has a greater probability to occur than a number that has just been drawn.[7]

This confusion makes the lottery business rather safe: a law of large numbers (the Central Limit Theorem, see Chapter 6) entails that a high number average sum of independent bets is quasi constant (with a very small variance related to the few winners). Contrary to the gamblers, the lottery and casino industries know precisely the probabilities and calculate the prizes and the related bets accordingly. This calculus is made in a static model that is relevant for them. Conversely, gamblers have a dynamic vision of the problem, which lets them hope that they will win the "big prize" one day. Paradoxically, the industry plays a quasi riskless game, while gamblers are attracted by the risks they take.

---

[4] It wouldn't be possible to calculate a subjective distribution from the observed bets of one gambler, because decisions violate the independence axiom. First, they depend on bet levels. Furthermore, bets do not qualify the event only, but also the perceived number of gamblers that are likely to bet on the same horses.

[5] Furthermore, expectations are not on bets but on gains: most are close to zero so the convex assumption applies. However, at least one gain is very high and for this the utility should be a different one and convexity should not hold.

[6] A clever, obviously not fallacious but very confusing advertisement of the French National Lottery said: "One hundred percent of the winners have played"!

[7] For instance in Dostoievsky's novel: *The Gambler*.

Simplifying, we can say that gambling supplies are made in accordance with a static approach, while demands are based on a dynamic one. This remark is valid for lotteries as well as for horse races.

In Chapter 12, we propose several instruments that can be applied to manage the risk of the gambling industry as well as those of any risky portfolios: they are founded on static models.

For the topic of this book, the relevance of chance games is not limited to the recent development of the gambling industry. Chance games are used as a training device to face risks. That was true in the past and still is at an individual level. Gambling is used also to analyse and observe individual behaviours in the face of risks. We have already evoked several times how chance games are at the root of probability theory and of decision theories. They are still used to observe and test individual behaviours in situations of risk.

For example, in a contingent valuation method on air pollution, Chanel *et al.* (2004) included a question on attitudes in the face of risks in the inquiry. Gambling or not in one of the National Lottery games was revealed to be a relevant characteristic of individuals (along with sex, age and income). This question was introduced because, in the preliminary inquiry, a simple game allowed us to compare the attitude in the face of bets, and answers to questions about the (objective) given probabilities to die from air pollution.

More generally, chance games are commonly used in experimental economics to get individuals to reveal their attitudes in the face of risks and to simulate market games where risky contracts are traded. The assumption underlying the use of chance games that are familiar to subjects in the experiments, is that their choices over several alternative decisions will be the same whether the outcomes concern a game, some situations they face in their day-to-day life, or in their professional activities. However, many experiences have shown that individual answers are conditioned by social norms, and psycho-sociologists have taught us that behaviours in a social environment are sensitive to different contexts.

Nevertheless, chance games are still a good way to get trained to face risks and they are a reference for individual attitudes in the face of risks. Furthermore, they yield a source of profit for an industry that can be used to raise funds, which could be reinvested to finance other risks, such as social ones.

9.2 RISKS AND INVESTMENTS

Risk-taking, i.e. originally “capital investment in an activity with uncertain returns”, is practised today in two essentially different ways. The first one is followed by individual investors, or entrepreneurs, that invest their own money and/or borrow some funds from a money lender (silent partner). The second one goes through institutionalised organisations: the financial trade markets. However, these two ways have many connections in industrial societies, as we shall see by examining different aspects of financial risks.

9.2.1 Financial Market Risks

Investing in stock exchange markets is an activity that has developed as from the 19th century in industrialised countries. Major societies split their capital between shareholders who detain securities and these securities are priced in financial markets. Some markets are related to a region, most function nowadays at national levels, many markets allow international trades. Pricing is done according to different systems that figure out the price at which securities can be traded among agents. Pricing can be done on a day-to-day basis or in continuous time.

Another way for firms to finance their activities is to issue bonds. A bond is a contract stipulating the amount of capital that is borrowed, the rate (defined by coupons, or annuities specifying how much is reimbursed each year) and the expiration date. Such contracts can then be traded among their holders in specialised bond markets.

Markets for bonds and for securities are meant to provide a place where investment risks can be shared and hedged. The buyer of a traded asset does not face the risk of lacking liquidity as would be the case with non-marketed shares of limited companies. On the other hand, buying traded assets is submitted to market risks. Indeed, the value of assets is determined by supplies and demands and it may increase or fall. Asset values depend on the intrinsic value of the societies that issue them, but they also depend on the "market moods": "bull" market, "bear" market, contagious effects, speculative bubbles, etc.

Market risk is characterised by the price variations of securities and of bonds. This is how it is perceived by the stock exchange investors (individuals, dealers and trading societies, financial institutions, companies, etc.). Present prices define the investment's value, while future possible prices correspond to the uncertain payoffs of this invested capital. There may be gains, but losses are possible also: that's where the risk lies. For securities, dividends must be added to these payoffs (capital gains and/or losses), they are uncertain because they depend on possible benefits or losses of the company, but they also depend on the decision of the company's general meeting of shareholders. Financial analysis is founded on scrutinising the companies' health so as to forecast its future value that will be reflected by its market prices. It is completed by an analysis of the evolution of relative market prices. Such expectations about price movements are based either on empirical methods based on graphs or form recognitions for trends. Others rely on statistical methods founded on stochastic calculus that yield reliable estimates of the prices' stochastic processes parameters. Last but not least, future prices of assets can be deduced from the financial market theory, as was introduced in the preceding parts of this book. This theory proposes different models and formalisations, we have already seen several of them and readers of this collection know others. This part concentrates on the distinction between the relevance of static vs dynamic models. Relevance depends on the type of asset to be valued and on the way the model is meant to be used. If the problem is to look for the theoretical value of a given security so as to compare it to other securities' prices in order to decide whether to buy or to sell it, a static model is satisfying. A dynamic model is necessary to relate the different stochastic prices' processes, if one considers an option (or any derivative asset) so as to build a complex strategy: hedging a portfolio for a given period or set of values; arbitrage between several securities; speculation about some parameters' sensibilities, etc.

9.2.2 Interest Rate Risk

A fundamentally dynamic risk is the risk linked to time: credit, i.e. borrowing or lending, up to some expiration date (maturity). For bonds issued by private companies or collective institutions, there is a double risk: the market risk we already mentioned, and the risk due to the different interest rates that are applied for the same maturity. Consider the case of an individual borrower (or lender). If he considers a real estate loan with a long-term expiration date at an interest rate fixed by a bank, for instance, he can be tied up after some years with a contract at a higher rate than he could obtain at present. If the bank is in a competitive context, it will have to revise the rate and reconsider the contract's terms. Another way is to adopt a variable rate: it varies as a function of the market rates, within an interval defined by the contract.

More generally, any investment's rate of return for a given expiration date must be compared to other borrowing or lending rates for the same date. It must as well be compared with other ways that could yield the same future payoffs according to a strategy based on the right time to buy the right available financial instruments. Then, several aspects of the interest rate risk can be compared:

- The maturity risk for a fixed interest rate.
- The risk to choose a fixed rate against a variable rate.
- The risk to put an end to the contract before its maturity.

Take the example of borrowing for a 10-year period (e.g. a real estate investment). If fixed interest rates are taken into consideration, rates today are known for all maturities going from 3 days to 10 years from the term structure of interest rates. Borrowing can be considered for the 10-year term or for shorter terms at which a refinancing could be done. For example, if one thinks interest rates will fall within 5 years, it may be wise to borrow up to five years and keep the opportunity to profit from lower rates later. The risk comes from the fact that rates may increase instead of decrease, as is expected today. However, it may be worth taking this risk if rates at 5 years are much lower than rates at 10 years. This way of reasoning can be extended to all maturities inferior to 10 years and the different risks compared with each other and with the immediate advantages.

Another method consists of borrowing at a variable rate. The rate varies as a function of the short-term interest rates. The risk comes from their variability: they may increase in such a way that the global final rate is much higher than the actual 10-year rate. This method will be chosen if the investor expects short-term rates to decrease steadily.

The risk of putting an end to a borrowing contract before maturity comes from an unexpected capital facility coming up during the borrowing period (unexpected heritage from a remote uncle, for example, or some good business opportunities). There will be a deductible to pay in order to break the borrowing contract. Furthermore, all depends on the rate at which the newly available capital can be invested. Hence there is a risk coming from the option to profit from such opportunities, an option the credit contract seldom includes. If institutionalised options are traded, competition between credit firms should integrate such risks into the proposed interest rates. Otherwise, these risks must be hedged by some adapted managing methods. The theory of financial markets, and more precisely the models of the term structure of interest rates, are called for in order to figure out the different possible future rates adapted to strategies meant to reduce the interest rate risk.

Given this risk is characterised by the evolution of the different interest rates, it is relevant to consider dynamic models of the term structure of interest rates, rather than working on a static one. In practice, the dynamics of interest rates, or even better the variations in time of the term structure of interest rates, is referred to. It is represented by the rates curve, and it indicates, at each future date, interest rates for all the zero-coupon government bonds for the different maturities.[8] If this curve is known, at least approximately at the time decisions have

[8] No intermediary reimbursements (coupons): interests and capital are paid off at the expiration date.

to be made, its future evolutions represent the uncertainty relevant to the decision problem. Several rate instruments (fixed rate bonds or variable rate bonds, futures, swaps, etc.) are considered as assets contingent on rates from the term structure. Such a model can then be referred to for valuation, arbitrage, speculation and hedging strategies.

9.2.3 Venture Capital

Venture Capital covers precisely the etymological meaning of the word risk! This way of calling specific firm financings, or their investment projects, spread in the USA during the 1940s. Nowadays a whole set of finance specialties belongs to venture capital: projects that are out of the usual for a given firm, or a new firm's plans. Small firms unable to open their capital directly, or to issue bonds, etc., look for such specialists. There is risk, in the original meaning, however this risk is to be taken in a world where financial markets exist, which opens many a possibility for risk-takers to invest at a "lower risk level".

Venture-capital societies have developed rapidly for financing innovations and for injecting capital into firms recently created or that are being restructured. At the beginning, such societies were mere departments of bigger institutions, however, the demand for such financings and their high yields have attracted individual investors so that venture-capital societies proliferate.

The demand faced by venture-capital companies is addressed by two types of firms:

- Ancient and important firms. They call for a venture-capital society when they have an innovative project or if they want to restructure an important part of their activities. They go through external financing if they can't risk their capital, or if they are reluctant to announce projects that could frighten their shareholders.
- Small or medium firms, most of them recently created. They present a new project that they cannot finance by way of the usual means (banking credit) because risks are too high. Many such firms' successful results were related to the development of the computer and electronic industry in the Silicon Valley in California. Young graduate students went to venture-capital firms to obtain funds they would not have been able to raise otherwise, thanks to which they were able to develop the processes and software programs we know today. In a more modest way, many a small local project could only be financed thanks to such means, even though they may have been partially supported by institutions. For instance, if venture capital takes the risk, it is easier for the project to be eligible for some public or banking support. Conversely, a little institutional support is often enough of a warrantee for the venture-capital society to invest in the project.

A venture-capital investment is different from investing the same amount in the stock exchange market. Indeed, it's a long-term investment because it cannot easily be cleared. Furthermore, venture-capital investors play an active role in the managing of the firm they invest in. Conversely, an investor in the stock exchange market only buys a share of an existing capital that can be resold at any time. Furthermore, unless the number of shares already held is great, the investor will not be prominent in the decision process of the board.

The venture-capital company takes into account the qualities of the entrepreneur, the managing team and the innovative project. All these are scrutinised and the demander must be open to answer and give all available information. The company will agree to finance only if it can influence the management of the firm, for instance by offering some formation to clerks or taking care of parts of the management itself. The contract takes all these aspects into account, as well as the global cost. Furthermore, the venture-capital society is interested in the benefits (if any!), besides the contract reimbursement of the debt. To put it simply, the venture capital "takes the risk" that the entrepreneur cannot take by itself. It seems obvious that in order to take such risks, the society requires that its management *savoir-faire* be used in order to offer some guarantee.

There are two ways to formalise the potential future gains of a venture-capital investment. Venture capital can be understood as a real investment in a project, a part of an internal investment of the firm. Then, the standard discounting of future returns, together with a project-specific risk premium, can be called for to figure out the project's net present value. This formalisation integrates time and the dynamics of the project is only captured by the discount factors in the NPV. Capital invested by a venture-capital society can be considered as any other type of capital investment, for instance on the stock market. In this case, a static model of market equilibrium is sufficient to arbitrage between its risks and its returns. The real investment theory that we shall present in Chapter 11 reconciles the two previous ways to look at such investments: it refers to market values but it does it in a dynamic setting.

9.2.4 Country Risk

If part of a firm's activities takes place in a foreign country, the returns on investment depend on some factors that are specific to this country. Similarly, the risks of bonds issued by a state depend on economic, political and climatic factors that are specific to that state. A well-known example are the "Russian bonds" issued by the tsarist regime, which were not honoured by the communist government that replaced it in 1917. More recently, we remember that many developing countries have been bound to renegotiate their liabilities in order not to default. Such negotiations are based on a decrease in the reimbursements, hence they induce a great variability in the returns to the lenders.

Wars and revolutions increase the risks relative to investments in a given country facing such instabilities. Similarly, natural catastrophes can have the same effects because of the high level of damages to properties, persons and infrastructure they may induce.

The notion of "country risk" is characterised in terms of potential losses, possible damages and bad events probabilities. Country risk is mentioned in warrantees and insurance contracts to hold before investing in a foreign country. As such, the notion covers all casualties that can happen in the host State as well as in the State of origin of the investor. For a firm that is based in a foreign country, there is a risk that sovereign acts from the host State may interfere with the subsidiary firms' control and management. For export, the main risk is related to difficulties to collect outstanding debts, another one is related to market closures. For import, price currency variations induce the main risk, another one comes from difficulties to find available supply at the right time. Banks are concerned by payments incidents due to interferences with State regulations and political decisions.

Any sovereign State introduces a political context into the country risk. Political risk is caused by the internal structure of the State and by its relationships with its neighbours. It increases when ideologies rise (xenophobia, revolution, terrorism, etc.) and/or during sociological and economic evolutions. The economic and financial status depends mostly on political decisions (price fixing, exchange limitations, salaries, etc.), but it relies also on the internal trading system (monopoly power, for instance) and the trades with other countries (competition, notably).

Some private and public insurance companies propose contracts to cover buying credits (opened by a bank to a foreign buyer) as well as selling liabilities (from the foreign seller to the buyer). Insurance premiums are a function of the country-risk valuation, they are based on political and economic analyses. They integrate to a great extent the institutional context (for instance, countries' agreement to exchange raw materials against manufactured products). Country-risk valuation is done on the basis of a forecast on the evolution of the country under concern. A way to foresee such evolutions consists of comparing observable factors (strike frequencies, GNP, demography, number of *coups d'états*, etc.). Analytical methods also refer to observable factors that they use in hazard signals, decision criteria and the rating system. Valuation must integrate the global solvency of the country, it is measured by two factors:

- the ratio between interest rate payments and exports, and
- the ratio between external debt and exports.

The (secondary) market for State bonds yields an important indication of the confidence the market puts into a State capacity to honour its debts. For instance, it has been observed that highly indebted developing countries have seen their liabilities priced well under their facial value when they meet difficulties to face their reimbursements. Banks who own such liabilities prefer to sell at a low price and face an immediate loss, rather than holding an asset that may incur a higher loss. Buyers of such liabilities are societies that are interested to invest in the issuing country, they can do it at a low cost by holding such claims.

Once a firm's exposition to a country risk has been determined (an import or export company, or a financial institution), and after this risk has been valued, the problem is to find the means to hedge this risk. The first thing is to integrate the "risk price" into the bought (or sold) commodities and into the interest rates. However, competition leaves little degree of freedom in that matter. Then, the risk can be externalised through financial institutions. For instance, this is what is done in the case of suppliers for credit insurance. Also, banks or specialised establishments can warrant payments through the factoring technique. Factoring, for an export company, consists of selling its foreign liabilities to a "factor" who is in charge of managing and collecting the debt. The export company is paid cash at the cost of a fee depending on the country risk.

Country risk can also be covered by a hedging portfolio. Such a portfolio is designed in a way that its payoffs should compensate potential losses due to the country defaults. A hedging portfolio is mainly based on derivative assets, which have known a great development as from the 1970s. For instance, call options or put options on some assets are instruments meant to limit losses in the case where the underlying security price gets too low or too high. An option on the external debt of a country (or on an index formed with its liabilities) is an instrument to hedge part of the losses induced by the country difficulty to meet its debts. More generally, a way to eliminate part of the investment risk relies on its diversification among countries with a low correlation risk.

Most often, country risk will be modelled in a way similar to credit risk (see the following section). When the problem is to determine the premium on a State's bonds, the similarity is obvious given the risk is linked to the default of the issuer. Causes for default are different, but consequences are the same for the investor. Rating is then the adapted measure instrument common to both credit and country risk. If the investment is meant to be held until maturity and risk is expected to occur only at that date, then a static management can be used. If one wants to measure (and eventually replenish) losses of market values due to rating decreases, then a dynamic management will avoid waiting for default to become unavoidable.

9.3 CREDIT RISK

Risk control is one of the essential activities of banks. Without it, neither the performances nor the perenniality of the firm could be assured. Nowadays, the governance of credit firms is founded on risk control. Some recent examples have shown that even important banking companies can collapse if they take risks beyond their capacities. Such a hazard entailed an inflation in regulation requirements for banking and other credit companies.

The main difficulty encountered in risk control is that the risk cannot be separated from the banking activity nor from its returns. Given a financial position, the risk is not that there may be some losses: there may be losses but there may be gains as well. Indeed, it is because there may be gains that the risk is taken: we are back to the etymology of the word. Therefore, risk control does not mean not to take risks, otherwise no activity would prevail, it means to take some risks that are "acceptable" in a sense that we shall make precise.

Risks that are taken by banks, whether in their credit activities or on the financial markets, are rewarded by payoffs, which correspond to *ex ante* risk premiums. The risk counterpart is to be found in the returns. Among two activities, if one is riskier, it must have a higher return to be chosen.[9] Therefore, risk management can be seen from two points of view:

1. The choice of activities with an optimal risk–return couple.
2. The definition of a risk level that must not be overturned in order not to put the survival of the firm in danger.

The two points of view are compatible and they correspond to expected returns maximisation under the constraint that the risk be inferior or equal to some defined level. In order to achieve this, it must be assumed that expected returns and risk can be characterised and measured.

All banking activities entail some risks, let us distinguish three classes of activities.

- Transformation: this is the classical indirect financing of an economy. The bank plays the role of an intermediary between agents requiring a financing and those who have an available capital: it transforms debts of the first into sparing instruments for the others.
- Market activities: refinancing, arbitrage, speculation and client orders services.
- Financial engineering: design and organisation of mergers and acquisitions.

The different types of risks faced by banks are the following.

- Market risk, *stricto sensu*: risks to lose generated by a position if the market evolution is unfavourable.
- Clearing risk linked to impossibilities to undo a position at an acceptable price.
- Counterpart risk: no counterpart can be found to the position.
- Legal and fiscal risk: modifications in regulations and laws.
- Operational risks, which are due to errors, frauds or technical problems.

[9] Assuming there are no constraints on the quantities of possible activities.

Credit risk is a counterpart risk that can occur if the borrower cannot pay (part of) the interest and/or the capital. It is present in intermediary credits as well as in bonds.

The factor that explains credit risk is the firms' default. Default of a firm entails that it is unable to clear its positions, whether partially or integrally, and this affects all of the creditors. The amounts at stake are important. It appears that the main default factors are related to the size (as measured by the turnover or the number of employees), the age and the legal form of the firm. Obviously, these characteristics are not independent. They play on the consequences of a default given that filing in bankruptcy is followed by compulsory liquidation for most small firms but less so for big ones.

