

Copyrighted Material

Jean-Pierre Danthine and John B. Donaldson

INTERMEDIATE FINANCIAL THEORY

THIRD EDITION



Copyrighted Material

Intermediate Financial Theory

Intermediate Financial Theory

Third Edition

Jean-Pierre Danthine
Swiss National Bank
Bundesplatz 1
Bern, Switzerland

John B Donaldson
Columbia Business School
New York, NY



AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

Copyright © 2015, 2005 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system
or transmitted in any form or by any means electronic, mechanical, photocopying, recording
or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights
Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333;
email: permissions@elsevier.com. Alternatively you can submit your request online by visiting
the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission
to use Elsevier material*

First edition 2001

The first edition of this book was published by Pearson Education, Inc.

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or
property as a matter of products liability, negligence or otherwise, or from any use or operation
of any methods, products, instructions or ideas contained in the material herein. Because of
rapid advances in the medical sciences, in particular, independent verification of diagnoses and
drug dosages should be made

British Library Cataloguing-in-Publication Data

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

Library of Congress Catalog Number

A catalog record for this book is available from the Library of Congress

ISBN-13: 978-0-12-386549-6

For information on all Academic Press publications
visit our website at <http://store.elsevier.com/>

Typeset by MPS Limited, Chennai, India
www.adimps.com

Printed and bound in the USA



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Preface

For all the suffering that it has caused, the global financial crisis provides a unique opportunity to review what we know or thought we knew about finance. It will challenge and enliven the teaching of finance for years to come. The third edition of Intermediate Financial Theory is affected accordingly. While our own goals for the text have not changed, many new topics have been added and many examples have been taken from recent experience. The task of reviewing the entire material in light of the financial crisis is, however, a work in progress and one that cannot be adequately undertaken within the context of periodic revisions of a textbook of reasonable length. Accordingly, it will be pursued on an ongoing basis on the text's website.

The third edition of Intermediate Financial Theory features 2 entirely new chapters and very substantial revisions to 11 more. With respect to the latter changes, there is greater emphasis on “behavioral finance” and many of the latest developments in portfolio theory are fully featured. The chapter on the consumption capital asset pricing model has been similarly expanded and brought to the theoretical frontier. Integral to the print version of the text are four Web Chapters. The intent of these chapters is to expand on the basic ideas presented in the text in ways that link them more directly to applied practice. Our review and assessment of the recent “financial crisis” is a particular case in point. Lastly, the third edition attempts to strengthen the link between theory and “the data”; that is, of evaluating a particular theory as regards its ability to replicate the associated data patterns, the so-called financial stylized facts.

While the market for financial textbooks is crowded at both the introductory and doctoral levels, it remains much thinner at the intermediate level. Teaching opportunities at this level, however, have greatly increased with the advent of masters of science programs in finance (master's degree programs in computational finance, in mathematical finance, and the like) and the continuing demand for higher-level courses in MBA programs.

The Master in Banking and Finance Program at the University of Lausanne, which admitted its first class in the fall of 1993 is a program of the aforementioned type. One of the first such programs of its kind in Europe, its objective is to provide advanced training to finance specialists in the context of a 1-year theory-based degree program. In designing the curriculum, it was felt that students should be exposed to an integrated course that would

introduce the full range of topics typically covered in financial economics courses at the doctoral level. Such exposure could, however, ignore the detailed proofs and arguments and concentrate on the larger set of issues and concepts to which any advanced practitioner should be exposed. This latest edition of our text retains this philosophy.

Accordingly, our ambition for this third edition is unchanged from earlier ones: first to review rigorously and concisely the main themes of financial economics (those that students should have encountered in prior courses) and, second, to introduce a number of *frontier* ideas of importance for the evolution of the discipline and of relevance from a practitioner's perspective. We want our readers not only to be at ease with the main concepts of standard finance (MPT, CAPM, etc.) but also to be aware of the principal new ideas that have marked the recent evolution of the discipline. Contrary to introductory texts, we aim at depth and rigor; contrary to higher-level texts, we do not emphasize generality. Whenever an idea can be conveyed through an example, this is the approach we choose. We continue to ignore proofs and detailed technical matters unless a reasonable understanding of the related concept mandates their inclusion.

Intermediate Financial Theory is intended primarily for master level students with a professional orientation, a good quantitative background, and who have completed at least one introductory finance course (or have read the corresponding text). As such, the book is targeted especially for masters students in finance, while remaining accessible and appropriate for an advanced MBA class in financial economics, one with the objective of introducing students to the precise modeling of many of the concepts discussed in their capital markets and corporate finance classes. In addition, we believe the book can be a useful reference for doctoral candidates in finance, particularly those whose lack of prior background might prevent them from drawing the full benefits of the very abstract material typically covered at that level. Finally, we hope it will be a useful refresher for well-trained practitioners. Although the mathematical requirements of the book are not great, some confidence in the use of calculus as well as matrix algebra is helpful.

In preparing this third edition, we maintain our earlier emphasis on the valuation of risky cash flows. This subject—asset pricing—constitutes the main focus of modern finance, and its shortcomings have come powerfully to the fore in the recent financial crisis. We also emphasize the distinction between valuation procedures that rely on general equilibrium principles and those based on arbitrage considerations.

At present there are four Web Chapters that are available to readers. These represent substantial extensions of ideas introduced in the print version of the book. Web Chapter A translates the Consumption CAPM model into a fully dynamic production setting so that the mutual influence of the financial markets on and by the macroeconomy can be made more explicit. Web Chapter B goes beyond using the martingale measure to price options to an exploration of the use of options concepts in the evaluation of complex securities, real

investment projects, and strategies of portfolio management, while Web Chapter C returns to a more general treatment of differential information. Lastly, Web Chapter D explores the origins and evolution of the recent financial crisis and does so in a way that is intended to assess the strengths and weaknesses of the current state of financial theory. By placing these chapters on the Web, we can more easily update and add to the material presented therein. More chapters will be added in future years.

Over the years, we have benefited from numerous discussions with colleagues over issues related to the material included in this book. We are especially grateful to Rajnish Mehra, Arizona State University; Elmar Mertens, IMF; Paolo Siconolfi, Lars Lochstoer, Kent Daniel, Andrew Ang and Tano Santos, all of Columbia Business School; and Erwan Morellec, University of Lausanne, the latter for his contribution to the corporate finance review of Chapter 2. We are also indebted to several generations of teaching assistants including François Christen, Philippe Gilliard, Tomas Hricko, Aydin Akgun, Paul Ehling, Oleksandra Hubal, and Lukas Schmid—and of MBF students at the University of Lausanne—who have participated in the shaping up of this material. Teaching and research assistants from the “other side of the Atlantic” have been hugely helpful as well, most especially, J.K. Auh, Mattia Landoni, and Zhongjin Lu. Their questions, corrections, and comments have led to a continuous questioning of the approach we have adopted and have dramatically increased the usefulness of this text. Finally, we reiterate our thanks to the Fondation du 450ème of the University of Lausanne for providing “seed financing” for this project.

Jean-Pierre Danthine
Bern, Switzerland

John B. Donaldson
New York, NY

*N'estime l'argent ni plus ni moins qu'il ne vaut:
c'est un bon serviteur et un mauvais maître
(Value money neither more nor less than it is worth:
It is a good servant and a bad master)*

Alexandre Dumas, fils, *La Dame aux Camélias* (Préface)

Cover

by Renée-Paule Danthine

« Petite fugue »; 2011 Oil and pastel on Japanese paper

Dedication

J.B. Donaldson wishes to dedicate his involvement in this book enterprise to his parents, Brown and Rachel Donaldson, his wife Charissa Asbury, Mario Gabelli who offered financial support through the provision of his academic chair and, most especially, his physicians, Kareem M. Abu-Elmagd, MD, and Guilherme Costa, MD, of the University of Pittsburgh Medical Center. In various ways, all these persons were vital to the completion of the book.

On the Role of Financial Markets and Institutions

Chapter Outline

1.1 Finance: The Time Dimension	3
1.2 Desynchronization: The Risk Dimension	6
1.3 The Screening and Monitoring Functions of the Financial System	7
1.4 The Financial System and Economic Growth	8
1.5 Financial Markets and Social Welfare	12
1.6 Financial Intermediation and the Business Cycle	18
1.7 Financial Crises	19
1.8 Conclusion	22
References	23
Complementary Readings	24
Appendix: Introduction to General Equilibrium Theory	24
Pareto Optimal Allocations	25
Competitive Equilibrium	27

1.1 Finance: The Time Dimension

Why do we need financial markets and institutions? We choose to address this question as our introduction to this text on financial theory. In doing so, we touch on some of the most difficult issues in finance and introduce concepts that will eventually require extensive development. Our purpose here is to phrase this question as an appropriate background for the study of the more technical issues that will occupy us at length. We also want to introduce some important elements of the necessary terminology. We ask the reader's patience as most of the sometimes difficult material introduced here will be taken up in more detail in the following chapters.

Fundamentally, a financial system is a set of institutions and markets permitting the exchange of contracts and the provision of services for the purpose of allowing the income and consumption streams of economic agents to be desynchronized—i.e., made less similar. It can, in fact, be argued that indeed the *primary* function of the financial system is to

permit such desynchronization. There are two dimensions to this function: the time dimension and the risk dimension. Let us start with time. Why is it useful to disassociate consumption and income across time? Two reasons come immediately to mind. First, and somewhat trivially, income is typically received at discrete dates, say monthly, while it is customary to wish to consume continuously (i.e., every day).

Second, and more importantly, consumption spending defines a *standard of living*, and most individuals find it difficult to alter their standard of living from month to month or even from year to year. There is a general, if not universal, desire for a *smooth* consumption stream. Because it deeply affects everyone, the most important manifestation of this desire is the need to save (consumption smaller than income) for retirement so as to permit a consumption stream in excess of income (dissaving) after retirement begins. The *life-cycle* patterns of income generation and consumption spending are not identical, and the latter must be created from the former. The same considerations apply to shorter horizons. Seasonal patterns of consumption and income, for example, need not be identical. Certain individuals (car salespersons, department store salespersons, construction workers) may experience variations in income arising from seasonal events (e.g., most new cars are purchased in the spring and summer; construction activity is much reduced in winter), which they do not like to see transmitted to their ability to consume. There is also the problem created by temporary layoffs due to variation in aggregate economic activity that we refer to as business cycle fluctuations. While they are temporarily laid off and without substantial income, workers do not want their family's consumption to be severely reduced ([Box 1.1](#)).

Furthermore, and this is quite crucial for the growth process, some people—entrepreneurs, in particular—are willing to accept a relatively small income (but not necessarily

BOX 1.1 Representing Preference for Smoothness

The preference for a smooth consumption stream has a natural counterpart in the form of the utility function, $U(\)$, which is typically used to represent the relative benefit a consumer receives from a specific consumption bundle. Suppose the representative individual consumes a single consumption good (or a basket of goods) in each of two periods, now and tomorrow. Let c_1 denote today's consumption level and c_2 tomorrow's, and let $U(c_1) + U(c_2)$ represent the level of utility (benefit) obtained from a given consumption stream (c_1, c_2) .

Preference for consumption smoothness must mean, for instance, that the consumption stream $(c_1, c_2) = (4, 4)$ is preferred to the alternative $(c_1, c_2) = (3, 5)$, or

$$U(4) + U(4) > U(3) + U(5)$$

Dividing both sides of the inequality by 2, this implies

$$U(4) > \frac{1}{2}U(3) + \frac{1}{2}U(5)$$

(Continued)

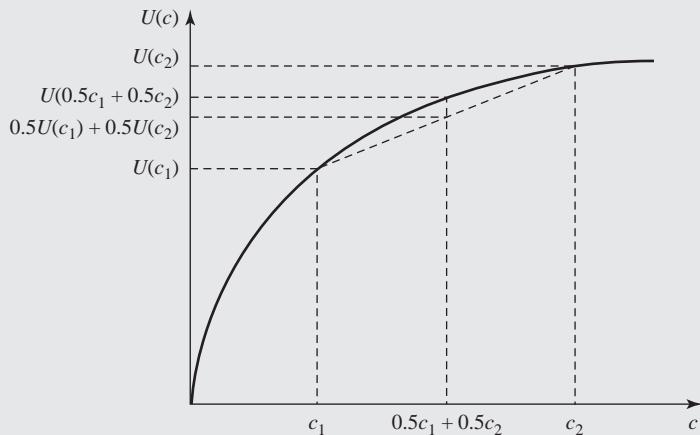
BOX 1.1 Representing Preference for Smoothness (Continued)


Figure 1.1
A strictly concave utility representation.

As shown in [Figure 1.1](#), when generalized to all possible alternative consumption pairs, this property implies that the function $U(\cdot)$ has the rounded shape that we associate with the term *strict concavity*.

consumption!) for an initial period of time in exchange for the prospect of high returns (and presumably high income) in the future. They are operating a sort of arbitrage over time. This does not disprove their desire for smooth consumption; rather, they see opportunities that lead them to accept what is formally a low-income level initially against the prospect of a much higher income level later (followed by a zero income level when they retire). They are investors who, typically, do not have enough liquid assets to finance their projects and, as a result, need to raise capital by borrowing or by selling shares.

Indeed, the first key element in finance is **time**. In a timeless world, there would be no assets, no financial transactions (although money would be used, it would have only a transaction function), and no financial markets or institutions. The very notion of a security (a financial contract) implies a time dimension.

Asset holding permits the desynchronization of consumption and income streams. The peasant putting aside seeds, the miser burying his gold, or the grandmother putting a few hundred dollar bills under her mattress are all desynchronizing their consumption and income, and in doing so, presumably seeking a higher level of well-being for themselves. A fully developed financial system should also have the property of fulfilling this same function *efficiently*. By that we mean that the financial system should provide versatile and

diverse instruments to accommodate the widely differing needs of savers and borrowers insofar as size (many small lenders, a few big borrowers), timing, and maturity of loans (how to finance long-term projects with short-term money), and the liquidity characteristics of instruments (precautionary saving cannot be tied up permanently). In other words, the elements composing the financial system should aim at *matching* the diverse financing needs of different economic agents as perfectly as possible.

1.2 Desynchronization: The Risk Dimension

We have argued that time is of the essence in finance. When we talk of the importance of time in economic decisions, we think in particular of the relevance of choices involving the present versus the future. But the future is, by its very nature, uncertain: financial decisions with implications (payouts) in the future are necessarily risky. Time and risk are inseparable. This is why **risk** is the second key word in finance.

For the moment, let us compress the time dimension into the setting of a “Now and Then” (present versus future) economy. The typical individual is motivated by the desire to smooth consumption between “Now” and “Then.” This implies a desire to identify consumption opportunities that are as similar as possible among the different possibilities that may arise “Then.” In other words, *ceteris paribus*—most individuals would like to guarantee their family the same standard of living whatever events transpire tomorrow: whether they are sick or healthy, unemployed or working, confronted with bright or poor investment opportunities, fortunate or hit by unfavorable accidental events.¹ This characteristic of preferences is generally described as “aversion to risk.”

A productive way to start thinking about this issue is to introduce the notion of *states of nature* or *states of the world*. A state of nature is a complete description of a possible scenario for the future across all the dimensions relevant for the problem at hand. In a “Now and Then” economy, all possible future events can be represented by an exhaustive list of states of nature. We can thus extend our former argument for smoothing consumption across time by noting that the typical “risk averse” individual would also like to experience similar consumption levels across all future states of nature, whether good or bad.

An efficient financial system offers ways for savers to reduce or eliminate, at a fair price, the risks they are not willing to bear (risk shifting). Fire insurance contracts eliminate the financial risk of fire, while put options contracts can prevent the loss in wealth associated with a stock’s price declining below a predetermined level, to mention but two examples. The financial system also makes it possible to obtain relatively safe aggregate returns from a large number of small, relatively risky investments. This is the process of diversification. By permitting economic agents to *diversify*, to *insure*, and to *hedge* their risks, an efficient

¹ *Ceteris paribus* is the Latin phrase for “everything else maintained equal.” It is an expression commonplace in the language of economics.

financial system fulfills the function of redistributing purchasing power not only over time, but also across states of nature.²

1.3 The Screening and Monitoring Functions of the Financial System

The business of desynchronizing consumption from income streams across time and states of nature is often more complex than our initial description may suggest. If time implies uncertainty, uncertainty may imply not only risk, but often *asymmetric information* as well. By this term, we mean situations where the agents involved have different information, with some being potentially better informed than others. How can a saver be assured that he will be able to find a borrower with a good ability to repay—the borrower himself knows more about this, but he may not wish to reveal all he knows—or an investor (an entrepreneur or a firm) with a good project, yielding the most attractive return for him and hopefully for society as well? Again, the investor is likely to have a better understanding of the project's prospects and of his own motivation to carry it through. What do “good” and “most attractive” mean in these circumstances? Do these terms refer to the highest potential return? What about risk?

What if the anticipated return is itself affected by the actions of the investors themselves (a phenomenon labeled “moral hazard”)? How does one share the risks of a project in such a way that both investors and savers are willing to proceed, taking actions acceptable to both? It is the task of financial intermediaries—banks, venture capital firms, and private equity firms—to answer these questions and to do so in such a way that brings the socially beneficial projects to fruition. An efficient financial system, and the financial institutions that define it, not only assists in these information and monitoring tasks, but also provides a range of instruments (contractual arrangements) suitable for the largest number of savers and borrowers, thereby contributing to the channeling of savings toward the most efficient projects.³

In the words of the preeminent economist, [Joseph Schumpeter \(1934\)](#), “Bankers are the gatekeepers of capitalist economic development. Their strategic function is to screen potential innovators and advance the necessary purchasing power to the most promising.”

² Both insurance and hedging are risk-reduction strategies but with one critical difference. In the case of insurance, the investor pays money—the insurance premium—to guarantee against a loss in value of some asset that he owns (a house, shares of stock). In the case of hedging, an investor adds to his portfolio, usually at very little cost, another asset (the “hedging asset”) with a price pattern that is opposite to that of his original portfolio: if the original portfolio declines in value, the newly added asset increases in value by an equal and offsetting amount (this is the case of a “perfect hedge”). The opposite is true, however, if the investor’s original portfolio increases in value: the hedging asset loses an equal and offsetting amount. The investor thus sacrifices potential gains to his portfolio’s value in exchange for protection against losses.

In the case of insurance, upward potential was not sacrificed, but the investor had to pay the premium.

³ If the extent of the information asymmetry between buyers and sellers in a market becomes too great, the market may shut down: no trades occur. This exact event occurred at the start of the financial crisis when the investment banks that had been packaging pools of US home mortgages into mortgage-backed securities (MBS) discovered that they could find no buyers. The natural buyers of these securities had become suspicious that their quality was not as advertised. The forced sale of a nearly insolvent Bear Stearns to JPMorgan Chase and the Lehman Brothers bankruptcy ensued.

For highly risky projects, such as the creation of a new firm exploiting a new technology, venture capitalists largely provide this function today.

1.4 The Financial System and Economic Growth

The performance of the financial system matters at several levels. We shall argue that it matters for growth, that it impacts the characteristics of the business cycle, and, most importantly, that it is a significant determinant of economic welfare. We tackle growth first. Channeling funds from savers to investors efficiently is obviously important.

Whenever more efficient ways are found to perform this task, society can achieve a greater increase in tomorrow's consumption for a given sacrifice in consumption today. As a result, savings becomes a more attractive alternative to current consumption, and households save more. Intuitively, more savings should lead to greater investment and thus greater future wealth. [Figure 1.2](#) indeed suggests that, for 90 developing countries over the period 1971–1992, there was a strong positive association between saving rates and growth rates. When looked at more carefully, however, the evidence is usually not as strong.⁴

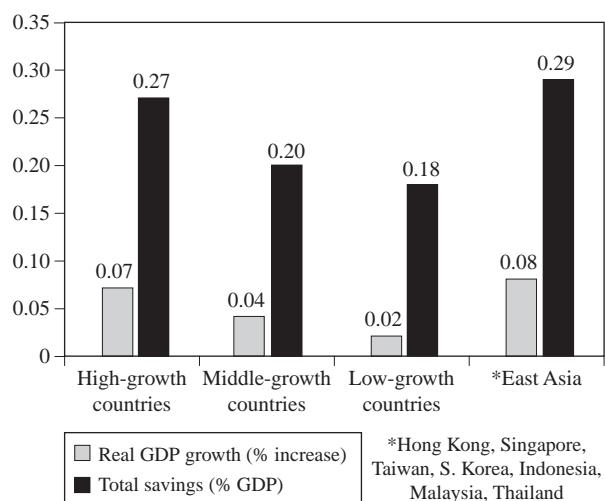


Figure 1.2
Savings and growth in 90 developing countries.

⁴ In a straightforward regression in which the dependent variable is the growth rate in real per capita gross national product (GNP), the coefficient on the average fraction of real GNP represented by investment (I/Y) over the prior 5 years is positive but insignificant. Together with other results, this is interpreted as suggesting a possible reverse causation from real per capita GNP growth to investment spending. See [Barro and Sala-i-Martin \(1995\)](#), Chapter 12, for a full discussion. There is also a theoretically important distinction between the effects of increasing investment (savings) (as a proportion of national income) on an economy's *level* of wealth and its *growth rate*. Countries that save more will *ceteris paribus* be wealthier, but they need not grow more rapidly. The classic growth model of [Solow \(1956\)](#) illustrates this distinction. See Web Chapter A.

One important reason may be that the hypothesized link is, of course, dependent on a *ceteris paribus* clause: It applies only to the extent savings are invested in appropriate ways. The economic performance of the former USSR reminds us that it is not enough only to save; it is also important to invest judiciously. Historically, the investment/GDP (gross domestic product, the measure of a nation's aggregate economic output) ratio in the USSR was very high in international comparisons, suggesting the potential for very high growth rates.⁵ After 1989, however, experts realized that the value of the existing stock of capital was not consistent with the former levels of investment. A great deal of the investment had been effectively wasted—in other words, allocated to poor or even worthless projects. Equal savings rates can thus lead to investments of widely differing degrees of usefulness from the viewpoint of future growth.

A more contemporary version of this same overinvestment phenomenon may be present in China. In 2012, investment in China represented 46% of output which is astonishingly high by international standards (e.g., the United States averages 15%; Switzerland averages around 20%), and a lively debate has arisen as to what fraction of China's investment will ultimately be useful.⁶

Let us go further than these general statements in the analysis of the savings and growth nexus and of the role of the financial system. Following [Barro and Sala-i-Martin \(1995\)](#), one can view the process of transferring funds from savers to investors in the following way.⁷ The least efficient system would be one in which all investments are made by the savers themselves. This is certainly inefficient because it requires a sort of “double coincidence” of intentions: good investment ideas occurring in the mind of someone lacking past savings will not be realized. Funds that a nonentrepreneur saves would not be put to productive use. Yet, this unfortunate situation is a clear possibility if the necessary confidence in the financial system is lacking, with the consequence that savers do not entrust the system with their savings. One can thus think of circumstances where savings never enter the financial system, or where only a small fraction does. When it does, it will typically enter via some sort of depository institution. In an international setting, a similar problem arises if national savings are primarily invested abroad, a situation that may reach alarming proportions in the case of underdeveloped

⁵ More precisely, GDP refers to the value, at market prices, of all final goods and services (those sold to end users) produced within a nation's geographical boundaries during a specific time period (usually a year).

⁶ Excessive investment comes at the price of lower household consumption. [Lee et al. \(2012\)](#) estimate this loss in consumption to have averaged 4% of total Chinese output.

⁷ For a broader perspective and a more systematic connection with the relevant literature on this topic, see [Levine \(1997\)](#).

countries.⁸ Let FS/S represent, then, the fraction of aggregate savings (S) being entrusted to the financial system (FS).

At a second level, the functioning of the financial system may be more or less costly. While funds transferred from a saver to a borrower via a direct loan are immediately and fully made available to the end user, the different functions of the financial system discussed above are often best fulfilled, or sometimes can only be fulfilled, through some form of intermediation, which typically involves some cost. Let us think of these costs as administrative costs, on the one hand, and costs linked to the reserve requirements of banks, on the other. Different systems will have different operating costs in this large sense, and, as a consequence, the amount of resources transferred to investors will also vary. Let us think of BOR/FS as the ratio of funds transferred from the financial system to borrowers and entrepreneurs.

Borrowers themselves may make diverse use of the funds borrowed. Some, for example, may have pure liquidity needs (analogous to the reserve needs of depository institutions), and if the borrower is the government, it may well be borrowing for consumption! For the savings and growth nexus, the issue is how much of the borrowed funds actually result in productive investments. Let I/BOR represent the fraction of borrowed funds actually invested. Note that BOR stands for borrowed funds whether private or public. In the latter case, a key issue is what fraction of the borrowed funds are used to finance public investment as opposed to public consumption.

Finally, let EFF denote the efficiency of the investment projects undertaken in society at a given time, with EFF normalized at unity; in other words, the average investment project has $EFF = 1$, the below-average project has $EFF < 1$, and conversely for the above average project (a project consisting of building a bridge leading nowhere would have an $EFF = 0$); K is the aggregate capital stock and Ω the depreciation rate. We may then write

$$\dot{K} = EFF \cdot I - \Omega K \quad (1.1)$$

or, multiplying and dividing I with each of the newly defined variables

$$\dot{K} = EFF \cdot (I/BOR) \cdot (BOR/FS) \cdot (FS/S) \cdot (S/Y) \cdot Y - \Omega K \quad (1.2)$$

⁸ The problem is slightly different here, however. Although capital flight is a problem from the viewpoint of building up a country's domestic capital stock, the acquisition of foreign assets may be a perfectly efficient way of building a national capital stock. The effect on growth may be negative when measured in terms of GDP, but not necessarily so in terms of national income or GNP. Switzerland is an example of a rich country investing heavily abroad and deriving a substantial income flow from it. It can be argued that the growth rate of the Swiss GNP (but probably not GDP) has been enhanced rather than diminished by this fact.

where our notation is meant to emphasize that the growth of the capital stock at a given savings rate is likely to be influenced by the levels of the various ratios introduced above.⁹ Let us now review how this might be the case.

One can see that a financial system performing its matching function efficiently will positively affect the savings rate (S/Y) and the fraction of savings entrusted to financial institutions (FS/S). This reflects the fact that savers can find the right savings instruments for their needs. In terms of overall services net of inconvenience, this acts like an increase in the return to the fraction of savings finding its way into the financial system. The matching function is also relevant for the I/BOR ratio. With the appropriate instruments (like flexible overnight loan facilities), a firm's cash needs are reduced and a larger fraction of borrowed money can actually be used for investment.

By offering a large and diverse set of possibilities for spreading risks (insurance and hedging), an efficient financial system will also positively influence the savings ratio (S/Y) and the FS/S ratio. Essentially this works through improved return/risk opportunities, corresponding to an improved trade-off between future and present consumption (for savings intermediated through the financial system). Furthermore, in permitting entrepreneurs with risky projects to eliminate unnecessary risks by using appropriate instruments, an efficient financial system provides, somewhat paradoxically, a better platform for undertaking riskier projects. If, on average, riskier projects are also the ones with the highest returns, as most of financial theory reviewed later in this book leads us to believe, one would expect that the more efficiently this function is performed, the higher (*ceteris paribus*) the value of EFF ; in other words, the higher, on average, the efficiency of the investment undertaken with the funds made available by savers.

Finally, a more efficient system may be expected to screen alternative investment projects more effectively and to monitor more thoroughly and more cost efficiently the conduct of the investments (efforts of investors). The direct impact is to increase EFF . Indirectly this also means that, on average, the return/risk characteristics of the various instruments offered savers will be improved and one may expect, as a result, an increase in both S/Y and FS/S ratios.

The previous discussion thus tends to support the idea that the financial system plays an important role in permitting and promoting the growth of economies.¹⁰ Yet growth is not an objective in itself. There is such a thing as excessive capital accumulation typically

⁹ $\dot{K} = dK/dt$, i.e., the change in K as a function of time.

¹⁰ There is statistical support asserting the beneficial consequences of financial development for economic growth at the country (King and Levine, 1993a, b), industry (Rajan and Zingales, 1998), and firm levels (Aghion et al., 2007).

funded in some way by financial repression directed at households. In the case of Italy, Jappelli and Pagano (1994) suggest that household borrowing constraints,¹¹ in general a source of inefficiency and the mark of a less than perfect financial system, may have led to more savings than desired in the 1980s.