Firms' bankruptcy statistics show that there is a specificity for the risk of small and medium size firms. It results from a higher vulnerability, but also from an under-capitalisation or from a too high indebtedness, mainly from banks. Indebtedness is one of the main difficulties for small and medium size firms, because it is a fixed charge in the accounting and is heavy in case sales or margins decrease. This is the reason why credits to small and medium size firms, which are the mainstream of credits to professionals, are among the riskier ones.

There are some means for creditors to limit the default risk or its consequences:

- Warrantees can bear on a value superior to the credited financing in order to compensate the discount from part of the asset on which the credit is based. Another way consists of doubling the warrantee on the deposit of the individual managers.
- Contractual clauses can, for instance, limit asset transfers or dividend payoffs in order to impoverish counterparts to the liabilities.
- In a way similar to the one we described for the insurance business, pooling is a way to share risks among several partners. The Banking Union is a formal pooling that shares debts by a contract between banks. Banking partnership is a simple coordination of banks on the conditions they impose on credits.

Provisions for bad debts can be formed when a non-payment, or a delay, or a payment incident is acknowledged. They can be put forth when the bank judges that the solvency of the firm is undermined or when it is doubtful. Provisions are on the capital and/or on the outstanding interests.

Most of the time, a static approach is relevant for credit risk valuation. It corresponds to a possible default at a given maturity. During a given period, the probability distribution of the binary variable: default or not, is estimated. This results in the credit spread: the (positive) spread between the rate afforded to the firm and the corresponding Treasury bill's (riskless) rate. This is an accountant's version of the default risk given that it is valued at the credit origin and that the credit value is the contract's one during the credit's life.

A market approach can be favoured where the claim is valued at the price at which it could be sold, even though it is held until the expiration date. The advantage of this method is that it can take information arrivals into account in order to adjust the claim values. There is a disadvantage, though: it renders more volatile the statement of affairs, which may influence negatively the market value of the credit firm. Indeed, volatility generates a risk premium.

A dynamic conception of the credit risk must take the evolution of the default probability into account, because it is a condition on the movements of the market value of the credit. Consequences of a static or a dynamic approach in terms of risk management will be scrutinised in Chapter 12.

10 Valuation without Flexibilities

Present valuation of future consequences can be achieved in a static model as long as information arrivals in the future do not matter. In the first section, we start with the most famous concept: the Net Present Value (NPV). Discounting methods and their pragmatic or theoretical foundations are discussed. We mentioned in Chapter 9 that there are two approaches to risk management:

1. The choice of activities with an optimal risk–return couple.
2. The definition of a risk level that must not be overturned in order not to put the survival of the firm in danger.

We address market-based valuation by presenting the two main static models of financial asset pricing. They are based on the Markowitz mean–variance efficient behaviour in the face of financial risks, which we have evoked several times during the previous chapters. If management is turned towards limiting losses, the adapted measure of risk is the VaR, to which we referred at length in Chapter 6 to characterise risk increases and that we shall present as a management instrument in Chapter 12.

10.1 THE NET PRESENT VALUE

The basic criterion used in capital budgeting is the NPV of a project. It is, originally, an accountancy assessment over past cash flows to draw up a balance sheet at the present date. If $F(t) = L(t) - C(t)$ is the net cash flow (incomes minus costs) at date t of a given activity A observed over past times: $-T, \ldots, t, \ldots, 0$, where 0 stands for present time, the NPV of this past activity is:[1]

$$NPV(A) = \sum_{t=-T}^{t=0} \frac{F(t)}{(1 + R_A)^t}$$

where R_A is an interest rate relative to the activity over this period. To be more precise, interest rates should be calculated for each date ($R_{A,t}$) and for each net cash flow ($R_{F(t)}$), e.g. the rate at which $F(t)$ has been or could have been invested, or borrowed, depending on its sign. Then, instead of $(1 + R_A)^t$ one would have (or one could define R_A as a solution of the equation):

$$(1 + R_A)^t = (1 + R_{F(-T)})(1 + R_{F(-T+1)}) \ldots (1 + R_{F(t-1)})(1 + R_{F(t)})$$

[1] We give the formula assuming time is represented by a discrete variable, a similar one is obtained for a continuous version: $1/(1 + R_A)^t$ is then replaced by $\exp(-r_A t)$ and the sum is an integral over the corresponding interval of times.

Applying this formula as a criterion to decide whether or not to invest in a project P, consists of forecasting future net cash flows, using some risk-adjusted interest rate R_P over a given period of time: $0, \ldots, t, \ldots, T$, and defining:

$$NPV(P) = \sum_{t=0}^{t=T} \frac{F(t)}{(1+R_P)^t}$$

NPV is referred to for an isolated projected activity as well as for a global valuation (and decision criterion) of an investment project: the (net) investment is $F(0)$ in the formula; its rate is 0. A project is feasible if and only if its NPV is positive, and among projects with positive NPVs, the higher the better. We are back to the principle of cost–benefit analysis and back also to the same problems regarding assignments of values for future cash flows and interest rates. Plus a major one that concerns the representation of the future and to which we shall come back later:

- Future cash flows are uncertain, hence $F(t)$ must be some summary of the possible net incomes that may be obtained at date t. They can also represent some individual certainty equivalent of these cash flows, taking risk aversion into account, for instance.
- Interest rates for future expiration dates are usually obtained from the market term structure of interest rates, they must be adjusted to the project's risk. They can also represent an individual preference for present (over future) consumption.

For both problems, the concern is about the risk of the project and the attitude of the decision-maker in the face of this risk.

The major problem we have evoked is about how the future is apprehended (i.e. understood and represented). There are at least two alternative ways:

- If future time is represented as we did for the past, then uncertain states at each date must be described (and eventually measured) so as to work up a summary of possible values, e.g. $F(t)$ is a mean value. Then, these mean values can be discounted the same way as we did for past observed values.
- If a list of possible scenarios (i.e. possible trajectories) of a given sequence of events indexed by time represents the future, then each future cash flow must be discounted into present values, and then these alternative present values can be summarised by one mean value.

But then, is it the same to figure out mean values first and then discount and sum them up, or do the reverse according to the second way? We shall come back to this problem at the end of Section 10.2: the answer depends on the assumptions made in order to average and to discount time.

In order to apply the NPV, two major problems are to be solved. The first one is to determine what are the future cash flows. We shall assume here this problem is solved and cash flows are well anticipated. The second one is to take the project's risks into account, we shall analyse it further on in this section. There are several possibilities:

- We can use one among the individual criteria (expected utility or mean–variance, for example).
- Or we can refer to instruments derived from financial theory (CAPM, for example).

Let us assume future cash flows are given by different scenarios that correspond to the uncertain states: $s = 1, \ldots, S$, and probabilities p_s are assigned to the states. We can figure out the rate of return of the project for each scenario by the marginal efficiency on investment (internal rate of return) $\rho_s(P)$, which is given by a solution of each of the equations:

$$\forall_S = 1, \ldots, S \sum_{t=0}^{t=T} \frac{F_s(t)}{(1+\rho_s(P))^t} = 0$$

If there is a unique solution to the system of equations, then we obtain a non-random internal rate $\rho(P)$, but in general, the internal rate is random: $\rho(P) = (\rho_s(P))_{s=1,\ldots,S}$. It is possible to figure out the mean and the variance of such a rate:

$$E[\rho(P)] = \sum_{s=1}^{s=S} p_s \rho_s(P), \qquad Var[\rho(P)] = \sum_{s=1}^{s=S} p_s \{\rho_s(P) - E[\rho(P)]\}^2$$

We can therefore choose between two projects by eliminating the one that is dominated in terms of mean–variance: strictly lower (expected) return for a higher or equal risk (variance), or strictly higher risk for a lower or equal return.

However, this method is less convenient than the NPV for two reasons:

1. The internal rate of return assumes that intermediary payoffs can be reinvested at that rate, but this is rarely feasible in practice.
2. Mean–variance only defines a partial order, so that many projects cannot be compared.

10.2 DISCOUNTING

In order to confront these difficulties, criteria from individual decision theory can be called for. For instance, the most well-known criterion, the decision-maker's expected utility of a project P payoffs is:

$$EU(P) = \sum_{s=1}^{s=S} p_s u[F_s(0), \ldots, F_s(T)]$$

If, furthermore, we assume utility to be separable in time and additive, the expectation becomes:

$$EU(P) = \sum_{s=1}^{s=S} p_s \sum_{t=0}^{t=T} u_t[F_s(t)]$$

Let R be a given riskless interest rate, then the net present EU is:

$$NPEU(P) = \sum_{s=1}^{s=S} p_s \sum_{t=0}^{t=T} \frac{u[F_s(t)]}{(1+R)^t} = \sum_{t=0}^{t=T} \frac{1}{(1+R)^t} EU[F(t)]$$

This is the traditional result of discounted expected utility for future cash flows.

The two previous techniques, mean–variance and expected utility, share two drawbacks:

1. They require knowing the probability distribution over the different scenarios for cash flows.
2. They rely on a given riskless interest rate.

In order to avoid them, subjective probability, capacity or subjective deformation functions can be referred to, according to one of the generalisations of expected utility presented in Chapter 2. Similarly, the interest rate can be subjective, i.e. it would represent the decision-maker's preferences for present over future consumption. More generally, discount factors can represent preferences that do not satisfy the separability in time assumption (and then the criterion's additivity with respect to time). Such models are relevant for investors that are averse to (non-random) payoffs variations from period to period. Indeed, additivity with respect to time relies on the investor's indifference for hedging possibilities: hedging, in this case, means that another investment smoothes the variability from period to period. For example, a cash flow paying 1, 2, 3, 4 is perfectly hedged by another one paying 4, 3, 2, 1. The same argument that has been invoked to relax the independence axiom about uncertain payoffs can be called for here to relax the axiom of separability in time: comonotonic (in time) non-random cash flows cannot hedge each other, hence separability will hold for them, but the present value of the sum of two non-comonotonic (in time) non-random cash flows may not be the sum of their discounted payoffs.

Several works have achieved such results: Gilboa (1989), Shalev (1997), De Waegenaere and Wakker (2001), Chateauneuf and Rébillé (2004) notably. De Waegenaere and Wakker (2001) propose a non-additive measure of time (discount factors) that is not necessarily positive either: negative discount factors represent a preference for future over present consumption. Such a measure may be relevant for long-term consequences such as nuclear waste or climate change effects.

However, the future could be looked at the other way around:

- First, time could be measured by some economically relevant factor (a discount factor, say, valid for a given state of the world). This corresponds to a trajectory of a random process.
- Then, uncertainty about which state of the world obtains could be measured (by, say, a joint probability). This corresponds to summing up a set of trajectories by a mean trajectory.

It so happens that if separability, both over time and over uncertainty, is assumed, the two hierarchies entail the same formula: the NPV is obtained as the expected discounted payoffs (or payoff utilities). This is not true any more if any one of the two separability assumptions are not satisfied, as would be the case if the decision-maker's valuation formulae are expressed in terms of Choquet instead of the usual Lebesgue (additive) integrals.

According to the usual hierarchy, uncertainty (Ω, F) is valued and then time (T) is integrated to obtain the NPV.

Value of a random variable at time t, X_t: $\forall t \in T$, $U(X_t) = \int X_t(\omega)\mathrm{d}\upsilon(\omega)$
Value of process X: $UU(X) = \int U(X_t)\mathrm{d}\rho(t)$

If we proceed the other way around, we get

Value of trajectory $X(\omega)$: $\forall \omega \in \Omega$, $V[X(\omega)] = \int X_t(\omega)\mathrm{d}\rho(t)$
Value of process X: $W(X) = \int V[X(\omega)]\mathrm{d}\upsilon(\omega)$

If all integrals are Lebesgue's, $W(X) = UU(X)$ because of Fubini's theorem: the hierarchy is irrelevant. But let us assume that the decision-maker is averse to time variations in payoffs and uncertainty averse, both being captured by the measures ρ and υ being convex capacities.

Then the integrals above are Choquet's and integrating the following two processes will show why the hierarchy matters.

Let time and uncertainty be represented by $T = \{1, 2\}, \Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and the σ-algebra (information) at time 1 be $F_1 = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \varnothing, \Omega\}$. Let us consider two risks X and Y with payoffs:

$$X_1(\{\omega_1\}) = X_1(\{\omega_2\}) = 0,\ X_1(\{\omega_3\}) = X_1(\{\omega_4\}) = 1$$

$$X_2(\{\omega_1\}) = X_2(\{\omega_2\}) = 0,\ X_2(\{\omega_3\}) = 1, X_2(\{\omega_4\}) = 2$$

$$Y_1(\{\omega_1\}) = Y_1(\{\omega_2\}) = 1,\ Y_1(\{\omega_3\}) = Y_1(\{\omega_4\}) = 0$$

$$Y_2(\{\omega_1\}) = Y_2(\{\omega_2\}) = 0,\ Y_2(\{\omega_3\}) = 1,\ Y_2(\{\omega_4\}) = 2$$

Suppose that the decision-maker's discount factors are: $\rho(1,\ 2) = 1.8,\ \rho(1) = 0.9,\ \rho(2) = 0.8$. They show a preference for present consumption: $\rho(1) > \rho(2)$ and aversion for time variations: $\rho(1,\ 2) > \rho(1) + \rho(2)$(convexity). The decision-maker's subjective capacities on uncertain states are: $v(\{\omega_1,\ \omega_2,\ \omega_3,\ \omega_4\}) = 1,\ v(\{\omega_1,\ \omega_2,\ \omega_4\}) = 0.7,\ v(\{\omega_1,\ \omega_2\}) = v(\{\omega_3,\ \omega_4\}) = 0.4,\ v(\{\omega_4\}) = 0.2$. This corresponds to risk aversion because $v(\{\omega_1,\ \omega_2,\ \omega_3,\ \omega_4\}) > v(\{\omega_1,\ \omega_2\}) + v(\{\omega_3,\ \omega_4\}),\ v(\{\omega_1,\ \omega_2,\ \omega_4\}) > v(\{\omega_1,\ \omega_2\}) + v(\{\omega_4\})$ (convexity).

Then we have:

$$V[X(\omega_1)] = V[X(\omega_2)] = 0,\ V[X(\omega_3)] = \rho(1,\ 2), V[X(\omega_4)] = \rho(1,\ 2) + \rho(2)$$

$$W(X) = \rho(1,\ 2)v(\{\omega_3,\ \omega_4\}) + \rho(2)v(\{\omega_4\}) = 0.88$$

$$V[Y(\omega_1)] = V[Y(\omega_2)] = \rho(1),\ V[Y(\omega_3)] = \rho(2), V[Y(\omega_4)] = 2\rho(2)$$

$$W(Y) = \rho(2) + [\rho(1) - \rho(2)]v(\{\omega_1, \omega_2, \omega_4\}) + [2\rho(2) - \rho(1)]v(\{\omega_4\}) = 1.01$$

$$U(X_1) = v(\{\omega_3,\ \omega_4\}), U(X_2) = v(\{\omega_3,\ \omega_4\}) + v(\{\omega_4\})$$

$$UU(X) = \rho(1,\ 2)v(\{\omega_3,\ \omega_4\}) + \rho(2)v(\{\omega_4\}) = 0.88$$

$$U(Y_1) = v(\{\omega_1,\ \omega_2\}), U(Y_2) = v(\{\omega_3,\ \omega_4\}) + v(\{\omega_4\})$$

$$UU(Y) = \rho(1,\ 2)v(\{\omega_1,\ \omega_2\}) + \rho(2)v(\{\omega_4\}) = 0.88$$

The inequality $W(Y) > W(X)$ expresses the fact that Y hedges more than X against payoff variations in time and that the decision-maker is averse to these variations. According to the usual hierarchy, the criterion UU cannot capture this effect: $UU(X) = UU(Y)$.

In a first approximation, such considerations may be neglected and an additive valuation can be used to figure out the net present value of an investment. Such methods' limitations are relative to the difficulty to estimate the utility function, the subjective measure of uncertainty and the subjective discount factors. In practice, analysis is done using functions and factors known for their practicability for calculus, a poor argument in favour of their relevance!

Furthermore, such methods are founded on an individual criterion and, as such, are not relevant for a collective decision problem or for the management of a risky investment. These limitations are the same as those we detailed in Chapter 2 about decision-making under uncertainty in a static framework. The same arguments we proposed then, can be invoked in favour of methods founded on the theory of financial markets. Indeed, in this case, valuations are based on market prices that are obtained through a collective organisation to aggregate individual behaviours. Furthermore, they rely on existing instruments, which can then be used to manage and achieve the optimal decisions.

Finance theory tells us that a project valuation can be obtained by its discounted expected payoffs, if a market risk premium is integrated in the expectation (fundamental finance formula, see Chapter 7). In the CAPM (see below for the theory), for instance, the risk premium is directly expressed in the interest rate. The NPV of a project P can be figured out as:

$$NPV(P) = \sum_{t=0}^{t=T} \frac{E[F(t)]}{(1+R)^t}$$

where:

$R = R_0 + \beta(R_M - R_0)$ (CAPM formula, see below)
$E[F(t)]$ is the expected cash flow at time t
R_0 is the riskless rate
R_M is the mean rate of return of the market portfolio
β measures the sensitivity of the project to the market risk.

Referring to this method does not require us to define different scenarios for cash flows nor a utility function: only expected values of the project's payoffs are necessary. However, several problems have to be faced in order to apply it, because the risk premium is that of the firm (or the public sector) involved in the project, and not the risk premium of the project itself. Using one for the other is justified only if the project is in line with the usual activity of the firm. The firm's β is calculated using its market value, this means that the firm must have marketed shares with prices historically observed. In order to avoid these difficulties, it is possible to refer to sector β' instead of market ones (especially in the case of a public project), or sensitivity to the market risk of firms with activities related to the project. However, all such approximations are controversial. We have mentioned in Chapter 8 and we shall see in Chapter 12 that a way out of such controversies is to construct a portfolio of assets such that it mimics the project's risk.

10.3 STATIC MODELS OF FINANCIAL MARKET EQUILIBRIUM PRICING

In this section, the future is described as a whole by a set of states of nature about which no information will arrive until the given expiration date, as in the static decision theoretic models of Chapters 2 and 6. Therefore, concepts we have defined, such as risk aversion, etc., are underlying the models we present here. The first one is the famous Capital Asset Pricing Model (CAPM) that we have already mentioned. It has been proposed by Sharpe (1964) and by Lintner (1965) as an equilibrium model of the mean–variance behaviour formalised by Markowitz (1952). This minimal representation of the agents' behaviours in a financial market was the beginning of a theoretical splitting between economic models of decision-making under uncertainty and market finance which concentrates on valuation models able to help and manage investments. The first one favoured expected utility maximisation (Chapter 2) as an analytical tool. It is easy to show that the double criterion mean–variance can be considered, under some assumptions, as a special case of expected utility maximisation which makes it possible to present CAPM as a microeconomic model (with only one agent), this was done by Mossin (1966). We do not favour this approach because it concentrates on individual demand and loses the specificity of Markowitz' model. Furthermore, agents' utilities are not observable and they cannot be aggregated. Meanwhile, it is very reasonable to assume and check that agents who deal in a financial market do behave in accordance with the double criterion mean–variance.

In order to keep even closer to the spirit of finance, which favours valuation models on the basis of observable data and estimates, we present the CAPM as a special case of the static arbitrage pricing model proposed by Ross (1976): the Arbitrage Pricing Theory (APT). The concept of "no arbitrage" has been at the root of the development and spreading of financial market theory during the 1970s. This is because it is less constraining than the classical economic concept of equilibrium. It can be shown, however, that in most models the two concepts are equivalent under some restrictive assumptions on agents' behaviours and clearing conditions.

A static model assumes there is an expiration date (three months is often the reference, but any date that is relevant for the investment problem can do). From the present date at which valuation is done and decisions are made until this expiration date, all assets are described by their returns or by their rates of return. Returns of assets are assumed to be random, with probability distributions that are known or at least with parameters that can be estimated from past economic data. This assumption is similar to the one made in macroeconomic models where relevant variables are expressed as a function of some observable and significant factors.

Uncertainty that prevails on the financial assets market is described by an econometric model where returns are "explained" by some indexes and corrected by error terms corresponding to the idiosyncratic (or intrinsic) asset risks. Such indexes are obtained as an aggregation of relevant data such as production sector indexes, general economic indexes (GNP) and commodities and financial market indexes (S&P 500, Dow Jones, Euronext 50, etc.). In the CAPM, only one financial market index is referred to.[2] We shall detail the model further after we have explained the results.

In the theory, agents are price-takers (they can't have an influence on price formation and the market is a perfectly competitive one). Their expectations on assets returns are assumed to be homogeneous, meaning that the econometric representation of uncertainty is relevant and parameter estimations are known and available (means and variances, merely). Agents' rationality is such that prices must not offer them arbitrage opportunities, otherwise they would use them and prices would not be in equilibrium. In this context, no arbitrage can be expressed in the following way:

If an asset can be formed without investment costs[3] *and if it is riskless, then it cannot yield a strictly positive return.*

In order to have the possibility to construct a riskless asset (in the model this means that its return has a zero variance), two assumptions have to be made:

1. There exists a referential riskless asset with the considered expiration date (a government bond is usually chosen). This assumption is common to all no arbitrage models.
2. It is possible to form a portfolio of marketed assets such that its variance is (approximately) zero. This assumption relies on the number of marketed assets being sufficiently large for both extrinsic and intrinsic risks to be diversified away. Obviously, intrinsic risk, which is measured by the error term variance is only converging towards zero when the number of assets increases (Law of Large Numbers).

[2] Let us note that most indexes have been defined before the CAPM model existed. Furthermore some of them, the Dow Jones in particular, do not meet the requirements of the CAPM market portfolio.

[3] In practice, this means that long positions are compensated by short ones.

The APT model is developed in terms of returns: the expected return of an asset must be equal to a constant (corresponding to the riskless return) to which are added prices corresponding to the different indexes, weighted by the sensitivity factors of the asset to each index. The formula can also be expressed in terms of rate of returns,[4] so that we see better the similarity with the CAPM as a special case.

Formally, if i is an asset and r_i its rate of return, the APT yields:

$$E(r_i) = \lambda_0 + \sum_{K=1}^{K} \lambda_k \beta_{i.k} \qquad \text{(APT)}$$

where λ_0 is the rate of return of the riskless asset, $\lambda_1, \ldots, \lambda_k$ are the excess rates of return of the marketed risks corresponding to the indexes and the $\beta_{i,k}$ are the (estimated) sensitivity of asset i to index k.

In the special case where only a market index (a portfolio), say M, is sufficient to describe the relevant risk, if r_M is the rate of return of the market portfolio and if r_0 is the riskless rate of return, the formula collapses into the CAPM:

$$E(r_i) = r_0 + [E(r_M) - r_0]\beta_i \qquad \text{(CAPM)}$$

Here, the β_i is explicitly the ratio between the covariance of r_i and r_M and the variance of r_M. It measures the sensitivity of asset i to the market (portfolio) variations; this parameter is an important risk management instrument, as we shall see in Chapter 12. An easy way to visualise the CAPM formula is the Security Market Line (Figure 10.1), i.e. the set of efficient portfolios composed of the riskless asset and the market portfolio.

Let us now detail the APT model. Assume an asset is defined by its payoffs at some date: x, it is a random variable. Let $q(x)$ be its (present) price so that its return is $R(x) = x - q(x)$. If $q(x) \neq 0$, $r(x) = R(x)/q(x)$ is the asset's rate of return.

The rate of return of each marketed asset is expressed by a regression equation on indexes: $I_1, \ldots, I_K$. They are assumed to be non-correlated and we have:

$$r(x) = \beta_0 + \sum_{k=1}^{K} \beta_{x,k} I_k + \varepsilon_x$$

where ε_x is the idiosyncratic error term.

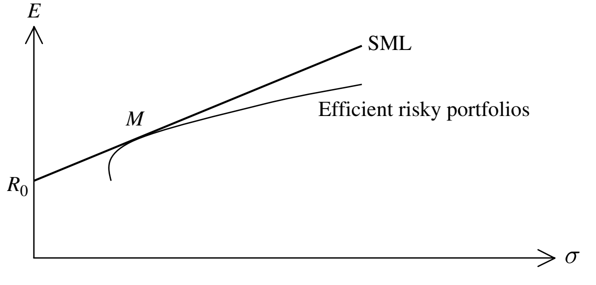


Figure 10.1 The Security Market Line

[4] Under the condition that present prices are not zero.

Let $x_1, \ldots, x_N$ be marketed assets, their prices are known and are: $q(x_1), \ldots, q(x_N)$ we assume they are not zero $(q(x_i) > 0)$. A portfolio is formed of quantities $\theta(x_i)$ of each of the marketed assets. Quantities are counted positively for a long position, negatively for a short one. A portfolio defines a new (marketable) asset defined by its payoffs:

$$x_\theta = \sum_{i=1}^{N} \theta(x_i) x_i$$

Assume it is possible to form such a portfolio at zero cost:

$$K(\theta) = \sum_{i=1}^{N} \theta(x_i) q(x_i) = 0$$

This is possible if long positions, $\theta(x_i) \geq 0$, are compensated by short ones, $\theta(x_i) \leq 0$.

Assume furthermore that this portfolio is built in such a way that it is riskless (i.e. its variance is zero). This is a way to express that the market is complete (the riskless asset can be replicated by marketed assets, and in fact, it is easy to show that any asset can be replicated by marketed assets). In practice, the result is only obtained as a limit when the number of marketed assets in the portfolio is large enough:

$$Var(x_\theta) = \sum_{i=1}^{N} \sum_{j=1}^{N} \theta(x_i)\theta(x_j) cov(x_i x_j) = 0$$

No arbitrage is then expressed as:

$$[K(\theta) = 0 \text{ and } Var(x_\theta) = 0] \Rightarrow E(r(x_\theta)) = 0$$

A geometrical argument is then invoked[5] to show that there exist $K + 1$ numbers: $\lambda_0, \lambda_1, \ldots, \lambda_k$ such that:

$$\forall i = 1, \ldots, N, \quad E(r(x_i)) = \lambda_0 + \sum_{k=1}^{K} \lambda_k \beta_{i,k}$$

We have seen how this formula (the CAPM in fact, but the argument works the same in the general APT case) relates to the Fundamental Finance Formula (FFF) in Chapter 7: the price of an asset in a complete market can be expressed as the expectation of its discounted payoffs with respect to the risk-adjusted measure.

10.4 THE VALUE AT RISK

We have defined the VaR already in Chapter 6 in order to characterise risk increases. The VaR can be preferred to the variance as a risk measure for two reasons:

1. Variance is symmetrical so that it measures the risk as a whole, not the risk about losses.
2. Many financial assets rates of returns have been shown empirically to have an infinite instantaneous variance.[6]

[5] It is based on orthogonals, as expressed by the equations, and parallels, thus implying a linear combination.

[6] Instead of Gaussian, Lévy-stable distributions are estimated. Their cumulative distribution functions have the same shape, but their tails are thicker.

Referring to variance, or to models based on it (CAPM), entails an under-weighting of the risk of losses. If the left "tails" of the cumulative distribution functions are considered, losses are easily measured by the left (negative) percentiles: this is the VaR.

Choosing a probability level, say 5%, we can directly read on the cumulative function graph the losses that can be incurred with this probability: this is the $VaR_{5\%}$. Say that a portfolio has a $VaR_{5\%} = €10m$, a more prudent manager will want to stop losses at a lower probability level, for instance choose a $VaR_{1\%}$ and construct a new portfolio in such a way that its $VaR_{1\%} = €10m$.

VaR is widely used to measure the risk level that credit firms must not overcome. It is also used to value performances of portfolio managements. As a managing instrument, it is directly called for in order to select assets in a portfolio. It must be noticed that, instead of referring to the mean–variance analysis, a mean–VaR analysis can be conducted. In the case of Gaussian distributions, both yield the same efficient frontier.

In Chapter 12 we shall come back to the VaR in more detail as an instrument to manage (i.e. to limit) losses.

11 Valuation with Options

The valuation models that have been successful since the 1970s are based on a dynamic measure of risks.

11.1 GENERAL THEORY OF A DYNAMIC MEASURE OF RISKS

In Chapter 7, the risk valuation model we presented relied on a static representation of the economy: no decisions were to be taken after the initial date. This approach is based on the perfect forecast assumption: future prices of commodities are perfectly forecasted at equilibrium by agents, at the initial date. There is no need to have a probability distribution over the future uncertain states,[1] because agents can hedge any risks thanks to the marketed assets if the financial market is complete.

If the number of marketed assets is not sufficient to yield any chosen income level in a future state, then the risk is not hedged and decisions may be modified if some states occur. In this case, agents will need to trade again in order to adjust their incomes in some situations. As an illustration of the problem, let us assume we face a risk of flood and there are no insurance companies that are willing to insure it. During the winter, assume an unexpected lot of snow falls on the surrounding mountains, then we know that a spring flood is more likely than if it had not snowed. In order to be prepared to face a terrible flood and in the absence of insurance, there is a need to put some money aside, or invest in some protection device if any exists. This is a dynamic behaviour, it entails some adjustments of the decisions as a function of information arrivals. The difference between a static market model and a dynamic one comes from the lack of possibility to hedge completely any risk without adjusting financial positions to information arrivals.

The dynamic model of market equilibrium we shall present, shares some common characteristics with the static one:

- There are two types of markets: one for assets where financial risks can be hedged, and one for commodities, in each future state.
- Future prices of commodities and of assets are perfectly forecasted at equilibrium.

As we shall see, there will be no need of as many assets in order to hedge risks in the dynamic model as in the static one. The model is closer to real situations in the sense that risks management in real life is indeed adapted to information arrivals.[2]

[1] At least in the finite set of states model.
[2] See Duffie (1988), (1992), Dana and Jeanblanc-Picqué (2003).

Time is introduced, it can be represented as a discrete parameter: $t = 1, 2, \ldots T$, or as a continuous one: $t \in [0, T]$. Technically, the first representation is simpler than the second one, however both yield similar results as we shall see. The choice of one over the other relates to their relevance to applied problems, as we shall see in Chapter 12.

11.1.1 Discrete Time Valuation

The dynamic version of the perfect foresight equilibrium model is based on a representation of future uncertainty by an event "tree", or lattice, similar to decision lattices we saw in Chapter 4. See Figure 11.1.

The finite set of states of the world that represent uncertainty is given by the tree's nodes: $S = \{s_j(t)/j = 1, \ldots, J(t); t = 1, \ldots, T\}$. A financial asset y yields payoffs $y(s)$ in state $s \in S$, it is formalised by a function $y : S \to R_+$. Let the set of marketed assets be: $Y = \{y_1, \ldots, y_N\}$. The price of asset y in state $s \in S \cup \{s(0)\}$ is $q_s(y)$. An asset portfolio θ is defined by assets held in each state $s : \theta_s = \{\theta_s(y_1), \ldots, \theta_s(y_N)\}$. The formation cost of θ in s is:

$$K_s(\theta) = \sum_{y \in Y} \theta(y) q_s(y)$$

The payoff of portfolio θ in state s, is:

$$\Theta(s) = \sum_{y \in Y} \theta(y) y(s)$$

In this model, agents can adjust their portfolios, they will decide on a portfolio strategy instead of a static portfolio. A strategy depends on information arrivals, this introduces a chronological order according to which transactions can take place.

Consider a node, on the event tree:

- State s occurs.
- Assets are traded at prices $q_s(y)$.
- Assets pay (dividends) $y(s)$.

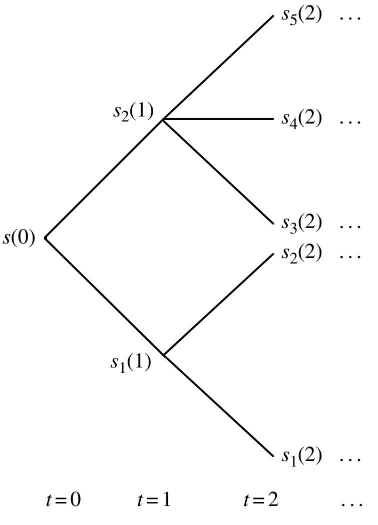


Figure 11.1 Event tree

A portfolio strategy assigns a portfolio to each state of the world: $\theta = (\theta_s)_{s \in S \cup \{s(0)\}}$. Payoffs are more complex for a strategy than in the static case: assets pay positive or negative amounts depending on their position, furthermore, adjustments may entail money coming in or out. Payoffs are the sum of two terms:

$$\forall s(t) \in S, \quad \Theta[s(t)] = \sum_{y \in Y} [\theta_{s(t-1)}(y) - \theta_{s(t)}(y)] q_{s(t)}(y) + \sum_{y \in Y} \theta_{s(t)}(y) y[s(t)]$$

where $s(t-1)$ is the unique predecessor of $s(t)$ in the event tree.

In a dynamic model, a strategy may start in any event node, it need not be defined as from the initial state. If the portfolio strategy θ starts in state s, its formation cost will be:

$$K_s(\theta) = \sum_{y \in Y} \theta_s(y) q_s(y)$$

A portfolio strategy starting in state s can be considered as a marketable asset, it is characterised by its payoffs in states that follow s. Let us denote by $S(s)$ the subtree following state s.

The general structure of the model is similar to the static one, however the no arbitrage assumptions must be adapted to intermediary states.

No arbitrage assumption 1 *If two marketable assets formed in a state s have the same payoffs in all states following s, then any portfolio strategy that yields these payoffs must have the same formation cost in s:*

$$\begin{aligned}
\forall \Theta &= [\Theta(\omega)]_{\omega \in S(s)}, \forall [\Theta'(\omega)]_{\omega \in S(s)}, \quad \{\forall \omega(t) \in S(s), \Theta[\omega(t)] \\
&= \sum_{y \in Y} [\theta_{\omega(t-1)}(y) - \theta_{\omega(t)}(y)] q_{\omega(t)}(y) + \sum_{y \in Y} \theta_{\omega(t)}(y) y[\omega(t)] \\
&= \sum_{y \in Y} [\theta'_{\omega(t-1)}(y) - \theta'_{\omega(t)}(y)] q_{\omega(t)}(y) + \sum_{y \in Y} \theta'_{\omega(t)}(y) y[\omega(t)] = \Theta'[\omega(t)]\} \\
&\Rightarrow K_s(\theta) = K_s(\theta')
\end{aligned}$$

where $\omega(t-1)$ is the unique predecessor of $\omega(t)$ in the event tree.

No arbitrage assumption 2 *If a marketable asset formed in state s has non-negative payoffs in all states following s and at least a strictly positive payoff in one state, then any portfolio strategy with these payoffs has a strictly positive formation cost in s:*

$$\forall \Theta = [\Theta(\omega)]_{\omega \in S(s)}, \quad \{\forall \omega(t) \in S(s), \Theta[\omega(t)] \geq 0, \exists \omega(t) \in S(s), \Theta[\omega(t)] > 0,$$
$$\Theta[\omega(t)] = \sum_{y \in Y} [\theta_{\omega(t-1)}(y) - \theta_{\omega(t)}(y)] q_{\omega(t)}(y) + \sum_{y \in Y} \theta_{\omega(t)}(y) y[\omega(t)]\} \Rightarrow K_s(\theta) > 0$$

These assumptions imply that the price functional is a positive linear form of marketable asset payoffs, as in the static case. Here, however, this is true also for intermediary dates and states.

Theorem (extension) *Under the two no arbitrage assumptions, the function $\bar{q}_s$ defined by $\bar{q}_s(\Theta) = K_s(\theta)$, where:*

$$\forall \omega(t) \in S(s), \quad \Theta[\omega(t)] = \sum_{y \in Y} [\theta_{\omega(t-1)}(y) - \theta_{\omega(t)}(y)] q_{\omega(t)}(y) + \sum_{y \in Y} \theta_{\omega(t)}(y) y[\omega(t)]$$

with $\omega(t-1)$ the unique predecessor of $\omega(t)$ in the event tree, is a positive linear form and its restriction to Y is q_s. Furthermore, there exists a positive function $\gamma_s : S(s) \to R$, such that:

$$\bar{q}_s(\Theta) = \sum_{\omega \in S(s)} \gamma_s(\omega)\Theta(\omega)$$

The result holds in particular for the initial state $s(0)$ and yields the initial price. Then the static result is a special case. We recognise in the valuation formula at each state, the Fundamental Finance Formula (FFF) we saw in Chapter 7. Indeed, $\bar{q}_s(\Theta)$ is the expectation of Θ's payoffs with respect to a measure defined by γ_s: the risk-adjusted measure in state s. The complete market assumption entails the uniqueness of the linear form and of this measure, as in the static case.

Definition (complete market) *The set Y of marketed assets forms a complete market if the portfolio strategies can replicate any payoff in each state in S:*

$$\forall Q : S \to R,\ \exists\, \theta \text{ s.t.}$$

$$\forall s(t) \in S, \quad \sum_{y \in Y} [\theta_{s(t-1)}(y) - \theta_{s(t)}(y)] q_{s(t)}(y) + \sum_{y \in Y} \theta_{s(t)}(y) y[s(t)] = Q[s(t)]$$

where $s(t-1)$ is the unique predecessor of $s(t)$ in the event tree.

The number of assets necessary to have a complete market need not be very high, compared to the number of states. The minimum necessary number is called the spanning number, it is equal to the maximum number of successors to any node in the event tree. In the special case where each node has only two successors (binomial tree), then two assets are sufficient to get a complete market. If assets are well chosen, any income level in all states can be obtained by a portfolio strategy. Otherwise stated: marketed assets span the vector set of payoffs.

Arbitrage valuation in a dynamic model generalises the results obtained in the static case. The FFF (expressed by the above theorem) extends to a stochastic process property: the martingale property.

A stochastic process[3] $(X_t)_{t \in T}$ is a martingale with respect to some probability distribution μ, if the value at time t, X_t is the expectation of the value at time $t+1$, conditional on information I_t available at date $t : X_t = E_\mu[X_{t+1}/I_t]$.

Let's illustrate the martingale property by a simple example: playing €1 on heads against €1 on tails with a fair coin. The cumulative gain process in this game is a martingale with respect to the empirical distribution. Indeed, let €$G(t)$ be the sum of the gains up to time t, given the gain expectation for the next draw is zero, the expectation of the cumulative gains in $t+1$ is still €$G(t)$.

The martingale property is obtained for the cumulative gain process of marketable assets, with respect to the risk-adjusted distribution and not with respect to some empirical distribution (which is not mentioned in the finite model, anyway). Indeed, if the martingale property was satisfied for the empirical distribution, it would mean that all assets would pay the riskless rate. In fact, risky assets pay the riskless rate plus a risk premium. The risk premium is integrated in the risk-adjusted distribution and this is why the martingale property is satisfied.

---

[3] That is, a sequence of random variables indexed by time.

Let $P(s)$ be the set of strict predecessors of a state s, the payoffs flow of an asset y held up to state s and then sold in s, is:[4]

$$z(s) = \sum_{\omega \in P(s)} y(\omega)$$

There is then a gain process resulting from buying y in 0 and selling it in state s:

$$Z(s) = z(s) + q_s(y)$$

If s occurs at date t, and assuming a constant riskless interest rate,[5] the discounted gain process is:

$$Z^a[s(t)] = \frac{Z[s(t)]}{(1+R_0)^t}$$

We can now state the martingale property for gain processes.