The excessive investment rates evident in China (noted earlier) are partially funded by high household savings (the household savings rate in China is presently approximately 25%) in part driven by financial repression in the form of government-mandated low interest rates that banks are permitted to offer depositors and prohibitions on the ownership of certain types of securities (e.g., stocks or bonds issued by foreign-based firms or foreign governments). In the absence of a significant social safety net (no government sponsored pensions or health care), households are thus induced to save a lot, much of which gets channeled into the investments of state-owned enterprises or local governments. Financial “repression” takes a different form in India: there are few banks in rural areas. As a result, many rural households do not commit their savings to the financial system at all, but prefer to buy “gold.” “Investment” in the form of gold purchases contributes nothing to India’s growth prospects.

While these examples are purely illustrative, they underscore the necessity of adopting a broader and more satisfactory viewpoint and of more generally studying the impact of the financial system on social welfare. This is best done in the context of the theory of general equilibrium, a subject to which we next turn.

1.5 Financial Markets and Social Welfare

Let us next consider the role of financial markets in the allocation of resources and, consequently, their effects on social welfare. The perspective provided here places the process of financial innovation in the context of the theory of general economic equilibrium whose central concepts are closely associated with the *Ecole de Lausanne* and the names of Léon Walras and Vilfredo Pareto.

Our starting point is the first theorem of welfare economics, which defines the conditions under which the allocation of resources implied by the general equilibrium of a decentralized competitive economy is efficient or optimal in the Pareto sense.

First, let us define the terms involved. Assume a timeless economy where a large number of economic agents interact. There is an arbitrary number of goods and services, n . Consumers possess a certain quantity (possibly zero) of each of these n goods (in particular, they have

¹¹ By “borrowing constraints,” we mean the limitations that the average individual or firm may experience in his or her ability to borrow, at current market rates, from financial institutions.

the ability to work a certain number of hours per period). They can sell some of these goods and buy others at prices quoted in markets. There are a large number of firms, each represented by a production function—i.e., a given ability (constrained by what is technologically feasible) to transform some of the available goods or services (inputs) into others (outputs)—for instance, combining labor and capital to produce consumption goods. Agents in this economy act selfishly: Individuals maximize their well-being (utility), and firms maximize their profits.

General equilibrium theory tells us that, thanks to the action of the price system, order will emerge out of this uncoordinated chaos, provided certain conditions are satisfied. In the main, these hypotheses (conditions) are as follows:

- H1: *Complete markets.* There exists a market on which a price is established for each of the n goods valued by consumers.
- H2: *Perfect competition.* The number of consumers and firms (i.e., demanders and suppliers of each of the n goods in each of the n markets) is large enough so that no agent is in a position to influence (manipulate) market prices; i.e., all agents take prices as given.
- H3: Consumers' preferences are convex.
- H4: Firms' production sets are convex as well.

H3 and H4 are technical conditions with economic implications. Somewhat paradoxically, the convexity hypothesis for consumers' preferences approximately translates into strictly concave utility functions. In particular, H3 is satisfied (in substance) if consumers display risk aversion, an assumption crucial for understanding financial markets, and one that will be made throughout this text. As already noted (Box 1.2), risk aversion translates into strictly concave utility functions (see Chapter 4 for details). H4 imposes requirements on

BOX 1.2 Representing Risk Aversion

Let us reinterpret the two-date consumption stream (c_1, c_2) of Box 1.1 as the consumption levels attained "Then" or "Tomorrow" in two alternative, equally likely, states of the world. The desire for a smooth consumption stream across the two states, which we associate with risk aversion, is obviously represented by the same inequality

$$U(4) > \frac{1}{2} U(3) + \frac{1}{2} U(5)$$

and it implies the same general shape for the utility function. In other words, assuming plausibly that decision makers are **risk averse**, an assumption in conformity with most of financial theory, implies that the utility functions used to represent agents' preferences are **strictly concave**.

the production technology. It specifically rules out increasing returns to scale in production. Although important, this assumption is nevertheless not at the heart of things in financial economics since for the most part we will abstract from the production side of the economy.

A **general competitive equilibrium** is a price vector p^* and an allocation of resources, resulting from the independent decisions of consumers and producers to buy or sell each of the n goods in each of the n markets, such that, at the equilibrium price vector p^* , supply equals demand in all markets simultaneously and the action of each agent is the most favorable to him or her among all those he can afford (technologically or in terms of his budget computed at equilibrium prices).

A **Pareto optimum** is an allocation of resources, however determined, where it is impossible to redistribute resources (i.e., to go ahead with further exchanges) without reducing the welfare of at least one agent. In a Pareto-efficient (or Pareto optimal—we will use the two terminologies interchangeably) allocation of resources, it is thus not possible to make someone better off without making someone else worse off. Such a situation may not be just or fair, but it is certainly efficient in the sense of avoiding waste.

Omitting some purely technical conditions, the main results of general equilibrium theory can be summarized as follows:

1. *The existence of a competitive equilibrium:* Under H1 through H4, a competitive equilibrium is guaranteed to exist. This means that there indeed exists a price vector and an allocation of resources satisfying the definition of a competitive equilibrium as stated above.
2. *First welfare theorem:* Under H1 and H2, a competitive equilibrium, if it exists, is a Pareto optimum.
3. *Second welfare theorem:* Under H1 through H4, any Pareto-efficient allocation can be decentralized as a competitive equilibrium.

The second welfare theorem asserts that, for any arbitrary Pareto-efficient allocation, there is a price vector and a set of initial endowments such that this allocation can be achieved as a result of the free interaction of maximizing consumers and producers interacting in competitive markets. To achieve a specific Pareto optimal allocation, some redistribution mechanism will be needed to reshuffle initial resources. The availability of such a mechanism, functioning without distortion (and thus waste), is, however, very much in question. Hence the dilemma between equity and efficiency that faces all societies and their governments.

The necessity of H1 and H2 for the optimality of a competitive equilibrium provides a rationale for government intervention when these hypotheses are not naturally satisfied. The case for antitrust and other “pro-competition” policies is implicit in H2; the case for

intervention in the presence of externalities or in the provision of public goods follows from H1, because these two situations are instances of missing markets.¹²

Note that so far there does not seem to be any role for financial markets in promoting an efficient allocation of resources. To restore that role, we must abandon the fiction of a timeless world, underscoring, once again, the fact that time is of the essence in finance! Introducing the time dimension does not diminish the usefulness of the general equilibrium apparatus presented above, provided the definition of a good is properly adjusted to take into account not only its intrinsic characteristics, but also the time period in which it is available. A cup of coffee available at date t is different from a cup of coffee available at date $t + 1$, and, accordingly, it is traded on a different market and it commands a different price. Thus, if there are two dates, the number of goods in the economy goes from n to $2n$.

It is easy to show, however, that not all commodities need be traded for future as well as current delivery. The existence of a spot and forward market for *one good only* (taken as the numeraire) is sufficient to implement all the desirable allocations, and, in particular, restore, under H1 and H2, the optimality of the competitive equilibrium. This result is contained in [Arrow \(1964\)](#). It provides a powerful economic rationale for the existence of credit markets, markets where money is traded for future delivery.

Now let us go one step further and introduce uncertainty, which we will represent conceptually as a partition of all the relevant future scenarios into separate *states of nature*. To review, a state of nature is an exhaustive description of one possible relevant configuration of future events. Using this concept, we can extend the applicability of the welfare theorems in a fashion similar to that used with time above, by defining goods according not only to the date but also to the state of nature at which they are (might be) available. This is the notion of contingent commodities. Under this construct, we imagine the market for ice cream decomposed into a series of markets: for ice cream today, ice cream tomorrow if it rains and the Dow Jones is at 10,000; if it rains and so on. Formally, this is a straightforward extension of the basic context: there are more goods, but this is not in itself restrictive¹³ ([Arrow, 1964](#); [Debreu, 1959](#)).

¹² Our model of equilibrium presumes that agents affect one another only through prices. If this is not the case, an economic externality is said to be present. These may involve either production or consumption. For example, there have been substantial negative externalities for fishermen associated with the construction of dams in the western United States: the catch of salmon has declined dramatically as these dams have reduced the ability of the fish to return to their spawning habitats. If the externality affects all consumers simultaneously, it is said to be a public good. The classic example is national defense. If any citizen is to consume a given level of national security, all citizens must be equally secure (and thus consume this public good at the same level). Both are instances of missing markets. Neither is there a market for national defense nor for rights to disturb salmon habitats.

¹³ In this context n can be as large as one needs without restriction.

The hypothesis that there exists a market for each and every good valued by consumers becomes, however, much more questionable with this extended definition of a typical good, as the example above suggests. On the one hand, the number of states of nature is, in principle, arbitrarily large and, on the other, one simply does not observe markets where commodities contingent on the realization of individual states of nature can routinely be traded. One can thus state that *if* markets are complete in the above sense, a competitive equilibrium is efficient, but the issue of completeness (H1) then takes center stage. Can Pareto optimality be obtained in a less formidable setup than one where there are complete contingent commodity markets? What does it mean to make markets “more complete”?

It was [Arrow \(1964\)](#), again, who took the first step toward answering these questions. Arrow generalized the result alluded to earlier and showed that it would be enough, in order to effect all desirable allocations, to have the opportunity to trade one good only across all states of nature. Such a good would again serve as the numeraire. The primitive security could thus be a claim promising \$1.00 (i.e., one unit of the numeraire) at a future date, contingent on the realization of a particular state, and zero under all other circumstances. We shall have a lot to say about such *Arrow–Debreu securities* (henceforth A–D securities), which are also called *contingent claims*. Arrow asserted that if there is one such contingent claim corresponding to each and every one of the relevant future date/state configurations, hypothesis H1 could be considered satisfied, markets could be considered complete, and the welfare theorems would apply. Arrow’s result implies a substantial decrease in the number of required markets.¹⁴ However, for a complete contingent claim structure to be fully equivalent to a setup where agents could trade a complete set of contingent commodities, it must be the case that agents are assumed to know all future spot prices, contingent on the realization of all individual states of the world. Indeed, it is at these prices that they will be able to exchange the proceeds from their A–D securities for consumption goods. This hypothesis is akin to the hypothesis of rational expectations.¹⁵

A–D securities are a powerful conceptual tool and are studied in depth in Chapters 9 and 11. They are not, however, the instruments we observe being traded in actual markets. Why is this the case, and in what sense is what we do observe an adequate substitute? To answer these questions, we first allude to a result (derived later on) which states that there is no single way to make markets complete. In fact, potentially a large number of alternative financial structures may achieve the same goal, and the complete A–D securities structure is only one of them. For instance, we shall describe, in Chapter 11, a context in which one might think of achieving an essentially complete market structure with options or derivative securities. We shall make use of this fact for pricing alternative instruments using arbitrage

¹⁴ Example: 2 dates, 3 basic goods, 4 states of nature: complete commodity markets require 12 contingent commodity markets plus 3 spot markets versus 4 contingent claims and 2×3 spot markets in the Arrow setup.

¹⁵ For an elaboration on this topic, see [Drèze \(1971\)](#).

techniques. Thus, the failure to observe anything close to A–D securities being traded is not evidence against the possibility that markets are indeed complete.

In an attempt to match this discussion on the role played by financial markets with the type of markets we see in the real world, one can identify the different needs met by trading A–D securities in a complete markets world. In so doing, we shall conclude that, in reality, different needs are met trading alternative specialized financial instruments (which, as we shall later prove, will all appear as portfolios of A–D securities).

As we have already observed, the time dimension is crucial for finance, and, correspondingly, the need to exchange purchasing power across time is essential. It is met in reality through a variety of specific *noncontingent* instruments, which are promised future payments independent of specific states of nature, except those in which the issuer is unable to meet his obligations (bankruptcies). Personal loans, bank loans, money market and capital market instruments, social security, and pension claims are all assets fulfilling this basic need for redistributing purchasing power in the time dimension. In a complete market setup implemented through A–D securities, the needs met by these instruments would be satisfied by a certain configuration of positions in A–D securities. In reality, the specialized instruments mentioned above fulfill the demand for exchanging income through time.

One reason for the formidable nature of the complete markets requirement is that a state of nature, which is a complete description of the relevant future for a particular agent, includes some purely personal aspects of almost unlimited complexity. Certainly the future is different for you, in a relevant way, if you lose your job, or if your house burns, without these contingencies playing a very significant role for the population at large. In a pure A–D world, the description of the states of nature should take account of these *individual contingencies* viewed from the perspective of each and every market participant! In the real world, insurance contracts are the specific instruments that deal with the need for exchanging income across purely individual events or states. The markets for these contracts are part and parcel of the notion of complete financial markets. Although such a specialization makes sense, it is recognized as unlikely that the need to trade across individual contingencies will be fully met through insurance markets because of specific difficulties linked with the hidden quality of these contingencies (i.e., the inherent asymmetry in the information possessed by suppliers and demanders participating in these markets). The presence of these asymmetries strengthens our perception of the impracticality of relying exclusively on pure A–D securities to deal with personal contingencies.

Beyond time issues and personal contingencies, most other financial instruments not only imply the exchange of purchasing power through time, but are also more specifically contingent on the realization of particular events. The relevant events here, however, are

defined on a collective basis rather than being based on individual contingencies; they are contingent on the realization of events affecting groups of individuals and observable by everyone. An example is the situation where a certain level of profits for a firm implies the payment of a certain dividend against the ownership of that firm's equity. Another is the payment of a certain sum of money associated with the ownership of an option or a financial futures. In the later cases, the contingencies (sets of states of nature) are dependent on the value of the underlying asset itself.

1.6 Financial Intermediation and the Business Cycle

Business cycles are the mark of all developed economies. According to much of current research, they are in part the result of external shocks with which these economies are repeatedly confronted. The depth and amplitude of these fluctuations, however, may well be affected by some characteristics of the financial system. This is at least the import of the recent literature on the financial accelerator. The mechanisms at work here are numerous, and we limit ourselves to giving the reader a flavor of the discussion.

The financial accelerator is manifest most straightforwardly in the context of monetary policy implementation. Suppose the monetary authority wishes to reduce the level of economic activity (inflation is feared) by raising real interest rates. The primary effect of such a move will be to increase firms' cost of capital and, as a result, to induce a decrease in investment spending as marginal projects are eliminated from consideration.

According to the financial accelerator theory, however, there may be further, substantial, secondary effects. In particular, the interest rate rise will reduce the value of firms' collateralizable assets. For some firms, this reduction may significantly diminish their access to credit, making them credit constrained. As a result, the fall in investment may exceed the direct impact of the higher cost of capital; tighter financial constraints may also affect input purchases or the financing of an adequate level of finished goods inventories. For all these reasons, the output and investment of credit-constrained firms will be more strongly affected by the action of the monetary authorities, and the economic downturn may be made correspondingly more severe. By this same mechanism, any economywide reduction in asset values may have the effect of reducing economic activity under the financial accelerator.

Which firms are most likely to be credit constrained? We would expect that small firms, those for which lenders have relatively little information about the long-term prospects, would be principally affected. These are the firms from which lenders demand high levels of collateral. [Bernanke et al. \(1996\)](#) provide empirical support for this assertion using US data from small manufacturing firms.

The financial accelerator has the power to make an economic downturn, of whatever origin, more severe. If the screening and monitoring functions of the financial system can be tailored more closely to individual firm needs, lenders will need to rely to a lesser extent on collateralized loan contracts. This would diminish the adverse consequences of the financial accelerator and perhaps the severity of business cycle downturns.

1.7 Financial Crises

A more radical version of the link between the financial markets and the business cycle is present in the experience of a financial crisis. It reflects the notion of a financial crisis as either a catalyst for, or the initial cause of, a severe and prolonged business cycle downturn.¹⁶ The Great Depression of the early 1930s and the recent “Great Recession” of 2007–2009 are only the most dramatic cases in point. Both were associated with a large decline in output, a dramatic decline in investment, and a large increase in the number of persons unemployed.

Financial crises of this magnitude are typically preceded by the ending of a price “bubble” in some important asset type, with ensuing price declines in it and many other asset categories. (The “Great Recession” in the United States, in particular, was preceded by the ending of a residential real estate price bubble chiefly in the states of the Southwest, California, Texas, and Florida. Since these homes had been purchased with large mortgages and very little equity, the prospect of numerous defaults immediately arose.) With bank assets (mortgages or MBSs in the Great Recession case) declining in value, many banks face the very real possibility of insolvency, and some collapse (Lehman Brothers).¹⁷ In either case, banks quickly become reluctant to extend any further risky loans for fear of making their own financial situations even worse if the loans cannot be repaid. As a consequence, a “credit crunch” ensues whereby many firms, and especially small and medium sized ones, find it essentially impossible to obtain loans to fund their investments and continuing operations. See [Figure 1.3](#) for some sense of the drop in US lending activity in the first years of the Great Recession. Aggregate investment spending “dries up.” Note that the resulting economic contraction parallels the one associated with the financial accelerator of [Section 1.6](#), except that it is typically more immediate and more intense. In effect, the financial system ceases to perform the functions assigned to it by society as described in [Section 1.4](#). The aforementioned drop in investment spending is often accompanied by a reduction in consumption spending as households react to their reduced wealth resulting from the decline in asset prices. Output contracts further.

¹⁶ For a detailed historical analysis, see [Reinhart and Rogoff \(2009\)](#).

¹⁷ Lehman Brothers was not technically a (commercial) bank because it was not permitted to take deposits. Neither did it have the right to receive loans from the US Federal Reserve, the lender of last resort. Lehman Brothers funded its assets (real estate, MBSs) with very short-term loans that had to be rolled over daily.

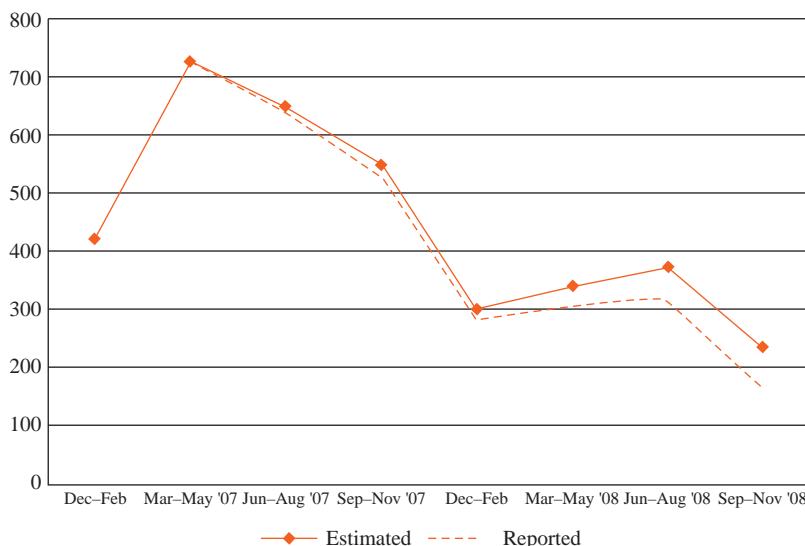


Figure 1.3: Total loan issuance, US Corporate Loans

Compiled from DealScan database of loan originals. Reported corresponds to loans reported in DealScan as of December 1, 2008. Source: [Ivashina and Scharfstein \(2008\)](#), Figure 1.

The financial crisis of 2007–2009 is estimated to have cost the US economy 22 trillion dollars, not only in the form of declines in asset values but also in the form of lost output.¹⁸ Worldwide losses have been estimated to be not less than 1 year’s world output, and as much as five times larger ([Haldane \(2010\)](#)). The cost of a malfunctioning financial services industry is clearly very great and, as of this writing, the world economy has yet to recover fully from the financial crisis of 2007–2009.¹⁹

The “Great Recession” was also the culmination of a large and rapid expansion of financial services along a number of dimensions. As a percentage of US GDP, the financial services sector rose from 4.9% in 1980 to 8.3% in 2006. The value of all United States issued private financial claims (stocks, bonds, etc.) rose from five times US GDP in 1980 to 10 times US GDP in 2007 ([Greenwood and Scharfstein, 2013](#)), a phenomenon observed in other well developed countries (see [Figure 1.4](#)). The provision of household credit similarly rose from an aggregate value equal to 0.48 GDP in 1980 to 0.99 GDP in 2007.

This enormous expansion in financial activity, in conjunction with its subsequent collapse, has led some to question whether the finance industry has grown too large. [Pagano \(2012\)](#), in particular, makes a strong case in this regard. For a large sample of countries,

¹⁸ Source: [General Accounting Office Report #GAO-13-180](#), “Financial Regulatory Reform: Financial Crisis Losses and Potential Impacts of the Dodd-Frank Act,” January 16, 2013.

¹⁹ We give a detailed overview of the 2007 financial crisis in Web Chapter D.

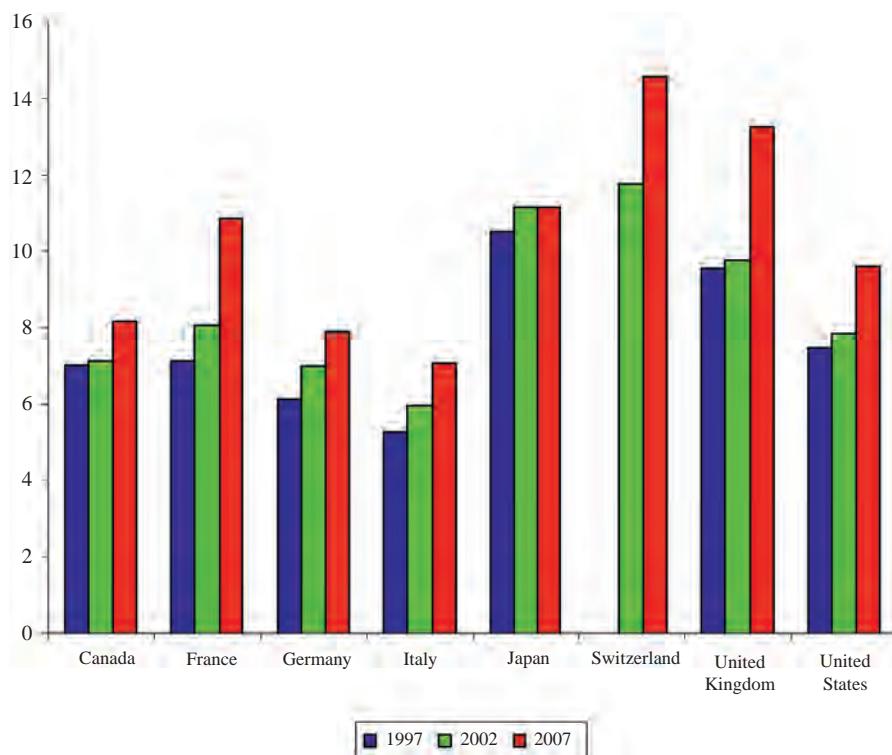


Figure 1.4: Financial deepening, advanced economies²¹
Ratio of total financial assets to GDP. Source: *Financial Accounts Statistics, OECD and Eurostat*.

he measures a country's degree of financial development either by the ratio of outstanding private credit to GDP or the ratio of aggregate stock market value to GDP (in both cases averages for the 1980–1995 period). Using these measures, he demonstrates a strong positive association between the growth in value added for industries that are highly dependent on external finance and either measure of financial development in non-OECD countries. For OECD countries, however, the association is small and not statistically different from zero. These results suggest that for countries in the earlier stages of their economic development, an expansion of the financial industry can enhance economic growth, but for countries with well-developed financial markets, this is no longer the case. [Pagano \(2012\)](#) concludes: “Beyond a certain point, financial development does not appear to contribute significantly to economic activity.”²⁰

²⁰ [Pagano \(2012\)](#), page 3.

²¹ This figure is taken from [Milesi-Ferretti and Tille \(2010\)](#).

He then goes on to study the relationship of bank credit-worthiness (as captured by a special index, which he constructs) and the ratio of private credit issued by deposit banks and other financial institutions to GDP as the measure of financial development. For developing countries where the private credit to GDP ratio is less than 50%, he finds a positive association (correlation) between these quantities; for developed countries with private credit to GDP ratios above 50%, however, the correlation turns negative. It becomes more negative for those countries where private credit as a fraction of GDP exceeds 100%, which is the case for the United States and the United Kingdom. While not formally conclusive, the [Pagano \(2012\)](#) results do suggest that in some countries the financial industry may have grown so large that it now has “a life of its own,” one with self interests that potentially compromise its primary role of matching savers to investors in an efficient way.

1.8 Conclusion

To conclude this introductory chapter, we advance a vision of the financial system progressively evolving toward the complete markets paradigm, starting with the most obviously missing markets and slowly, as technological innovation decreases transaction costs and allows the design of more sophisticated contracts, completing the market structure. Have we arrived at a complete market structure? Have we come significantly closer? There are opposing views on this issue. While a more optimistic perspective is proposed by [Merton \(1990\)](#) and [Allen and Gale \(1994\)](#), we choose to close this chapter on two healthily skeptical notes. [Tobin \(1984, p. 10\)](#), for one, provides an unambiguous answer to the above question:

New financial markets and instruments have proliferated over the last decade, and it might be thought that the enlarged menu now spans more states of nature and moves us closer to the Arrow–Debreu ideal. Not much closer, I am afraid. The new options and futures contracts do not stretch very far into the future. They serve mainly to allow greater leverage to short-term speculators and arbitrageurs, and to limit losses in one direction or the other. Collectively they contain considerable redundancy. Every financial market absorbs private resources to operate, and government resources to police. The country cannot afford all the markets the enthusiasts may dream up. In deciding whether to approve proposed contracts for trading, the authorities should consider whether they really fill gaps in the menu and enlarge the opportunities for Arrow–Debreu insurance, not just opportunities for speculation and financial arbitrage.

[Shiller \(1993, pp. 2–3\)](#) is even more specific with respect to missing markets:

It is odd that there appear to have been no practical proposals for establishing a set of markets to hedge the biggest risks to standards of living. Individuals and organizations could hedge or insure themselves against risks to their standards of living if an array of

risk markets—let us call them macro markets—could be established. These would be large international markets, securities, futures, options, swaps or analogous markets, for claims on major components of incomes (including service flows) shared by many people or organizations. The settlements in these markets could be based on income aggregates, such as national income or components thereof, such as occupational incomes, or prices that value income flows, such as real estate prices, which are prices of claims on real estate service flows.