Theorem (martingale property) *The discounted gain process of an asset is a martingale with respect to the risk-adjusted measure:*

$$E_{v_{s(t)}}[Z^a(\omega(t+1))/\{s(t)\}] = Z^a(s(t))$$

As a special case, consider an asset y that pays no dividends until the expiration date T. Its gain process is then identical to its price process, so that the martingale property becomes:

$$q_{s(t)}(y) = \frac{1}{(1+R_0)} E_{v_{s(t)}}[q_{\omega(t+1)}(y)/\{s(t)\}]$$

If the formula is expressed for the total time interval from 0 to T, we get:

$$q_0(y) = \frac{1}{(1+R_0)^T} E_v\{y[\omega(T)]\}$$

We recognise the FFF obtained in the static model with the riskless rate for time T obtained as a composition of instantaneous constant rates for each time interval. Indeed, if there are no intermediary payoffs, the static and the dynamic cases are not fundamentally different, we shall see this in some of the applications in Section 11.2. However, with intermediary payoffs, the problem can be seriously more complex.

This financial market model can be integrated into a perfect foresight equilibrium dynamic model in the same way as we did in Chapter 7 for the static case. The main difference lies in the budget constraints of consumers: they have to adapt to portfolio adjustments according to prices in each state and date. We obtain also an equivalence with the general equilibrium model, as is expressed in the following theorem.

Theorem (equivalence) *Under the assumptions of no arbitrage and complete market in an economy characterised by the consumption sets, preferences and initial endowments of agents: $(X_i, u_i, w_i)_{i=1,\ldots,I}$, if there exists a general equilibrium with price vector p^*, then there exists a perfect foresight equilibrium with price vectors $(\hat{p}, \hat{q})$ and reciprocally. They are such that agents' endowments are identical in both equilibrium and prices satisfy:*

$$\forall s \in S, \;\; p^*(s) = \gamma(s)\hat{p}(s)$$

---

[4] Following the chronological order we made precise.

[5] The case where the riskless rate varies from one period to the other does not introduce major difficulties.

where γ is defined by:

$$\forall y \in Y,\ \hat{q}_0(y) = \sum_{s \in S} \gamma(s) y(s)$$

$$\forall s \in S, \hat{q}_s(y) = \sum_{\omega \in S(s)} \gamma_s(\omega) y(\omega)$$

$$\forall s \in S, \forall \omega \in S(s), \quad \gamma_s(\omega) = \frac{\gamma(\omega)}{\gamma(s)}$$

Because agents' endowments are the same in both models, the two welfare theorems hold. This yields some justification to the use of market prices for risks as a collective valuation basis. All of these results can be extended to the case where the set S is not finite, without major difficulties.

11.1.2 Valuation in Continuous Time

The structure of an asset market in continuous time that is compatible with a general equilibrium model requires us to solve several conceptual and technical problems. We shall concentrate on the concepts.

In contrast with the previous model, valuation in continuous time requires a probabilised description of uncertainty. In fact, not any kind of uncertainty representation can be accommodated in the models: both in the general valuation approach of this subsection and the valuation of a specific asset in the next section, uncertainty will be assumed to be generated by Brownian motions,[6] the reason why will appear during the developments.

Definition (Brownian motion) *The standard Wiener process, also called Brownian motion, $(B(t))_{t \in [0,T]}$ is defined by:*

(1) $B(0) = 0$.
(2) $\forall t, \forall t', 0 \le t \le t' \le T, B(t') - B(t)$ *follows a zero mean and* $t' - t$ *variance Gaussian distribution.*
(3) *The process increases are independent:*

$\forall t \in [0, T], \forall \tau, 0 \le \tau \le T - t, \forall t_i = t_1, \ldots, t_n, 0 < t_1 < \ldots < t_{n+1} = t, B(t + \tau) - B(t)$ *is independent of* $B(t_{i+1}) - B(t_i)$

(4) *The process trajectories are continuous: for almost all states* $s \in S$*, the function* $t \to B(t, s)$ *is continuous.*

The Brownian motion has several interesting properties that make it a relevant way to describe price behaviours:

- It is a Markov process. This means that only the last information is relevant for the probability on future uncertainty: the probability that the process takes some future value in a given set, conditioned on past information, only depends on the last available information: the past trajectory does not count, all that counts is the last realisation: the process has no memory.
- Its increases are Gaussian and independent random variables.

[6] Very few general results can be obtained out of this setting for mathematical reasons.

The justification of using Brownian motions as an approximation is found in the Central Limit Theorem we have seen in Chapter 2. It is a characterisation of a random walk[7] resulting from the combination of many independent processes.

As we have seen, the discounted prices of assets do not satisfy the martingale property with respect to the empirical distribution, in general. In order to model asset price processes, it is necessary to add a trend to the Brownian motion: this defines the general Wiener process characterised by the stochastic differential equation:

$$\mathrm{d}W(t) = \mu\ \mathrm{d}t + \sigma\,\mathrm{d}B(t)$$

It is a Markov process with continuous trajectories (but differentiable nowhere) with independent increases such that:

$\forall\, t, \forall\, t', 0 \leq t \leq t' \leq T$, $W(t') - W(t)$ follows a Gaussian distribution: $N(\mu(t' - t), \sigma^2(t' - t))$

Even more general, an Ito process $(X(t))_{t\in[0,T]}$ is a Markov process with continuous trajectories (diffusion process) such that there exist two functions μ and σ satisfying:

$$\mu(x, t) = \lim_{h\to 0} \frac{E[X(t+h) - X(t)/X(t) = x]}{h}$$
$$\sigma^2(x, t) = \lim_{h\to 0} \frac{Var[X(t+h) - X(t)/X(t) = x]}{h}$$

μ is the instantaneous expectation, also called trend, or drift, and σ is the instantaneous standard error, also called diffusion coefficient or volatility.

The special case where μ and σ are constants is the general Wiener process, and in the case where $\mu = 0$ and $\sigma = 1$, we are back to the Brownian motion.

In the valuation model in continuous time, the set of states S is generated by M Brownian motions $(B_1, \ldots, B_M)$. This assumes that an empirical distribution describes uncertainty, it is assumed to be known and referred to by all agents.

A financial asset is defined by a function $y: S \to R_+$. The finite set of marketed assets is $Y = \{y_0, y_1, \ldots, y_N\}$, where y_0 is the riskless asset with price equal to 1, by definition. The price of asset y in state $s \in S$ is $q_s(y)$. Asset payoffs and prices are processes $[y(t)]_{t\in[0,\,T]}$ and $[q(y)]_{t\in[0,\,T]}$. The chronological order introduced in the preceding model is still referred to here, for each infinitesimal time interval $\mathrm{d}t$.

We keep the same structure for the financial asset market as the one we had in discrete time. The "law of unique price" (no arbitrage assumption 1) entails that if a payoffs process formed in a state s can be generated by two different portfolio strategies, both must have the same formation cost in s. The "no free lunch rule" (no arbitrage assumption 2) implies that a portfolio strategy with positive payoffs (and at least one strictly positive payoff) must have a strictly positive formation cost.

Both assumptions, expressed in continuous time, yield the extension result of the FFF obtained in the discrete time model:

The formation costs of portfolio strategies define a price functional on marketable assets that is positive and linear.

[7] Brown represented by this process the movement of pollen particles in water.

The complete market assumption says that any payoff process can be obtained through a portfolio strategy of marketed assets, it yields the uniqueness of the linear form, and then of the risk-adjusted measure it defines.

The necessary number of marketed assets to have a complete market is $M+1$, i.e. M[8] marketed risky assets plus the riskless asset. In the case we shall encounter in applications, there is only one Brownian motion. Then only one risky asset plus the riskless asset are sufficient to complete the market. This comes from the possibility (in the model and approximately in practice) to continuously modify the portfolio strategies in order to generate an infinity of payoff schemes.

In the continuous time model, the martingale property is obtained:[9] the payoff flow of an asset y held until date t and sold in t, is:

$$z(t) = \int_0^t \mathrm{d}y(\tau)$$

The gain process corresponding to the strategy: buying y at time 0 and selling it at time t is still:

$$Z(t) = z(t) + q_t(y)$$

With a constant riskless interest rate,[10] the gain process at date t is:

$$Z^a(t) = exp[-R_0 t]Z(t)$$

Theorem (martingale property) *The discounted gain process of an asset is a martingale with respect to the risk-adjusted distribution:*

$$\forall t, \forall t', 0 \le t \le t' \le T, \quad E_{v(t)}[Z^a(t')/I(t)] = Z^a(t)$$

where $I(t)$ is the information at date t.

In this model, the risk-adjusted distribution is obtained as a modification of the empirical one. The risk-adjusted distribution is absolutely continuous with respect to the empirical one, so that there exists a Radon–Nikodym derivative such that, if μ is the empirical distribution, $\mathrm{d}v/\mathrm{d}\mu = f$, where the function f integrates the risk premium. Then $E_v(y) = \int y.f\,\mathrm{d}\mu$.

In the special case where asset y does not pay any dividends before the expiration date T, its gain process is equal to its price process and the martingale property reduces to:

$$q_t(y) = exp[-R_0(T-t)]E_{v(t)}[y(T)/I(t)]$$

Valuation at date 0 is then:

$$q_0(y) = exp(-R_0 T)E_v[y(T)]$$

and we recognise the FFF obtained in the static case in Chapter 7.

The link between the perfect foresight equilibrium and the general equilibrium models is however slacker in this setting:[11] We have an "only if" instead of the "if and only if" obtained in the other models.

---

[8] M characterises the multiplicity of the martingale, i.e. the number of Brownian motions necessary to describe the relevant uncertainty in the model.

[9] See Harrison and Kreps (1979), Harrison and Pliska (1981).

[10] Discounting is done by an exponential in continuous time.

[11] See Duffie and Huang (1985).

Theorem *Under the assumptions of no arbitrage and complete market in an economy characterised by the consumption sets, preferences and initial endowments of agents, if there exists a general equilibrium, then there exists a perfect foresight equilibrium yielding the same endowments to the agents.*

As a consequence, the first welfare theorem tells us that a central planner can select a general equilibrium set of Pareto-optimal endowments for agents. This set of endowments could be obtained by a decentralisation through financial markets. Conversely, equilibrium endowments obtained by financial markets at a perfect foresight equilibrium may not be obtained by a general equilibrium model and then nothing proves that it would be Pareto-optimal.

11.2 APPLICATIONS TO RISK VALUATION

The theory and the foundations of the risk valuation in dynamic models are well established. From some well-defined marketed asset prices, a measure (the risk-adjusted measure) can be deduced and applied to other risks that are not traded but are marketable through portfolio strategies. This yields these risks a value, an arbitrage one and even an equilibrium value, that they would have if they were actually traded.

However, the applications of this method can encounter conceptual and technical problems. Among such applications, more difficulties are faced for investment valuation than for financial assets valuation. There are two reasons for this:

- The first one concerns asset payoffs.
- The second one is that financial markets abound in data about prices, while new investments can only be compared to similar, already existing ones.

Financial assets payoffs are well defined by contracts. For instance, fixed rate riskless bonds (government bonds) are defined without any ambiguity when they are issued. They pay a coupon, an annual one in general, which is a fixed percentage of the facial value, until maturity. At the expiration date, the reimbursement is added to the last coupon. All these are defined at the issuing date, hence the payoffs are known with certainty by all the agents.

In the case of a variable rate bond, the coupon is defined by a variable percentage of the nominal. It is yielded by a reference to a market rate, for instance the EURIBOR 3 months at the issuing date. Then payoffs are a deterministic function of this observable variable. Such a bond can then be considered as an asset contingent on this variable: a derivative asset on this underlying variable (e.g. the EURIBOR index). If this underlying variable is a rate, or a price, the arbitrage valuation can be directly applied: the risk-adjusted measure can be deduced from the underlying price, or rate, and it yields the value of the contingent asset.

Some assumptions have to be satisfied, they depend on the model chosen to represent the underlying variable. Furthermore, the method relies heavily on the deterministic function relating the bond's payoffs and the underlying variable. If the bond is actually issued on a market, then this function is institutionally defined.

These properties are usually not satisfied by a real investment. We have evoked the same difficulty in Part II about controversial risks. In the best of cases, it is possible to define a natural underlying variable, but the deterministic function has to be estimated. For instance, the applications to copper mines valuation that have often been studied in the finance literature, refer to future copper prices as the natural underlying variable to future profits. However, the relationships between copper prices and profits are not well known. It is then necessary to construct or to estimate such a relation between the two variables and this introduces some arbitrary choices. Furthermore, it is often essential to introduce some other explanatory variables of the profit (such as salaries). All these difficulties can be overcome by solving problems that are not simple indeed, before applying the theoretical results.

The other reason why application of arbitrage valuation to investments is difficult is that there are few available data, in general. This is in contrast with the case of financial assets, for which data are plenty. With few observed reliable data, it is difficult to formalise the uncertainty relevant to the valuation of an investment. We have seen in Part II that some methods based on discrete data are available. The interest of such an approach is that, in the case of discrete models, we do not need to estimate an empirical distribution.

If furthermore, the risk (investment) only pays at a unique date, then valuation is similar to the one obtained in a static model. But many risks pay at several dates. Then, valuation can only be achieved by backward induction. In such a dynamic model, if the underlying variable is not directly linked to a market value, it is often necessary to refer to optimal strategies of agents, instead of a purely objective valuation based on market prices. Indeed, optimal choices depend on available information and on expectations about future information (option values). These are not completely integrated by market values if the investment or the risk to be valued is not clearly related to marketed assets.

11.2.1 Financial Risks

The development of financial markets as from the 1970s went together with the creation of new financial instruments and the multiplication of the number of traded assets. We shall not look over all of these assets and the specificities of their markets and their uses, but we shall present some features of risk analysis that they have in common. To make things easier, we shall concentrate on the valuation of options. There are two motivations for this choice:

- Historically they were the first new assets to be priced by arbitrage methods in a dynamic model, as such they are representative of the whole class of derivatives.
- The other motivation is related to this type of instrument that can be generalised to real investment problems (often referred to as "real options" valuation).

Starting with the simplest case, we shall first consider the so-called European options that can be exercised at their expiration date only. Then, we shall evoke the supplementary difficulties that arise if the exercise can take place at intermediary dates (American options).

European option valuation will be studied in a discrete time model (Cox, Ross and Rubinstein (1979), binomial model), then in continuous time (Black–Scholes (1973) and Merton (1973b), Ito process models).

A call is a buying option (a put is a selling option), it gives the holder the right, without obligation, to buy (sell) a financial instrument: the underlying asset. This right can be exercised at a fixed date (the expiration date) for a European option, up to this date for an American one, at a fixed price: the strike price. The underlying asset can be of a different nature: security, index, rate, currency, future, commodity, etc. We shall consider the standard case of an underlying security, the other cases are similar but some complexities are added because the "price" of the underlying asset has first to be defined.

Options are used as instruments in many applications: hedging against the increase or the decrease of the underlying asset, speculation (lever effect), or for taking advantage of arbitrage opportunities. They can be integrated in complex strategies, whether static or dynamic ones, and their popularity comes partly from this flexibility.

An option price is called a premium, it cannot be negative because there are no negative payoffs (the option opens possibilities but entails no obligations to its holder). In order to figure this price out, some assumptions on the underlying asset price must be made. We shall assume that the underlying security yields no dividends during the maturity of the option, to make things simpler in our first approach. Furthermore, we assume the underlying security prices follow a binomial process that is described by Figure 11.2.

S is the price of the security at the initial date ($S > 0$), u and d are strictly positive parameters such that, if R_0 is the (constant per period) interest rate: $0 < d < 1 + R_0 = r < u$.[12] Let a call on this security be defined by its expiration date T and strike price K. This European call pays at date T: $\max[S(T) - K, 0]$, where $S(T)$ is the security price at T.

Because of the particular structure of the binomial lattice, we shall see that, at each date and in each state, it is possible to replicate the call's payoffs with a portfolio formed of the underlying security and the riskless bond (two linear equations and two unknowns). Hence the market formed of these two marketed assets is complete and the call can be valued by arbitrage.

Let's assume we are at date $t = 0, 1, \ldots, T - 1$, the price of the security is $u^j d^{t-j} S$, where $j = 0, 1, \ldots, t$ is the number of ups between 0 and t, and where $t - j$ is the number of downs. The price of a riskless unit asset (paying 1) is r^t, where the constant interest rate is $r - 1$. Let us denote by $s = u^j d^{t-j}$ the corresponding state and let's agree that date 0 corresponds to a state 0. A portfolio formed in state s is $\theta(s) = (\delta(s),\ \beta(s))$, where $\delta(s)$ is the quantity of the security and $\beta(s)$ the quantity of the riskless asset. If at date t and in state s, the security price is sS then, at date $t + 1$, in the successor states of s, it can take only two values: usS if it increases and dsS if it decreases.

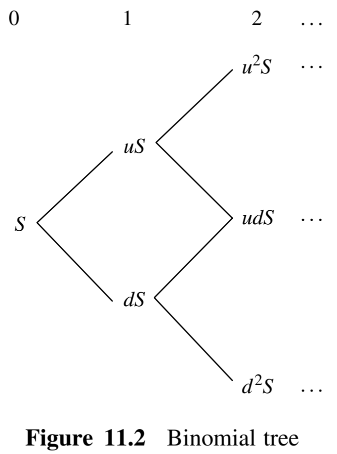


Figure 11.2 Binomial tree

[12] This inequality is in fact a no arbitrage condition: if the riskless rate was superior to $u - 1$ say, no agent would buy the security so that its price would be zero.

Portfolio $\theta(s)$ has a formation cost at date $t: K[\theta(s)] = \delta(s)sS + \beta(s)r^t$. Its possible values at date $t+1$ are:

$\delta(s)usS + \beta(s)r^{t+1}$, if the security increases
$\delta(s)dsS + \beta(s)r^{t+1}$, if not.

Let us write: $C(s)$, $C(us)$ and $C(ds)$ for the values of the call in the states s, us and ds. The portfolio $\theta(s)$ will be chosen to replicate the payoffs (prices) of the option at date $t+1$: this yields the two linear equations with unknowns $\delta(s)$ and $\beta(s)$ and a unique solution. Then, under no arbitrage, the price of the call must be equal to the formation cost of the replicating portfolio and we have:

$$C(s) = \delta(s)sS + \beta(s)r^t = \frac{1}{r}\left[\frac{r-d}{u-d}C(us) + \frac{u-r}{u-d}C(ds)\right]$$

Let $p = \frac{r-d}{u-d}$, it must be remarked that $0 < p < 1$ and $1 - p = \frac{u-r}{u-d}$. Then we have:

$$C(s) = \frac{1}{r}\left[pC(us) + (1-p)C(ds)\right] = \frac{1}{r}E_p\left[C(t+1)/\{s\}\right]$$

The martingale property is satisfied: the option's price $C(s)$ is the discounted expectation of its future payoffs, with respect to the risk-adjusted probability measure: $(p,\ 1-p)$.

This formula is simple because of two specificities:

- The call does not pay off before the expiration date, so that its gains are its prices.
- The riskless rate is assumed to be constant until the option's maturity.

An alternative way to obtain the call's price would have been to apply the martingale property instead of constructing the replicating portfolio. The martingale property implies that there exists a probability distribution $(p,\ 1-p)$ such that the price of a marketed asset (here, the security) satisfies:

$$S(s) = sS = \frac{1}{r}E_p\left[S(t+1)/\{s\}\right] = \frac{1}{r}\left[pusS + (1-p)dsS\right]$$

From this equation, p is obtained and the martingale property is applied again to the call price and yields the formula for $C(s)$. We have calculated the risk-adjusted distribution from the prices of the marketed assets (here only one is sufficient), it is unique because the market is complete with two assets over one period (and dynamically in this model over the whole time interval). Then we can use it to price any marketable asset: in particular, the call is marketable because it can be replicated by a portfolio.

If we apply the martingale property at successive dates, we get the option's price at the initial date: again, it is the discounted expectation of its final payoffs with respect to the risk-adjusted distribution. The risk-adjusted distribution is obtained from those we had used for one-time intervals (they are all the same because of the constant parameters). Indeed, at each date, the payoffs (prices) are independent random variables with the same distribution, so that the well-known result over T periods is a binomial distribution. The probability to reach state $s = u^j d^{T-j}$ at date T is $\binom{T}{j}\ p^j\ (1-p)^{T-j}$. The option's initial price is then:

$$C(0) = \frac{1}{r^T}\sum_{j=0}^{j=T}\binom{T}{j}p^j(1-p)^{T-j}\max(u^j d^{T-j}S - K,\ 0)$$

This is the Cox, Ross and Rubinstein (1979) formula. We remark that the result does not rely on any empirical distribution on the price process of the security (only the lattice described the relevant uncertainty), this level of generality will not be available in continuous time.

Let us take the same problem as before but in a continuous time setting: valuing a European call on an underlying security that does not pay dividends until the expiration date and with a constant rate riskless asset. But here, time is continuous: $t \in [0, T]$. The security price process $S : [0, T] \to R$ is assumed to be a geometric Brownian motion (a particular case of an Ito process):

$$\forall\ t \in [0, T], \quad \mathrm{d}S(t) = \mu S(t)\ \mathrm{d}t + \sigma S(t)\ \mathrm{d}B(t)$$

where $B(t)$ is the standard Wiener process (i.e. Brownian motion).

The option's payoffs at the exercise date are: $C(T) = \max[S(T) - K, 0]$, where K is the strike price. Hence, the option's payoffs are a function of the underlying security's payoffs at the expiration date; from this it can be deduced that, at each intermediary date, this is still true:[13]

$$\forall\ t \in [0, T], \quad C(t) = F\,[S(t), t]$$

If, furthermore, we assume F satisfies some regularity properties (differentiability, etc.), Ito's lemma can be applied. It yields the differential of the option's price as a function of its underlying security's price.

Ito's lemma *Let an Ito process be defined by the stochastic differential equation:*

$$\forall\ t \in [0, T], \quad \mathrm{d}X(t) = \mu[X(t), t]\ \mathrm{d}t + \sigma[X(t), t]\ \mathrm{d}B(t)$$

and let $Y(t) = F\,[X(t), t]$, where F is a deterministic, twice-continuously differentiable, be a process. Then Y is also an Ito process defined by the differential equation:

$$\mathrm{d}Y(t) = \left\{ \frac{\partial F}{\partial t} + \mu[X(t), t] \frac{\partial F}{\partial X(t)} + \frac{1}{2}\sigma^2[X(t), t] \frac{\partial^2 F}{\partial X(t)^2} \right\} \mathrm{d}t + \sigma[X(t), t] \frac{\partial F}{\partial X(t)}\ \mathrm{d}B(t)$$

Applying Ito's lemma to the call price process yields:

$$\forall\ t \in [0, T], \quad \mathrm{d}C(t) = m[S(t), t]\ C(t)\ \mathrm{d}t + s[S(t), t]\ C(t)\ \mathrm{d}B(t)$$

where

$$m[S(t), t] = \frac{\dfrac{\partial F}{\partial t} + \mu S(t) \dfrac{\partial F}{\partial S} + \dfrac{1}{2}\sigma^2 S^2(t) \dfrac{\partial^2 F}{\partial S^2}}{C(t)}$$

$$s[S(t), t] = \frac{\sigma S(t) \dfrac{\partial F}{\partial S}}{C(t)}$$

The uncertainty about the call, formalised by the Brownian motion B, is the same as that about the underlying security's prices. This is the reason why, at each date, it is possible

[13] It will be proven by the result of this model.

to replicate the option by a portfolio formed of the security and the riskless asset, or, equivalently, to replicate the riskless asset with the security and the call.

Let θ be this portfolio with the security quantity being a and the call quantity being b, its value at any date $t \in [0, T]$ is:

$$\Theta(t) = aS(t) + bC(t)$$

During an infinitesimal time interval dt, assume the portfolio composition is constant, then its value variation depends only on the price variations of the assets in it:

$$\mathrm{d}\Theta(t) = a\,\mathrm{d}S(t) + b\,\mathrm{d}C(t)$$

Replacing d$S(t)$ and d$C(t)$ by their values:

$$\mathrm{d}\Theta(t) = \{a\ \mu\ S(t) + b\ m[S(t), t]C(t)\}\ \mathrm{d}t + \{a\ \sigma\ S(t) + b\ s[S(t), t]C(t)\}\ \mathrm{d}B(t)$$

The parameters a and b can be chosen to be such that the portfolio is riskless during the infinitesimal time interval dt. Then, it must be the case that the random term, i.e. the factor on dB is zero. We get a first necessary equation that a and b must solve:

$$a\ \sigma\ S(t) + b\ s[S(t), t]\ C(t) = 0$$

A second equation is obtained by the no arbitrage condition: if the portfolio is riskless, its rate of return must be r:

$$\frac{\mathrm{d}\Theta(t)}{\Theta(t)} = r\mathrm{d}t \Rightarrow a(\mu - r)S(t) + \{[m[S(t), t] - r\}C(t) = 0$$

a and b are then obtained as a solution of these two equations. The trivial solution $a = b = 0$ does not obtain if and only if the determinant is zero:

$$\frac{m[S(t), t] - r}{s[S(t), t]} = \frac{\mu - r}{\sigma} = \lambda$$

This yields the expression of the market price of risk (measured by σ, the instantaneous volatility): the rate λ that must be added to the riskless rate to compensate the risk. This market price of risk is the same for all assets that depend on the same uncertainty representation: B.

The option's price is obtained by solving the differential equation where functions m and s have been replaced by their expressions above:

$$\frac{\partial F}{\partial t} + rS(t)\frac{\partial F}{\partial S} + \frac{1}{2}\sigma^2 S^2(t)\frac{\partial^2 F}{\partial S^2} - rC(t) = 0$$

Under the final condition: $C(T) = \max(S(T) - K,\ 0)$

It so happens that this differential equation is the same as the heat diffusion equation (after a change in the variables) that has a known analytical solution. This is the famous Black and Scholes (1973) solution:

$$C(t) = S(t)N(d_1) - Kexp[-r(T - t)]N(d_2)$$

where N is the cumulative normal (normalised Gaussian) distribution and:

$$d_1 = \frac{\ln\left[\frac{S(t)}{K}\right] + \left(r + \frac{\sigma^2}{2}\right)(T - t)}{\sigma(T - t)^{\frac{1}{2}}}$$

$$d_2 = d_1 - \sigma(T - t)^{\frac{1}{2}}$$

The same result can be obtained by a direct application of the martingale property, instead of forming the replicating portfolio strategy and the no arbitrage condition.

Under the assumptions that the riskless rate is constant and there are no intermediary payoffs for the security (no dividends) nor for the call (European call), the option's price is the discounted expectation of its final payoffs with respect to the risk-adjusted distribution:

$$C[S(t), t] = exp[-r(T-t)]\, E_v\{C[S(T), T]\}$$

In order to figure out the risk-adjusted distribution, let's apply Ito's lemma to the process Y defined by $Y(t) = ln[S(t)]$. We get:

$$\mathrm{d}Y(t) = \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma \mathrm{d}B(t)$$

The martingale property implies that all assets have the same expected returns with respect to the risk-adjusted distribution. Then $\mu = r$, and the previous equation is:

$$\mathrm{d}Y(t) = \left(r - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma \mathrm{d}B(t)$$

Y is then a general Wiener process with parameters: $\left(r - \frac{\sigma^2}{2}\right)$ and σ.

At date t, $Y(t)$ is known, $Y(T)$ follows a normal distribution with mean $Y(t) + \left(r - \frac{\sigma^2}{2}\right)(T-t)$ and variance $\sigma^2(T-t)$. The initial variable $S(T)$ follows a lognormal distribution with parameters $\ln[S(t)] + \left(r - \frac{\sigma^2}{2}\right)(T-t)$ and $\sigma^2(T-t)$. Let us call f its density function, then we have:

$$C[S(t), t] = exp[-r(T-t)] \int_K^{+\infty} (x-K) f(x)\mathrm{d}x$$

Figuring out the integral yields the Black and Scholes (1973) formula.

The Black and Scholes (B&S) and the Cox, Ross and Rubinstein (CRR) methods to value European call options are similar; they differ by the representation of the prevailing uncertainty, but the formula they obtain are related. If we consider that the discrete time intervals are converging towards zero, and if the parameters u and d are adjusted adequately, then the CRR formula converges towards the B&S one. If ρ is the continuous riskless rate and r the discrete one, the adjustments are the following:

$$u = 1 + \mu h + \sigma\sqrt{h}, \quad d = 1 + \mu h - \sigma\sqrt{h}, \quad r = 1 + \rho h$$

where $h = \frac{T-t}{n}$, n being the number of time intervals. The two formulas are not absolutely equivalent, but they can be related by these parameter adjustments.

We have seen the simplest case where there are no intermediary payoffs, either for the call (European) or for the underlying security. In this case, the martingale property directly applies on future prices (payoffs) to yield the asset valuation at any date. When there are intermediary payoffs, the asset values cannot be figured out without some backward induction. As a result, it is not possible to obtain a general explicit valuation formula in these cases. We shall present such a problem through the following example.

Take an American put option, on an underlying security paying no dividends up to the exercise date.[14] The actual price of the security is €100, the exercise price is $K = €95.5$ and

[14] We have chosen a put option this time because it is never advantageous to exercise a call option before the expiration date, if no dividends are paid.

the riskless rate is $R_0 = 3.5\%$. The model is over two periods ($T = 2$) and we assume the security price goes up or down each period by 5%

We can then apply these data to the binomial model:

$$u = 1.05, \quad d = 0.95, \quad r = 1.035, \quad p = 0.85.$$

We apply the martingale property at date 1 and get the price of the put option in both up and down states:

$$P(u) = \frac{1}{r}\left[p \max(K - u^2 S,\ 0) + (1 - p)\max(K - udS, 0)\right] = 0$$

$$P(d) = \frac{1}{r}\left[p \max(K - udS,\ 0) + (1 - p)\max(K - d^2 S, 0)\right] \cong €1.05$$

An American option can be exercised at any time. At date 1 in states up and down, the intrinsic[15] put values are:

$$E(u) = \max(K - uS,\ 0) = 0$$

$$E(d) = \max(K - dS,\ 0) = €2.5$$

In the down state, the option's intrinsic value is more than its theoretical value, hence the option should be exercised. The converse is true in the up state. The option's market price must be adjusted, otherwise there would be an arbitrage opportunity: the option could be bought for €1.05 and exercised immediately to get €2.5. Hence, the market price should be €2.5.

In order to figure out the initial price of the put, the martingale formula must be applied to prices at date 1 instead of the values that have been calculated with the same formula at date 1:

$$P(0) = \frac{1}{r}\left[pP(u) + (1 - p)E(d)\right] \cong €0.36^{16}$$

To calculate the market price of an American option, it is generally necessary to solve a dynamic program in order to take "option values" into account. Option values are the gains made possible thanks to some decision depending on information arrivals. We shall face this problem again to value investment decisions where flexibilities and the possibilities to delay a decision must be taken into account in the project valuation.

11.2.2 Investment Projects

All the methods founded on NPV are static. Even though their applications are relevant in many contexts, they entail a definitive choice to invest or not, when, in many circumstances, there are some dynamic choices to make. Dynamic choices are related to some decisions being reversible and others not, or some decisions can be delayed in time, etc. In order to decide on an irreversible choice or to decide to delay the decision for a future occasion, it is crucial to learn some information before making the decision. Flexibility consists of a decision being able to adapt to new information arrivals, so as to adjust to new contexts. The

[15] The value the put would have if it was exercised at that date.

[16] The option cannot be exercised at the initial date.

opportunity to invest, or reinvest, can be considered as an option, which adds a time value to the project coming from potential gains due to the asymmetry in the option.[17] Making an irreversible decision kills the option and hence decreases the project value. The "real options" theory aims to take these opportunities costs into account, something the traditional NPV approach cannot achieve.

Many examples of real options can be found when valuing flexibility: options to delay the investment, options to default during a plan construction, options to extend the project, options to limit activities, options to open or close some facilities, options to give up, options to use an alternative output, etc.

What is the basic problem in real options theory?[18] When realised, an investment project provides an output (we assume one unit without loss of generality) with no production cost. The selling price of this output is uncertain, let us assume that it follows a geometric Brownian motion:[19]

$$\forall t \in [0, T], \quad \mathrm{d}P(t) = mP(t)\ \mathrm{d}t + \sigma P(t)\ \mathrm{d}B(t)$$

With no production cost and for a unit of output, P represents the firm's profit. Let us assume that the project's value is a deterministic function of the output's price: $V = V(P)$. There are two ways to figure out the project's value: the contingent asset approach and dynamic programming. We shall develop the first approach here in the line of this section. Furthermore, dynamic programming has two drawbacks: it refers to an arbitrary constant discount rate, and it is founded on an individual criterion.

In order to apply the no arbitrage valuation method, the price process P must be replicated by a portfolio strategy of marketed assets. This means, in practice, that it is possible to find and gather such assets in the economy in order to have a portfolio paying off the same amounts as P. This is easily feasible for an investment to produce some commodity, because they are traded on cash or future markets. This is equally feasible for industrial produce if their prices are correlated to the stock values or to indexes on the relevant activity sector.

According to the arbitrage valuation method, let us form at date t, a portfolio composed of one unit of the project and a short position on n units of outputs. Its value is:

$$\Theta(t) = V(t) - nP(t)$$

If the portfolio components remain constant during $\mathrm{d}t$, its value variation during this time interval will be:

$$\mathrm{d}\Theta(t) = \mathrm{d}V(t) - n\ \mathrm{d}P(t) + [P(t) - n\ \delta\ P(t)]\ \mathrm{d}t$$

The first two terms correspond to a capital gain, while the third one is the flow of profits and in the last one, δ is the marginal convenience rate for storage,[20] i.e. the net benefit flow for a supplementary stored unit.

$\mathrm{d}V(t)$ is given by Ito's lemma:

$$\mathrm{d}V(t) = \left[mP(t)V'(P) + \frac{1}{2}\sigma^2 P^2(t)V''(P) \right] \mathrm{d}t + \sigma P(t)V'(P)\mathrm{d}B(t)$$

[17] Losses are limited by the option value, while possible gains are not limited.
[18] See Dixit and Pindyck (1994), Trigeorgis (1996).
[19] This is not a very restrictive assumption because the model can be discretised if such a representation of time is more relevant to the problem at hand.
[20] For a non-perishable commodity.

Then:

$$d\Theta(t) = \{mP(t)[V'(P) - n] + \frac{1}{2}\sigma^2 P^2(t)V''(P) - n\delta P(t) + P(t)\}dt + \{\sigma P(t)[V'(P) - n]\}dB(t)$$

$n = V'(P)$ is chosen so that the portfolio is riskless.

Then, according to the no arbitrage assumption, its rate of return must be equal to the riskless rate:

$$\frac{d\Theta(t)}{\Theta(t)} = r dt \Rightarrow \frac{1}{2}\sigma^2 P^2(t)V''(P) + (r - \delta)P(t)V'(P) + P(t) - rV(P) = 0$$

This stochastic differential equation has a known explicit solution, the process V can be expressed as a function of P and then the option value to invest in the project is obtained. An alternative way consists of finding the option value by using the solution to the previous differential equation as a border condition when exercising the option is optimal. Given the option is optimally exercised, its value is strictly positive. This shows that it is necessary to take this value into account in the project's valuation, the NPV cannot do it.

In applications, the problem is to determine the marginal convenience rate for storage δ, because it determines the time evolution of the value of detaining the output. Let μ be the market rate of return for the project, in a portfolio paying no dividends. This portfolio is perfectly correlated with V and hence with P. Under no arbitrage, the dividend of the project should be: $\delta = \mu - m$. The rate of return μ can be obtained by the CCAPM.[21] From the results of this model, only the idiosyncratic risk is paid for by the market: the excess mean return with respect to the riskless rate r relative to its risk (volatility) is equal to the market price of risk λ. The part of the non-diversifiable risk corresponds to the correlation coefficient between the output price and the market portfolio price $M: \rho(V, M) = \rho(P, M)$. The market price of risk is applied to obtain:

$$\frac{\mu - r}{\sigma} = \lambda \rho(P, M)$$

In the CCAPM the market price of risk is measured by the excess mean return of the market portfolio with respect to the riskless rate, relative to its risk:

$$\lambda = \frac{r_M - r}{\sigma_M}$$

This yields δ:

$$\delta = r + \sigma \rho(P, M)\frac{r_M - r}{\sigma_M} - m$$

In order to enrich the model, production flexibility can be introduced. With constant production costs C, the firm can stop temporarily the activity if the output price goes under the cost and restart it when it goes over it again. The profit is:

$$\pi(P) = \max(P - C, 0)$$

The project is valued by a portfolio formed of one unit of the installed project and a short position on $n = V'(P)$ output units. The stochastic differential equation obtained by no arbitrage is:

$$\frac{1}{2}\sigma^2 P^2(t)V''(P) + (r - \delta)P(t)V'(P) - rV(P) + \pi(P) = 0$$

[21] The dynamic version of CAPM in continuous time, Merton (1973a).

The option's value of the project is obtained as a solution.

The problem is much more complex if the quantity of output is a variable. Indeed, there is a new decision level for the producer. The profit can be written as:

$$\pi(P) = \max\{Pf(Y) - C(Y)\}$$

where:

Y is the input vector
$f(\cdot)$ the production function
$C(\cdot)$ the cost function.

With simple specifications (a unique input, a power production function and a competitive market for the input), the previous method can be applied and yield explicit solutions. The problem can be put in even more general terms if costs are stochastic. With two independent variables (the selling price and the production cost), it is seldom the case that explicit solutions can be found for the stochastic differential equations obtained from no arbitrage. But numerical methods are available to get approximations to solutions.

Extensions of the real options to the case of controversial risks will be presented in Chapter 12.

12

Static and Dynamic Risk Instruments

12.1 STATIC RISK MANAGEMENT INSTRUMENTS

In this section, we shall concentrate on two instruments that are adapted to a static vision of risk management: the beta (β) from the CAPM to take risk into account in the NPV, and the VaR, an alternative to variance as a measure of risk more adapted to manage risks of losses.

The β has been introduced in Chapter 10, it measures an asset's sensitivity to the market risk: the asset's and the market portfolio's covariance divided by the market risk (measured by the market portfolio's variance). Its use is theoretically founded on the CAPM, or by its extension (APT). Let us recall the CAPM formula:

$$E(r_i) = r_0 + [E(r_M) - r_0]\ \beta_i$$

In order to visualise this formula, it can be represented in the mean–beta $(E(r), \beta)$ plan: a straight line is obtained see Figure 12.1. It is the Security Market Line (SML) that was presented in Chapter 10 in the (E, σ) plan.

The β has been introduced in the firms' financial decisions analysis, after the success of the CAPM model made it a reference. Beforehand, the NPV of an investment project neglected the price of risk or introduced it as a somewhat arbitrary mark-up of the riskless discount rate. The sensitivity of a firm's risk to the market risk indicates in which proportion the market price of risk should be introduced in the discount rate. Hence, instead of r_0, its $E(r_i)$, given by the CAPM formula, will be used to figure out the NPV of firm i's project.