References

- Aghion, P., Fally, T., Scarpetta, S., 2007. Credit constraints as a barrier to entry and post-entry growth of firms. *Econ. Policy*. 22, 731–779.
- Allen, F., Gale, D., 1994. Financial Innovation and Risk Sharing. MIT Press, Cambridge, MA.
- Arrow, K.J., 1964. The role of securities in the allocation of risk. *Rev. Econ. Stud.* 31, 91–96.
- Barro, R.J., Sala-i-Martin, X., 1995. Economic Growth. McGraw-Hill, New York, NY.
- Bernanke, B., Gertler, M., Gilchrist, S., 1996. The financial accelerator and the flight to quality. *Rev. Econ. Stat.* 78, 1–15.
- Bernstein, P.L., 1992. Capital Ideas. The Improbable Origins of Modern Wall Street. The Free Press, New York, NY.
- Debreu, G., 1959. Theory of Value: An Axiomatic Analysis of Economic Equilibrium. John Wiley & Sons, New York, NY.
- Drèze, J.H., 1971. Market Allocation Under Uncertainty. *Eur. Econ. Rev.* 2, 133–165.
- General Accounting Office Report #GAO-13-180, Financial Regulatory Reform: Financial Crisis Losses and Potential Impacts of the Dodd-Frank Act, January 16, 2013.
- Greenwood, R., Scharfstein, D., 2013. The growth of finance. *J. Econ. Perspect.* 27, 3–28.
- Haldane, A., 2010. The \$100 Billion Question. Comments given at the Institute of Regulation and Risk, Hong Kong, March 30. <<http://bankofengland.co.uk/publications/documents/speeches/2010/speech433.pdf>>.
- Ivashina, V., Scharfstein, D., 2008. Bank lending during the financial crisis of 2008. Working paper, Harvard Business School. Published under the same title in *Journal of Financial Economics*, 2010 (97), 319–338.
- Jappelli, T., Pagano, M., 1994. Savings, growth, and liquidity constraints. *Q. J. Econ.* 109, 83–109.
- King, R., Levine, R., 1993a. Finance and growth: Schumpeter may be right. *Q. J. Econ.* 108, 713–737.
- King, R., Levine, R., 1993b. Finance, entrepreneurship and growth. *J. Monet. Econ.* 32, 513–542.
- Lee, I.H., Syed, M., Xueyau, L., 2012. Is China overinvesting and does it matter? IMF Working Paper WP/12/277.
- Levine, R., 1997. Financial development and economic growth: views and agenda. *J. Econ. Lit.* 35, 688–726.
- Merton, R.C., 1990. The financial system and economic performance. *J. Financ. Serv.* 4, 263–300.
- Milesi-Ferretti, G.-M., Tille, C., 2010. The great retrenchment: international capital flows during the global financial crisis. Working Paper, CEPR, Economic Policy.
- Pagano, M., 2012. Finance: economic lifeblood or toxin? Working Paper, CEPR and University of Naples Federico II.
- Rajan, R., Zingales, L., 1998. Financial dependence and growth. *Am. Econ. Rev.* 88, 559–587.
- Reinhart, C., Rogoff, K., 2009. This Time Is Different: Eight Centuries of Financial Folly. Princeton University Press, Princeton.
- Schumpeter, J., 1934. The Theory of Economic Development, Duncker & Humblot, Leipzig. Trans. R. Opie (1934), Harvard University Press, Cambridge, MA.
- Shiller, R.J., 1993. Macro Markets—Creating Institutions for Managing Society's Largest Economic Risks. Clarendon Press, Oxford.
- Solow, R.M., 1956. A contribution to the theory of economic growth. *Q. J. Econ.* 32, 65–94.
- Tobin, J., 1984. On the efficiency of the financial system. *Lloyds Bank Rev.* 1–15.

Complementary Readings

As a complement to this introductory chapter, the reader will be interested in the historical review of financial markets and institutions found in the first chapter of [Allen and Gale \(1994\)](#). [Bernstein \(1992\)](#) provides a lively account of the birth of the major ideas making up modern financial theory, including personal portraits of their authors.

Appendix: Introduction to General Equilibrium Theory

The goal of this appendix is to provide an introduction to the essentials of general equilibrium theory, thereby permitting a complete understanding of [Section 1.6](#) and facilitating the discussion of subsequent chapters (from Chapter 8 onward). To make this presentation as simple as possible, we will take the case of a hypothetical exchange economy (i.e., one with no production) with two goods and two agents. This permits using a very useful pedagogical tool known as the Edgeworth–Bowley box.

Let us analyze the problem of allocating efficiently a given economywide endowment of 10 units of good 1 and 6 units of good 2 among two agents, A and B. In [Figure A1.1](#), we measure good 2 on the vertical axis and good 1 on the horizontal axis. Consider the choice problem from the origin of the axes for Mr. A, and upside down (i.e., placing the origin in the upper right corner), for Ms. B. An allocation is then represented as a point in a rectangle of size 6×10 . Point E is an allocation at which Mr. A receives 4 units of

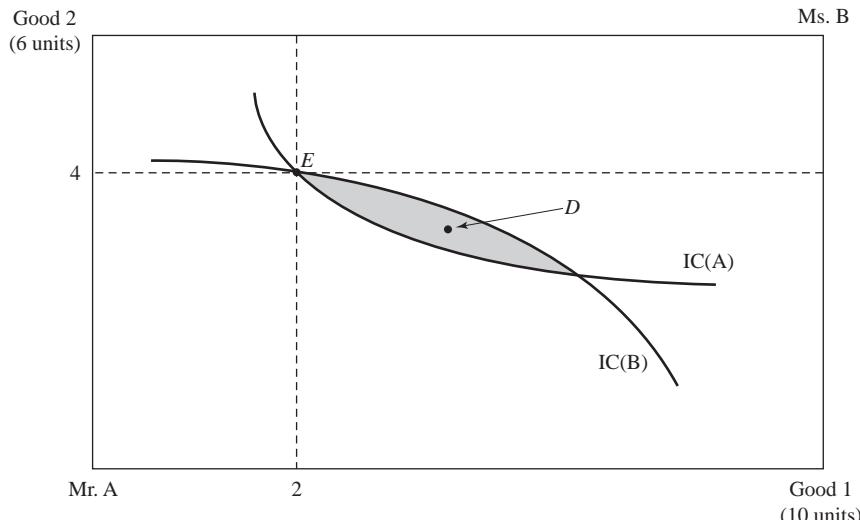


Figure A1.1
The Edgeworth–Bowley box: the set of Pareto superior allocations.

good 2 and 2 units of good 1. Ms. B gets the rest, 2 units of good 2 and 8 units of good 1. All other points in the box represent feasible allocations, i.e., alternative ways of allocating the resources available in this economy.

Pareto Optimal Allocations

In order to discuss the notion of Pareto optimal or efficient allocations, we need to introduce agents' preferences. They are fully summarized, in the graphical context of the Edgeworth–Bowley box, by indifference curves (IC) or utility level curves. Starting from the allocation E represented in [Figure A1.1](#), we can thus record all feasible allocations that provide the same utility to Mr. A. The precise shape of such a level curve is person specific, but we can at least be confident that it slopes downward. If we take away some units of good 1, we have to compensate him with some extra units of good 2 if we are to leave his utility level unchanged. It is easy to see as well that the ICs of a consistent person do not cross, a property associated with the notion of transitivity (and with rationality) in Chapter 3. And we have seen in [Boxes 1.1 and 1.2](#) that the preference for smoothness translates into a strictly concave utility function, or, equivalently, convex-to-the-origin level curves as drawn in [Figure A1.1](#). The same properties apply to the IC of Ms. B, of course viewed upside down with the upper right corner as the origin.

With this simple apparatus we are in a position to discuss further the concept of Pareto optimality. Arbitrarily tracing the level curves of Mr. A and Ms. B as they pass through allocation E (but in conformity with the properties derived in the previous paragraph), only two possibilities may arise: they cross each other at E or they are tangent to one another at point E. The first possibility is illustrated in [Figure A1.1](#), the second in [Figure A1.2](#). In the first case, allocation E cannot be a Pareto optimal allocation. As the picture illustrates clearly, by the very

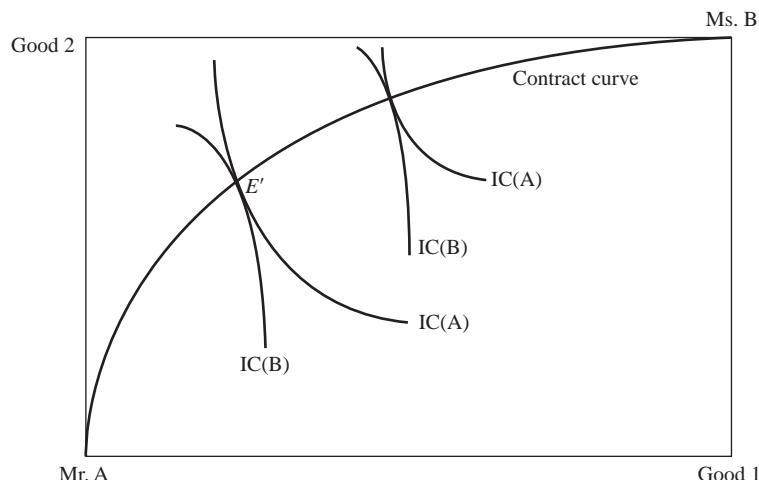


Figure A1.2
The Edgeworth–Bowley box: the contract curve.

definition of level curves, if the ICs of our two agents cross at point E, there is a set of allocations (corresponding to the shaded area in [Figure A1.1](#)) that both Mr. A and Ms. B simultaneously prefer to E. These allocations are Pareto superior to E, and, in that situation, it would indeed be socially inefficient or wasteful to distribute the available resources as indicated by E. Allocation D, for instance, is feasible and preferred to E by both individuals.

If the ICs are tangent to one another at point E' as in [Figure A1.2](#), no redistribution of the given resources exists that would be approved by both agents. Inevitably, moving away from E' decreases the utility level of one of the two agents if it favors the other. In this case, E' is a Pareto optimal allocation. [Figure A1.2](#) illustrates that it is not generally unique, however. If we connect all the points where the various ICs of our two agents are tangent to each other, we draw the line, labeled the contract curve, representing the infinity of Pareto optimal allocations in this simple economy.

An indifference curve for Mr. A is defined as the set of allocations that provide the same utility to Mr. A as some specific allocation; e.g., allocation E: $\{(c_1^A, c_2^A) : U(c_1^A, c_2^A) = U(E)\}$. This definition implies that the slope of the IC can be derived by taking the total differential of $U(c_1^A, c_2^A)$ and equating it to zero (no change in utility along the IC), which gives:

$$\frac{\partial U(c_1^A, c_2^A)}{\partial c_1^A} dc_1^A + \frac{\partial U(c_1^A, c_2^A)}{\partial c_2^A} dc_2^A = 0 \quad (1.3)$$

and thus

$$-\frac{dc_2^A}{dc_1^A} = \frac{\frac{\partial U(c_1^A, c_2^A)}{\partial c_1^A}}{\frac{\partial U(c_1^A, c_2^A)}{\partial c_2^A}} \equiv MRS_{1,2}^A \quad (1.4)$$

That is, the negative (or the absolute value) of the slope of the IC is the ratio of the marginal utility of good 1 to the marginal utility of good 2 specific to Mr. A and to the allocation (c_1^A, c_2^A) at which the derivatives are taken. It defines Mr. A's marginal rate of substitution (MRS) between the two goods.

[Equation \(1.4\)](#) permits a formal characterization of a Pareto optimal allocation. Our former discussion has equated Pareto optimality with the tangency of the ICs of Mr. A and Ms. B. Tangency, in turn, means that the slopes of the respective ICs are identical. Allocation E, associated with the consumption vector $(c_1^A, c_2^A)^E$ for Mr. A and $(c_1^B, c_2^B)^E$ for Ms. B, is thus Pareto optimal if, and only if,

$$MRS_{1,2}^A = \frac{\frac{\partial U(c_1^A, c_2^A)^E}{\partial c_1^A}}{\frac{\partial U(c_1^A, c_2^A)^E}{\partial c_2^A}} = \frac{\frac{\partial U(c_1^B, c_2^B)^E}{\partial c_1^B}}{\frac{\partial U(c_1^B, c_2^B)^E}{\partial c_2^B}} = MRS_{1,2}^B \quad (1.5)$$

Equation (1.5) provides a complete characterization of a Pareto optimal allocation in an exchange economy except in the case of a corner allocation, i.e., an allocation at the frontier of the box where one of the agents receives the entire endowment of one good and the other agent receives none. In that situation, it may well be that the equality could not be satisfied except, hypothetically, by moving to the outside of the box, i.e., to allocations that are not feasible since they require giving a negative amount of one good to one of the two agents.

So far we have not touched on the issue of how the discussed allocations may be determined. This is the viewpoint of Pareto optimality, which analysis is exclusively concerned with deriving efficiency properties of given allocations, regardless of how they were achieved. Let us now turn to the concept of competitive equilibrium.

Competitive Equilibrium

Associated with the notion of competitive equilibrium is the notion of markets and prices. One price vector (one price for each of our two goods), or simply a relative price taking good 1 as the numeraire, and setting $p_1 = 1$, is represented in the Edgeworth–Bowley box by a downward-sloping line. From the viewpoint of either agent, such a line has all the properties of the budget line. It also represents the frontier of their opportunity set. Let us assume that the initial allocation, before any trade, is represented by point I in Figure A1.3. Any line sloping downward from I does represent the set of allocations that Mr. A, endowed with I, can obtain by going to the market and exchanging (competitively, taking prices as given)

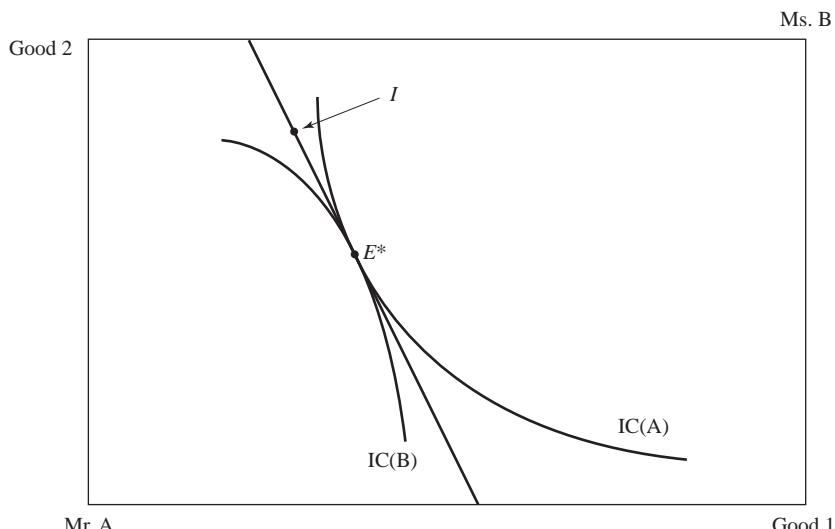


Figure A1.3
The Edgeworth–Bowley box: equilibrium achieved at E^* .

good 1 for 2 or vice versa. He will maximize his utility subject to this budget constraint by attempting to climb to the highest IC making contact with his budget set. This will lead him to select the allocation corresponding to the tangency point between one of his ICs and the price line. Because the same prices are valid for both agents, an identical procedure, viewed upside down from the upper right-hand corner of the box, will lead Ms. B to a tangency point between one of her ICs and the price line. At this stage, only two possibilities may arise: Mr. A and Ms. B have converged to the same allocation (the two markets, for good 1 and 2, clear—supply and demand for the two goods are equal and we are at a competitive equilibrium); or the two agents' separate optimizing procedures have led them to select two different allocations. Total demand does not equal total supply, and an equilibrium is not achieved. The two situations are described, respectively, in [Figures A1.3](#) and [A1.4](#).

In the disequilibrium case of [Figure A1.4](#), prices will have to adjust until an equilibrium is found. Specifically, with Mr. A at point A and Ms. B at point B, there is an excess demand for good 2 but insufficient demand for good 1. One would expect the price of 2 to increase relative to the price of good 1 with the likely result that both agents will decrease their net demand for 2 and increase their net demand for 1. Graphically, this is depicted by the price curve tilting with point I as the axis and looking less steep (indicating, for instance, that if both agents wanted to buy good 1 only, they could now afford more of it). With regular ICs, the respective points of tangencies will converge until an equilibrium similar to the one described in [Figure A1.3](#) is reached.

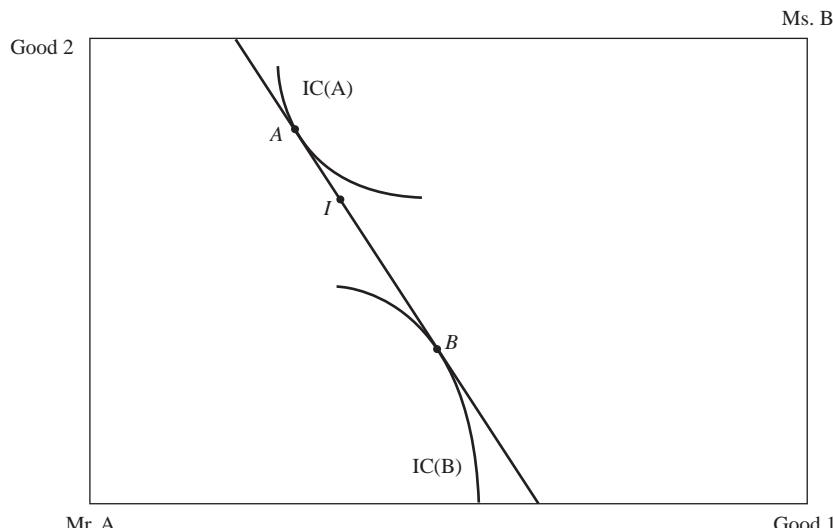


Figure A1.4

The Edgeworth–Bowley box: disequilibrium, excess demand for good 1.

We will not say anything here about the conditions guaranteeing that such a process will converge. Let us rather insist on one crucial necessary precondition: that an equilibrium exists. In the text we have mentioned that assumptions H1 to H4 are needed to guarantee the existence of an equilibrium. Of course, H4 does not apply here. H1 states the necessity of the existence of a price for each good, which is akin to specifying the existence of a price line. H2 defines one of the characteristics of a competitive equilibrium: that prices are taken as given by the various agents and the price line describes their perceived opportunity sets. Our discussion here can enlighten the need for H3. Indeed, in order for an equilibrium to have a chance to exist, the geometry of [Figure A1.3](#) makes clear that the shape of the two agents' ICs is relevant. The price line must be able to separate the "better than" areas of the two agents' ICs passing through a same point—the candidate equilibrium allocation. The better than area is simply the area above a given IC. It represents all the allocations providing higher utility than those on the level curve. This separation by a price line is not generally possible if the ICs are not convex, in which case an equilibrium cannot be guaranteed to exist. The problem is illustrated in [Figure A1.5](#).

Once a competitive equilibrium is observed to exist, which logically could be the case even if the conditions that guarantee existence are not met, the Pareto optimality of the resulting allocation is ensured by H1 and H2 only. In substance this is because once the common price line at which markets clear exists, the very fact that agents optimize taking prices as given leads them to a point of tangency between their highest IC and the common price line. At the resulting allocation, both MRS are equal to the same price line and, consequently, are identical. The conditions for Pareto optimality are thus fulfilled.

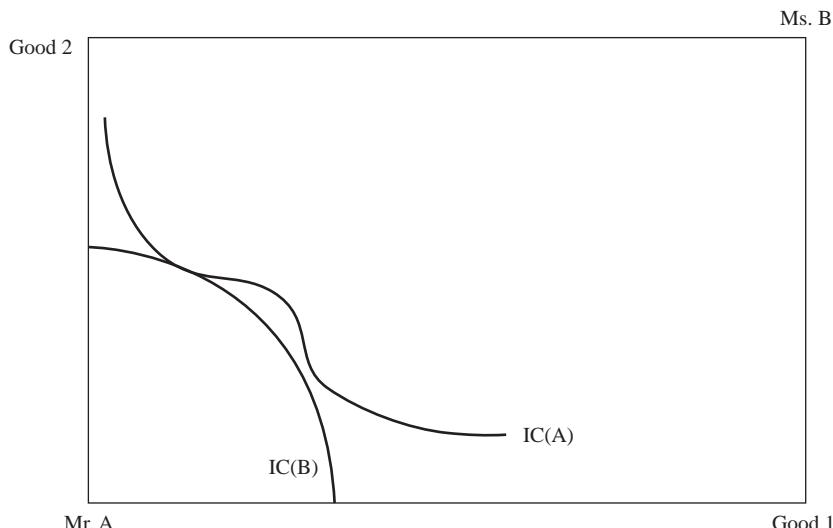


Figure A1.5
The Edgeworth—Bowley box: nonconvex indifference curves.

The Challenges of Asset Pricing: A Road Map

Chapter Outline

2.1 The Main Question of Financial Theory	31
2.2 Discounting Risky Cash Flows: Various Lines of Attack	33
2.3 Two Main Perspectives: Equilibrium versus Arbitrage	35
2.4 Decomposing Risk Premia	37
2.5 Models and Stylized Facts	39
2.5.1 The Equity Premium	40
2.5.2 The Value Premium	42
2.5.3 The Term Structure	43
2.6 Asset Pricing Is Not All of Finance!	44
2.6.1 Corporate Finance	44
2.6.2 Capital Structure	45
2.6.3 Taxes and Capital Structure	46
2.6.4 Capital Structure and Agency Costs	48
2.6.5 The Pecking Order Theory of Investment Financing	49
2.7 Banks	49
2.8 Conclusions	51
References	51

2.1 The Main Question of Financial Theory

Valuing risky cash flows or, equivalently, pricing risky assets is at the heart of financial theory.

Our discussion thus far has been conducted from the perspective of society as a whole, and it argues that a progressively more complete set of financial markets will generally enhance societal welfare by making it easier for economic agents to transfer income across future dates and states via the sale or purchase of individually tailored portfolios of securities. The desire of agents to construct such portfolios will depend as much upon the market prices of the relevant securities as on their strict availability, and this leads us to the main topic of the text.

Indeed, the major practical question in finance is, “How do we value a risky cash flow?” and the main objective of this text is to provide a complete and up-to-date treatment of how it can be answered. For the most part, this book is thus a text on asset pricing. Indeed, an asset is nothing else than the right to future cash flows, whether these future cash flows are the result of interest payments, dividend payments, insurance payments, or the resale value of the asset. Furthermore, when we compute a project’s risk-adjusted present value (PV), we are, in effect, asking the question: If this project’s cash flow were traded as though it were a security, at what price would it sell given that it should pay the prevailing rate on other securities with the same risk level? We compare its fair market value, estimated in this way, with its cost, P_0 . Evaluating a project is thus a special case of evaluating a security.

Viewed in this way and abstracting from risk for the moment, the key object of our attention, be it an asset or an investment project, can be summarized as in [Table 2.1](#).

In Table 2.1, $t = 0, 1, 2, \dots, \tau, \dots, T$ represents future dates. The duration of each period, the length of time between $\tau - 1$ and τ , is arbitrary and can be viewed as 1 day, 1 month, 1 quarter, or 1 year. The expression $\tilde{C}F_\tau$ stands for the possibly uncertain cash flows in period τ (whenever useful, we will identify random variables with a tilde), r_τ^f is the risk-free, per-period interest rate prevailing between date 0 and τ , and P_0 denotes the to-be-determined current price or valuation of the future cash flow. If the future cash flows will be available for sure, valuing the flow of future payments is easy. It requires adding the future cash flows after discounting them by the risk-free rate of interest, i.e., adding the cells in the last line of the table. The discounting procedure is indeed at the heart of our problem: it clearly serves to translate future payments into current dollars (those that are to be used to purchase the right to these future cash flows or in terms of which the current value of the future cash flow is to be expressed). In other words, the discounting procedure is what makes it possible to compare future dollars (i.e., dollars that will be available in the future) with current dollars.

If, however, the future cash flows will not be available for certain but are subject to random events—the interest payments depend on the debtor remaining solvent, the dividend payments depend on the financial strength of the equity issuer, the returns to the investment project depend on its commercial success—then the valuation question becomes trickier, so much so that there does not exist a universal way of proceeding that dominates all others.

In the same way that one dollar for sure tomorrow does not generally have the same value as one current dollar, one dollar tomorrow under a set of more or less narrowly defined

Table 2.1: Valuing a risk-free cash flow

$t = 0$ $P_0?$	$t = 1$ $\tilde{C}F_1$ $\frac{CF_1}{(1+r_1^f)}$	$t = 2$ $\tilde{C}F_2$ $\frac{CF_2}{(1+r_2^f)^2}$	$\dots \tau \dots$ $\tilde{C}F_\tau$ $\frac{CF_\tau}{(1+r_\tau^f)^\tau}$	$t = T$ $\tilde{C}F_T$ $\frac{CF_T}{(1+r_T^f)^T}$
-------------------	---	---	--	---

circumstances, i.e., in a subset of all possible states of nature, is also not worth even one current dollar discounted at the risk-free rate. Assume the risk-free rate of return is 5% per year, then discounting one dollar available in 1 year at the risk-free rate yields $(\$1/1.05) \cong \0.95 . This equality is exactly that: it states that \$1 tomorrow will have a market price of \$0.95 today when 1-year risk-free securities earn 5%. It is a market assessment to the extent that the 5% risk-free rate is an equilibrium market rate. Now if \$1 for sure tomorrow is worth \$0.95, it seems likely that \$1 tomorrow “possibly,” i.e., in a restricted subset of the states of nature, should certainly be worth less than \$0.95.

One can speculate, for instance, that if the probability of \$1 in a year is about $\frac{1}{2}$, then one should not be willing to pay more than $\frac{1}{2} \times \$0.95$ for that future cash flow. But we have to be more precise than this. To that end, several lines of attack will be pursued. Let us outline them.

2.2 Discounting Risky Cash Flows: Various Lines of Attack

First, as in the certainty case, it is plausible to argue (and it can be formally demonstrated) that the valuation process is additive: the value of a sum of future cash flows will take the form of the sum of the values of each of these future cash flows. Second, as already anticipated, we will work with probabilities, so that the random cash flow occurring at a future date τ will be represented by a random variable: $\tilde{C}F_\tau$, for which a natural reference value is its expectation $E\tilde{C}F_\tau$. Another reference value would be this expected future cash flow discounted at the risk-free rate: $E\tilde{C}F_\tau / ((1 + r_f^\tau)^\tau)$. Now the latter expression cannot generally be the solution to our problem, although it is intuitively understandable that it will be when the risk issue does not matter—i.e., when market participants can be assumed to be risk neutral. In the general case where risk must be taken into account, which typically means that risk-bearing behavior needs to be remunerated, alterations to that reference formula are necessary. These alterations may take any of the following forms:

1. The most common strategy consists of discounting at a rate that is higher than the risk-free rate, i.e., to discount at a rate that is the risk-free rate increased by a certain amount π (a risk premium) as in

$$\frac{E\tilde{C}F_\tau}{(1 + r_f^\tau + \pi_\tau)^\tau}$$

The underlying logic is straightforward: To price an asset equal to the present value of its expected future cash flows discounted at a particular rate is to price the asset in a manner such that, at its present value price, it is expected to earn that discount rate. The appropriate rate, in turn, must be the analyst’s estimate of the expected rate of return on other financial assets that represent title to cash flows similar in risk and timing to that

of the asset in question. This strategy has the consequence of pricing the asset to pay the prevailing competitive rate for its risk class. When we follow this approach, the key issue is to compute the appropriate risk premium.¹

2. Another approach, in the same spirit, consists of correcting the expected cash flow itself in such a way that one can continue discounting at the risk-free rate. The standard way of doing this is to decrease the expected future cash flow by a factor Π that once again will reflect some form of risk or insurance premium as in

$$\frac{E\tilde{C}F_\tau - \Pi_\tau}{(1+r_\tau^f)^\tau}$$

3. The same idea can take the form, it turns out quite fruitfully, of distorting the probability distribution over which the expectations operator is applied so that taking the expected cash flow with this modified probability distribution justifies once again discounting at the risk-free rate:

$$\frac{\hat{E}\tilde{C}F_\tau}{(1+r_\tau^f)^\tau}$$

Here \hat{E} denotes the expectation taken with respect to the modified probability distribution.

4. Finally, one can think of decomposing the future cash flow $\tilde{C}F_\tau$ into its state-by-state elements. Denote $(CF(\theta_\tau))$ the actual payment that will occur in the specific possible state of nature θ_τ . If one is able to find the price today of \$1 in period τ conditional on that particular state θ_τ being realized, say $q(\theta_\tau)$, then surely the appropriate current valuation of $\tilde{C}F_\tau$ is

$$\sum_{\theta_\tau \in \Theta_\tau} q(\theta_\tau)CF(\theta_\tau)$$

where the summation takes place over all the possible future states θ_τ . The quantity $q(\theta_\tau)$ is often referred to as a “state price,” and in important applications it will resemble the more traditional discount factor.

The procedures described above are alternative ways of attacking the difficult valuation problem we have outlined, but they can only be given content in conjunction with theories explaining how to compute the risk premia (cases 1 or 2), to identify the distorted

¹ Let us be sure we understand exactly what this expression says: the r_τ^f denotes the period by period rate of return on a default-free security which pays an amount of money τ periods in the future while π_τ is the return risk premium expected to prevail over this same time horizon. We thus discount each future cash flow “individually.” In a typical calculation where there are cash flows in many future periods, we frequently assume $r_\tau^f = r_f$ and $\pi_\tau = \pi$ for all τ ; i.e., the risk-free rate and risk premium are constant looking forward. All future cash flows thus end up being discounted at the same rate.

probability distribution (case 3) or to price future dollars state by state (case 4). For strategies 1 and 2, this can be done using the capital asset pricing model (CAPM), the consumption capital asset pricing model (CCAPM), or the arbitrage pricing theory (APT); strategy 3 is characteristic of the Martingale approach; strategy 4 describes the perspective of Arrow–Debreu (AD) pricing.

2.3 Two Main Perspectives: Equilibrium versus Arbitrage

There is another, even more fundamental way of classifying alternative valuation theories. All the known valuation theories cited above employ one of two main methodologies: the equilibrium approach or the arbitrage approach.

The traditional equilibrium approach consists of an analysis of the factors determining the supply and demand for the cash flow (asset) in question. The arbitrage approach attempts to value a cash flow on the basis of observations made on the values of the various elements making up that cash flow.

Let us illustrate this distinction with an analogy. You are interested in pricing a bicycle. There are two ways to approach the question. If you follow the equilibrium approach, you will want to study the determinants of supply and demand. Who are the producers? How many bicycles are they able to produce? What are the substitutes, including probably the existing stock of old bicycles potentially appearing on the second-hand market? After dealing with supply, turn to demand: Who are the buyers? What are the forecasts of the demand for bicycles? And so on. Finally, you will turn to the market structure. Is the market for bicycles competitive? If so, we know how the equilibrium price will emerge as a result of the matching process between demanders and suppliers. The equilibrium perspective is a sophisticated, complex approach, with a long tradition in economics, one that has also been applied in finance, at least since the 1950s. We will follow it in the first part of this book, adopting standard assumptions that simplify, without undue cost, the supply and demand analysis for financial objects: the supply of financial assets at any point in time is assumed to be fixed, and financial markets are viewed as competitive. Our analysis can thus focus on the determinants of the demand for financial assets.

This requires that we first spend some time discussing the preferences and attitudes toward risk of investors, those who demand the assets (Chapters 3 and 4), before modeling the investment process, i.e., how the relative demands for the various financial assets are determined (Chapters 5–7). Armed with these tools, we will review the three main equilibrium theories, the CAPM in Chapter 8, AD pricing in Chapter 9, and the CCAPM in Chapter 10.²

² In Web Chapter A we generalize the CCAPM by making more explicit the macroeconomic setting that underlies the CCAPM. An understanding of these macrofinancial linkages has become more important given the recent financial crisis and the advent of the “Great Recession.”

The arbitrage approach to valuing bicycles starts from observing that a bicycle is not (much) more than the sum of its parts. Accordingly, if you know the price of all the necessary components—frame, handlebar, wheel, tire, saddle, brake, and gearshift—you can determine relatively easily the market value of the bicycle. The knowledge of how to assemble the bicycle and the time required to do so, however, are not in infinite supply. These considerations suggest that the arbitrage approach may hold only as an approximation, one that may be rather imprecise in circumstances where the time and intellectual ability required to “assemble the bicycle” from the necessary spare parts are nontrivial; i.e., when the remuneration of the necessary “engineers” matters.³

The arbitrage approach is, in a sense, much more straightforward than the equilibrium approach. It is also more robust: if the arbitrage relationship between the price of the bicycle and the price of its parts does not hold, anyone with a little time could become a bicycle manufacturer and make good money. If too many people exploit that idea, however, the prices of parts and the prices of bicycles will start adjusting and be forced into line. This very idea is especially powerful for the object at hand, financial assets, because if markets are complete in the sense discussed in Section 1.6, then it can easily be shown that *all* the component prices necessary to value *any* arbitrary cash flow are available. Furthermore, little time and few resources (relative to the global scale of product markets) are needed to exploit arbitrage opportunities in financial markets.

There is, however, an obvious limitation to the arbitrage approach. Where do we get the price of the parts if not through an equilibrium approach? That is, the arbitrage approach is much less ambitious and more partial than the equilibrium approach. Even though it may be more practically useful in the domains where the price of the parts is readily available, it does not make up for a general theory of valuation and, in that sense, has to be viewed as a complement to the equilibrium approach. In addition, the equilibrium approach, by forcing us to rationalize investors’ demand for financial assets, provides useful lessons for the practice of asset management. The foundations of this inquiry will be put in place in Chapters 3–7—which together make up Part II of the book—while Chapter 16 will extend the treatment of this topic beyond the case of the traditional one-period static portfolio analysis and focus on the specificities of long-run portfolio management.

Finally, the arbitrage and equilibrium approaches can be combined. In particular, one fundamental insight that we will develop in Chapter 11 is that any cash flow can be viewed

³ In a similar vein, “financial engineers” seek to create new securities by cleverly packaging existing ones, or seek to design arbitrage portfolios which increase in value as the relative prices of their constituent securities (by analogy, the bicycle, and its independently traded constituent parts) come into better alignment. While a pure arbitrage portfolio is strictly risk free, most arbitrage portfolios arising from financial engineering will have positive payoffs provided certain low probability (“Black Swan”) events do not occur. At the start of the financial crisis, it was these very low probability events that came to pass with disastrous consequences for the institutions themselves (e.g., AIG).

Table 2.2: The road map

	Equilibrium	Arbitrage
<i>Preliminaries</i>		
<i>Computing risk premia</i>	Utility theory—Chapters 3–4 Investment demand—Chapters 5–7 CAPM—Chapter 8 CCAPM—Chapter 10	APT—Chapter 14
<i>Identifying distorted probabilities</i>		Martingale measure—Chapters 12–13
<i>Pricing future dollars state by state</i>	AD pricing I—Chapter 9	AD pricing II—Chapter 11

as a portfolio of AD securities, i.e., it can be replicated with AD securities. This makes it very useful to start using the arbitrage approach with AD securities as the main building blocks for pricing assets or valuing cash flows. Conversely, the same chapter will show that options can be very useful in completing the markets and thus in obtaining a full set of prices for “the parts that will then be available to price the bicycles.” In other words, the AD equilibrium pricing theory is a good platform for arbitrage valuation. The link between the two approaches is indeed so tight that we will use our acquired knowledge of equilibrium models—reviewed in Part III—to understand one of the major arbitrage approaches, the Martingale pricing theory (Chapters 12 and 13).⁴ We will then propose an overview of the APT in Chapter 14. Chapters 11 through 15 together make up Part IV of this book. This outline is summarized in [Table 2.2](#).

Part V will focus on three extensions. As already mentioned, Chapter 16 deals with long-run asset management. Chapter 17 focuses on some implications of incomplete markets whose consequences are illustrated from the twin viewpoints of the equilibrium and arbitrage approaches. We will use it as a pretext to review the Modigliani–Miller theorem and, in particular, to understand why it depends on the hypothesis of complete markets. Finally, in Chapter 18, we will open up, just a little, the Pandora’s box of heterogeneous beliefs. Our goal is to understand a number of issues that are largely swept under the rug in standard asset management and pricing theories and, in the process, restate the efficient market hypothesis.⁵

2.4 Decomposing Risk Premia

Method 1 in [Section 2.2](#) represents the standard methodology for real investment project evaluation and, if only for this reason, deserves a bit more attention. Generalizing

⁴ Web Chapter B presents additional illustrations and applications of Martingale pricing theory.

⁵ Web Chapter D concludes with a discussion of the recent “financial crisis.” While we describe the causes (in our view) and consequences of the crisis, our principal objective is to relate the event to the concepts introduced in this text.

the present value expression to a many period cash flow timing typical of real projects yields

$$V_{\text{project}} = \sum_{t=1}^T \frac{E\tilde{C}F_t}{(1+r_f+\pi_p)^t}$$

where we simplify the discussion below by assuming a constant risk-free rate of interest yields ($r_t^f \equiv r_f$ for all t) and a constant project risk premium π_p . Recall that an objective, market-based valuation of the project requires that π_p represents the risk premium on a stock (or portfolio of stocks) whose cash-flow timing and risk characteristics resemble (in a manner to be made formal in later chapters) those of the project; i.e.

$\pi_p = \pi_i = E\tilde{r}_i - r_f$, where $E\tilde{r}_i$ is the expected return to the “approximating stock or stock portfolio i .”

Work in empirical asset pricing suggests is that the return premium $\tilde{r}_i - r_f$ on a stock i may be intertemporally decomposed as a linear combination of fundamental stochastic factors \tilde{F}_t^j , $j = 1, 2, \dots, J$:

$$\tilde{r}_{i,t} - r_f = \alpha_i + \beta_i^1 \tilde{F}_t^1 + \beta_i^2 \tilde{F}_t^2 + \dots + \beta_i^J \tilde{F}_t^J + \tilde{\varepsilon}_{i,t} \quad (2.1)$$

where the β_i^j measures the sensitivity of stock i 's return to the specific underlying factor j , and $\tilde{\varepsilon}_{i,t}$ is an i.i.d. mean zero random component. Under representation (2.1), all stock returns are thus determined by the same J factors. These factors affect different stocks to differing degrees as measured by a stock's specific factor sensitivities.

The impact of these factors subsumes the entire risk premium except for a “white noise” residual.

What are these factors? Some factors define macroeconomic conditions such as the inflation rate or the state of the business cycle as measured by the GDP growth rate. All firms are affected, to varying degrees, by the business cycle. Chen et al. (1986) consider factors such as industrial production, inflation expectations, and oil prices. Other factors are believed to measure various permanent psychological biases on the part of investors, biases that appear, in some cases, to have a permanent influence on equity return patterns.

Why the focus is on estimating risk premia, when our ultimate objective is asset pricing? Given cash-flow estimates, prices and returns are, of course, “dual” to one another in the sense that knowing a project's price determines the implied risk premium and vice versa via the present value relationship. Another motivation is that most “investors” are not undertaking real investment projects but are buying portfolios of individual securities. Rather than thinking of next period's price of a security relative to its price today, it is more natural for investors to think in terms of what the security is expected to earn above r_f —its risk premium—over their chosen time horizon.

2.5 Models and Stylized Facts

Financial economics has as its goal the understanding of financial market behavior. While this statement may seem obvious, it has no real content until we clarify what an “understanding” would mean. For asset pricing phenomena, in particular, it means that the event under study can be explained in a model economy where self-interested economic agents determine their demands for various securities based on certain first principles/axioms of economic behavior.⁶ The model may also go on to specify how these demands interact with the supplies of the various securities to determine equilibrium prices and returns. Chapters 3–10 of this text essentially construct the basic model paradigms of finance.

A model is necessarily an abstraction or (dramatic) simplification of reality. Many economic mechanisms are ignored with only the most critical retained. As a result, there will inevitably be some aspects of reality which it will be unable to explain. What characteristics, then, describe a good in contrast to a poor economic model? We propose the three criteria listed below:

- i. A good model must be simple enough to enrich our intuition. In other words, the principal economic mechanisms within the model that allow it to explain the phenomenon under study must be readily apparent.
- ii. The abstraction which the model represents must be tailored to (or rich enough to address meaningfully) the questions being asked of it. Researchers would not seek to understand observed patterns in the US distributions of income and wealth, for example, in a representative (single) agent model.⁷
- iii. The model should be able to give precise answers to questions we pose concerning the behavior of the real economy.

How does one acquire confidence in a model that ostensibly satisfies the above criteria? The answer is straightforward: the more historical phenomena the model is able to explain successfully, the more confidence researchers have in its ability to provide answers to current and future real-world questions. It is here that the financial “stylized facts” enter the scene. These “stylized facts” are simply well-documented price, quantity, or return patterns that have been present in financial market data over long periods of time. In the latter sense, they are said to be “secular.”⁸ At a minimum, it is imperative that a good and trustworthy asset pricing model be able to replicate (reproduce) them.

⁶ As such economic models differ from scientific ones (e.g., of the atom). Scientific models describe the laws of nature, which are invariant to human activity. In economic models, everything is the result of human activity.

⁷ Requirement (ii) is not intended to suggest that radically different models should be proposed to explain different data regularities. The underlying principals must be the same and the results mutually consistent in areas of overlap.

⁸ That is, “secular” as in “existing or continuing through ages and centuries” (Webster’s Collegiate Dictionary, Ninth Edition).

Although his focus is on the mechanisms underlying business cycles, we adopt the general modeling perspective of R.E. Lucas, Jr. as expressed in [Lucas \(1980\)](#):

One of the functions of theoretical economics is to provide fully articulated, artificial economic systems that can serve as laboratories in which policies that would be prohibitively expensive to experiment with in actual economies can be tested out at much lower cost. To serve this function well, it is essential that the artificial “model” economy be distinguished as sharply as possible in discussion from actual economies. Insofar as there is confusion between statements of opinion as to the way we believe actual economies would react to particular policies and statements of verifiable fact as to how the model will react, the theory is not being effectively used to help us to see which opinions about the behavior of actual economies are accurate and which are not. This is the sense in which insistence on the “realism” of an economic model subverts its potential usefulness in thinking about reality. Any model that is well enough articulated to give clear answers to the questions we put to it will necessarily be artificial, abstract, patently “unreal.”

At the same time, not all well-articulated models will be equally useful. Though we are interested in models because we believe they may help us to understand matters about which we are currently ignorant, we need to test them as useful imitations of reality by subjecting them to shocks for which we are fairly certain how actual economies, or parts of economies, would react. The more dimensions on which the model mimics the answers actual economies give to simple questions, the more we trust its answers to harder questions. This is the sense in which more “realism” in a model is clearly preferred to less.

On this general view of the nature of economic theory then, a “theory” is not a collection of assertions about the behavior of the actual economy but rather an explicit set of instructions for building a parallel or analogue system – a mechanical, imitation economy. A “good” model, from this point of view, will not be exactly more “real” than a poor one, but will provide better imitations. Of course, what one means by a “better imitation” will depend on the particular questions to which one wishes answers.

Accordingly we next highlight a few quantitative and qualitative properties of financial markets that serve as basic stylized facts against which the financial models we will propose in the remainder of this text should be measured. We focus exclusively on capital market phenomena rather than empirical regularities related to the firm’s corporate financing activities.

2.5.1 The Equity Premium

A broadly diversified portfolio of stocks (e.g., the S&P₅₀₀, the DAX, the CAC) consistently earns average returns substantially in excess of the risk-free rate (normally proxied by the return on short-term debt securities issued by a nation’s national treasury authority). Tables 2.3 and 2.4 give some ideas as to the magnitudes involved.

Table 2.3: US returns: 1889–2010^a

Time period	Real Return on a Market Index ^b	Real Return on a Relatively Riskless Security	% Risk Premium
	Mean	Mean	Mean
1889–2010	7.5%	1.1%	6.4%
1889–1978	7.0%	0.8%	6.2%
1926–2010	8.0%	0.8%	7.2%
1946–2010	7.5%	0.8%	6.7%

^aData from [Mehra \(2012\)](#); annualized returns.^bThe S&P₅₀₀ and its antecedents.**Table 2.4: The equity premium: the principal capital markets^a**

Country	Time Period	% Risk Premium	Country	Time Period	% Risk Premium
Belgium	1900–2010	5.5%	Sweden	1900–2010	6.6%
Holland	1900–2010	6.5%	UK	1900–2010	6.0%
France	1900–2010	8.7%	Australia	1900–2010	8.3%
Germany	1900–2010	9.8%	Canada	1900–2010	5.6%
Ireland	1900–2010	5.3%	India	1991–2004	11.3%
Italy	1900–2010	9.8%	Japan	1900–2010	9.0%

^aSource and details: [Dimson et al. \(2010\)](#); annualized returns.

Note that in the US statistics ([Table 2.3](#)), the premium for long horizons never falls below 6%. With minor exceptions, the same results usually carry over to 10-year horizons (i.e., the pattern is secular). Similar if not stronger results carry over to international data ([Table 2.4](#)), despite two wars which led to substantial capital destruction and, in some cases, the cessation of organized competitive stock trading (e.g., Germany, France). The experiences of the United Kingdom and Canada compare most closely with the United States in that there was no interruption to trading and the respective governments did not seek to control this form of financial market activity. India's stock market experience is more recent and Japan's stellar performance is largely a post-World War II phenomenon prior to 1990. Across all markets, a robust equity premium over long horizons is an empirical fact.⁹

Within the current class of “rational economic models,” those that we are about to describe in the chapters to follow, it is extremely difficult to replicate these statistics, so much so

⁹ A prominent exception is the experience of recent history. For the US stock market, the average annual return for the period 2000–2010 was –0.39%, while US Treasury bonds of time to maturity exceeding 10 years paid on average 7.18% over the same period.

Table 2.5: Average annualized excess returns for 10 portfolios sorted on BE/ME^a

Lowest		→ Increasing (BE/ME) →							Highest	
Port 1	Port 2	Port 3	Port 4	Port 5	Port 6	Port 7	Port 8	Port 9	Port 10	
6.76	7.64	7.89	7.65	8.43	8.92	9.02	10.88	11.65	12.75	

^aBased on monthly data for the period 1963.1 through 2011.7. These (value weighted) portfolios are reconstructed (i.e., all the Compustat stocks are reassigned to one of the 10 portfolios) at the end of June of each year based on the end of the previous year's BE and ME values. We thank Tano Santos for making this data available to us.

that they are often described as constituting the “equity premium puzzle.”¹⁰ Furthermore, despite the substantial risk premia evident in [Tables 2.3 and 2.4](#), most of the US population owns very little stock: the 2007 Survey of Consumer Finance reports that about one-half of US households own no stock at all; for high-income households 23% own no stock. This fact constitutes a second “puzzle,” at least as regards the modern theory of portfolio composition ([Chapters 6 and 7](#)).

The equity premium is a secular time series property of stock returns.

2.5.2 The Value Premium

The value premium is a statement about the cross section of stock returns. It is the empirically robust observation that stocks with a higher (book value of equity)/(market value of equity) (BE/ME) ratio have, on average, higher excess returns than stocks with low (BE/ME) values. Consider [Table 2.5](#), which describes the annualized average excess returns on 10 portfolios ($E r_{port\ i} - r_f$, $i = 1, 2, \dots, 10$) of Compustat stocks sorted on the basis of their (BE/ME) ratio.¹¹

Notice that the highest (BE/ME) portfolio has nearly twice the average excess returns of the lowest (BE/ME) portfolio. The return pattern observed in [Table 2.5](#) is known as the “value premium.” It is characteristic of international stock markets and all historical time periods. The value premium becomes the “value premium puzzle” because financial theory has no all-encompassing explanation as to why it should be observed. Traditional risk-based theories—the idea that investors dislike risky returns and must therefore be compensated with higher average returns in order to hold, willingly, higher risk assets—cannot explain the pattern of [Table 2.5](#). In particular, the CAPM, which we discuss in

¹⁰ In the models we will consider the supply of equities and risk-free bonds is typically assumed to be constant. This is generally an innocuous assumption: the supply of IBM equity shares outstanding, for example, does not change from year to year. The focus of these models is thus principally to characterize security demands by investors, from which follow (given fixed supplies) equilibrium prices and rates of return. The resulting equity premium usually falls far short of what is manifest in the data.

¹¹ Compustat is a large publicly accessible database of historical stock returns (and much other information).

Chapter 8, and which is the most widely cited risk-based theory of returns, cannot explain [Table 2.5](#).

The value premium is a secular cross-sectional property of stock returns.

2.5.3 The Term Structure

In what follows we will also explore various features of the bond market and, in particular, the market for default free government securities (e.g., US Treasury securities). In this specific market, the fundamental notion is that of the term structure of nominal interest rates: the family of interest rates on zero-coupon, default-free nominal bonds of progressively greater maturity.¹² More specifically, at any time t , it is the collection of interest rates $\{r_{t,1}, r_{t,2}, \dots, r_{t,J}\}$, where $r_{t,j}$ is the period t interest rate on a default-free security with cash-flow pattern:

T	$t+1$	$t+2\dots$	$t+j-1\dots$	$t+j$	$t+j+1\dots$	$t+J$
$-q_{t,j}^b$	0	0	0	\$1000	0	0

where $q_{t,j}^b = \$1000 / ((1 + r_{t,j})^j)$ with $q_{t,j}^b$ denoting the security's period t market price.

With no uncertainty in the payments, these rates reflect the pure time value of money and, as such, constitute one of the building blocks of any discounting procedure (recall [Section 2.2](#)). In this sense, the term structure is a fundamental concept. Accordingly, the basic features of the bond market are usually expressed as properties of the term structure. In particular, we will seek to explain the following:

- i. The term structure of interest rates is typically upward sloping: default free discount securities with longer times to maturity typically command higher rates. We want to understand not only why this is true but also what drives the exceptions.
- ii. The term structure of interest rates generally moves up or down for all maturities simultaneously. This means that an increase in the short rate is accompanied by an increase in rates for bonds of all maturities. For US Treasury securities, 99% of the variation in returns at any maturity is related to shifts in the entire term structure. This fact stands in stark contrast to its analogous relationship for stocks: roughly 80–90% of the variation in returns to any particular stock is generally *unrelated* to aggregate market movements.

This observation suggests that there may be one macroeconomic quantity (factor) affecting all default-free rates similarly. Can this factor be identified?

¹² By a “nominal” bond we mean one that pays prefixed dollar (CHF, Euro) amounts, that are not adjusted for inflation.

- iii. Lastly, there is the phenomenon of relative volatility: properly assessed, returns on long term US Treasury securities are more volatile than the returns on short-term bonds, even when normalized by their higher expected returns (see topic (i)). Such a result is puzzling in that our immediate intuition would suggest otherwise: long-term bond prices and returns should not be as sensitive to business cycle or other macroeconomic events, as these are generally of shorter duration (than the bond's time to maturity) and thus their consequences for bond returns should tend to “average out” over the long bond's time to maturity. [Shiller \(1979\)](#) refers to this phenomenon as the “bond volatility puzzle.”¹³

Taken together, the equity premium (a quantitative assessment) and the three qualitative properties of the term structure detailed above illustrate important stylized facts that good models should be able to replicate.

2.6 Asset Pricing Is Not All of Finance!

2.6.1 Corporate Finance

Intermediate financial theory focuses on the valuation of risky cash flows. Pricing a future (risky) dollar is the dominant ingredient in most financial problems. But it is not all of finance! Our capital markets perspective in particular sidesteps many of the issues surrounding how the firm generates and protects the cash-flow streams to be priced. It is this concern that is at the core of corporate financial theory or simply *corporate finance*.

In a broad sense, corporate finance is concerned with decision making at the firm level whenever it has a financial dimension, has implications for the financial situation of the firm, or is influenced by financial considerations. In particular, it is a field concerned, first and foremost, with the investment decision (what projects should be accepted), the financing decision (what mix of securities should be issued and sold to finance the chosen investment projects), the payout decision (how should investors in the firm, and in particular the equity investors, be compensated), and risk management (how corporate resources should be protected against adverse outcomes). Corporate finance also explores

¹³ Strictly speaking, these are statements about default-free coupon bonds. In fact, the US Treasury, for instance, does not issue zero coupon bonds of more than 6 months time to maturity. It is possible, however, to extract the implied default-free zero coupon bond prices from the prices and cash flows associated with default-free coupon bonds, if a sufficient number of distinct types (different cash flows) are issued. The set of IRRs (internal rates of return) on default free coupon bonds of successively greater maturity is referred to as the “yield curve.” Its qualitative properties do not differ significantly from those of the term structure (e.g., both move in tandem). Typically there is little quantitative difference as well, and the expressions “term structure” and “yield curve” are often used (incorrectly) interchangeably. See Chapter 11 for a more detailed discussion. The percentages are from [Ang \(2012\)](#).

issues related to the size and the scope of the firm, e.g., mergers and acquisitions and the pricing of conglomerates, the internal organization of the firm, the principles of corporate governance, and the forms of remuneration of the various stakeholders.¹⁴

All of these decisions individually and collectively influence the firm's free cash-flow stream and, as such, have asset pricing implications. The decision to increase the proportion of debt in the firm's capital structure, for example, increases the riskiness of its equity cash-flow stream and the standard deviation of the equilibrium return on equity.

Of course, when we think of the investment decision itself, the solution to the valuation problem is of the essence. Indeed, many of the issues typically grouped under the heading of capital budgeting are intimately related to the focus of the present text. We will be silent, however, on most of the other issues listed above, which are better viewed as arising in the context of bilateral (rather than market) relations and, as we will see, in situations where asymmetries of information play a dominant role.

The goal of this section is to illustrate the difference in perspectives by reviewing, selectively, the corporate finance literature, particularly as regards the capital structure of the firm and contrasting it with the capital markets perspective that we will be adopting throughout this text. In so doing, we also attempt to give the flavor of an important research area while reminding the reader of the many important topics this text elects not to address.

2.6.2 Capital Structure

We focus on the capital structure issue in Chapter 17 where we explore the assumption underlying the famous Modigliani–Miller irrelevance result: in the absence of taxes, subsidies, and contracting costs, the value of a firm is independent of its capital structure if the firm's investment policy is fixed and financial markets are complete. Our emphasis will concern how this result fundamentally rests on the complete markets assumption.

The corporate finance literature has not ignored the completeness issue but rather has chosen to explore its underlying causes, most specifically information asymmetries between the various agents concerned, managers, shareholders, and so forth.¹⁵ While we touch on

¹⁴ The recent scandals (Hewlett-Packard, AIG) in the United States, Europe (gigantic trading losses in JPM Chase, UBS, and Société Générale, due to rogue or unsupervised trading) and in Japan (Olympus Optical) place in stark light the responsibilities of boards of directors for ultimate firm oversight as well as their frequent failure to provide it. The large question here is what sort of board structure is consistent with superior long-run firm performance?

¹⁵ Tax issues have tended to dominate the corporate finance capital structure debate until recently, and we will review this arena shortly. The relevance of taxes is not a distinguishing feature of the corporate finance perspective alone. Taxes also matter when we think of valuing risky cash flows, although we will have very little to say about it except that all the cash flows we consider are to be thought of as after-tax cash flows.

the issue of heterogeneity of information in a market context, we do so only in Chapter 18, emphasizing there that heterogeneity raises a number of tough modeling difficulties. These difficulties justify the fact that most of capital market theory either is silent on the issue of heterogeneity (in particular, when it adopts the arbitrage approach) or explicitly assumes homogeneous information on the part of capital market participants.

In contrast, the bulk of corporate finance builds on asymmetries of information and explores the various problems they raise. These are typically classified as leading to situations of “moral hazard” or “adverse selection.” An instance of the former is when managers are tempted to take advantage of their superior information to implement investment plans that may serve their own interests at the expense of those of shareholders or debtholders. An important branch of the literature concerns the design of contracts, which take moral hazard into account. The choice of capital structure, in particular, will be seen potentially to assist in their management (see, for example, [Zwiebel, 1996](#)).

A typical situation of adverse selection occurs when information asymmetries between firms and investors make firms with “good” investment projects indistinguishable to outside investors from firms with poor projects. This suggests a tendency for all firms to receive the same financing terms (a so-called pooling equilibrium where firms with less favorable prospects may receive better than deserved financing arrangements). Firms with good projects must somehow indirectly distinguish themselves in order to receive the more favorable financing terms they merit. For instance, they may want to attach more collateral to their debt securities, an action that firms with poor projects may find too costly to replicate (see, for example, [Stein, 1992](#)). Again, the capital structure decision may sometimes help in providing a resolution of the “adverse selection” problem. Below we review the principal capital structure perspectives.

2.6.3 Taxes and Capital Structure

Understanding the determinants of a firm’s capital structure (the proportion of debt and equity securities it has outstanding in value terms) is the classical problem in corporate finance. Its intellectual foundations lie in the seminal work of [Modigliani and Miller \(1958\)](#), who argue for capital structure irrelevance in a world without taxes and with complete markets (an hypothesis that excludes information asymmetries).

The corporate finance literature has also emphasized the fact that when one security type receives favored tax treatment (typically, this is debt via the tax deductibility of interest), then the firm’s securities become more valuable in the aggregate if more of that security is issued, since to do so is to reduce the firm’s overall tax bill and thus enhance the free cash flow to the security holders. Since the bondholders receive the same interest and principal

payments, regardless of the tax status of these payments from the firm's perspective, any tax-based cash-flow enhancement is captured by equity holders. Under a number of further specialized assumptions (including the hypothesis that the firm's debt is risk-free), these considerations lead to the classical relationship

$$V_L = V_U + \tau D$$

The value of a firm's securities under partial debt financing (V_L , where L denotes leverage in the capital structure) equals its value under all equity financing (V_U , where U denotes unlevered or an all-equity capital structure) plus the present value of the interest tax subsidies. This latter quantity takes the form of the corporate tax rate (τ) times the value of debt outstanding (D) when debt is assumed to be perpetual (unchanging capital structure).