The price of risk obtained by the β is particularly relevant for firms that are close to market valued societies, or for portfolio management. For a firm that is poorly related to the market, or a new project that is not in the line of the usual activities of the firm, there are means to compare its risk to the market risk. The same means are called for, in order to measure the sensitivity of a public project's risk to the market risk. We shall see these methods in Section 12.2, they require us to consider the risk in a dynamic approach. Take a public investment to prevent a hazard: a dam to prevent floods, for example. The risk at stake is not related to the security market. However, the risks that the dam prevents, such as a production site, are related to marketed firms. Then it is by analysing these risks over past data that a hedging portfolio of marketed assets can be constructed and the portfolio's β will be used for valuing the project. This can only be done in a dynamic approach and using past data series. Even though the risk premium obtained this way is not as reliable as it would be for a firm similar to a marketed one, it is relevant to get a better approximation of the risk's value, or at least the value of the part of it that could be diversified away by market instruments.

Another application of the β is in the performance analysis of risk management. All performance measures in portfolio management are founded on CAPM, and it can be shown that they are equivalent. They help to verify what is the level of risk that is faced when some expected return is aimed at. The same performance instruments can be applied in any other risk management problem, be it a private or a public investment. Let us take the example of three preventive measures against some hazard (flood): dam, flooded areas, derivative canal. The three of them aim at the same expected return (in terms of casualty reduction) and define the rate of return:

(expected prevented loss – investment cost)/ investment cost

In each case, the risk resulting from the hazard and the prevention measure is simulated on the basis of the results of similar investments. Thanks to these data, the sensitivity of these risks to some market risk can be estimated: $\beta_1, \beta_2, \beta_3$. The three investment risks are represented in the $(E(r), \beta)$ plane by three points on the SML curve. The investment risks' efficiency is seen directly on Figure 12.1: it is measured by the distance between the points and the SML line.[1] In our example, the second investment has a zero distance: it is efficient as compared to the market. This means that the same returns could be obtained through a portfolio of marketed assets that replicate the risk resulting from the prevention investment 2 (flooded areas). Otherwise stated, this preventive investment is equivalent, in monetary terms, to investing in a portfolio that would hedge the casualties.

The negative distance for investment 3 (canal) indicates that it would be more efficient to invest in a hedging portfolio.

The first investment (dam) shows a positive performance, this means that it does better than a perfect risk insurance (perfectly hedging portfolio), maybe because of some induced gains. In general, this is an argument in favour of this decision against the others.

The different applications of β in risk management are particularly adapted to risks that are close to market risk, such as some risk ventures or country risks if the country in question is in connection with other countries' markets. Such applications are valid in a static approach,

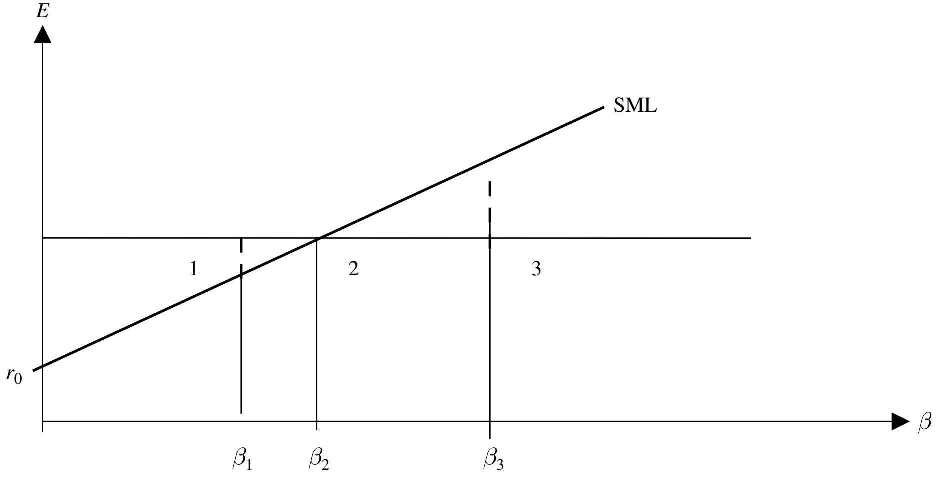


Figure 12.1 Security Market Line and investment efficiency

[1] In portfolio management, this distance is the Jensen performance index.

i.e. for a fixed time horizon from a given initial date, under the condition that no information arrival during this period may influence investment decisions.

The same remark is valid for the next tool we present here, VaR, the measure of losses which is referred to as a safety requirement for risk management. For a given time period, or a given time horizon, the VaR measures the maximal potential loss that can occur for a given probability level: 1%, 2%, 5%, etc. The probability level is chosen according to the reliability that is relevant to the managing problem, it seldom goes over 5%. The VaR is seen directly on the cumulative distribution function of the portfolio's risk: on the left tail of the function, the probability levels p yield the VaR_p: the p-percentile (see Figure 12.2).

Several software packages were developed to achieve reliable VaR calculations,[2] since this risk measure was officially recommended by the Basle Committee (1996).

In most risk-managing problems, the VaR is a relevant measure because it takes correlations into account. This is obvious in portfolio management, where correlations between assets are used to diversify risks or to increase returns. We shall come back to the use of VaR in such situations in Section 12.2.

But this is true also in public risk management such as environment problems, where many random variables interfere and correlations are also important. However, in such situations, it is the global risk that is measured directly because the analyses of the different causes and impacts are too complex to handle. VaR yields such a global measure of the potential losses, it takes into account possible loss increases or, conversely, the hedging of possible effects of precautionary and preventive measures, for example.

In a static management problem, tools such as the β and the VaR are used to make decisions after an objective level (expected profit, for example) has been fixed. At that level, they discriminate between alternative projects according to the measure of the risks they provide. The comparative advantage of these instruments over other ones derived from individual decision theory is that they are objective, instead of subjective, and collective in the sense of referring to market values and observable data. For example, a risk-aversion coefficient is purely individual and derived from a utility function that is hardly observable.

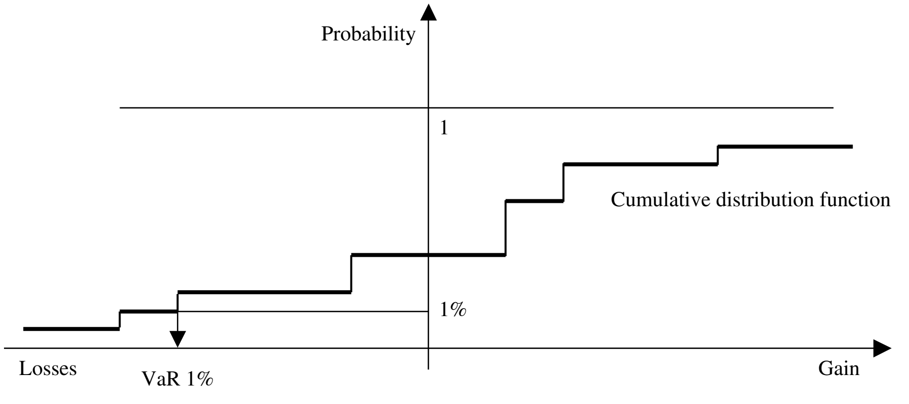


Figure 12.2 Cumulative distribution function and VaR

[2] Notably *Riskmetrics*®.

Furthermore, it is bound to be controversial if used in deciding about a public project to prevent some risk. Similarly, an individual discount factor does integrate a risk premium, but this premium is related to the individual preference for present consumption and not to the price to pay for borrowing money nor to insure a risk with marketed instruments.

However, all these instruments face the same obvious limitations that users must be well aware of: they are used in a static vision of the problem. In fact, no decision problem can be completely limited to a given time period. The demand for insurance is relevantly analysed on a given time period, a year in general. However, the data that are used to estimate probabilities, for instance, are related to past time periods and integrate dynamics. Insurance companies do manage their risk portfolio by dynamic instruments. Similarly, the gambling industry manages the risks in a static framework, but it addresses clients that see the risk they buy in a dynamic way. In both cases, the static approach is relevant only if it is well understood that the initial date corresponds to a given information state. Correspondingly, the finite horizon is relevant because, at that date, new information may arrive. For instance, a market risk premium is defined at a given time and reflects all the relevant information that is aggregated by market prices. It is bound to change when changes in information modify demands and supplies. The static approach is relevant only if the risk premium is not likely to change before the time horizon. Otherwise stated, the static instruments are conditional on available information.

Conditioning introduces time and a dynamic version of the decision problem. As long as probabilities are concerned, conditioning is achieved by the use of Bayes' rule. Let us just mention here that, if non-additive risk measures are involved (e.g. capacities), conditioning is not straightforward. We will mention the problem in the General Conclusion. The reason why conditioning is simple with probabilities and not with capacities, is that separability of the measure with respect to uncertain states is satisfied with the former and not with the latter.

This remark introduces another difficulty that is overlooked in most static approaches linked to NPV: separability with respect to time is also an assumption. Discount factors, individuals as well as market ones, are weights that are assigned to payoffs (or payoff utilities) in an additive form. This additive form is obtained because discount factors are assumed to be independent of the payoff level. This is not true in practice: different amounts are not discounted at the same rate. As a consequence, the separability assumption only holds for regular cash payoffs with similar levels and close enough expiration dates. But discount factors, as well as all other tools in a static framework, are conditional on available information. As such, they depend on the initial date and the information that is available then. They should be revised according to information arrivals, and this information may concern amounts as well as other characteristics of the risks that have to be managed during the fixed time period.

12.2 MANAGING FLEXIBILITIES

We present dynamic management methods for two applications. In contrast with traditional approaches, modern credit risk management integrates market instruments that we present here. The second application addresses the problem of controversial risks. Present in most public and some private investment projects, they are characterised by no agreed upon representations and measures of the prevailing uncertainty. We propose a new approach founded on the teachings of real investments theory adapted to these risks valuation.

12.2.1 Managing Credit Risk

Rating consists of rank credit issuers thanks to a number (mark) representing a synthesis of their characteristics. This is more a qualitative than a quantitative approach. It can be done by internal auditing but more generally it is achieved by rating agencies. It applies to claims that are traded on a market and is a managing tool for banks. Thanks to the rating, firms are ranked according to their risk profile (risk classes).

The rating of a credit issuer, or of a particular issue, is related to the default probability. In 2005, ratings of the Standard & Poor's agency generated the default rates in Table 12.1. The results show that the rate yields reliable information only if the time horizon is large enough. Over 5 years, the default frequencies decrease significantly with the firms' rates.

Banks can refer to these data in order to figure out the expected losses on their credits. For each risk class, the expected loss is obtained by the credit amounts multiplied by two parameters: the default posterior probability and the expected non-collection coefficient. The total risk is obtained by the sum of the risks for each class. The unexpected loss is figured out by the same method, where the maximal posterior default probability for each class is referred to. This is a static use of the rating instrument: the time horizon is arbitrarily fixed for the calculation and the possible evolutions of the firms are not taken into account. Therefore, their dynamic hedging strategies are neglected. It can be applied in the case where claims must be held until the expiration date and losses are counted *ex post*.

The traditional accounting practice values the debts by discounting their book values. The new international accounting norms change this by referring instead to market values. In some countries, banking lobbies resist the new practice. Their argument is that the high market volatility will mechanically affect the credit firms' balance sheet. During economic crisis, the firms' default increases and banks are penalised twice:

- Through the classic way the claims on defaulting firms lose some or all of their value.
- Lower firms' rates entail a decrease in their market values and the credit firms' balance sheet falls down.

If prudential norms are satisfied, activities will be slowed down and such credit limitations are bound to reinforce the business's doldrums.

Table 12.1 N.R.-Removed Cumulative Average Default Rates, 1981 to 2005 (%)

	Time horizon														
Rating	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12	Y13	Y14	Y15
AAA	0.00	0.00	0.03	0.07	0.11	0.20	0.30	0.47	0.53	0.60	0.60	0.60	0.60	0.60	0.60
AA	0.01	0.03	0.08	0.16	0.25	0.38	0.54	0.68	0.79	0.92	1.03	1.16	1.28	1.40	1.50
A	0.04	0.13	0.25	0.42	0.64	0.86	1.11	1.34	1.62	1.90	2.11	2.29	2.48	2.63	2.91
BBB	0.27	0.81	1.40	2.25	3.11	3.97	4.67	5.35	5.93	6.63	7.34	7.96	8.68	9.52	10.29
BB	1.20	3.71	6.86	9.94	12.74	15.57	18.02	20.27	22.39	24.04	25.66	27.00	28.09	28.85	29.93
B	5.91	13.60	20.55	26.23	30.48	34.01	37.23	40.15	42.36	44.75	46.82	48.49	50.25	52.15	53.72
CCC/C	30.41	40.02	46.13	50.55	56.04	58.53	59.63	60.43	64.38	67.72	67.72	67.72	67.72	69.19	69.19
Investment Grade	0.11	0.32	0.56	0.89	1.24	1.59	1.91	2.22	2.50	2.81	3.08	3.32	3.57	3.84	4.14
Speculative Grade	5.05	10.32	15.33	19.61	23.12	26.20	28.90	31.34	33.50	35.46	37.17	38.57	39.86	41.13	42.35
All Rated	1.67	3.36	4.91	6.24	7.34	8.28	9.07	9.76	10.36	10.92	11.40	11.80	12.19	12.58	12.98

Source: Standard & Poor's Global Fixed Income Research; Standard Poor's CreditPro® 7.02, reproduced with permission of Standard & Poor's, a division of The McGraw-Hill Companies, Inc.

However, these arguments can be reversed. Indeed, a decrease in asset values is part of the regulation mechanism and should not be ignored. If banks go on supplying credits in order to smooth difficulties for some time, they cumulate bad debts up to the point where they could burst out. The example of Japan shows, whatever the determination of political decision-makers, that a painful purge must be done at some point. This experience enforces the market values as a reference to credit firms' regulation.

According to this point of view, all claims can be considered as if they were marketable and could be sold before their expiration date. Then, the default risk has two components:

- The loss of the capital (or part of it) and/or of the outstanding interests.
- The spread between the market premium for this risk and a government bond's rate with the same maturity, increases. This entails a decrease of the claim value and the market value of the credit is lessened.

In order to manage the risk in a dynamic way, transition probabilities between two ratings can be referred to. For Standard & Poor's, changes in rating from year to year[3] are given (in %) by Table 12.2, where date t ratings are rows and date $t+1$ are columns.

Credit risk cannot be managed debt by debt, it must be done globally. Default probabilities are aggregated in order to obtain a probability on credit losses. To achieve this, credit institutions can use either a static model: default mode, or a dynamic model: mark to market.

Default mode integrates a credit loss only if the borrower defaults at a fixed time horizon. The potential loss is contingent on three variables:

- The bank's exposure to credit.
- A 0/1 indicator on default.
- The loss rate in the case of default.[4]

The default probability is given by the last column of the table, it is a function of the rating (internal or external) of the claim. Hence, data are aggregated according to ratings.

If conditioning is not taken into account, the default probability is estimated over several credit cycles, but then the business cycle effects are not integrated. In the so-called "conditional" approaches, the transition probabilities matrix is a function of the state of the

Table 12.2 Average One-Year N.R.-Removed Transition Rates, 1981 to 2005 (%)

From/To	AAA	AA	A	BBB	BB	B	CCC/C	D
AAA	91.42	7.92	0.51	0.09	0.06	0.00	0.00	0.00
AA	0.61	90.68	7.91	0.61	0.05	0.11	0.02	0.01
A	0.05	1.99	91.43	5.86	0.43	0.16	0.03	0.04
BBB	0.02	0.17	4.08	89.94	4.55	0.79	0.18	0.27
BB	0.04	0.05	0.27	5.79	83.61	8.06	0.99	1.20
B	0.00	0.06	0.22	0.35	6.21	82.49	4.76	5.91
CCC/C	0.00	0.00	0.32	0.48	1.45	12.63	54.71	30.41

Source: Standard & Poor's Global Fixed Income Research; Standard & Poor's CreditPro® 7.02, reproduced with permission of Standard & Poor's, a division of The McGraw-Hill Companies, Inc.

[3] Until year 2005.
[4] In case the debt can be partly recovered after default.

economy. The likelihood of a rating increment (decrement) increases (decreases) in the case of a high conjuncture (low conjuncture). However, the conjuncture's inversions are hard to anticipate. There are correlations between the three variables on which risk is contingent: they are not taken into account and this is a limitation to these models. Another limitation is due to the lack of historical data on the credit performances over several cycles. Strong assumptions are made in order to make calculations. This difficulty arises because many instruments are not directly market valued and because defaults are rare events.

Mark to market is founded on a dynamic representation of risks in order to represent the fluctuations of the financial markets. The claims portfolio is valued at the issuing date and at maturity, and the difference between the two figures measures the loss: the rating deterioration is interpreted as a loss. All the columns of the transition probability matrix are called for. There are two ways to figure out the values:

- The value of a debt without default can be calculated by the present value of the claim's future cash flows: cash flows are discounted. The spread that is used to discount is given by the market for bonds with the same rating and the same expiration date. Future cash payoffs are not known, they depend on two uncertainty sources: the future ratings and the term structure of credit spreads. According to this method, the loss rate for non-default claims is not taken into account. This assumes that senior[5] and subordinate claims should have the same spread, which is not true.
- The claim valuation by the risk-adjusted distribution method. A firm defaults if its assets value goes under the level necessary to support its outstanding debt. In this method, contingent payoffs, instead of the claim's ones, are discounted according to Figure 12.3 (where l is the loss rate).

A claim is then considered as a set of contracts contingent on the borrower's assets. The debt value is the sum of the present values of these derivative contracts. The discount rate that is used is given by the term structure of interest rates and by the risk-adjusted measure (the premium of the default risk). The risk adjustment depends on the volatility of the borrower's assets.

Risk management must avoid the debtors' defaults involving the bank's solvency. To avoid this, the bank must have capital stocks sufficient to support these defaults. The possibility is that the bank defaults cannot be completely eliminated because no capital stock would be

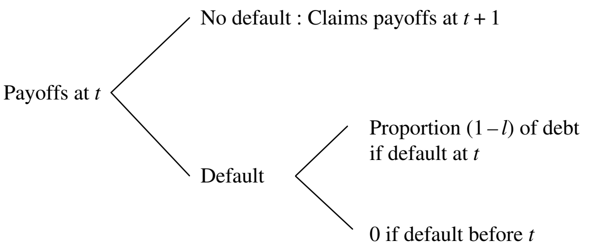


Figure 12.3 Payoffs tree

[5] They are reimbursed in priority in the case of default.

sufficient to support the simultaneous defaults of all of its debtors. Then, a default probability level must be fixed so as to calculate the capital necessary not to outstand it. This level is chosen in accordance with the rating of the bank. In the case it is AA, then the default probability is the historical default rate for private bonds rated AA with the same maturity.

The VaR yields the maximum potential loss for a given probability and expiration date. The allowance of capital necessary to hedge the maximum potential loss for the chosen probability level is figured out thanks to the VaR.

Capital is allocated as if each branch were a subsidiary. The amounts are fixed in accordance with the global allowance, and they depend on the branches. Capital profitability is an important instrument to determine the bank's strategy. Several problems are to be solved in order to figure out profitability:

- Should it be based on the capital allowance or on those that are spent?
- How are risks integrated in the different branches? A way out is to adjust the risk thanks to the Risk Adjusted Return On Capital (RAROC).

The main regulations define the minimal norms to hedge risks: credit risk, market risk, liquidity risk. As far as credit risk is concerned, the Basle Committee defined the Cooke ratio and the European Union another one: the Solvency European Ratio (SER). The ratio of capital over risk must be equal to 8%. The risk is defined by a linear combination of the balance sheet's claims and the other outstanding claims.

Securitisation[6] brought credit and market activities closer. Since the 1980s, this financial innovation has changed the intermediation business. Indeed, it takes apart credit acceptance and administration from the financing system. Securitisation works the same way for credit institutions as for insurance (see Part II). A firm sells some of its financial assets to a Special Purpose Vehicle (SPV) that finances this operation by issuing new assets: the Asset-Backed Securities (ABS). The default risk goes from the first assets to the second ones because the SPV are riskless.