In return terms, this value relationship can be transformed into a relationship between levered and unlevered equity returns:

$$r_L^e = r_U^e + (1 - \tau)(D/E)(r_U^e - r_f)$$

i.e., the return on levered equity, r_L^e , is equal to the return on unlevered equity, r_U^e , plus a risk premium due to the inherently riskier equity cash flow that the presence of the fixed payments to debt creates. This premium, as indicated, is related to the tax rate, the firm's debt/equity ratio (D/E), a measure of the degree of leverage, and the difference between the unlevered equity rate and the risk-free rate, r_f . Immediately we observe that capital structure considerations influence not only expected equilibrium equity returns via

$$Er_L^e = Er_U^e + (1 - \tau)D/E(Er_U^e - r_f)$$

where E denotes the expectations operator, but also the variance of returns since

$$\sigma_{r_L^e}^2 = (1 - (1 + \tau)D/E)^2 \sigma_{r_U^e}^2 > \sigma_{r_U^e}^2$$

under the mild assumption that r_f is constant in the very short run. These relationships illustrate but one instance of corporate financial considerations affecting the patterns of equilibrium returns as observed in the capital markets.

The principal drawback to this tax-based theory of capital structure is the natural implication that if one security type receives favorable tax treatment (usually debt), then if the equity share price is to be maximized the firm's capital structure should be composed exclusively of that security type—i.e., all debt, which is not observed. More recent research in corporate finance has sought to avoid these extreme tax-based conclusions by balancing

the tax benefits of debt with various costs of debt, including bankruptcy and agency costs.¹⁶ Our discussion broadly follows [Harris and Raviv \(1991\)](#).

2.6.4 Capital Structure and Agency Costs

This important segment of the literature seeks to explain financial decisions by examining the conflicts of interests among claimholders within the firm. Although agency conflicts can take a variety of forms, most of the literature has focused on manager's incentives to increase investment risk—the asset substitution problem—or to reject positive Net Present Value (NPV) projects—the underinvestment problem. Both of these conflicts increase the cost of debt and thus reduce the firm's value-maximizing debt ratio.

Another commonly discussed determinant of capital structure arises from manager–stockholder conflicts. Managers and shareholders have different objectives. In particular, managers tend to value investment more than shareholders do. Although there are a number of potentially powerful internal mechanisms to control managers, the control technology normally does not permit the costless resolution of this conflict between managers and investors. Nonetheless, the cash-flow identity implies that constraining financing, hedging, and payout policy places indirect restrictions on investment policy. Hence, even though investment policy is not contractible, by restricting the firm in other dimensions, it is possible to limit the manager's choice of an investment policy. For instance, [Jensen \(1986\)](#) argues that debt financing can increase firm value by reducing the free cash flow. This idea is formalized in more recent papers by [Stulz \(1990\)](#) and [Zwiebel \(1996\)](#). Also, by reducing the likelihood of both high and low cash flows, risk management can control not only shareholders' underinvestment incentives but managers' ability to overinvest as well.

More recently, the corporate finance literature has put some emphasis on the cost that arises from conflicts of interest between controlling and minority shareholders. In most countries, publicly traded companies are not widely held but rather have controlling shareholders. Moreover, these controlling shareholders have the power to pursue private benefits at the expense of minority shareholders, within the limits imposed by investor protection. The recent “law and finance” literature following [Shleifer and Vishny \(1997\)](#) and [La Porta et al. \(1998\)](#) argues that the expropriation of minority shareholders by the controlling shareholder

¹⁶ There are many other proposed capital structure theories. [Lee \(2014\)](#), for example, proposes that firms eschew present and future available tax benefits to debt financing and rather maintain large cash balances in order to finance unexpected future investments and as a precaution against bad times (periods of low cash-flow generation). Such firms are unable to raise cash at times of critical need by selling existing assets because of severe capital adjustment costs.

is at the core of agency conflicts in most countries. While these conflicts have been widely discussed in qualitative terms, the literature has largely been silent on the magnitude of their effects.

2.6.5 The Pecking Order Theory of Investment Financing

The seminal reference here is [Myers and Majluf \(1984\)](#) who again base their work on the assumption that investors are generally less well informed (asymmetric information) than insider-managers vis-à-vis the firm's investment opportunities. As a result, new equity issues to finance new investments may be so underpriced (reflecting average project quality) that NPV positive projects from a societal perspective may have a negative NPV from the perspective of existing shareholders and thus not be financed. [Myers and Majluf \(1984\)](#) argue that this underpricing can be avoided if firms finance projects with securities that have more assured payout patterns and thus are less susceptible to undervaluation: internal funds and, to a slightly lesser extent, debt securities, especially risk-free debt. It is thus in the interests of shareholders to finance projects first with retained earnings, then with debt, and lastly with equity. An implication of this qualitative theory is that the announcement of a new equity issuance is likely to be accompanied by a fall in the issuing firm's stock price since it indicates that the firm's prospects are too poor for the preferred financing alternatives to be accessible.

The pecking order theory has led to a large literature on the importance of security design. For example, [Stein \(1992\)](#) argues that companies may use convertible bonds to get equity into their capital structures "through the backdoor" in situations where informational asymmetries make conventional equity issues unattractive. In other words, convertible bonds represent an indirect mechanism for implementing equity financing that mitigates the adverse selection costs associated with direct equity sales. This explanation for the use of convertibles emphasizes the role of the call feature—that will allow good firms to convert the bond into common equity—and costs of financial distress—that will prevent bad firms from mimicking good ones. Thus, the announcement of a convertible bond issue should be greeted with a less negative—and perhaps even positive—stock price response than an equity issue of the same size by the same company.

2.7 Banks

Banks merit our attention at the present juncture because the concepts we have been discussing (e.g., firm leverage) apply as well to banks but to a "unique" degree. To organize our banking discussion, let us first compare the (simplified) balance sheets of a typical industrial firm and a commercial bank.

Balance sheets (all entries measured in value terms)

Industrial Firm		Bank	
Assets	Liabilities	Assets	Liabilities
<ul style="list-style-type: none"> • Tangible assets (factories, machinery, inventories) • Intangible assets (patents, process technology, trademarks, etc.) 	<ul style="list-style-type: none"> • Debt issued by the firm • Equity of shareholders 	<ul style="list-style-type: none"> • Loans • Cash • Default-free government securities • Other securities (mortgage backed securities, for example) 	<ul style="list-style-type: none"> • Deposits taken by the bank • Debt securities issued by the bank • Equity of shareholders

Note that the deposits of the banks constitute a (large) portion of its liabilities. Pure investment banks, as distinct from commercial banks, are financial companies that extend loans and make equity investments, but they cannot take deposits as a source of funding. The great general-purpose international banks (e.g., Deutsche, UBS, Société Générale, Citi) are both; i.e., they are holding companies with investment banking and commercial banking divisions. In periods of financial distress, either division effectively becomes responsible for the debts of the other since the overall bank is a single legal entity.

For nonfinancial Eurozone corporations, the average (D/E) ratio prior to the financial crisis (2008) was about 0.8; for banks in the Eurozone, the average (D/E) ratio was about 30. Banks were, and continue to be, much more highly leveraged than other firms. In the past, pure investment banks were even more highly leveraged than bank-holding companies in general. Lehman Brothers, for example, was leveraged to a (D/E) ratio exceeding 50. A high leverage ratio is the first unique feature of bank corporations.

A second distinction, already alluded to, concerns the funding banks receive in the form of deposits. Nonbank corporations that take out bank loans or issue long-term bonds know exactly the timing and magnitude of the future interest and principal payments they must make. To the extent they borrow in the form of taking deposits, banks, however, may be required, suddenly and unexpectedly, to repay the “loans” if depositors collectively initiate large aggregate withdrawals. Such an event may occur if depositors suspect that the bank’s investments (its loan and securities portfolio) have not maintained their value, leading to the possibility of bankruptcy, and a delayed return of deposits or potential deposit losses. Economists call these circumstances “bank runs.”¹⁷ The phenomenon by which some of a bank’s funding may suddenly disappear is the second special feature of banks.¹⁸

¹⁷ To forestall bank runs, governments now provide deposit insurance against deposit losses at least up to a maximum. In the United States, the relevant entity is the Federal Deposit Insurance Corporation (FDIC), which is funded by a small tax on banks.

¹⁸ Anticipating our discussions of portfolio theory, banks can be viewed as very highly leveraged equity portfolios: a large long position in long-term assets financed by a short position in short-term assets.

Lastly, banks are the principle providers of credit to smaller businesses which find it costly to issue their own bonds directly in the capital markets. This notion of credit is very broad: banks provide loans for investment projects, loans for inventories and wage payments, temporary trade financing, etc. When these services become deficient because certain prominent banks are in financial difficulty (a so-called credit crunch), the macroeconomic effects are both negatively severe and long lasting. The ongoing Great Recession is only the most recent case in point. While the bankruptcy of a nonfinancial firm can have devastating effects on the local economy in which it operates, the consequences of bankruptcies or near-bankruptcies among the large players in the international banking system can have adverse national and international consequences. This feature of the banking system constitutes its third unique characteristic. It is an aspect of the “too big to fail” debate.

We return to the unique features of the banking system in web chapter D.

2.8 Conclusions

We have presented four general approaches and two main perspectives on the valuation of risky cash flows. This discussion was meant to provide an organizing principle and a road map for the extended treatment of a large variety of topics on which we are now embarking. We then went on to present some stylized facts of the financial markets and used these to define what a good financial (valuation) model should be. Our brief excursion into corporate finance was intended to suggest some of the agency issues that are part and parcel of a firm’s cash-flow determination. That we have elected to focus on pricing issues surrounding those cash-flow streams does not diminish the importance of the many issues surrounding their creation.

References

- Ang, A., 2012. Fixed Income. Columbia Business School, Mimeo.
- Chen, N.F., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. *J. Bus.* 59, 383–404.
- Dimson, E., Marsh, P., Staunton, M., 2010. *Triumph of the Optimists: 101 Years of Global Investment Returns*. Princeton University Press, Princeton, NJ.
- Harris, M., Raviv, A., 1991. The theory of capital structure. *J. Finan.* 46, 297–355.
- Jensen, M., 1986. Agency costs of free cash flow, corporate finance, and takeovers. *Am. Econ. Rev.* 76, 323–329.
- Lee, J.H., 2014. Debt Servicing Costs and Capital Structure, Working Paper, Columbia University, Department of Economics.
- Lucas Jr., R.E., 1980. Methods and problems in business cycle theory. *J. Money Credit Bank.* 12, 696–715.
- Mehra, R., 2012. Consumption-based asset pricing models. *Annu. Rev. Finan. Econ.* 4, 13.1–13.25.
- Modigliani, F., Miller, M., 1958. The cost of capital, corporate finance, and the theory of investment. *Am. Econ. Rev.* 48, 261–297.
- Myers, S., Majluf, N., 1984. Corporate financing and investment decisions when firms have information that investors do not have. *J. Finan. Econ.* 13, 187–221.

- La Porta, R., Lopes de Silanes, F., Shleifer, A., Vishny, R., 1998. Law and finance. *J. Polit. Econ.* 106, 1113–1155.
- Shiller, R., 1979. The volatility of long-term interest rates and expectations models of the term structure. *J. Polit. Econ.* 87, 1190–1219.
- Shleifer, A., Vishny, R., 1997. A survey of corporate governance. *J. Finan.* 52, 737–783.
- Stein, J., 1992. Convertible bonds as backdoor equity financing. *J. Finan. Econ.* 32, 3–23.
- Stulz, R., 1990. Managerial discretion and optimal financial policies. *J. Finan. Econ.* 26, 3–27.
- Zwiebel, J., 1996. Dynamic capital structure under managerial entrenchment. *Am. Econ. Rev.* 86, 1197–1215.

Making Choices in Risky Situations

Chapter Outline

3.1 Introduction	55
3.2 Choosing Among Risky Prospects: Preliminaries	56
3.3 A Prerequisite: Choice Theory Under Certainty	61
3.4 Choice Theory Under Uncertainty: An Introduction	63
3.5 The Expected Utility Theorem	66
3.6 How Restrictive Is Expected Utility Theory? The Allais Paradox	72
3.7 Behavioral Finance	75
3.7.1 Framing	76
3.7.2 Prospect Theory	78
3.7.2.1 <i>Preference Orderings with Connections to Prospect Theory</i>	83
3.7.3 Overconfidence	84
3.8 Conclusions	85
References	85

3.1 Introduction

The first stage of the equilibrium perspective on asset pricing consists of developing an understanding of the determinants of the **demand** for securities of various risk classes. Individuals demand securities (in exchange for current purchasing power) in their attempt to redistribute income across time and states of nature. This is a reflection of the consumption-smoothing and risk-reallocation function central to financial markets. Our endeavor requires an understanding of three building blocks:

1. How financial risk is defined and measured.
2. How an investor's attitude toward or tolerance for risk is to be conceptualized and then measured.
3. How investors' risk attitudes interact with the subjective uncertainties associated with the available assets to determine an investor's desired portfolio holdings (demands).

In this and the next chapter, we give a detailed overview of points 1 and 2; point 3 is treated in succeeding chapters.

3.2 Choosing Among Risky Prospects: Preliminaries

When we think of the “risk” of an investment, we are typically thinking of uncertainty in the future cash-flow stream to which the investment represents title. Depending on the state of nature that may occur in the future, we may receive different payments and, in particular, much lower payments in some states than others. That is, we model an asset’s associated cash flow in any future time period as a **random variable**.

Consider, for example, the investments listed in [Table 3.1](#), each of which pays off next period in either of two equally likely states. We index these states by $\theta = 1, 2$ with their respective probabilities labeled π_1 and π_2 .

First, this comparison serves to introduce the important notion of **dominance**. Investment 3 clearly dominates both investments 1 and 2 in the sense that it pays as much in all states of nature and strictly more in at least one state. The **state-by-state dominance** illustrated here is the strongest possible form of dominance. Without any qualification, we will assume that all rational individuals would prefer investment 3 to either of the other two. Basically, this means that we are assuming the typical individual to be nonsatiated in consumption: she desires more rather than less of the consumption goods these payoffs allow her to buy.

In the case of dominance, the choice problem is trivial and, in some sense, the issue of defining risk is irrelevant. The ranking defined by the concept of dominance is, however, very incomplete. If we compare investments 1 and 2, we see that neither dominates the other. Although it performs better in state 2, investment 2 performs much worse in state 1. There is no ranking possible on the basis of the dominance criterion. The different prospects must be characterized from a different angle. The concept of risk enters necessarily.

On this score, we would probably all agree that investments 2 and 3 are comparatively riskier than investment 1. Of course, for investment 3, the dominance property means that the only risk is an upside risk. Yet, in line with the preference for smooth consumption discussed in Chapter 1, the large variation in date 1 payoffs associated with investment 3 is to be viewed as undesirable in itself. When comparing investments 1 and 2, the qualifier

Table 3.1: Asset payoffs (\$)

	Cost at $t = 0$	Value at $t = 1$	
		$\pi_1 = \pi_2 = 1/2$	
		$\theta = 1$	$\theta = 2$
<i>Investment 1</i>	−1000	1050	1200
<i>Investment 2</i>	−1000	500	1600
<i>Investment 3</i>	−1000	1050	1600

“riskier” undoubtedly applies to the latter. In the worst state, the payoff associated with 2 is much lower; in the best state it is substantially higher.

These comparisons can alternatively, and often more conveniently, be represented if we describe investments in terms of their performance on a per dollar basis. We do this by computing the state-contingent rates of return (ROR) that we will typically associate with the symbol r . In the case of the above investments, we obtain the results given in [Table 3.2](#).

One sees clearly that all rational individuals should prefer investment 3 to the other two and that this same dominance cannot be expressed when comparing 1 and 2.

The fact that investment 2 is riskier, however, does not mean that all rational risk-averse individuals would necessarily prefer 1. Risk is not the only consideration, and the ranking between the two projects is, in principle, preference dependent. This is more often the case than not; dominance usually provides a very incomplete way of ranking prospects. This fact suggests we must turn to a description of preferences, the main objective of this chapter.

The most well-known approach at this point consists of summarizing such investment return distributions (i.e., the random variables representing returns) by their mean (Er_i) and variance (σ_i^2), $i = 1, 2, 3$. The variance (or its square root, the standard deviation) of the rate of return is then naturally used as the measure of “risk” of the project (or the asset). For the three investments just listed, we have:

$$Er_1 = 12.5\%; \quad \sigma_1^2 = \frac{1}{2}(5 - 12.5)^2 + \frac{1}{2}(20 - 12.5)^2 = (7.5)^2, \quad \text{or} \quad \sigma_1 = 7.5\%$$

$$Er_2 = 5\%; \quad \sigma_2 = 55\% \quad (\text{similar calculation})$$

$$Er_3 = 32.5\%; \quad \sigma_3 = 27.5\%$$

If we decided to summarize these return distributions by their means and variances only, investment 1 would clearly appear more attractive than investment 2: it has both a higher mean return and a lower variance. In terms of the mean–variance criterion, investment 1 dominates investment 2; 1 is said to *mean–variance dominate* 2. Our previous discussion makes it clear that **mean–variance dominance** neither implies nor is implied by state-by-state dominance. Investment 3 mean–variance dominates 2 but not 1, although it dominates

Table 3.2: State-contingent ROR (r)

	$\theta = 1$	$\theta = 2$
<i>Investment 1</i>	5%	20%
<i>Investment 2</i>	-50%	60%
<i>Investment 3</i>	5%	60%

them both on a state-by-state basis! This is surprising and should lead us to be cautious when using any mean–variance return criterion. Later, we will detail circumstances where it is fully reliable. At this point, let us anticipate that it is not generally so and that restrictions will have to be imposed to legitimize its use.

The notion of mean–variance dominance, which plays a prominent role in modern portfolio theory, can be expressed in the form of a criterion for selecting investments of equal magnitude:

1. For investments of the same Er , choose the one with the lowest σ .
2. For investments of the same σ , choose the one with the greatest Er .

In the framework of modern portfolio theory, one could not understand a rational agent choosing investment 2 rather than investment 1.

We cannot limit our inquiry to the concept of dominance, however. Mean–variance dominance provides only an incomplete ranking among uncertain prospects, as [Table 3.3](#) illustrates.

When we compare these two investments, we do not clearly see which is best; there is no dominance in either state-by-state or mean–variance terms. Investment 5 is expected to pay 1.25 times the expected return of investment 4, but, in terms of standard deviation, it is also 3 times riskier. The choice between 4 and 5, when restricted to mean–variance characterizations, would require specifying the terms at which the decision maker is willing to *substitute* expected return for a given risk reduction. In other words, what decrease in expected return is the decision maker willing to accept for a 1% decrease in the standard deviation of returns? Or, conversely, does the 1 percentage point additional expected return associated with investment 5 adequately compensate for the (3 times) larger risk? Responses to such questions are preference dependent (i.e., they vary from individual to individual).

Suppose, for a particular individual, the terms of the trade-off are well represented by the index E/σ . Since $(E/\sigma)_4 = 4$ while $(E/\sigma)_5 = 5/3$, investment 4 is better than investment 5 for that individual. Of course, another investor may be less risk averse; i.e., he may be willing to accept more extra risk for the same expected return. For example, his preferences may be

Table 3.3: State-contingent ROR (r)

	$\theta = 1$	$\theta = 2$
<i>Investment 4</i>	3%	5%
<i>Investment 5</i>	2% $\pi_1 = \pi_2 = \frac{1}{2}$ $ER_4 = 4\%; \sigma_4 = 1\%$ $ER_5 = 5\%; \sigma_5 = 3\%$	8%

adequately represented by $(E - 1/3\sigma)$ in which case he would rank investment 5 (with an index value of 4) above investment 4 (with a value of $3\frac{2}{3}$).¹

All these considerations strongly suggest that we have to adopt a more general viewpoint for comparing potential return distributions. This viewpoint is part of utility theory, to which we now turn after describing some of the problems associated with the empirical characterization of return distributions in [Box 3.1](#).

BOX 3.1 Computing Means and Variances in Practice

Useful as it may be conceptually, calculations of distribution moments such as the mean and the standard deviation are difficult to implement in practice: we rarely know what the *future* states of nature are, let alone their probabilities. We also do not know the returns in each state. A frequently used proxy for a future return distribution is its historical distribution. This amounts to selecting a historical time period and a periodicity, say monthly prices for the past 60 months, and computing the historical (net) returns as follows:

- a. Discrete compounding

$$r_j^e = (\text{net}) \text{ return to stock ownership in month } j = ((q_j^e + d_j)/q_{j-1}^e) - 1$$

where q_j^e is the price of the stock in month j , and d_j its dividend, if any, that month; $1 + r_j^e$ is referred to as the gross return. We then summarize the past distribution of stock returns by the average historical return and the variance of the historical returns. By doing so, we, in effect, assign an equal probability of $\frac{1}{60}$ to each past observation or event.

- b. Continuous compounding

To understand how to compute period-by-period returns “under continuous compounding,” we must first explain what this convention entails. Conceptually, continuous compounding is the result of discrete compounding when the corresponding time interval becomes infinitesimally small. Suppose an investor’s wealth is Y_0 , which he invests at a rate r for one period (let us say a month as in (a)). If the rate r is continuously compounded over this single period, the cumulative wealth consequence is as follows:

$$Y_0 \mapsto Y_0 \lim_{n \rightarrow \infty} \left(1 + \frac{r}{n}\right)^n = Y_0 e^r$$

(Continued)

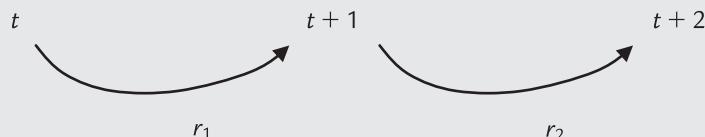
¹ Observe that the proposed index is not immune to the criticism discussed above: investment 3 ($E/\sigma = 1.182$) is inferior to investment 1 ($E/\sigma = 1.667$). Yet, we know that three dominates one because it pays a higher return in every state. This problem is pervasive with the mean–variance investment criterion: whatever the terms of the trade-off between mean and variance or standard deviation, one can produce a paradox such as the one illustrated above. Accordingly, this criterion is not generally applicable without additional restrictions. The index E/σ resembles, but is not identical to, the Sharpe ratio, $(E\hat{r} - r_f)/\sigma_{\hat{r}}$, where r_f denotes the risk-free rate.

BOX 3.1 Computing Means and Variances in Practice (Continued)

which exceeds the cumulative effect under discrete compounding ($Y_0 e^r > Y_0(1 + r)$ if $r > 0$). It also follows from this identification that if a one period rate r is continuously compounded for a succession of J periods, the cumulative wealth effect will be

$$Y_0 \xrightarrow{t} Y_0 \lim_{n \rightarrow \infty} \underbrace{\left\{ \left(1 + \frac{r}{n}\right)^n \left(1 + \frac{r}{n}\right)^n \cdots \left(1 + \frac{r}{n}\right)^n \right\}}_{J \text{ terms}} = Y_0 e^{Jr}$$

Lastly, if wealth Y_0 is invested successively at the continuously compounded discrete rates r_1 and r_2 , the cumulative effect will be



$$Y_0 \xrightarrow{t} Y_0 \lim_{n \rightarrow \infty} \left\{ \left(1 + \frac{r_1}{n}\right)^n \left(1 + \frac{r_2}{n}\right)^n \right\} = Y_0 e^{r_1 + r_2};$$

i.e., under continuous compounding rates of return may simply be added to give their cumulative effect. As we will see in subsequent chapters, this additive feature will allow various calculations to be simplified if the continuous compounding convention is assumed. Note also that our notation assigns the “age interpretation” to time periods: period 10, say, corresponds to the end of the 10th time interval just as a child’s 10th birthday is celebrated at the conclusion of his 10th year of life.

Given this setting, how are returns computed from discrete data under the continuous compounding convention? Again, let us assume the following data for some stock is presented to us:

$$t+j-1 \quad t+j$$

$$q_{j-1}^e \quad q_j^e + \text{div}_j$$

(Continued)

BOX 3.1 Computing Means and Variances in Practice (Continued)

We may now ask the question: what discrete rate of return $r_j^{e,\text{cont.}}$, *when continuously compounded*, was earned by this stock in the course of period j ? Equivalently, what rate of return $r_j^{e,\text{cont.}}$ satisfies:

$$q_{j-1}^e e^{r_j^{e,\text{cont.}}} = q_j^e + \text{div}_j?$$

Thus,

$$r_j^{e,\text{cont.}} = \ln\left(\frac{q_j^e + \text{div}_j}{q_{j-1}^e}\right) = \ln(1 + r_j^e)$$

In what follows in this book, the unspoken assumption is that reported return data is computed under the continuous compounding convention as per the above calculation. Accordingly, we generally do not employ the “cont.” superscript.²

Using the pattern of historical returns (however measured) to infer properties of the future return distribution makes sense if we think the “mechanism” generating these returns is “stationary”: that the future will in some sense closely resemble the past. In practice, this hypothesis is rarely fully verified and, at the minimum, it requires careful checking.³

² When returns are small, $r_j \approx \ln(1 + r_j)$. This is a standard approximation. Note that continuously compounded returns are the natural logarithm of discrete gross returns and, in this sense, are the continuously compounded cointegral to discrete gross returns.

³ The accuracy of the mean and variance estimates from historical data as stand-ins to the underlying return distribution’s true (future) mean and variance is a topic of great significance and to which we will return (c.f. Chapter 7).

3.3 A Prerequisite: Choice Theory Under Certainty

A good deal of financial economics is concerned with how people make choices. The objective is to understand the systematic part of individual behavior and to be able to predict (at least in a loose way) how an individual will react under specific economic circumstances. Economic theory describes individual behavior as the result of a process of optimization under constraints, the objective to be reached being determined by individual preferences, and the constraints being a function of the person’s income or wealth level and of market prices. This approach, which defines the *homo economicus* and the notion of **economic rationality**, is justified by the fact that individual behavior is predictable only to the extent that it is systematic, which must mean that there is an attempt to achieve a well-defined objective. It is not to be taken literally or normatively.⁴

⁴ By this we mean that economic science does not *prescribe* that individuals maximize, optimize, or simply behave as if they were doing so. It just finds it productive to summarize the systematic behavior of economic agents with such tools.

To develop this sense of rationality systematically, we begin by summarizing the objectives of investors in the most basic way: we postulate the existence of a preference relation, represented by the symbol \succeq , describing investors' ability to compare various bundles of goods and services. For two bundles a and b , the expression

$$a \succsim b$$

is to be read as follows: For the investor in question, bundle a is either strictly preferred to bundle b , or he is indifferent between them. Pure indifference is denoted by $a \sim b$, strict preference by $a > b$.

The notion of economic rationality can then be summarized by the following assumptions:

- A.1 Every investor possesses such a preference relation and it is *complete*, meaning that he is able to decide whether he prefers a to b , b to a , or both, in which case he is indifferent with respect to the two bundles. That is, for any two bundles a and b , either $a \geq b$ or $b \geq a$, or both. If both hold, we say that the investor is indifferent with respect to the bundles and write $a \sim b$.
- A.2 This preference relation satisfies the fundamental property of transitivity: For any bundles a , b , and c , if $a \geq b$ and $b \geq c$, then $a \geq c$.

A further requirement is also necessary for technical reasons:

- A.3 The preference relation \succeq is continuous in the following sense: Let $\{x_n\}$ and $\{y_n\}$ be two sequences of consumption bundles such that $x_n \mapsto x$ and $y_n \mapsto y$.⁵ If $x_n \geq y_n$ for all n , then the same relationship is preserved in the limit $x \geq y$.

A key result can now be expressed in the following proposition.

Theorem 3.1 Assumptions A.1 through A.3 are sufficient to guarantee the existence of a continuous, time-invariant, real-valued utility function⁶ u , such that for any two objects of choice (consumption bundles of goods and services),

$$\begin{aligned} a \succsim b &\quad \text{if and only if} \\ u(a) &\geq u(b). \end{aligned}$$

Proof See, for example, [Mas-Colell et al. \(1995\)](#), Proposition 3.C.1.

This result asserts that to endow decision makers with a utility function (which they are assumed to maximize) is, in reality, no different than to assume their preferences among objects of choice define a relation possessing the (weak) properties summarized in A.1 through A.3.

⁵ We use the standard sense of (normed) convergence in R^N .

⁶ In other words, $u: R^N \rightarrow R^+$.

Note that [Theorem 3.1](#) implies that if $u(\cdot)$ is a valid representation of an individual's preferences, any increasing transformation of $u(\cdot)$ is also valid since such a transformation, by definition, will preserve the ordering induced by $u(\cdot)$. Note also that the notion of a consumption bundle is, formally, very general. Different elements in a bundle may represent the consumption of the same good or service in different time periods. One element might represent a vacation trip in the Bahamas this year; another may represent exactly the same vacation next year. We can further expand our notion of different goods to include the same good consumed in mutually exclusive states of the world. Our preference for hot soup, for example, may be very different if the day is warm rather than cold. These thoughts suggest that [Theorem 3.1](#) is really quite general and can, formally at least, be extended to accommodate uncertainty. Under uncertainty, however, ranking bundles of goods (or vectors of monetary payoffs, see below) involves more than pure elements of taste or preferences. In the hot soup example, it is natural to suppose that our preferences for hot soup are affected by the probability we attribute to the day being hot or cold. Disentangling pure preferences from probability assessments is the subject to which we now turn.

3.4 Choice Theory Under Uncertainty: An Introduction

Under certainty, the choice is among consumption baskets with known characteristics. Under uncertainty, however, our emphasis changes. The objects of choice are typically no longer consumption bundles but vectors of state-contingent money payoffs (we will reintroduce consumption in Chapter 5). Such vectors are formally what we mean by an investment or an *asset* available for purchase. When we purchase a share of a stock, for example, we know that its sale price in one year will differ depending on what events transpire within the firm and in the world economy. Under financial uncertainty, therefore, the choice is among alternative investments leading to different possible income levels and, hence, ultimately different consumption possibilities. As before, we observe that people do make investment choices, and if we are to make sense of these choices, there must be a stable underlying order of preference defined over different alternative investments. The spirit of [Theorem 3.1](#) will still apply. With appropriate restrictions, these preferences can be represented by a utility index defined on investment possibilities, but obviously something deeper is at work. It is natural to assume that individuals have no intrinsic taste for the assets themselves (IBM stock as opposed to Royal Dutch Petroleum stock (hereafter RDS), for example). Rather, they are interested to know what payoffs these assets will yield and with what likelihood (see [Box 3.2](#), however).

BOX 3.2 Investing Close to Home

Although the assumption that investors only care for the final payoff of their investment without any trace of “romanticism” is standard in financial economics, there is some evidence to the contrary and, in particular, for the assertion that many investors, at the margin at least, prefer to purchase the claims of firms whose products or services are familiar to them.

In particular, [Huberman \(2001\)](#) examines the stock ownership records of the seven regional Bell operating companies (RBOCs) (due to a series of mergers, these seven firms have combined into presently two entities, Verizon and AT&T). He discovered that, with the exception of residents of Montana, Americans were more likely to invest in their local RBOC than in any other. When they did, their holdings averaged \$14,400. For those who ventured farther from home and hold stocks of the RBOC of a region other than their own, the average holding is only \$8246. Considering that every local RBOC cannot be a better investment choice than all of the other six, Huberman interprets his findings as suggesting investors’ psychological need to feel comfortable with where they put their money.

One may further hypothesize that investor preferences are indeed very simple after uncertainty is resolved: They prefer a higher monetary payoff to a lower one or, equivalently, to earn a higher return rather than a lower one. Of course they do not know *ex ante* (i.e., before the state of nature is revealed) which asset will yield the higher payoff. They have to choose among prospects, or probability distributions representing these payoffs. And, as we saw in [Section 3.2](#), typically, no one investment prospect will strictly dominate the others. Investors will be able to imagine different possible scenarios, some of which will result in a higher return for one asset, with other scenarios favoring other assets. For instance, let us go back to our favorite situation where there are only two states of nature; in other words, two conceivable scenarios and two assets, as seen in [Table 3.4](#).

There are two key ingredients in the choice between these two alternatives. The first is the probability of the two states. All other things being the same, the more likely is state 1, the more attractive IBM stock will appear to prospective investors. The second is the *ex post* (once the state of nature is known) level of utility provided by the investment. In [Table 3.4](#), IBM yields \$100 in state 1 and is thus preferred to RDS, which yields \$90 if this scenario is realized. RDS, however, provides \$160 rather than \$150 in state 2. Obviously, with

Table 3.4: Forecasted price per share in one period

	State 1	State 2
IBM	\$100	\$150
RDS	\$90	\$160
Current price of both assets is \$100.		

unchanged state probabilities, things would look different if the difference in payoffs were increased in one state as in [Table 3.5](#).

Here even if state 1 is slightly more likely, the superiority of RDS in state 2 makes it look more attractive. A more refined perspective is introduced if we go back to our first scenario but now introduce a third contender, Sony, with payoffs of \$90 and \$150, as seen in [Table 3.6](#).

Sony is dominated by both IBM and RDS. But the choice between the latter two can now be described in terms of an improvement of \$10 over the Sony payoff, either in state 1 or in state 2. Which is better? The relevant feature is that IBM adds \$10 when the payoff is low (\$90), while RDS adds the same amount when the payoff is high (\$150). Most people would think IBM more desirable, and with equal state probabilities, would prefer IBM. Once again this is an illustration of the preference for smooth consumption (smoother income allows for smoother consumption).⁷ In the present context, one may equivalently speak of risk aversion or of the well-known microeconomic assumption of decreasing marginal utility (the incremental utility steadily declines when adding ever more consumption or income).

The expected utility theorem provides a set of hypotheses under which an investor's preference ranking over investments with uncertain money payoffs may be represented by a utility index combining, in the most elementary way (i.e., linearly), the two ingredients just

Table 3.5: Forecasted price per share in one period

	State 1	State 2
IBM	\$100	\$150
RDS	\$90	\$200
Current price of both assets is \$100.		

Table 3.6: Forecasted price per share in one period

	State 1	State 2
IBM	\$100	\$150
RDS	\$90	\$160
Sony	\$90	\$150
Current price of all assets is \$100.		

⁷ Of course, for the sake of our reasoning, one must assume that nothing else important is going on simultaneously in the background, and that other things, such as income from other sources, if any, and the prices of the consumption goods to be purchased with the assets' payoffs, are unchanged irrespective of what the payoffs actually are.

discussed—the preference ordering on the ex post money payoffs and the respective probabilities of these payoffs.

We first illustrate this notion in the context of the two assets considered earlier. Let the respective probability distributions on the price per share of IBM and RDS be described, respectively, by $\tilde{p}_{\text{IBM}} = p_{\text{IBM}}(\theta_i)$ and $\tilde{p}_{\text{RDS}} = p_{\text{RDS}}(\theta_i)$ together with the probability π_i that the state of nature θ_i will be realized. In a two-state context, the expected utility theorem provides sufficient conditions on an agent's preferences over uncertain asset payoffs, denoted \succeq , such that there exists a function $\mathbb{U}(\)$, defined over uncertain asset payoffs, and an associated utility-of-money function $U(\)$ such that

- i. $\tilde{p}_{\text{IBM}} \succcurlyeq \tilde{p}_{\text{RDS}}$ if and only if $\mathbb{U}(\tilde{p}_{\text{IBM}}) \geq \mathbb{U}(\tilde{p}_{\text{RDS}})$ where
 - ii. $\mathbb{U}(\tilde{p}_{\text{IBM}}) = EU(\tilde{p}_{\text{IBM}}) = \pi_1 U(p_{\text{IBM}}(\theta_1)) + \pi_2 U(p_{\text{IBM}}(\theta_2))$
- $$> \pi_1 U(p_{\text{RDS}}(\theta_1)) + \pi_2 U(p_{\text{RDS}}(\theta_2)) = EU(\tilde{p}_{\text{RDS}}) = \mathbb{U}(\tilde{p}_{\text{RDS}})$$

More generally, for these preferences, the utility of any asset A with payoffs $p_A(\theta_1), p_A(\theta_2), \dots, p_A(\theta_N)$ in the N possible states of nature with probabilities $\pi_1, \pi_2, \dots, \pi_N$ can be represented as

$$\mathbb{U}(\tilde{p}_A) = EU(p_A(\theta_i)) = \sum_{i=1}^N \pi_i U(p_A(\theta_i))$$

In other words, by the weighted mean of ex post utilities using the state probabilities as weights. $\mathbb{U}(\tilde{p}_A)$ is a real number. Its precise numerical value, however, has no more meaning than if you are told that the temperature is 40° when you do not know if the scale being used is Celsius or Fahrenheit. It is useful, however, for comparison purposes.

By analogy, if it is 40° today, but it will be 45° tomorrow, you at least know it will be warmer tomorrow than it is today. Similarly, the expected utility number is useful because it permits attaching a number to a probability distribution and this number is, under appropriate hypotheses, a good representation of the relative ranking of a particular member of a family of probability distributions (assets under consideration).

3.5 The Expected Utility Theorem

We elect to discuss this theorem in the simple context where objects of choice take the form of simple lotteries. A generic lottery will be denoted (x, y, π) ; it offers payoff (consequence) x with probability π and payoff (consequence) y with probability $1 - \pi$. This notion of a lottery is actually very general and encompasses a huge variety of possible payoff structures. For example, x and y may represent specific monetary payoffs as in [Figure 3.1](#), or x may be a payment while y is a lottery as in [Figure 3.2](#), or even x and y may both be lotteries as in [Figure 3.3](#). Extending these possibilities, some or all of the x_i 's

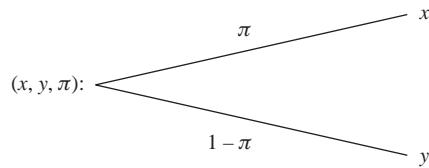


Figure 3.1
A simple lottery (x, y are monetary payoffs).

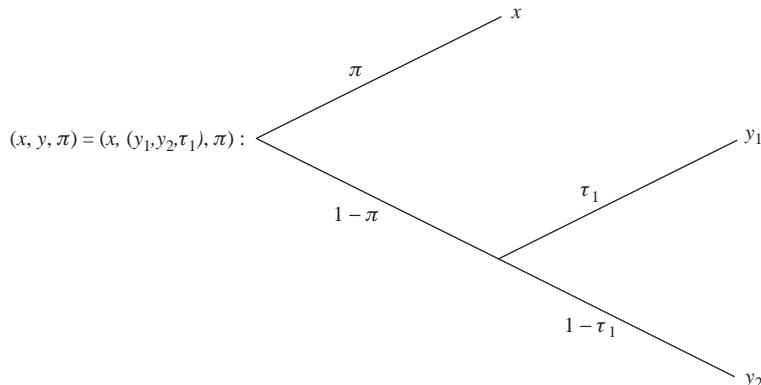


Figure 3.2
A compound lottery (y is itself a lottery).

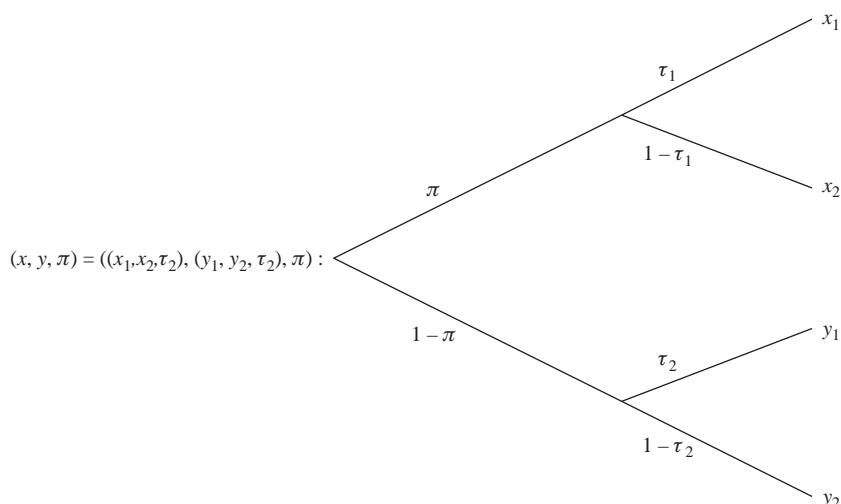


Figure 3.3
A compound lottery (both x and y are themselves lotteries).

and y_i 's may themselves be lotteries, and so on. We also extend our choice domain to include individual payments, lotteries where there is one, certain, monetary payoff; for instance,

$$(x, y, \pi) = x \text{ if (and only if) } \pi = 1 \text{ (see axiom C.1)}$$

Moreover, the theorem holds as well for assets paying a continuum of possible payoffs, but our restriction to discrete payoffs makes the necessary assumptions and justifying arguments easily accessible. Our objective is conceptual transparency rather than absolute generality. All the results extend to much more general settings.

Under these representations, we will adopt the following **axioms and conventions**:

- C.1. a. $(x, y, 1) = x$
- b. $(x, y, \pi) = (y, x, 1 - \pi)$
- c. $(x, z, \pi) = (x, y, \pi + (1 - \pi)\tau) \text{ if } z = (x, y, \tau)$

C.1c informs us that agents are concerned with the net cumulative probability of each outcome. Indirectly, it further accommodates lotteries with multiple outcomes; see [Figure 3.4](#), for an example with lotteries (x, y, π') , and $(z, w, \hat{\pi})$, where $\pi_1 = \pi'$, $\pi = \pi_1 + \pi_2$, etc.

- C.2. There exists a preference relation \succeq , defined on lotteries, which is complete and transitive.
- C.3. The preference relation is continuous in the sense of A.3 in [Section 3.3](#).

By C.2 and C.3 alone, we know ([Theorem 3.1](#)) that there exists a utility function, which we will denote by $\mathbb{U}()$, defined both on lotteries and on specific payments

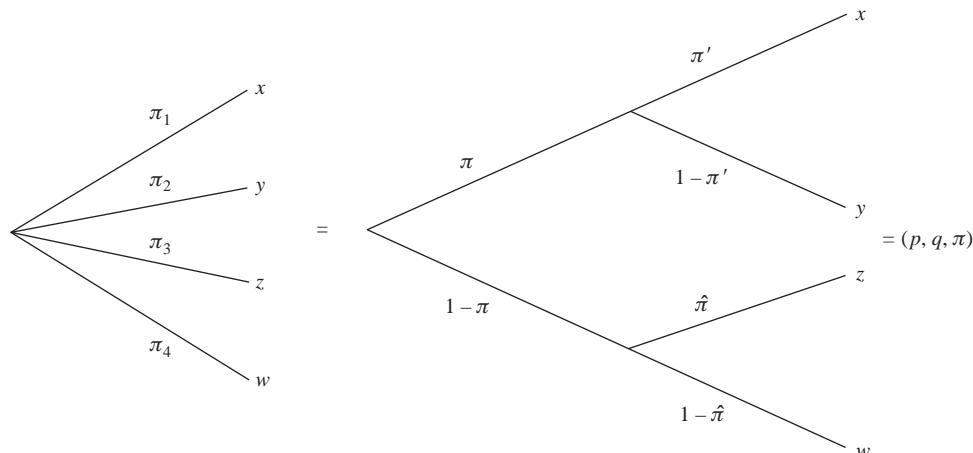


Figure 3.4

A lottery with multiple outcomes reinterpreted as a compound lottery.

since, by assumption C.1a, a payment may be viewed as a (degenerate) lottery. For any payment x , we identify

$$U(x) = \mathbb{U}((x, y, 1)) \quad (3.1)$$

Our remaining assumptions are thus necessary only to guarantee that \mathbb{U} assumes the expected utility form.

- C.4. Independence of irrelevant alternatives. Let (x, y, π) and (x, z, π) be any two lotteries; then, $y \succeq z$ if and only if $(x, y, \pi) \succeq (x, z, \pi)$.
- C.5. For simplicity, we also assume that there exists a best (i.e., most preferred lottery), b , as well as a worst, least desirable, lottery w .

In our argument to follow (which is constructive, i.e., we explicitly exhibit the expected utility function), it is convenient to use relationships that follow directly from these latter two assumptions. In particular, we will use C.6 and C.7:

- C.6. Let x, k, z be consequences or payoffs for which $x > k > z$. Then there exists a π such that $(x, z, \pi) \sim k$.
- C.7. Let $x > y$. Then $(x, y, \pi_1) \succeq (x, y, \pi_2)$ if and only if $\pi_1 > \pi_2$. This follows directly from C.4.

Theorem 3.2 Consider a preference ordering, defined on the space of lotteries, that satisfies axioms C.1 to C.7. Then there exists a utility function \mathbb{U} defined on the lottery space, with associated utility-of-money function $U()$, such that:

$$\mathbb{U}((x, y, \pi)) = \pi U(x) + (1 - \pi)U(y) \quad (3.2)$$

Proof We outline the proof in a number of steps:

By [Theorem 3.1](#) we know that $\mathbb{U}()$ exists with associated $U()$ as its restriction to certain monetary payments, defined as per [Eq. \(3.1\)](#). We must now show that $\mathbb{U}()$ and $U()$ are related by [Eq. \(3.2\)](#).

1. Without loss of generality, we may normalize $\mathbb{U}()$ so that $\mathbb{U}(b) = 1$, $\mathbb{U}(w) = 0$.
2. For all other lotteries z , define $\mathbb{U}(z) = \pi_z$ where π_z satisfies $(b, w, \pi_z) \sim z$
Constructed in this way $\mathbb{U}(z)$ is well defined since
 - a. by C.6, $\mathbb{U}(z) = \pi_z$ exists, and
 - b. by C.7, $\mathbb{U}(z)$ is unique. To see this latter implication, assume, to the contrary, that $\mathbb{U}(z) = \pi_z$ and also $\mathbb{U}(z) = \pi'_z$ where $\pi_z > \pi'_z$. By assumption C.4,
$$z \sim (b, w, \pi_z) \succ (b, w, \pi'_z) \sim z, \text{ a contradiction}$$
3. It follows also from C.7 that if $m > n$, $\mathbb{U}(m) = \pi_m > \pi_n = \mathbb{U}(n)$. Thus, $\mathbb{U}()$ has the property of a utility function.

4. Lastly, we want to show that $\mathbb{U}(\cdot)$ has the required property. Let x, y be monetary payments, π a probability. By C.1a, $U(x), U(y)$ are well-defined real numbers. By C.6,

$$\begin{aligned}(x, y, \pi) &\sim ((b, w, \pi_x), (b, w, \pi_y)), \pi \\ &\sim (b, w, \pi\pi_x + (1 - \pi)\pi_y), \text{ by C.1c.}\end{aligned}$$

Thus, by definition of $\mathbb{U}(\cdot)$,

$$\mathbb{U}((x, y, \pi)) = \pi\pi_x + (1 - \pi)\pi_y = \pi U(x) + (1 - \pi)U(y)$$

Although we have chosen x, y as monetary payments, the same conclusion holds if they are lotteries.

Before going on to refine our understanding of the expected utility theorem, it is important to be absolutely clear on terminology: First, the overall $\mathbb{U}(\cdot)$ is defined over lotteries. It is referred to as the von Neumann–Morgenstern (VNM) utility function, so named after the originators of the theory, the justly celebrated mathematicians John von Neumann and Oskar Morgenstern. In the construction of a VNM utility function, it is customary first to specify its restriction to certainty monetary payments, the so-called utility-of-money function $U(\cdot)$ or simply (and hereafter) the *utility function*. Note that the VNM utility function and its associated utility function are not the same. The VNM utility function is defined over uncertain asset payoff structures, while its associated utility function is defined over individual money payments.

The key identifier of the “expected utility” construct is that these two concepts are linearly related: either the utility function is linearly related to the VNM function via the probability weights or the VNM function is linearly related to the state probabilities, with weights being the state-by-state utility-of-money values. The second interpretation leads to the common expression that VNM-expected utility preferences are “linear in the probabilities.”

Given the objective specification of probabilities (thus far assumed), it is the utility function that uniquely characterizes an investor. As we will see shortly, different additional assumptions on $U(\cdot)$ will identify an investor’s tolerance for risk. We do, however, impose the maintained requirement that $U(\cdot)$ be increasing for all candidate utility functions (more money is preferred to less). Note also that the expected utility theorem confirms that investors are concerned only with an asset’s final payoffs and the cumulative probabilities of achieving them. For expected utility investors, the structure of uncertainty resolution is thus irrelevant (Axiom C.1c).⁸

Although the introduction to this chapter concentrates on comparing rates of return distributions, our expected utility theorem in fact gives us a tool for comparing different asset

⁸ See Section 5.7.1 for a generalization on this score.

payoff distributions. Without further analysis, it does not make sense to think of the utility function as being defined over a rate of return. This is true for a number of reasons.

First, returns are expressed on a per unit (per US\$, Swiss CHF, etc.) basis and do not identify the magnitude of the initial investment to which these rates are to be applied. We thus have no way to assess the implications of a return distribution for an investor's wealth position. It could, in principle, be anything. Second, the notion of a rate of return implicitly suggests a time interval: the payout is received after the asset is purchased. So far we have only considered the atemporal evaluation of uncertain investment payoffs. In Chapter 4, we generalize the VNM representation to preferences defined over rates of returns.

As in the case of a general order of preferences over bundles of commodities, the VNM-expected utility representation is preserved under linear transformations. If $\mathbb{U}(\cdot)$ is a von Neuman–Morgenstern utility function, then $\mathbb{V}(\cdot) = a\mathbb{U}(\cdot) + b$, where $a > 0$, is also such a function. To verify this assertion, let (x, y, π) be some uncertain payoff, and let $U(\cdot)$ be the utility-of-money function associated with \mathbb{U} .

$$\begin{aligned}\mathbb{V}((x, y, \pi)) &= a\mathbb{U}((x, y, \pi)) + b = a[\pi U(x) + (1 - \pi)U(y)] + b \\ &= \pi[a\mathbb{U}(x) + b] + (1 - \pi)[a\mathbb{U}(y) + b] = \pi\mathbb{V}(x) + (1 - \pi)\mathbb{V}(y)\end{aligned}$$

Every linear transformation of an expected utility function is thus also an expected utility function. The utility-of-money function associated with \mathbb{V} is $[aU(\cdot) + b]$; $\mathbb{V}(\cdot)$ represents the same preference ordering over uncertain payoffs as $\mathbb{U}(\cdot)$. On the other hand, a nonlinear transformation does not always respect the preference ordering. It is in that sense that utility is said to be **cardinal**.

Lastly, we need to clarify the direct connection between $U(\cdot)$ and $u(\cdot)$ (c.f., [Theorem 3.1](#)). Economic science recognizes that money has no value *per se*; its significance lies in the consumption goods that may be purchased with it. Accordingly, consider some financial asset (portfolio) that pays $(Y(\theta_1), \dots, Y(\theta_N))$, with $Y(\theta_i)$ denoting its money payoff in state θ_i , $i = 1, 2, \dots, N$. Suppose also that the investor who acquires this asset has available to him J distinct consumption goods in each of the states. The proportions of the consumption goods the investor elects to consume may differ across states reflecting potentially different state-contingent prices as denoted by $(P_1(\theta_i), P_2(\theta_i), \dots, P_J(\theta_i))$, where $P_j(\theta_i)$ is the price of good j in state θ_i .

Presuming the investor wishes to spend his money as wisely as possible irrespective of what state may be realized, we can define the utility-of-money function $U(Y(\theta_i))$ as the maximum level of consumption utility he may achieve in state i given his income $Y(\theta_i)$ and the consumption goods prices noted above; in effect, we define $U(Y(\theta_i))$ by

$$U(Y(\theta_i)) \equiv \underset{\{c_1(\theta_i), \dots, c_J(\theta_i)\}}{\text{def}} \max u(c_1(\theta_i), \dots, c_J(\theta_i)) \quad (3.3)$$

$$\text{s.t. } c_1(\theta_i)P_1(\theta_i) + \dots + c_J(\theta_i)P_J(\theta_i) \leq Y(\theta_i)$$

where $c_j(\theta_i)$ is the consumption of good j in state θ_i . The constraint is referred to as the investor's budget constraint. Note that $U(Y(\theta_i))$ subsumes three important quantities: the investor's relative preference for the different goods available in state θ_i (as per $u(\cdot)$), the relative prices of these goods, and the investor's state θ_i income, $Y(\theta_i)$.

A fuller treatment of this identification would also acknowledge that investors typically save some portion of their income. This consideration requires a multiperiod setting and comes to the fore beginning in Chapter 4.

3.6 How Restrictive Is Expected Utility Theory? The Allais Paradox

Although apparently innocuous, the above set of axioms has been hotly contested as representative of rationality. In particular, it is not difficult to find situations in which investor preferences violate the independence axiom. Consider the following four possible asset payoffs (lotteries):

$$\begin{aligned} L^1 &= (10,000, 0, 0.1) & L^2 &= (15,000, 0, 0.09) \\ L^3 &= (10,000, 0, 1) & L^4 &= (15,000, 0, 0.9) \end{aligned}$$

When investors are asked to rank these payoffs, the following ranking is frequently observed:

$$L^2 \succ L^1$$

(presumably because L^2 's positive payoff in the favorable state is much greater than L^1 's while the likelihood of receiving it is only slightly smaller) and

$$L^3 \succ L^4$$

(Here it appears that the certain prospect of receiving 10,000 is worth more than the potential of an additional 5000 at the risk of receiving nothing.)

By the structure of compound lotteries, however, it is easy to see that:

$$\begin{aligned} L^1 &= (L^3, L^0, 0.1) \\ L^2 &= (L^4, L^0, 0.1) \quad \text{where } L^0 = (0, 0, 1) \end{aligned}$$

By the independence axiom, the ranking between L^1 and L^2 , on the one hand, and L^3 and L^4 , on the other, should thus be identical!

This is the Allais paradox.⁹ There are a number of possible reactions to it.

1. Yes, my choices were inconsistent; let me think again and revise them.
2. No, I'll stick to my choices. The following kinds of things are missing from the theory of choice expressed solely in terms of asset payoffs:
 - the pleasure of gambling, and/or
 - the notion of regret.

The idea of regret is especially relevant to the Allais paradox, and its application in the prior example would go something like this. L^3 is preferred to L^4 because of the regret involved in receiving nothing if L^4 were chosen and the bad state ensued. We would, at that point, regret not having chosen L^3 , the certain payment. The expected regret is high because of the nontrivial probability (0.10) of receiving nothing under L^4 . On the other hand, the expected regret of choosing L^2 over L^1 is much smaller (the probability of the bad state is only 0.01 greater under L^2 , and in either case the probability of success is small), and insufficient to offset the greater expected payoff. Thus L^2 is preferred to L^1 .

The Allais paradox is but the first of many phenomena that appear to be inconsistent with standard preference theory. Another prominent example is the general pervasiveness of *preference reversals*, events that may approximately be described as follows. Individuals participating in controlled experiments were asked to choose between two lotteries, (4, 0, 0.9) and (40, 0, 0.1). More than 70% typically chose (4, 0, 0.9). When asked at what price they would be willing to sell the lotteries if they were to own them, however, a similar percentage demanded the higher price for (40, 0, 0.1). At first appearances, these choices would seem to violate transitivity. Let x , y be, respectively, the sale prices of (4, 0, 0.9) and (40, 0, 0.10). Then this phenomenon implies

$$x \sim (4, 0, 0.9) \succ (40, 0, 0.1) \sim y, \text{ yet } y \geq x$$

Alternatively, it may reflect a violation of the assumed principle of procedure invariance, which is the idea that investors' preference for different objects should be indifferent to the manner by which their preference is elicited. Surprisingly, more narrowly focused experiments, which were designed to force a subject with expected utility preferences to behave consistently, gave rise to the same reversals. The preference reversal phenomenon could thus, in principle, be due either to preference intransitivity or to a violation of the independence axiom, or of procedure invariance.

Through a series of carefully constructed experiments, some researchers have attempted to assign responsibility for preference reversals to procedure invariance violations. But this is a particularly alarming conclusion as [Thaler \(1992\)](#) notes. It suggests that “*the context and*

⁹ Named after the Nobel Prize-winner Maurice Allais who was the first to uncover the phenomenon. See [Allais \(1964\)](#).

procedures involved in making choices or judgements influence the preferences that are implied by the elicited responses. In practical terms this implies that (economic) behavior is likely to vary across situations which economists (would otherwise) consider identical.”