Even though the method could be applied to any fixed returns assets (or even quasi-fixed returns assets), it only worked out for large claim groups for which the credit risk can be examined. The major success in securitisation has been observed for mortgages: Mortgage-Backed Securities (MBS).

Securitisation is organised when the intermediary is more efficient in the production and the management of credits than in their financing.[7] Another favourable situation is when the risk allocation is integrated in a general strategy, or when it is imposed by prudential regulations.

12.2.2 Construction of a Hedging Portfolio

Controversial risk valuation requires a dynamic approach. This is clear from the precautionary principle statements, for instance, which put forward the relevance of flexibilities and the possibilities to postpone decisions until new information is available.

We have seen in Chapter 8 that arbitrage valuation was relevant for controversial risks as long as we were able to construct virtual (marketable) assets underlying these risks. This was done within the static theoretical framework, the challenge here is to extend the principle to

[6] See Thomas (1999).

[7] Automobile credit institutions or credit cards issuers, for example.

the dynamic framework of the models presented in this part. More precisely, we come back to the methods presented in Chapter 8 to look for the existence of a deterministic function between the risk to be valued and a portfolio of marketed assets. Such a deterministic function defines the risk to be valued as a derivative asset of a marketable asset, and then the theories of Chapter 11 apply. It can be found by a functional form, or through comonotonicity, between the risk and the portfolio.

(1) Looking for a functional form. The risk to be valued, C, will occur at time T. A portfolio, with initial formation cost S, is related by a functional form to C:

$$C(T) = f[S(T)]$$

It is common practice to refer to the CAPM in a first approach to a valuation problem. In this case, the market portfolio is used as underlying security and C's sensitivity is estimated. In the more general APT model, assets are described by a list of explicative independent random variables (econometric model). This corresponds in our setting to f being linear. However, in the APT, explicative variables are not marketed assets or market indexes only, they can be any relevant economic factors. Risk premiums are calculated as excess returns with respect to the riskless rate of a marketed asset that is "assimilated" to the risk in the explicative variable. In practice, this requires constructing a portfolio of marketed assets that replicates the random variable, i.e. doing what we are explaining how to do!

A general function f can be obtained by the method of the functional coefficient (Chapter 8). The coefficients in the portfolio yielding $S(T)$ are such that $f[S(T)]$ minimises the difference between 1 (perfect correlation) and its FCC with C. Another method is to construct a portfolio such that its coefficients minimise the tracking error between C and the portfolio.

Let us assume that on the time interval $[0, T]$, the portfolio's price process is a geometric Brownian motion:

$$\forall t \in [0, T], \quad \mathrm{d}S(t) = \mu S(t)\,\mathrm{d}t + \sigma S(t)\,\mathrm{d}B(t)$$

Following the classical option valuation approach, we assume that there is a function F such that at each intermediary date:

$$\forall t \in [0, T], \quad C(t) = F(S(t), t)$$

Assuming the same regularity assumptions on F, we can apply Ito's lemma and the Black and Scholes method to obtain a stochastic differential equation that can be solved explicitly or by some numerical method.

An alternative way to solve the problem is that the price of a marketable asset is its discounted payoffs' expected value with respect to the risk-adjusted probability distribution (FFF). As for an option, the logarithmic change of variables and Ito's lemma yield that the $S(T)$ distribution is lognormal in t, and $C(t)$ is obtained.

(2) Looking for comonotonicity. Maximising Kendall's index (Chapter 8) determines a portfolio with payoffs comonotonic to those of the risk to be valued. If C and S are strictly comonotonic, i.e. if, for any two s and s': $[C(s) - C(s')]\ [S(s) - S(s')] > 0$, then there exists an increasing function F such that $C = F(S)$. This property opens the way to apply the replication principle for contingent assets in a binomial model, even though the function F between final payoffs of the asset to be valued and the portfolio's is not

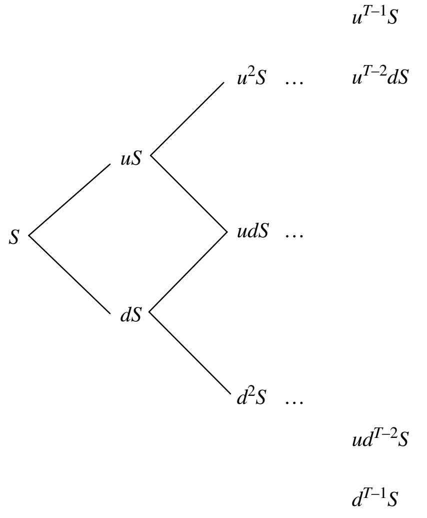


Figure 12.4 Binomial tree

known. Comonotonicity relates the final payoffs of the two assets because they are ranked in the same order. The final payoffs of the contingent asset can be put on the binomial tree representing the uncertainty about the underlying asset's payoffs. Notice that, in this approach, there are no intermediary payoffs, so that payoffs and prices are identified at each date-state of information (i.e. node of the tree).

Without loss of generality, let us rank the T final values of the risk in increasing order: $C_1 < \cdots < C_T$. The portfolio that maximises Kendall's index is assumed to be strictly comonotonic to the risk to be valued in the sequel. From the portfolio's observed values, we can construct the relevant binomial tree with final outcome T (i.e. over $T-1$ periods), where $u > d$ without loss of generality as in Figure 12.4.

Estimation of u and d parameters can be done according to two methods:

- Directly by estimation on the sample of a discrete distribution:

$$(u^{T-1}S, u^{T-2}\mathrm{d}S, \ldots, u\mathrm{d}^{T-2}S, \mathrm{d}^{T-1}S)$$

- By estimation of the volatility and then by discretisation of the continuous process:

$$\mu = \frac{\sum_{t=1}^{T-1} \ln\left[\frac{S(t+1)}{S(t)}\right]}{T-1}, \sigma^2 = \frac{\sum_{t=1}^{T-1} \left\{\ln\left[\frac{S(t+1)}{S(t)}\right] - \mu\right\}^2}{T-2}$$

Once these parameters are determined, the risk-adjusted probability is given by: $P = (p, 1-p),\ p = \frac{(1+R)-d}{u-d}$, where R is the riskless interest rate.

Because C and S are comonotonic, C's payoffs can be ranked in the same order as those of S, then they can be placed on the final nodes of the binomial tree designed for S and assigned the corresponding risk-adjusted probabilities (Table 12.3).

Table 12.3 Risk-adjusted probabilities

$u^{T-1}S$	$\to$	C_T	$\to$	p^{T-1}
$u^{T-2}\mathrm{d}S$	$\to$	C_{T-1}	$\to$	$p^{T-2}(1-p)$
...		...		...
$u\mathrm{d}^{T-2}S$	$\to$	C_2	$\to$	$p(1-p)^{T-2}$
$\mathrm{d}^{T-1}S$	$\to$	C_1	$\to$	$(1-p)^{T-1}$

Then we can apply the FFF: the value of the risk is its discounted payoffs' expectation with respect to the risk-adjusted probability. The present value of C is then:

$$C(0) = \frac{1}{(1+R)^{T-1}} \sum_{i=0}^{i=T-1} \frac{(T-i)!}{i!(T-1-i)!} p^i (1-p)^{T-1-i} C_{i+1}$$

Concerning controversial risks, this method is appealing because it does not require to use empirical probabilities on the future payoffs: such an empirical distribution is typically a source of questions and/or of controversies among experts.

CONCLUDING COMMENTS ON PART III

It is a general tendency to use static models to analyse risks, at least in a first approach. In order to go over the limitations of such a representation of the future, the relevance of its domain of application must be clear.

Beta and VaR are instruments that require a reliable probability distribution, hence their use assumes that historical data are available and satisfy the usual assumptions for statistical inference. For a new risk: a capital venture, a country risk in the case of war, the insurance of a natural catastrophe, or a public project for a nuclear plant, such historical data are missing or not reliable. The theory can still be referred to if the objective distribution is replaced by a subjective one. However, there are other limitations about the use of such tools under that concept: they do not rely on collective data but on an individual forecast of what they could be in the future.

Let us recall why VaR may be preferred to variance (variability) for safety requirements:

- It always exists, while an asset variance may be infinite.
- It measures the probability of losses instead of total variability.
- Thanks to its parameter (the probability level or, equivalently, the loss level) it can be adapted to the risk situation of the individual decision-maker.
- It is present in all the general risk-aversion characterisations that we have evoked in Chapter 6.

As a risk measure, it must not be confused with a decision criterion, even though it may act as pushing up the decision because the constraint it defines is reached. A decision criterion must take gains or returns, and not losses only, into account (e.g. a double criterion such as mean–variance or mean–VaR).

The main problem in a static representation of the future is that it concentrates all outcomes in one criterion. This is achieved by assuming that uncertain states and future time periods can be summarised into a present value. We pointed out that the usual hierarchy between time and uncertainty in the integration of future values may be contradicted by a decision-maker's behaviour in the face of payoff variations in time. Indeed, discount factors may take into account the variability of cash payoffs over time. In the same way that random variables may hedge each other and amount to a lower risk when combined together, certain cash payoffs could smooth their variability over time: when one goes up, the other goes down so that, combined together, the two cash payoffs vary little while each of them had many ups and downs. This hedging over time effect cannot be taken into account if time separability is assumed. This is in contrast with the actual behaviours of risk managers who do care about clearing problems that too much variability in cash payoffs may raise. It can be shown that such a hedging over time effect implies that discounting is not additive, or otherwise stated, that separability is not satisfied.

Most of these problems are avoided, however, if discount rates are market ones, but static market models are founded on some strong assumptions:

- Means and variances (and hence covariances) of all assets are known to agents.
- No arbitrage opportunities, expressed here in terms of means and variances, refers to an equilibrium of the demands for assets.
- Complete market underlies the possibility to form a riskless portfolio (approximately).

The first assumption is shared in common with the dynamic models: it is not as reductive as it seems at first glance. We insisted in the previous parts of the book on the restrictive assumption that all agents know a probability distribution over a set of future states and that they use it in the criterion of their preferences. This is because there are very few situations where agents can share such a knowledge about the future that concerns their risks. When a market for assets is organised, however, the situation is different: past prices and payoffs are known and form a huge database from which estimations can be done. Furthermore, if rates of returns, instead of payoffs are used, the i.i.d. assumption, necessary for using past data as statistics, is approximately satisfied. This is why all models in this part rely on rates: estimations are reliable. They yield means and variances for the static models and the instantaneous trends and variability in the dynamic ones. Furthermore, the role of the probability distribution is limited, as it only provides a description of the uncertainty: the relevant economic measure is the risk-adjusted one. The main concern about the use of past data to represent the future is that they are past ones! Assuming they are consistent with the perception of the future can be justified in the case of statistical regularities, but we know from experience that krachs and speculative bubbles are not uncommon in financial markets.

The last two assumptions are similar to those in the perfect foresight equilibrium model. However here, they take an approximate form, which makes them much easier to satisfy in applications.

In the dynamic case, complete market becomes quite an acceptable assumption, when it was clearly unrealistic in the perfect foresight equilibrium and even more so in the general equilibrium model. Indeed, thanks to flexibilities in portfolio strategies, replication of a position does not require that marketed assets hedge every position. The appealing Black and Scholes, and Merton formulas show that an option can be replicated by only two assets. However, one must be conscious in applications that the result cannot be applied without caution: the spanning number may be much greater than two, depending on the underlying asset and the representation of uncertainty. For instance, in many applied work, uncertainty cannot be described by only one Brownian motion because it must take into account several sources of risk. Then the spanning number increases, and the calculus difficulties as well.

The main difficulties encountered in extending the FFF to real investments, to public projects or to uninsurable risks is related to the distance between such risks and the marketed ones. The "real option theory" has been applied with success to a number of real investments, however these were clearly linked to marketed assets: stocks of similar firms, futures on the product, etc. In most cases, the CAPM could be invoked as a first approximation because the beta of the firm was not zero. This is not the case with risks, including public concerns such as environmental, social and political ones. We have indicated, however, that even such risks could be replicated if one uses statistical methods to construct an adapted portfolio of marketed assets. As promising as these methods can be, many problems have to be solved, both at the theoretical level and at the practical one:

- There must be some available and relevant data for the risk to be valued.
- In most cases, the probability distributions have to be known (estimated).
- The data must correspond to data on financial markets.
- The marketed assets that will be included in the portfolio must be chosen *a priori*.

The last difficulty has not yet been solved: within a market place, different marketed asset bases yield replicating portfolios that give consistent valuations. But there are no reasons to choose one market place over the other in most cases, and even no reasons why not to pick up assets in several market places.

We shall come back in the General Conclusion to the paths that actual research opens to enlarge the applicability domain of valuing risks outside the range of classical risk economics.

General Conclusion

Risk analysis has been renewed by the consciousness that some risks require other management methods than the traditional ones. Several features in the perception of a new risk problem may show that no traditional instruments are adapted to that particular risk. Such features can be: dynamics, no available probability distribution, controversies about what is the relevant probability distribution, collective instead of individual decision problems, a time horizon that concerns future generations, etc.

There is always an individual or a collective decision problem behind risks:

- Taking a risk is a decision.
- The study of a risk necessitates to make precise who is going to bear the consequences, who is willing to insure, who can take prevention measures and so on: each of these "who's" are making a decision.
- Managing a risk is a sequence of decision problems.

Any choice requires some valuation, at least implicitly. For practical as well as theoretical reasons, most modern economies turn to markets to value commodities as well as risks. Then, the main problem that remains to be solved is to give a value to non-marketed risks. We have proposed several ways to address the difficulty, we can summarise them through three main "lessons".

1 How to Deal With Controversies

When the choice concerns several people, each of them may have a different appreciation of risks. The public decision-maker may want to integrate these individual perceptions and valuations, instead of imposing its own. A capacity, instead of a probability, is recommended by decision theory to take a set of different probability distributions into account. When experts differ among scientists about which theory, and hence which probability, is relevant for a phenomenon causing a risk, then a capacity is a proper way to take the whole knowledge into consideration, instead of eliciting one of the theories or choosing one at random.

Controversies may concern the time horizon at which a risk's consequences are expected. A dynamic model may take such controversies into account: flexible prevention measures can be taken, the option values can be integrated in the risk valuation, etc.

Foresights of the future may be different among agents: sometimes opposite, often complementary, then different corresponding scenarios can be presented in the contingent valuation method to value non-marketed consequences. Aggregating individual valuations by market games is a way to complete the market simulation mechanism.

2 Look for Market Valuation

Methods such as contingent valuation, hedonistic prices, etc. are available to give a monetary value to qualitative consequences. Then a risk with non-marketed consequences can be formalised as a financial asset.

Securitisation transforms non-marketed financial assets into marketed ones through the portfolio that is traded on the financial market. Obviously this direct method can only be considered when non-marketed risks are homogenous enough to form a fund. In all other cases, markets yield an aggregated price of risk that can be referred to in order to value non-marketed risks. Real option theory can be applied to value a project when there is a natural underlying security for the project. It can be extended to controversial risks by constructing a portfolio of marketed assets with a risk similar to the project's risk. Then, the project and the portfolio should be valued the same: the portfolio's formation cost would be the price of the risk if it were marketed.

Real option theory teaches us how to take a project's flexibilities into account so as to use financial market instruments to manage such risks.

3 Measuring Time

Precaution is inspired by several characteristics of the future: time horizons, the importance of future generations or more simply, information arrivals. Dynamic decision-making can be achieved thanks to dynamic programming methods applied to the replication of risks by portfolio strategies. Then, financial markets methods and financial engineering are a relevant way to take flexibilities into account, however they rely on a parametric representation of time. Such a parametric time is inherited from classical mechanics and is the universal social reference, it does not take the relative importance of time periods into account. The relative weights that must be assigned to time periods cannot be neglected in decision-making. Economics is among the first science to have put forward other relevant time measures: discount factors, based on marketed financial instruments.

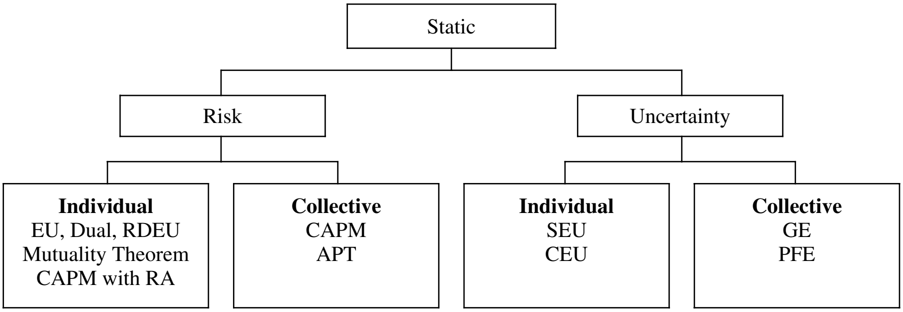
Economic theory also investigated discount factors based on individual behaviours and preferences for present consumption. When there are no relevant markets for some risks, then individual decision theory is called for to analyse intertemporal preferences.

We can summarise these lessons and the different models on which they are founded in Figures C.1 and C.2, taking the distinctions that have formed the three parts of this book from the opposite angle.

From this summary of the book, we can extract three main traps that should be avoided when valuing or managing risk.

TRAPS

1. The representative agent fiction. Chapter 3 has shown that such a representation of a public decision-maker was not founded on solid theoretical grounds, even though it can help to analyse some general problems and develop useful concepts. In equilibrium theory, the limitations are worse: the mere existence of markets is founded on agents behaving differently when they choose commodities and risky assets. Collective choice can be founded on market values when they exist, taking the limitations of the theory as a relative way



APT: Arbitrage Pricing Theory

CAPM: Capital Asset Pricing Model

CEU: Choquet Expected Utility

EU: Expected Utility

GE: General Equilibrium

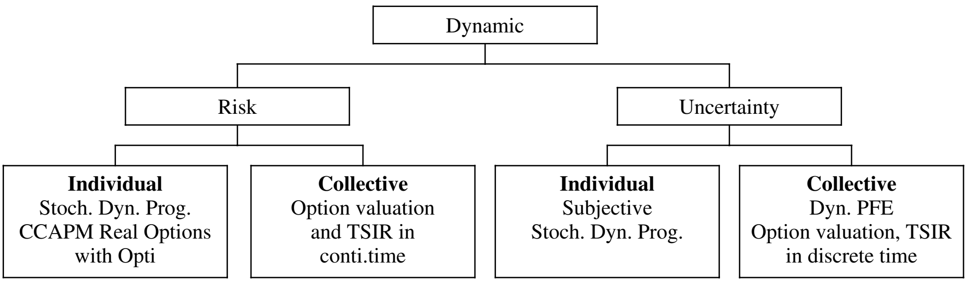
PFE: Perfect Foresight Equilibrium

RA: Representative Agent

RDEU: Rank Dependent Expected Utility

SEU: Subjective Expected Utility

Figure C.1 Static models



CCAPM: Consumption-Based CAPM

Continuous time

Discrete time

PFE: Perfect Foresight Equilibrium

Real option with individual utility optimisation

Stochastic dynamic programming

TSIR: Term Structure of Interest Rate

Figure C.2 Dynamic models

to get approximations. One of these is the satisfaction of the independence axiom and its corresponding axiom on time separability.

2. Past observations that do not yield frequencies. In order to apply a law of large numbers, observations must satisfy some assumptions on independence of random variables in order to apply statistical inference theory. Most new risk problems come from the violation of this assumption, excluding classical methods based on the insurance principles and on a common prior known and referred to by all agents. Market data are an exception, especially as far as rates of return are concerned. We must not forget that past data can only predict the future to some extent.