This is tantamount to the assertion that the notion of a preference ordering is not well defined. While investors may be able to express a consistent (and thus mathematically representable) preference ordering across television sets with different features (e.g., size of the screen and quality of the sound), this may not be possible with lotteries or consumption baskets containing widely diverse goods.

[Grether and Plott \(1979\)](#) summarize this conflict in the starker possible terms:

Taken at face value, the data demonstrating preference reversals are simply inconsistent with preference theory and have broad implications about research priorities within economics. The inconsistency is deeper than the mere lack of transitivity or even stochastic transitivity. It suggests that no optimization principles of any sort lie behind the simplest of human choices and that the uniformities in human choice behavior which lie behind market behavior result from principles which are of a completely different sort from those generally accepted.

At this point it is useful to remember, however, that the ultimate goal of financial economics is not to describe individual, but rather market, behavior. There is a real possibility that occurrences of seeming individual irrationality essentially “wash out” when aggregated at the market level. On this score, the proof of the pudding is in the eating and we have little alternative but to see the extent to which the basic theory of choice we are using is able to illuminate financial phenomena of interest. All the while, the discussion above should make us alert to the possibility that unusual phenomena might be the outcome of deviations from the generally accepted preference theory articulated above. While there is, to date, no preference ordering that accommodates preference reversals—and it is not clear there will ever be one—more general constructs than expected utility have been formulated to admit other, seemingly contradictory, phenomena. Further complications arise under collective choice; see [Box 3.3](#).

BOX 3.3 On the Rationality of Collective Decision Making

Although the discussion in the text pertains to the rationality of individual choices, it is a fact that many important decisions are the result of collective decision making. The limitations to such a process are important and, in fact, better understood than those arising at the individual level. It is easy to imagine situations in which transitivity is violated once choices result from some sort of aggregation over more basic preferences.

Consider three portfolio managers who decide which stocks to add to the portfolios they manage by majority voting. The stocks currently under consideration are General Electric

(Continued)

BOX 3.3 On the Rationality of Collective Decision Making (Continued)

(GE), Daimler (DAI), and Sony (S). Based on his fundamental research and assumptions, each manager has rational (i.e., transitive) preferences over the three possibilities:

Manager 1: $GE \succeq_1 DAI \succeq_1 S$

Manager 2: $S \succeq_2 GE \succeq_2 DAI$

Manager 3: $DAI \succeq_3 S \succeq_3 GE$

If they were to vote all at once, they know each stock would receive one vote (each stock has its advocate). So they decide to vote on pairwise choices: (GE versus DAI), (DAI versus S), and (S versus GE). The results of this voting (GE dominates DAI, DAI dominates S, and S dominates GE) suggest an intransitivity in the aggregate ordering. Although an intransitivity, it is one that arises from the operation of a collective choice mechanism (voting) rather than being present in the individual orders of preference of the participating agents. There is a large literature on this subject that is closely identified with Arrow's "Impossibility Theorem". See [Arrow \(1963\)](#) for a more exhaustive discussion.

3.7 Behavioral Finance

"The political man of the Greeks, the religious man of the Hebrews and Christians, the enlightened economic man of eighteenth century Europe (the original of that mythical present day character the 'good European') [have] been superseded by a new model for the conduct of life. Psychological man is... more native to American culture than the Puritan sources of that culture would indicate."¹⁰

The notion of a "rational investor" underlies most of financial theory and, indeed, most of what is presented in this book. A rational investor, very simply, is one with two essential attributes: (i) his preferences over random money payoffs are VNM-expected utility, as just described, and (ii) the probabilities he assigns to these payoffs are objective in that they incorporate all past and present information available to the investor in a manner that respects correct statistical procedure.¹¹ Unless specified otherwise, VNM-expected utility will represent our default context going forward. But despite the elegant and straightforward nature of the rational investor construct, there remain numerous empirical choice phenomena, which it cannot rationalize. How are they to be understood? One answer to this question lies in the domain of "behavioral finance".

¹⁰ Reiff (2006, page 48). Reiff (2006) is speaking of social trends, but it is not surprising that economic science would be similarly swept along.

¹¹ In equilibrium contexts, we will go one step further and strengthen the latter requirement to one of "rational expectations." This means that the probabilities objectively computed for the random payoffs in fact coincide with the payoffs' true probability distribution.

Behavioral finance is a theory-in-progress which seeks to fill this gap by departing from the rational investor assumptions in ways that are thought to better reflect various findings in experimental psychology. Most of the assumptions in behavioral finance have not been axiomatized in the context of choices-over-lotteries. Rather, they are supported by circumstantial empirical evidence. We give illustrations of several behavioral notions below. The difficulty in evaluating these concepts in the present context lies precisely in the absence of a formal axiomatic basis underlying them. In that sense we do not yet fully grasp “what they imply.”

3.7.1 **Framing**

“Framing” is simply the notion that individuals’ choices may be substantially influenced by the context in which they are presented. As a very simple illustration, would your decision to purchase a steak versus fish for dinner be different if the steak is advertised as:

- “90% fat free,” or
- “10% fat content”?

While *ex post* recognizing that these alternatives convey the same information, it seems apparent that the former description is more likely to elicit a positive “steak” decision for the majority of shoppers.¹²

The same phenomena appear to be present in investment choices. In a classic study, Kahneman and Tversky (1979) explore individual choices across the following lotteries:

- i. In the context of first being given \$1000, participants were asked to choose between the following lotteries A and B:
 - A: (\$1000, 0, .5)
 - B: (\$500, 0, 1)
- ii. In a context of first being given \$2000, these same participants were asked to choose between
 - C: (-\$1000, 0, .5)
 - D: (-\$500, 0, 1).¹³

These lotteries are summarized in Figure 3.5.

¹² Procedure invariance (prior section) and “framing” are not the same notion. Procedure invariance requires that an investor’s preference over lotteries be the same irrespective of whether they are directly compared or their certainty equivalents compared; i.e., the ranking is preserved under any methodology as to how the comparison is to be rendered (informally, irrespective of “how the problem is to be solved”). Framing concerns the context of the comparison, once a method of comparison has been chosen.

¹³ Kahneman and Tversky (1979) actually conduct their study with payoffs denominated in terms of Israeli currency (lira). At the time, the average monthly income was approximately 3000 lira.

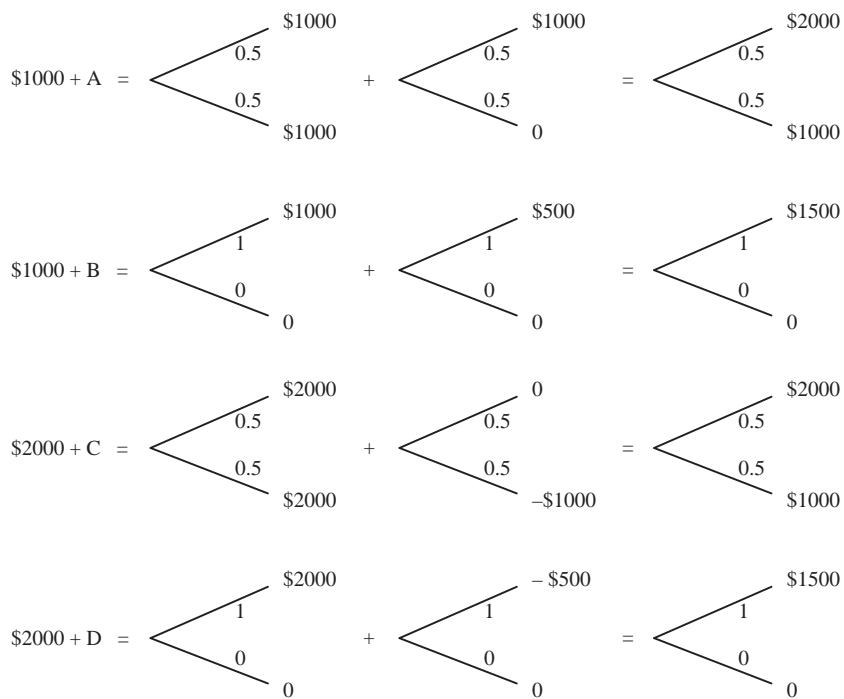


Figure 3.5
Four lotteries with preceding initial payments.

For a majority of those participating in the experiment, $B > A$ and $C > D$ despite the fact that A and C are equivalent as are B and D when taking full account of the differing initial payments.

How do we interpret these (inconsistent) choices? Apparently, it mattered to most participants that the choice between lotteries A and B was presented as a choice of *gains relative to \$1000* while in the second case the choices were presented as *losses relative to \$2000*. This distinction is viewed as a manifestation of the phenomenon of framing. Note that under VNM-expected utility, framing, as illustrated above, is irrelevant since only total wealth payoffs matter.

Framing is often cited as one factor potentially contributing to the failure of investors to diversify; i.e., to invest their wealth in portfolios of more than a few (one) assets. It is the idea that when investors consider the acquisition of various assets, they frame the decision on the basis of bilateral comparisons alone, without considering the interaction of multiple asset return patterns and the benefits that may follow. To illustrate this notion in the simple context of three assets and two states of nature θ_1 and θ_2 , consider the assets in Figure 3.6.

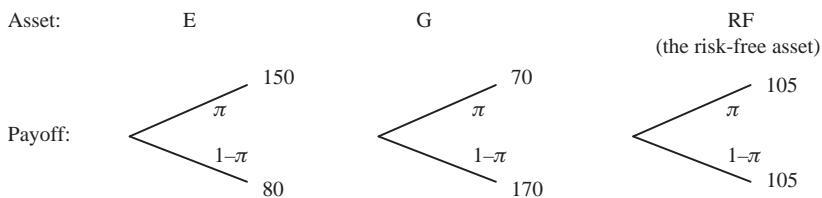


Figure 3.6
Three candidate assets.

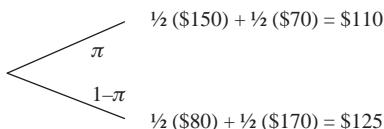


Figure 3.7
Payoff outcomes, portfolio $\{\frac{1}{2}E, \frac{1}{2}G\}$.

Assume $\pi = \frac{1}{2}$, and let the prices of the assets be $q_E = q_G = q_{RF} = \$100$. Under narrow framing, an investor may individually compare E and G to RF, find each individually less desirable (both E and G have large payoff variances and a substantial probability of significant loss), and end up with a portfolio composed exclusively of asset RF. Nevertheless, we see from Figure 3.7 that RF is clearly state by state *dominated* by the portfolio $\{\frac{1}{2}E, \frac{1}{2}G\}$.

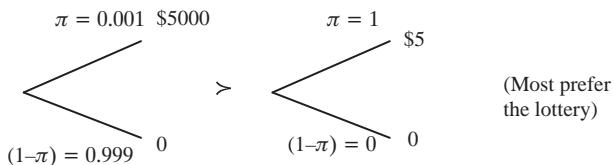
It is unclear what utility specification would eliminate all framing phenomena.

3.7.2 Prospect Theory

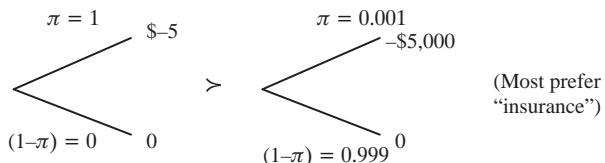
At present, Kahneman and Tversky's (1979, 1992) Prospect Theory is the most highly developed behavioral theory of choice.¹⁴ It rests on a number of experimental observations:

- i Consider the random payoff $(\$110, -\$100, \frac{1}{2})$. In experimental settings, a majority of participants declined to accept this lottery, irrespective of their level of personal wealth, even if it was offered to them at zero cost.
- ii Consider the four basic lotteries from Section 3.7.1
 - A: $(\$1000, 0, .5)$
 - B: $(\$500, 0, 1)$
 - C: $(-\$1000, 0, .5)$
 - D: $(-\$500, 0, 1)$.

¹⁴ Barberis (2013) provides a detailed overview of the theory and applications that have followed from it. This section owes much to him.



(in either case the expected payoff is \$5) and,



(In either case the expected payoff is -\$5.)

Figure 3.8
Preferences for lotteries and insurance.

When offered for comparison without any prior wealth distinctions (no issues of framing), a majority of participants displayed the following preference:

$$B > A \text{ and } C > D.$$

- iii Participants generally displayed a preference for both lotteries and insurance when offered in closely related choice settings; in particular, see [Figure 3.8](#).

Building on these and other observations [Kahneman and Tversky \(1979\)](#) propose a theory of choice under uncertainty (Prospect Theory) with four principal ingredients.

1. Investors ultimately derive utility not from their absolute wealth levels (as in the VNM-expected utility case) but from gains or losses relative to some reference or benchmark value. This (critical) element in their theory is suggested by observation (i): since at very high wealth levels, the acceptance or rejection of (\$110, \$-100, $\frac{1}{2}$) is really of little consumption consequence, its rejection at all wealth levels suggests that it is the gains or losses themselves, possibly relative to some preconceived benchmark, that really matter to investors. They thus propose a utility-of-money function of the form $U(Y - \bar{Y})$ where \bar{Y} is the benchmark. The benchmark can be thought of as either a minimally acceptable wealth level or, under the proper transformations, a cutoff rate of return. It can be changing through time reflecting prior experience. Unfortunately, Prospect Theory does not offer a general guide as to how the benchmark should be selected in any specific choice setting.
2. Since (\$110, \$-100, $\frac{1}{2}$) has a positive expected value, its rejection at all wealth levels also suggests that agents feel losses more acutely than gains (of greater comparative

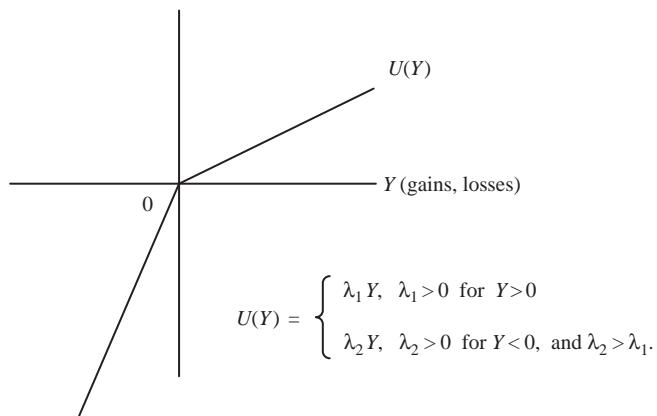


Figure 3.9
Loss-averse utility function.

magnitude). This is the sense of loss-averse preferences or simply “loss aversion.” The simplest illustration of a loss-averse utility-of-money function is seen in [Figure 3.9](#), where $\bar{Y} = 0$, a zero benchmark.

3. As a further refinement, consider the choices in observation (ii). The fact that most participants prefer lottery B to lottery A suggests dislike of risk over positive gain lotteries. The preference for lottery C over lottery D, however, suggests “risk loving” behavior over losses. As we will see in the next chapter these features imply that $U(Y - \bar{Y})$ is concave for $Y > \bar{Y}$ but convex for $Y < \bar{Y}$.

An illustration of a utility representation satisfying (1)–(3) is as follows: Let \bar{Y} denote the benchmark payoff and define the investor’s utility-of-money function $U(Y)$ by

$$U(Y) = \begin{cases} \frac{(|Y - \bar{Y}|)^{1-\gamma_1}}{1 - \gamma_1}, & \text{if } Y \geq \bar{Y} \\ -\lambda(|Y - \bar{Y}|)^{1-\gamma_2}, & \text{if } Y \leq \bar{Y} \end{cases}$$

where $\lambda > 1$ captures the extent of the investor’s aversion to “losses” relative to the benchmark, and $\gamma_1 > 0$ and $\gamma_2 > 0$ need not coincide (but $\gamma_1 \neq 1$, $\gamma_2 \neq 1$). In other words, the curvature of the function may differ for deviations above or below the benchmark. See [Figure 3.10](#) for an illustration. Clearly, all three features can have a large impact on the relative ranking of uncertain lottery payoffs.

4. There is a fourth attribute of Prospect Theory that also has its origins in observation. (iii) One interpretation of the choices found in that observation is that investors overweight low probability tail events, both favorable and unfavorable. Following on

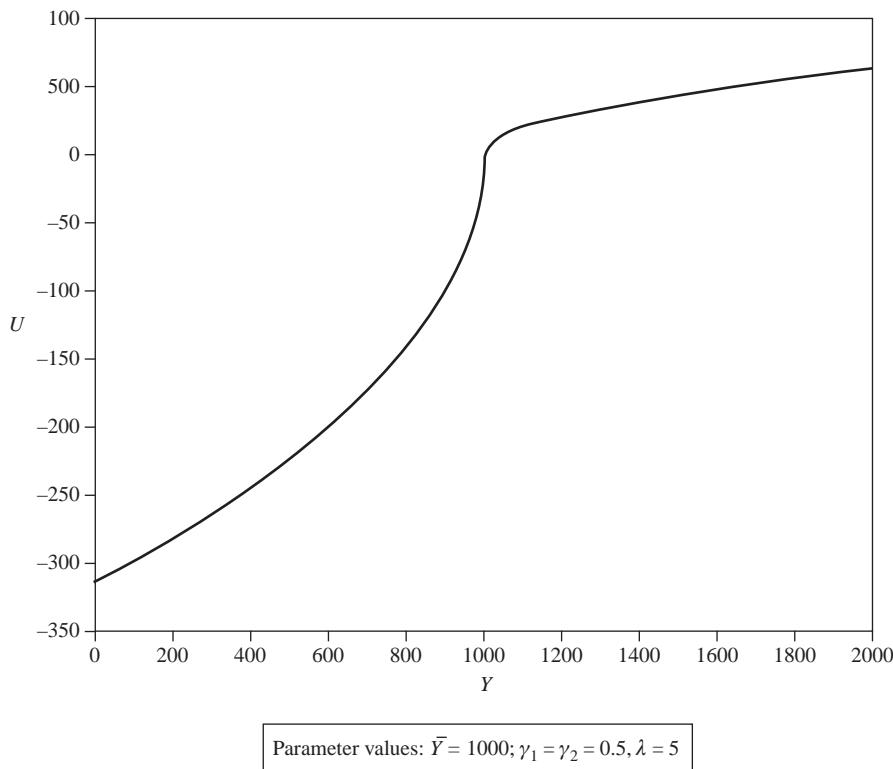
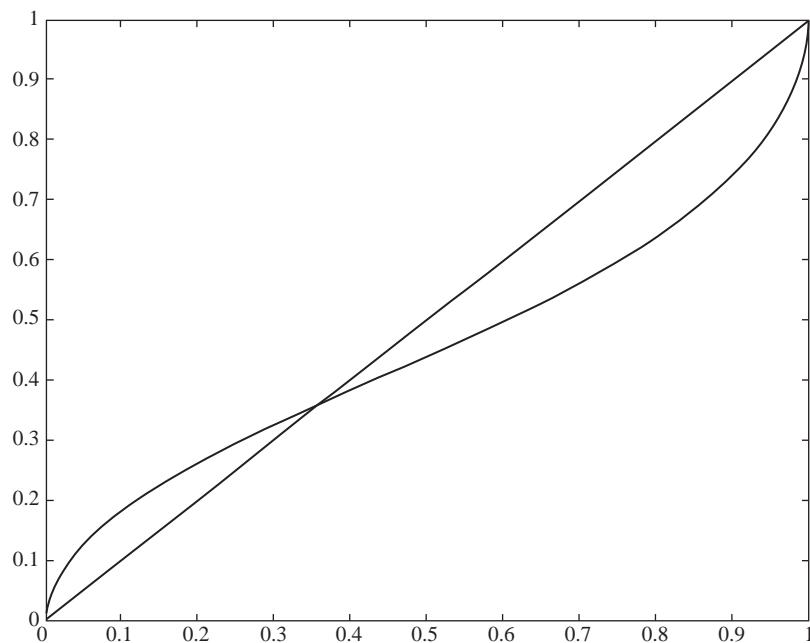


Figure 3.10
Utility function for Prospect Theory.

this possibility, Kahneman and Tversky (1979) elect to weigh the utilities of the various outcomes using a nonlinear function of the true probabilities which is asymmetric in a manner that gives a high weighting to tail events. These weightings are not necessarily to be interpreted as erroneous probability estimates, but as perhaps reflecting relative welfare consequences of the outcomes for the investor that are not observable.

See [Tversky and Kahneman \(1992\)](#) for a full discussion. A sample weighting function taken from [Tversky and Kahneman \(1992\)](#) is found in [Figure 3.11](#). Note that this weighting function resembles a probability distribution function where there is a large likelihood of “extreme” events.

As a paradigm for rationalizing laboratory observations (i)–(iv), Prospect Theory has no equal at the moment. But does it have any direct advantage over VNM-expected utility in explaining equilibrium market phenomena? There are at least two relevant works in this regard, Barberis and Huang (2008) and Benartzi and Thaler (1995). In the first paper, the authors show that the skewness in the return distribution of a common stock can have important pricing implications when investors’ loss-averse preferences are defined over

**Figure 3.11**

The probability weighting function. *Source: This figure is taken from Tversky and Kahneman (1992) as presented in Barberis (2013).*

changes in the value of their portfolios. In particular, securities with positively skewed return distributions are priced higher (and have lower average returns) and negatively skewed securities lower (and have higher average returns) than would be the case in a VNM-expected utility environment, a consequence that follows from the asymmetric weighting function fundamental to Prospect Theory. It is a feature that appears to be present in the cross section of security returns (see, for example, Boyer et al. (2010) and Conrad et al. (2013)).

The basic intuition in the Benartzi and Thaler (1995) paper is that loss-averse investors will dislike equity securities because stock market returns are much more highly dispersed relative to bond returns. This fact should lead, they argue, to a higher equity premium in an equilibrium setting where investors have loss-averse preferences than would be possible in the same environment where investors are VNM-expected utility. In a formal dynamic equilibrium quantitative model where investors have loss-averse preferences, Barberis and Huang (2001) go on to confirm the assertions of Benartzi and Thaler (1995), subject to qualifications.

There are many other utility forms that have been proposed over the past 30 years which are linked in some way to Prospect Theory. Some are widely employed in the research literature. We conclude this section by enumerating a few of them below.

3.7.2.1 Preference Orderings with Connections to Prospect Theory

- i. Survival benchmark: Investors evaluate lotteries according to $\mathbb{U}(\tilde{Y}) = EU(\tilde{Y} - \bar{Y})$ where \bar{Y} represents a minimum level of income for a decent lifestyle.
- ii. Habit formation preferences: Investors become accustomed to a particular income level and measure their utility by the extent to which their present income realization departs from it. For example,

$$\mathbb{U}(\tilde{Y}_t) = EU(\tilde{Y}_t - Y_{t-1})$$

where the habit is identified by the prior period's income (and associated consumption) level, thereby introducing a time dimension. See [Constantinides \(1990\)](#) and [Sundaresan \(1989\)](#).

- iii. “Keeping up with the Joneses” or relative income status: Once again investors evaluate lotteries according to

$$\mathbb{U}(\tilde{Y}) = EU(\tilde{Y} - \bar{Y})$$

except that \bar{Y} represents the average income level in the investor's reference community. See [Abel \(1990\)](#).¹⁵

- iv. “Disappointment aversion”; [Gul \(1991\)](#): Here we will need to be a bit more detailed in our representation of the expectations operator E . A disappointment averse investor evaluates lotteries according to

$$\mathbb{U}(\tilde{Y}) = \int_{\underline{Y}}^B u(Y)dF(Y) + AY \int_B^{\bar{Y}} u(Y)dF(Y)$$

where \underline{Y}, \bar{Y} denote, respectively, the minimum and maximum payoffs associated with \tilde{Y} , A is a number $0 < A < 1$, and B is a certain payment (a certainty equivalent) in exchange for which the investor would be willing to sell the lottery \tilde{Y} . With $A < 1$, the investor effectively weighs payments above this benchmark level less heavily than ones below it—a sort of indirect “loss” aversion. He is, essentially, more concerned with low-value outcomes, low in the sense of falling short of the amount for which the investor would have been willing to sell the lottery. Routledge and Zinn (2003) modify the original [Gul \(1991\)](#) representation in a way that endogenizes the construction of B so as to make low payoff realizations even more painful than in [Gul \(1991\)](#).

- v. The notion of regret; [Loomis and Sugden \(1982\)](#): The idea here is that if an investor selects one lottery over another, then his utility benefit of a particular state's payment will be diminished had the payoff to the rejected asset in that same state been higher; i.e., given the realized state, the investor experiences regret for not having chosen the

¹⁵ In all these three cases, the benchmark is calculated so that the utility function $U(\cdot)$ is always defined over a positive quantity.

other asset. More formally, in deciding which of lotteries \tilde{Y}_1 and \tilde{Y}_2 to choose, the investor computes (for \tilde{Y}_1 ; \tilde{Y}_2 is evaluated symmetrically):

$$\begin{aligned}\mathbb{U}(\tilde{Y}_1) &= EU(\tilde{Y}_1) + ER(\tilde{Y}_1 - \tilde{Y}_2) \\ &= \sum_{i=1}^N \pi(\theta_i) u(Y_1(\theta_i)) + \delta \sum_{i=1}^N \pi(\theta_i) R(Y_1(\theta_i) - Y_2(\theta_i))\end{aligned}$$

where i indexes the states, $\delta > 0$ and, like $U(\cdot)$, $R(\cdot)$, the regret function, is a monotone, strictly increasing function which satisfies $R(0) = 0$ and $-R(-\xi) = R(\xi)$ for any $\xi > 0$. $R(\cdot)$ ¹⁶ diminishes expected utility in those states where \tilde{Y}_2 has the higher outcome. In this sense the alternative asset's payoff serves as the benchmark on a state-by-state basis. [Loomes and Sugden \(1982\)](#) demonstrate that regret preferences can rationalize many of the choice anomalies presented in [Kahneman and Tversky \(1979\)](#). As one might expect, however, regret preferences do not satisfy the VNM transitivity axiom which has the implication that, *per se*, they cannot necessarily be used to isolate the best (highest expected utility) of a collection of eligible lotteries.

Preference representations (i)–(v) are a representative sample of “what’s out there”. Except for disappointment aversion, all lack a full axiomatic basis grounded in choices-over-lotteries. It is also not clear what the corresponding “reverse engineered” preferences over consumption goods would look like, a problem that is typically sidestepped by assuming one composite consumption good with a normalized price of one so that income and consumption are always numerically equal. In all cases, however, the takeaway is the same: investors evaluate gains and losses relative to a benchmark and do so asymmetrically in a way that tends to “overweight losses” relative to expected utility.

3.7.3 Overconfidence

A variety of studies find evidence that suggests pervasive overconfidence among physicians, nurses, attorneys, engineers, and others.¹⁷ [Monitor \(2007\)](#) finds, in a survey, that more than 70% of professional portfolio managers regard the service they provide as “above average.” These various studies measure overconfidence in different ways, consistent with the difficulties inherent in making the notion precise in individual contexts. A more formal study again suggestive of overconfidence is [Barber and Odean \(2000\)](#); see also [Odean \(1998, 1999\)](#). Using data from 78,000 individually managed accounts at a large discount brokerage company, these authors find that these accounts substantially underperform various commonplace benchmarks largely because of the large transaction

¹⁶ Alternatively, $R(\cdot)$ could assume the form $R(Y_1(\theta_i) - Y_2(\theta_i)) = \min\{0, (Y_1(\theta_i) - Y_2(\theta_i))\}$, etc.

¹⁷ Some references are [Baumann et al. \(1991\)](#), [Wagenaar and Kern \(1986\)](#), [Russo and Schoemaker \(1992\)](#), and [De Bont and Thaler \(1990\)](#) for, respectively, nurses, attorneys, high-level managers, and portfolio managers.

costs attendant to frequent trading. They interpret these findings as suggesting that individual investors are overconfident in their ability to pick “winning stocks.”