3. Forget option values. The main difference between static and dynamic representations of the future comes from taking into account flexibilities and option values. An option has a positive value and hence always increases a project's present value. Neglecting options opportunities amounts to undervaluing an investment and hence to being deprived of opportunities that would be worthy otherwise. Flexibilities are a main concern for implementing the precautionary principle, and more generally risks that are related to some scientific uncertainties. They imply taking future information arrivals into account and thus can only be included in a risk analysis within a dynamic framework. Financial market theory has taught us how to integrate flexibilities and options into risk valuation, but they can also be integrated in an individual optimisation problem when no marketed assets are underlying the risk to be valued.

These traps have appeared in the economic literature that put forward three paradoxical results when confronted with experiments or available data sets:

- The observed individual choices violate (in 60% of the cases) the independence axiom, even when an objective probability distribution is accepted (Allais' and Ellsberg's paradoxes). This questions the use of an expected utility maximisation in private as well as in public choice.
- The observed risky interest rates are inconsistent with "reasonable" risk aversion coefficients, still assuming objective probabilities and the existence of a representative agent satisfying expected utility and more assumptions (risk premium puzzle).
- The observed term structure of interest rates is "inverted" with respect to the theoretical one, under the same assumptions.

Such results force us to reconsider the models, however many textbooks on the economics of time and uncertainty present these theories (and some mention the "paradoxes", e.g. Gollier (2001)).

There are ways to avoid these traps and their related paradoxes. Current research focuses on some paths that should enlarge the domains of the theories' applications to the management of risks.

RESEARCH PATHS

1. The representation of time and of the future. An asset is contingent on two characteristics at least: time and uncertainty. An individual has the choice between three options to value it:

- Time is measured for a given uncertain state: state-dependent measure of time.
- Uncertainty is measured for a given time state: time-dependent measure of uncertainty.
- Time and uncertainty are measured together by the asset's present value.

As one can expect, none of these ways to measure time or uncertainty will yield the same result. An example was given in Chapter 10. The relevance of each of these hierarchies has to be put under scrutiny.

2. Conditioning non-additive measures. In individual decision theory, extension of probabilities to non-additive measures (e.g. capacities) questions the consistence of updating rules in a dynamic setting.[1] The economic measure of time: discount factors, can be justified

[1] Chateauneuf *et al*. (2001), Eichberger *et al*. (2005).

on individual or on market behaviours. There are ways to include aversion to variations in time and aversion to uncertainty.[2] More generally, the problem is to condition preferences both on time and on uncertain events (information) because each of them may influence the other. Then the difficulty arises from a general measure on the future not being decomposable into its components.

3. Replicating risks by priced risky assets. Economic theory assumes perfect markets to obtain its fundamental results, but concepts can be extended to take some relevant imperfections into account.[3] However, much is left to be done in this field.

We pushed forward a method to replicate assets by constructing a marketed asset portfolio on the basis of past available data. This method has two main drawbacks:

- First, as all methods based on statistical inference, it is limited by regularity properties to forecast the future.
- Second, it doesn't indicate on which grounds the marketed asset basis should be chosen.

Obviously, available and relevant data are important, but they cannot do all the work by themselves. Some intuition about correlations, or at least explanatory factors, would help to choose assets in one market place or in the other. They could also be selected among a set of economic activities related to the risk at stake. Alternatively, economic analysis could indicate how to pick up here and there economically relevant indexes that may influence or be influenced by the risk to be valued.

The APT is founded on an *a priori* description of the relevant uncertainty by economic indexes used as explanatory variables in an econometric model. In order to use it as a valuation device, it had to pursue similar replication problems by marketed asset portfolios for the economic indexes. Much applied work has been conducted and results obtained,[4] and can be adapted to the general setting of valuing non-marketed risks.

The capacity of economics and finance to design new models for risks has been put forward in this book. The limitations in the valuation and the management of risks we have encountered are more related to the applied domains.

[2] Kast and Lapied (2005).
[3] Wang (2000), Jouini and Kallal (1995). See also De Waegenaere *et al*. (2003) and Chateauneuf *et al.* (1996).
[4] Ross (1978), Roll and Ross (1980).

References

Allais, M. (1953) "Le comportement rationnel de l'homme devant le risque: critique des postulats et axiomes de l'école américaine", *Econometrica*, 21, pp. 503–546.

Anscombe, F. and Aumann, R. (1963) "A definition of subjective probability", *Annals of Mathematical Statistics*, 34, pp. 199–205.

Aris, R. (1964) *Discrete Dynamic Programming*. Blaisdell, New York.

Arrow, K. (1950) "A difficulty in the concept of social welfare", *Journal of Political Economy*, 58, pp. 328–346.

Arrow, K. (1953) "Le rôle des valeurs boursières dans l'allocation optimale des risques", *Econométrie*, 40, pp. 41–47 Cahiers du CNRS, reprint in "The role of securities in the optimal allocation of risk-bearing", *Review of Economic Studies*, 31, 1964, pp. 91–96.

Arrow, K. (1971) *Essays in the Theory of Risk Bearing*. Markham, Chicago.

Arrow, K. and Lind, R. (1970) "Uncertainty and the evaluation of public investment decisions", *American Economic Review*, 60, pp. 364–378.

Baumol, W. (1961) *Economic Theory and Operation Analysis*. Prentice Hall, Englewood Cliffs, NJ.

Bellman, R. (1957) *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bellman, R. and Dreyfus, S. (1963) *Applied Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bergson, A. (1938) "A reformulation of certain aspects of welfare economics", *The Quarterly Journal of Economics*, 68, pp. 233–252.

Bernoulli, D. (1738) "Specimen theoriae novae de mensura sortis", *Comentarii academiae scientiarum imperialis petropolitanae*.

Bernstein, P. (1998) *Against the Gods: The Remarkable Story of Risk*. John Wiley & Sons, Chichester.

Bickel, P. and Lehmann, E. (1979) "Descriptive statistics for nonparametric model IV. Spread", in J. Juresckova (ed.), *Contributions to Statistics*. Reidel Publishing, Dordrecht.

Black, F. and Scholes, M. (1973) "The pricing of options and corporate liabilities", *Journal of Political Economy*, 81, pp. 637–654.

Borch, K. (1962) "Equilibrium in insurance markets", *Econometrica*, 30, pp. 424–444.

Borch, K. (1990) *Economics of Insurance*. North-Holland, London.

Briys, E. and de Varenne, F. (2000) *The Fisherman and the Rhinoceros*. John Wiley & Sons, Chichester.

Caillois, R. (1967) *Les jeux et les hommes*. Gallimard, Paris.

Chambers, R. and Quiggin, J. (2000) *Uncertainty, Production and Agency: The State Contingent Approach*. Cambridge University Press, Cambridge.

Chanel, O., Faugere, E., Geniaux, G., Kast, R., Luchini, S. and Scapecchi, P. (2004) "Valorisation économique des effets de la pollution atmosphérique: Résultats d'une enquête contextuelle", *Revue Économique*, 55, 1, pp. 65–84.

Chateauneuf, A. (1991) "On the use of capacities in modeling uncertainty aversion and risk aversion", *Journal of Mathematical Economics*, 20, pp. 343–369.

Chateauneuf, A. and Cohen, M. (1994) "Risk seeking with diminishing marginal utility in a non expected utility model", *Journal of Risk and Uncertainty*, 9, pp. 77–91.

Chateauneuf, A., Cohen, M. and Meilijson, I. (1997) "New tools to better model behavior under risk and uncertainty", *Finance*, 18, pp. 25–46.

Chateauneuf, A., Kast, R. and Lapied, A. (1994) “Market preferences revealed by prices: non-linear pricing in slack markets”, in M. Machina and B. Munier (eds), *Models and Experiments in Risk and Rationality*. Kluwer Academic, Amsterdam.

Chateauneuf, A., Kast, R. and Lapied, A. (1996) “Choquet pricing in financial markets with frictions”, *Mathematical Finance*, 6, pp. 323–330.

Chateauneuf, A., Kast, R. and Lapied, A. (2001) “Conditioning capacities and Choquet integrals: the role of comonotony”, *Theory and Decision*, 51, pp. 367–386.

Chateauneuf, A. and Rébillé, Y. (2004) “Some characterizations of non-additive multi-period models”, *Mathematical Social Sciences*, 48, pp. 235–250.

Choquet, G. (1953) “Théorie des capacités”, *Annales de l’institut Fourier*, 5, pp. 131–295.

Cox, J., Ross, S. and Rubinstein, M. (1979) “Option pricing: a simplified approach”, *Journal of Financial Economics*, 7, pp. 229–263.

Crawford, H. (1991) *Sumer and the Sumerians*. Cambridge University Press, Cambridge.

Dana, R.-A. and Jeanblanc-Picqué, M. (2003) *Financial Markets in Continuous Time*, Springer-Verlag, New York.

Davis, R. (1963) “Recreation planning as an economic problem”, *Natural Resources Journal*, pp. 239–249.

Debreu, G. (1954) “Representation of a preference ordering by a numerical funcion”, reprint in R. Thrall, C. Comb and R. Davis (eds), *Decision Process*. John Wiley & Sons, Chichester, 1970.

Debreu, G. (1959) *Theory of Value*. Yale University Press, New Haven, CT.

De Finetti, B. (1930) “Fondamenti logici del raggionamento probabilistico”, *Bolletino dell’unione matematica italiana*, 9.

De Waegenaere, A., Kast, R. and Lapied, A. (2003) “Choquet pricing and equilibrium”, *Insurance: Mathematics and Economics*, 32/3, pp. 359–370.

De Waegenaere, A. and Wakker, P. (2001) “Non comonotonic Choquet integrals”, *Journal of Mathematical Economics*, 36, pp. 45–60.

Diecidue, E. and Wakker, P. (2002) “Dutch books: Avoiding strategic and dynamic complications and a comonotomic extension”, *Mathematical Social Sciences*, 43, pp. 135–149.

Dixit, A. and Pindyck, R. (1994) *Investment under Uncertainty*. Princeton University Press, Princeton, NJ.

Doherty, N. and Eeckhoudt, L. (1995) “Optimal insurance without expected utility: The dual theory and the linearity of insurance contracts”, *Journal of Risk and Uncertainty*, 10, 2, pp. 157–179.

Doherty, N. and Schlessinger, H. (1983) “Optimal insurance in incomplete markets”, *Journal of Political Economy*, 91, pp. 1045–1054.

Drèze, J. and Stern, N. (1987) “The theory of cost–benefit analysis”, in A. Auerbach and M. Feldstein (eds), *Handbook of Public Economics*. Elsevier Sciences, Amsterdam.

Duffie, D. (1988) *Security Markets – Stochastic Models*. Academic Press, Orlando, FL.

Duffie, D. (1992) *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, NJ.

Duffie, D and Sonnenschein, H. (1989) “Arrow and general equilibrium theory”, *Journal of Economic Literature*, 27, pp. 565–598.

Duffie, D. and Huang, C.-F. (1985) “Implementing Arrow–Debreu equilibria by continuous trading of few long-lived securities”, *Econometrica*, 53, pp. 1337–1356.

Eichberger, J., Grant, S. and Kelsey, D. (2005) “CEU preferences and dynamic consistency”, *Mathematical Social Sciences*, 49, 2, pp. 143–151.

Froot, K., Murphy, B., Stern, A. and Usher, S. (1995) “The emerging asset class: insurance risk”, Special report from Guy Carpenter and Company, Inc.

Gilboa, I. (1989) “Expectation and variation in multi-period decisions”, *Econometrica*, 57, pp. 1153–1169.

Gilboa, I. and Schmeidler, D. (1995) “Case-based decision theory”, *The Quarterly Journal of Economics*, 110(3), pp. 605–639.

Gollier, C. (2001) *The Economics of Risk and Time*. MIT Press, Cambridge, MA.

Gonzalez, R. and Wu, G. (1999) “On the shape of the probability weighting function”, *Cognitive Psychology*, 38, pp. 129–166.

Grinols, E. (1985) "Public investment and social risk-sharing", *European Economic Review*, 29, pp. 303–321.

Harrison, J. and Kreps, D. (1979) "Martingales and arbitrage in multiperiod securities markets", *Journal of Economic Theory*, 20, pp. 381–408.

Harrison, J. and Pliska, S. (1981) "Martingales and stochastic integrals in the theory of continuous trading", *Stochastic Processes and their Applications*, 11, pp. 215–260.

Hirschleifer, J. (1966) "Investment decisions under uncertainty: applications of the state-preference approach", *The Quarterly Journal of Economics*, 80, pp. 252–277.

Jaffray, J-Y. (1989) "Linear utility theory for belief functions", *Operation Research Letters*, 8, pp. 107–112.

Jewitt, I. (1989) "Choosing between risky prospects: the characterization of comparative statics results, and location independent risk", *Management Science*, 35, pp. 60–70.

Jouini, E. and Kallal, H. (1995) "Martingales, arbitrage and equilibrium in securities markets with transaction costs", *Journal of Economic Theory*, 66, pp. 178–197.

Kahneman, D. and Tversky, A. (1979) "Prospect theory: an analysis of decision under risk", *Econometrica*, 47, pp. 263–291.

Karni, E. (1992) "Optimal insurance: a non expected utility analysis", in G. Dionne (ed.), *Insurance Economics*. Kluwer Academic, Amsterdam.

Kast, R. (1991) "A foundation of Bayesian statistics", *Theory and Decision*, 31, pp. 175–197.

Kast, R. and Lapied, A. (2001) "Gestion de l'assurance chômage par des actifs contingents aux taux de chômage", WP GREQAM.

Kast, R. and Lapied, A. (2003) "Comonotonic bookmaking and attitudes to uncertainty", *Mathematical Social Sciences*, 46, pp. 1–7.

Kast, R. and Lapied, A. (2005) "Updating Choquet valuation and discounting information arrivals", WP GREQAM-IDEP.

Kast, R., Lapied, A. and Pardo, S. (2004) "Construction d'un portefeuille sous-jacent virtuel", *Revue Economique*, 55, pp. 407–418.

Kast, R., Lapied, A., Pardo, S. and Protopopescu, C. (2001) "Evaluation de risques controversés par la théorie des options réelles", *Economie et Prévision*, 149, pp. 51–63.

Kast, R. and Luchini, S. (2002) "Calcul économique et risques socio-politiques: une procédure d'évaluation des projets publics", *Economie et Prévision*, 156(5), pp. 73–84.

Kast, R. and Luchini, S. (2004) "Value of statistical life revisited within rank dependent expected utility", WP GREQAM.

Keynes, J. (1921) *A Treatise on Probability*. Macmillan, London.

Kihlstrom, R., Romer, D. and Williams, S. (1981). "Risk aversion with random initial wealth", *Econometrica*, 49, pp. 911–920.

Knight, F. (1921) *Risk, Uncertainty and Profit*. Houghton Mifflin, Boston.

Kreps, D. (1982) "Multiperiod securities and the efficient allocation of risk: A comment on the Black-Scholes option pricing model", in J. MacCall (ed.), *The Economics of Uncertainty and Information*, University of Chicago Press, Chicago.

Laffont, J.-J. and Martimort, D. (2002) *The theory of incentives: The principal-agent model*. Princeton University Press, Princeton, NJ.

Landsberger, M. and Meilijson, I. (1994) "Co-monotone allocations, Bickel–Lehmann dispersion and the Arrow–Pratt measure of risk aversion", *Annals of Operations Research*, 52, pp. 97–105.

Lintner, J. (1965) "The valuation of risky assets and the selection of risky investments in stock portfolios and capital budget", *Review of Economics and Statistics*, 47, pp. 13–37.

Litzenberger, R., Beaglehole, D. and Reynolds, C. (1996) "Assessing catastrophe reinsured-linked securities as a new asset class", *Journal of Portfolio Management*, 22, pp. 76–86.

Luce, R. and Raiffa, H. (1957) *Games and Decisions*. John Wiley & Sons, Chichester.

Luciano, E. and Kast, R. (2001) "A value at risk approach to background risk", *The Geneva Papers on Risk and Insurance*, 26, pp. 91–116.

Markowitz, H. (1952) "Porfolio selection", *Journal of Finance*, 7, pp. 77–91.
Merton, R. (1973a) "An intertemporal capital asset pricing model", *Econometrica*, 41, pp. 867–888.
Merton, R. (1937b) "The theory of rational option pricing", *Bell Journal of Economics and Management Science*, 4, pp. 141–183.
Mossin, J. (1966) "Equilibrium in a capital market", *Econometrica*, 34, pp. 768–783.
Myles, G. (1995) *Public Economics*. Cambridge University Press, Cambridge.
Osborne, J. and Rubinstein, A. (1994) *A course in game theory*. MIT Press, Cambridge, Mass.
Plott, C. and Sunder, S. (1988) "Rational expectations and the aggregation of diverse information in laboratory security markets", *Econometrica*, 56, pp. 1085–1118.
Pratt, J. (1964) "Risk aversion in the small and in the large", *Econometrica*, 32, pp. 122–136.
Quiggin, J. (1982) "A theory of anticipated utility", *Journal of Economic Behavior and Organisations*, 3(4), pp. 323–242.
Quiggin, J. (1992) "Increasing risk, another definition", in A. Chikun (ed.), *Progress in Decision, Utility and Risk Theory*. Kluwer Academic, Amsterdam.
Ramsey, F. (1926) "Truth and probability", reprint in R. Braithwaite (ed.), *The Foundations of Mathematics and Other Logical Essays*. Routledge, London, 1931.
Roll, R. and Ross, S. (1980) "An empirical investigation of the Arbitrage Pricing Theory", *The Journal of Finance*, 35, pp. 1079–1103.
Ross, S. (1976) "The arbitrage theory of capital asset pricing", *Journal of Economic Theory*, 13, pp. 341–360.
Ross, S. (1978) "A simple approach to the valuation of risky streams", *The Journal of Business*, 51, pp. 453–475.
Rothschild, M. and Stiglitz, J. (1970) "Increasing risk: I. A definition", *Journal of Economic Theory*, 2, pp. 245–243.
Sandmo, A. (1972) "Discount rates for public investment under uncertainty", *International Economic Review*, 13, pp. 287–302.
Savage, L. (1954) *The Foundations of Statistics*. John Wiley & Sons, Chichester.
Schmeidler, D. (1986) "Integral representation without additivity", *Proceedings of the American Mathematical Society*, 97, pp. 255–261.
Schmeidler, D. (1989) "Subjective probability and expected utility without additivity", *Econometrica*, 57, pp. 571–587.
Shalev, J. (1997) "Loss aversion in a multi-period model", *Mathematical Social Sciences*, 33, pp. 203–226.
Sharpe, W. (1964) "Capital asset prices: a theory of market equilibrium under conditions of risks", *Journal of Finance*, 19, pp. 425–442.
Stapleton, R. and Subrahmanyam, M. (1978) "Capital market equilibrium in a mixed economy, optimal public sector investment decision rules, and the social rate of discount", *The Quarterly Journal of Economics*, 92, pp. 399–412.
Sunder, S. (1995) "Experimental asset markets: a survey", in J. Kagel and E. Roth (eds), *Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ.
Thomas, H. (1999) "A preliminary look at gains from asset securitization", *Journal of International Financial Markets, Institutions and Money*, 9, pp. 321–333.
Trigeorgis, L. (1996) *Real Options*. MIT Press, Cambridge, MA.
Tversky, A. and Kahneman, D. (1981) "Loss aversion in riskless choice: a reference-dependent model", *The Quarterly Journal of Economics*, 106, pp. 1039–1061.
Tversky, A. and Kahneman, D. (1992), "Advances in prospect theory: cumulative representation of uncertainty", *Journal of Risk and Uncertainty*, 5, pp. 297–323.
Von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. Princeton Universtity Press, Princeton, NJ.
Wang, S. (2000) "A class of distortion operators for pricing financial and insurance risks", *Journal of Risk and Insurance*, 67, pp. 15–36.
Yaari, M. (1987) "Dual theory of choice under uncertainty", *Econometrica*, 55(1), pp. 95–115.
Yonnet, P. (1985) *Jeux, modes et masses*. Gallimard, Paris.

Index

Index compiled by Terry Halliday