It is not clear how the notion of overconfidence can be modeled in a framework where agent preferences are represented by utility functions of some type. One possible approach is to borrow from Prospect Theory as regards subjective weightings on the various possible outcomes. In particular, we might expect that an overconfident investor is one who assigns lower weightings to negative outcomes than does the typical investor or who assumes the information he has collected regarding a stock’s future return distribution is more precise than it actually is. This latter approach is taken by Daniel et al. (1998), who study the consequences of overconfidence in an equilibrium security pricing model. These authors find that overconfidence, modeled in this way, can lead to “momentum”-like effects: price increases today followed, on average, by further price increases tomorrow, and vice versa.

3.8 Conclusions

The expected utility theory is the workhorse of choice theory under uncertainty. It will be put to use systematically in this book, as it is in most of financial theory. We have argued in this chapter that the expected utility construct provides a straightforward, intuitive mechanism for comparing uncertain asset payoff structures. As such, it offers a well-defined procedure for ranking the assets themselves.

Two ingredients are necessary for this process:

1. An estimate of the probability distribution governing the asset’s uncertain payments. While it is not trivial to estimate this quantity, it must also be estimated for the much simpler and less flexible mean/variance criterion.
2. An estimate of the agent’s utility-of-money function; it is the latter that fully characterizes his preference ordering. How this can be identified is one of the topics of the next chapter.

Behavioral theories represent plausible, experimentally based deviations from the axioms underlying expected utility. They make us aware of departures from “rationality” that have potential implications for explaining phenomena that are difficult to account for in the standard VNM framework.

References

- Abel, A., 1990. Asset pricing under habit formation and catching up with the Joneses. *Am. Econ. Rev. Pap. Proc.* 80, 38–42.
- Allais, M., 1964. Le comportement de l’homme rationnel devant le risque: Critique des postulats de l’école Américaine. *Econometrica*. 21, 503–546.
- Arrow, K.J., 1963. *Social Choice and Individual Values*. Yale University Press, New Haven, CT.

- Barberis, N., 2013. Thirty years of prospect theory in economics: a review and assessment. *J. Econ. Perspect.* 27, 173–196.
- Barberis, N., Huang, M., 2001. Mental accounting, loss aversion, and individual stock returns. *J. Finan.* 56, 1247–1292.
- Barberis, N., Huang, M., 2008. Stocks as lotteries: the implications of probability weighting for security prices. *Am. Econ. Rev.* 98, 2066–2100.
- Barber, B., Odean, T., 2000. Trading is hazardous to your wealth: the common stock performance of individual investors. *J. Finan.* 55, 773–806.
- Baumann, A., Deber, R., Thompson, G., 1991. Overconfidence among physicians and nurses: the macro-certainty, micro-uncertainty phenomenon. *Soc. Sci. Med.* 32, 167–174.
- Benartzi, S., Thaler, R., 1995. Myopic loss aversion and the equity premium puzzle. *Q. J. Econ.* 110, 73–92.
- Boyer, B., Mitton, T., Vorkink, K., 2010. Expected idiosyncratic skewness. *Rev. Financ. Stud.* 23, 169–202.
- Conrad, J., Dittmar, R., Ghysels, E., 2013. Ex ante skewness and expected stock returns. *J. Finan.* 68, 85–124.
- Constantinides, G., 1990. Habit formation: a resolution of the equity premium puzzle. *J. Polit. Econ.* 98, 519–543.
- Daniel, K., Hirshleifer, D., Subrahmanyam, A., 1998. Investor psychology and security market under- and overreactions. *J. Finan.* 53, 1839–1885.
- DeBont, W., Thaler, R., 1990. Do security analysts overreact? *Am. Econ. Rev.* 80, 52–57.
- Grether, D., Plott, C., 1979. Economic theory of choice and the preference reversal phenomenon. *Am. Econ. Rev.* 75, 623–638.
- Gul, F., 1991. A theory of disappointment aversion. *Econometrica*. 59, 667–686.
- Huberman, G., 2001. Familiarity breeds investment. *Rev. Financ. Stud.* 14, 659–680.
- Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica*. 47, 263–291.
- Loomes, G., Sugden, R., 1982. Regret theory: an alternative theory of rational choice under uncertainty. *Econ. J.* 92, 805–824.
- Mas-Colell, A., Whinston, M.D., Green, J.R., 1995. Microeconomic Theory. Oxford; New York: Oxford University Press.
- Monitor, J., 2007. Behavioral Investing: A Practitioner's Guide to Applying Behavioral Finance. John Wiley & Sons, West Sussex.
- Odean, T., 1998. Are investors reluctant to realize their losses? *J. Finan.* 53, 1775–1798.
- Odean, T., 1999. Do investors trade too much? *Am. Econ. Rev.* 89, 1279–1298.
- Reiff, P., 1966. The Triumph of the Therapeutic, New York: Harper and Row, 1966; quote taken from the reprinted edition, ISI Books: Wilmington, Delaware (2006).
- Routledge, B., Zin, S., 2003. Generalized appointment aversion and asset prices, NBER Working Paper no. 10107.
- Russo, J., Schoemaker, P., 1992. Managing overconfidence. *Sloan Manage. Rev.* 33, 7–17.
- Sundaresan, S., 1989. Intertemporally dependent preferences and the volatility of consumption and wealth. *Rev. Finan. Stud.* 2, 73–89.
- Thaler, R.H., 1992. The Winner's Curse. Princeton University Press, Princeton, NJ.
- Tversky, A., Kahneman, D., 1992. Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* 5, 297–323.
- Wagenaar, W., Keren, G., 1986. Does the expect know? the reliability of predictions and confidence ratios of experts. In: Hollnagel, E., Mancini, G., Woods, D.D. (Eds.), Intelligent Decision Support in Process Environments. Springer, Berlin.

Measuring Risk and Risk Aversion

Chapter Outline

4.1. Introduction	87
4.2. Measuring Risk Aversion	87
4.3. Interpreting the Measures of Risk Aversion	90
4.3.1. Absolute Risk Aversion and the Odds of a Bet	90
4.3.2. Relative Risk Aversion in Relation to the Odds of a Bet	92
4.3.3. Risk Neutral Investors	93
4.4. Risk Premium and Certainty Equivalence	94
4.5. Assessing the Degree of Relative Risk Aversion	97
4.6. The Concept of Stochastic Dominance	98
4.7. Mean Preserving Spreads	102
4.8. An Unsettling Observation About Expected Utility	105
4.9. Applications: Leverage and Risk	106
4.9.1. An Example	108
4.9.2. Is Leverage a Good Thing?	109
4.9.3. An Application to Executive Compensation	111
4.10. Conclusions	112
References	113
Appendix: Proof of Theorem 4.2	113

4.1 Introduction

We argued in Chapter 1 that the desire of investors to avoid risk, i.e., to smooth their consumption across states of nature and, for that reason, to avoid variations in the value of their portfolio holdings, is one of the primary motivations for financial contracting. But we have not thus far imposed restrictions on the VNM (von Neumann–Morgenstern) expected utility representation of investor preferences, which necessarily guarantee such behavior. For that to be the case, it must be further specialized. This is the subject of the present chapter to define, precisely, the notion of risk aversion and then to discuss its implications for the form of $U(\cdot)$.

4.2 Measuring Risk Aversion

What does the term *risk aversion* imply about an agent's utility function? Consider a financial contract where the potential investor either receives an amount of money h with

probability $\frac{1}{2}$ or must pay an amount h with probability $\frac{1}{2}$. Our most basic sense of risk aversion must imply that for any level of personal wealth Y , a risk-averse investor would not wish to own such a security. In utility terms, this must mean

$$U(Y) > \left(\frac{1}{2}\right)U(Y+h) + \left(\frac{1}{2}\right)U(Y-h) = EU(\tilde{Y})$$

where the expression on the right-hand side of the inequality sign is the VNM-expected utility associated with the random wealth levels:

$$Y+h, \text{ probability } = \frac{1}{2}$$

$$Y-h, \text{ probability } = \frac{1}{2}$$

This inequality can only be satisfied for all wealth levels Y if the agent's utility function has the form suggested in [Figure 4.1](#). When this is the case, we say the utility function is strictly concave.

The important characteristic implied by this and similarly shaped utility functions is that the slope of the graph of the function decreases as the agent becomes wealthier (as Y increases). That is, the marginal utility (MU), represented by the derivative $d(U(Y))/d(Y) \equiv U'(Y)$, decreases with greater Y . Equivalently, for twice differentiable utility functions, $d^2(U(Y))/d(Y)^2 \equiv U''(Y) < 0$. For this class of functions, the latter is indeed a necessary and sufficient condition for risk aversion.

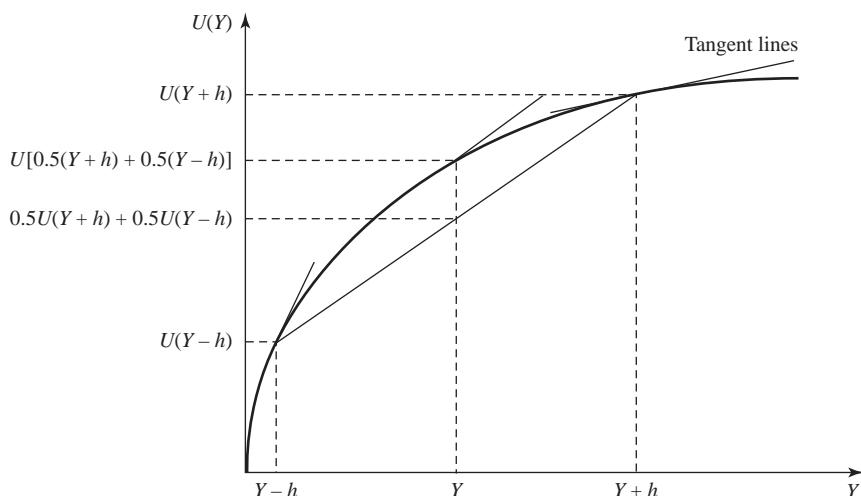


Figure 4.1
A strictly concave utility function.

As the discussion indicates, both consumption smoothing and risk aversion are directly related to the notion of decreasing MU of wealth. Whether it is envisaged across time or states, decreasing MU basically implies that income (or consumption) deviations from a fixed average level diminish rather than increase utility. Essentially, the positive deviations do not help as much as the negative ones hurt.

Risk aversion can also be represented in terms of indifference curves. [Figure 4.2](#) illustrates the case of a simple situation with two states of nature. If consuming c , i.e., c_1 in state 1 and c_2 in state 2, represents a certain level of expected utility $EU = k_1$ and consuming c^* , i.e., c_1^* in state 1 and c_2^* in state 2, permits achieving the same level of expected utility, then the convex-to-the-origin indifference curve that is the appropriate translation of a strictly concave utility function indeed implies that the expected utility level generated by the average consumption is $(c + c^*)/2$ in both states (in this case a certain consumption level) is larger ($=k_2$) than k_1 .

We would like to be able to measure the degree of an investor's aversion to risk. This will allow us to compare whether one investor is more risk averse than another and to understand how an investor's risk aversion affects his investment behavior (e.g., the composition of his portfolio).

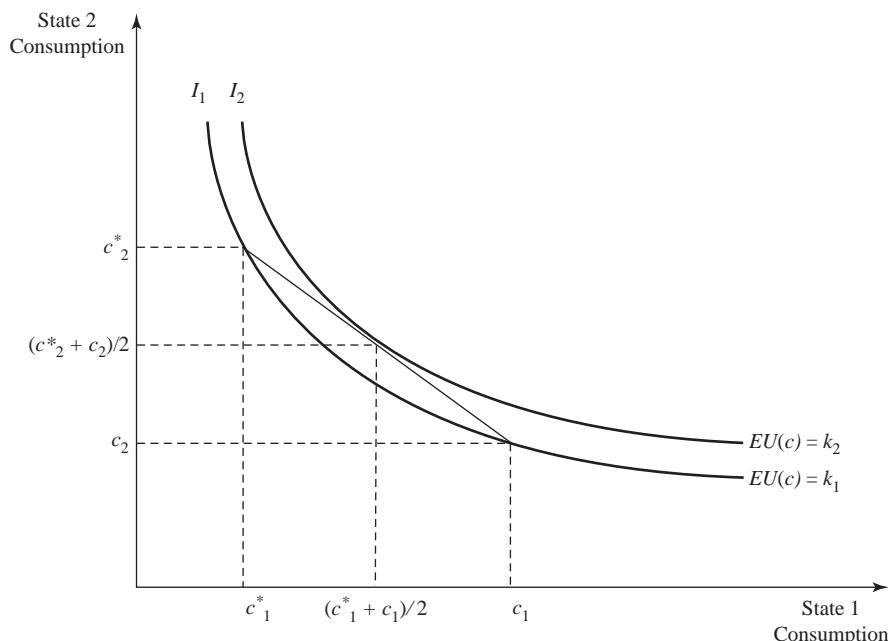


Figure 4.2
Indifference curves.

As a first attempt toward this goal, and since $U''(\cdot) < 0$ implies risk aversion, why not simply say that investor A is more risk averse than investor B , if and only if $|U''_A(Y)| \geq |U''_B(Y)|$, for all income levels Y ? Unfortunately, this approach leads to the following inconsistency. Recall that the preference ordering described by a utility function is invariant to linear transformations. In other words, suppose $U_A(\cdot)$ and $\bar{U}_A(\cdot)$ are such that $\bar{U}_A(\cdot) = a + bU_A(\cdot)$ with $b > 0$. These utility functions describe the identical ordering and thus must display identical risk aversion. Yet, if we use the above measure, we have

$$|\bar{U}''_A(Y)| > |U''_A(Y)|, \text{ if } b > 1$$

This implies that investor A is more risk averse than he is himself, which must be a contradiction.

We therefore need a measure of risk aversion that is invariant to linear transformations. Two widely used measures of this sort have been proposed by, respectively, [Pratt \(1964\)](#) and [Arrow \(1971\)](#):

- (i) absolute risk aversion = $U''(Y)/U'(Y) \equiv R_A(Y)$
- (ii) relative risk aversion = $YU''(Y)/U'(Y) \equiv R_R(Y)$.

Both of these measures have simple behavioral interpretations. Note that instead of speaking of risk aversion, we could use the inverse of the measures proposed above and speak of risk tolerance. This latter terminology may be preferable on some occasions.

4.3 Interpreting the Measures of Risk Aversion

4.3.1 Absolute Risk Aversion and the Odds of a Bet

Consider an investor with wealth level Y who is offered—at no charge—an investment involving winning or losing an *amount* h , with probabilities π and $1 - \pi$, respectively. Note that any investor will accept such a bet if π is high enough (especially if $\pi = 1$) and reject it if π is small enough (surely if $\pi = 0$). Presumably, the willingness to accept this opportunity will also be related to his level of current wealth, Y . Let $\pi = \pi(Y, h)$ be that probability at which the agent is indifferent between accepting or rejecting the investment. It is shown below that

$$\pi(Y, h) \cong \frac{1}{2} + \left(\frac{1}{4}\right)hR_A(Y) \tag{4.1}$$

where the symbol \cong represents “is approximately equal to.”

The higher his measure of absolute risk aversion, the more favorable odds the investor will demand in order to be willing to accept the investment. If $R_A^1(Y) \geq R_A^2(Y)$, for agents 1 and

2, respectively, then investor 1 will always demand more favorable odds than investor 2, and in this sense investor 1 is more risk averse.

It is useful to examine the magnitude of this probability. Consider, for example, the family of utility-of-money functions of the form

$$U(Y) = -\frac{1}{v} e^{-vY} \quad (4.2)$$

where v is a parameter.

For this case,

$$\pi(Y, h) \cong \frac{1}{2} + \left(\frac{1}{4}\right)hv$$

In other words, the odds requested are independent of the level of initial wealth (Y). On the other hand, the more wealth at risk (h), the greater the odds of a favorable outcome demanded. This expression advances the parameter v as the appropriate measure of the degree of *absolute* risk aversion for these preferences.

Let us now derive Eq. (4.1). By definition, $\pi(Y, h)$ must satisfy

$$\underbrace{U(Y)}_{\text{utility if he foregoes the bet}} = \underbrace{\pi(Y, h)U(Y + h) + [1 - \pi(Y, h)]U(Y - h)}_{\text{expected utility if the investment is accepted}} \quad (4.3)$$

By an approximation (Taylor's Theorem), we know that:

$$U(Y + h) = U(Y) + hU'(Y) + \frac{h^2}{2}U''(Y) + H_1$$

$$U(Y - h) = U(Y) - hU'(Y) + \frac{h^2}{2}U''(Y) + H_2$$

where H_1, H_2 are remainder terms of order higher than h^2 . Substituting these quantities into Eq. (4.3) gives

$$\begin{aligned} U(Y) &= \pi(Y, h) \left[U(Y) + hU'(Y) + \frac{h^2}{2}U''(Y) + H_1 \right] \\ &\quad + (1 - \pi(Y, h)) \left[U(Y) - hU'(Y) + \frac{h^2}{2}U''(Y) + H_2 \right] \end{aligned} \quad (4.4)$$

Collecting terms gives

$$U(Y) = U(Y) + (2\pi(Y, h) - 1)[hU'(Y)] + \frac{h^2}{2}U''(Y) + \underbrace{\pi(Y, h)H_1 + (1 - \pi(Y, h))H_2}_{=_{\text{def}} H(\text{small})}$$

Since the remainder term H is small—it is a weighted average of terms of order higher than h^2 and is, thus, itself of order higher than h^2 —it can be ignored.

Solving for $\pi(Y, h)$ then yields

$$\pi(Y, h) = \frac{1}{2} + \frac{h}{4} \left[\frac{-U''(Y)}{U'(Y)} \right] \quad (4.5)$$

Utility functions for which $R_A(Y)$ is constant are referred to as displaying constant absolute risk aversion (CARA), with Eq. (4.2) being a principal case in point: $R_A(Y) = \nu$.

4.3.2 Relative Risk Aversion in Relation to the Odds of a Bet

Consider now an investment opportunity similar to the one just discussed except that the amount at risk is a *proportion* of the investor's wealth. In other words, $h = \theta Y$, where θ is the fraction of wealth at risk. By a derivation almost identical to the one presented above, it can be shown that

$$\pi(Y, \theta) \cong \frac{1}{2} + \frac{1}{4}\theta R_R(Y) \quad (4.6)$$

If $R_R^1(Y) \geq R_R^2(Y)$, for investors 1 and 2, then investor 1 will always demand more favorable odds, for any level of wealth, when the fraction θ of his wealth is at risk.

It is also useful to illustrate this measure by an example. Another popular family of VNM utility functions (for reasons to be detailed in the next chapters) has the form¹

$$U(Y) = \frac{Y^{1-\gamma}}{1-\gamma}, \text{ for } 0 < \gamma \text{ and } \gamma \neq 1 \quad (4.7)$$

$$U(Y) = \ln Y, \text{ if } \gamma = 1$$

¹ Utility-of-money functions $U(Y)$ arise as the solution to problems of the form (3.3). Consider a utility of consumption function $u(c_1, c_2)$, defined over two goods c_1 and c_2 of the form $u(c_1, c_2) = -c_1^{-\gamma_1}c_2^{-\gamma_2}$, where $\gamma_1 > 0$ and $\gamma_2 > 0$. The resulting $U(Y)$ function has the form of (4.7): $U(Y) = \kappa(p_1, p_2, \gamma_1, \gamma_2)1/(1-\gamma)Y^{1-\gamma}$, where $\gamma = 1 + \gamma_1 + \gamma_2$ and $\kappa()$ is a function dependent on the consumption goods prices and the γ_1, γ_2 parameters. It is a positive constant from the investor's perspective (he takes prices as "given"). See the web notes to this chapter for details.

In the latter case, the probability expression becomes

$$\pi(Y, \theta) \approx \frac{1}{2} + \frac{1}{4}\theta$$

In this case, the requested odds of winning are not a function of initial wealth (Y) but depend upon θ , the fraction of wealth that is at risk: the lower the fraction θ , the more investors are willing to consider entering into a bet that is close to being fair (a risky opportunity where the probabilities of success or failure are both $\frac{1}{2}$). In the former, more general, case the analogous expression is

$$\pi(Y, \theta) \approx \frac{1}{2} + \frac{1}{4}\theta\gamma \quad (4.8)$$

Since $\gamma > 0$, these investors demand a higher probability of success. Furthermore, if $\gamma_2 > \gamma_1$, the investor characterized by $\gamma = \gamma_2$ will always demand a higher probability of success than will an agent with $\gamma = \gamma_1$, for the same fraction of wealth at risk. In this sense, a higher γ denotes a greater degree of *relative* risk aversion for this investor class.

Utility-of-money functions for which $R_R(Y)$ is constant are said to display constant relative risk aversion (CRRA). Utility form Eq. (4.7) is the work horse representative of this class ($R_R(Y) = \gamma$).

4.3.3 Risk Neutral Investors

One class of investors deserves special mention at this point. They are significant, as we shall later see, for the influence they have on the financial equilibria in which they participate. This is the class of investors who are risk neutral and who are identified with utility functions of a linear form

$$U(Y) = cY + d$$

where c and d are constants and $c > 0$.

Both of our measures of the degree of risk aversion, when applied to this utility function give the same result:

$$R_A(Y) \equiv 0 \text{ and } R_R(Y) \equiv 0$$

Whether measured as a proportion of wealth or as an absolute amount of money at risk, such investors do not demand better than even odds when considering risky investments of the type under discussion. They are indifferent to risk and are concerned only with an asset's expected payoff.

4.4 Risk Premium and Certainty Equivalence

The context of our discussion thus far has been somewhat artificial because we were seeking especially convenient probabilistic interpretations for our measures of risk aversion. More generally, a risk-averse agent ($U''(\cdot) < 0$) will always value an investment at something less than the expected value of its payoffs. Consider an investor, with current wealth Y , evaluating an uncertain risky payoff \tilde{Z} . For any distribution function F_z ,

$$U(Y + E\tilde{Z}) \geq E[U(Y + \tilde{Z})]$$

provided that $U''(\cdot) < 0$. This is a direct consequence of a standard mathematical result known as Jensen's inequality.

Theorem 4.1 (Jensen's inequality) Let $g(\cdot)$ be a concave function on the interval (a,b) , and \tilde{x} be a random variable such that $\text{Prob}\{\tilde{x} \in (a, b)\} = 1$. Suppose the expectations $E(\tilde{x})$ and $Eg(\tilde{x})$ exist, then

$$E[g(\tilde{x})] \leq g[E(\tilde{x})]$$

Furthermore, if $g(\cdot)$ is strictly concave and $\text{Prob}\{\tilde{x} = E(\tilde{x})\} \neq 1$, then the inequality is strict.

This theorem applies irrespective of whether the interval (a,b) on which $g(\cdot)$ is defined is finite or infinite. If a and b are both finite, the interval can be open or closed at either endpoint. If $g(\cdot)$ is convex, the inequality is reversed. See [De Groot \(1970\)](#).

To put it differently, if an uncertain payoff is available for sale, a risk-averse agent will only be willing to buy it at a price less than its expected payoff. This statement leads to a pair of useful definitions. The (maximal) certain sum of money a person is willing to pay to acquire an uncertain opportunity defines his **certainty equivalent** (CE) for that risky prospect; the difference between the CE and the expected value of the prospect is a measure of the uncertain payoff's **risk premium**. It represents the maximum amount the agent would be willing to pay to avoid the investment or gamble.

Let us make this notion more precise. The context of the discussion is as follows. Consider an agent with current wealth Y and utility function $U(\cdot)$ who has the opportunity to acquire an uncertain investment \tilde{Z} with expected value $E\tilde{Z}$. The CE of the risky investment \tilde{Z} , $\text{CE}(Y, \tilde{Z})$, and the corresponding risk or insurance premium, $\Pi(Y, \tilde{Z})$, are the solutions to the following equations:

$$EU(Y + \tilde{Z}) = U(Y + \text{CE}(Y, \tilde{Z})) \tag{4.9a}$$

$$= U(Y + E\tilde{Z} - \Pi(Y, \tilde{Z})) \tag{4.9b}$$

which implies

$$\text{CE}(Y, \tilde{Z}) = E\tilde{Z} - \Pi(Y, \tilde{Z}) \quad \text{or} \quad \Pi(Y, \tilde{Z}) = E\tilde{Z} - \text{CE}(Y, \tilde{Z})$$

These concepts are illustrated in [Figure 4.3](#).

It is intuitively clear that there is a direct relationship between the size of the risk premium and the degree of risk aversion of a particular individual. This link can be made quite easily. For simplicity, the derivation that follows applies to the case of an actuarially fair prospect \tilde{Z} , one for which $E\tilde{Z} = 0$. Using Taylor series approximations, we can develop the left-hand side (LHS) and right-hand side (RHS) of the definitional [Eqs \(4.9a\)](#) and [\(4.9b\)](#).

LHS:

$$\begin{aligned} EU(Y + \tilde{Z}) &= EU(Y) + E[\tilde{Z}U'(Y)] + E\left[\frac{1}{2}\tilde{Z}^2U''(Y)\right] + EH(\tilde{Z}^3) \\ &= U(Y) + \frac{1}{2}\sigma_{\tilde{Z}}^2U''(Y) + EH(\tilde{Z}^3) \end{aligned}$$

RHS:

$$U(Y - \Pi(Y, \tilde{Z})) = U(Y) - \Pi(Y, \tilde{Z})U'(Y) + H(\Pi^2)$$

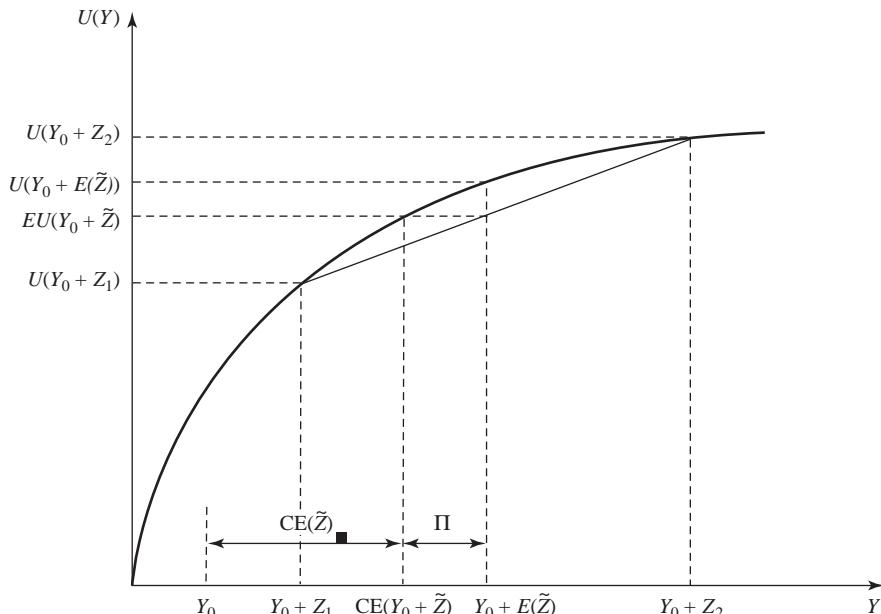


Figure 4.3
CE and the risk premium: an illustration.

or, ignoring the terms of order Z^3 or Π^2 or higher ($EH(\tilde{Z}^3)$ and $H(\Pi^2)$),

$$\Pi(Y, \tilde{Z}) \cong \frac{1}{2} \sigma_{\tilde{Z}}^2 \left(\frac{-U''(Y)}{U'(Y)} \right) = \frac{1}{2} \sigma_{\tilde{Z}}^2 R_A(Y)$$

To illustrate, consider our earlier example in which $U(Y) = (Y^{1-\gamma})/(1-\gamma)$, and suppose $\gamma = 3$, $Y = \$500,000$, and

$$\tilde{Z} = \begin{cases} \$100,000 & \text{with probability } \frac{1}{2} \\ -\$100,000 & \text{with probability } \frac{1}{2} \end{cases}$$

For this case, the approximation specializes to

$$\Pi(Y, \tilde{Z}) = \frac{1}{2} \sigma_{\tilde{Z}}^2 \frac{\gamma}{Y} = \frac{1}{2} (100,000)^2 \left(\frac{3}{500,000} \right) = \$30,000$$

To confirm that this approximation is a good one, we must show that

$$U(Y - \Pi(Y, \tilde{Z})) = U(500,000 - 30,000) \cong \frac{1}{2} U(600,000) + \frac{1}{2} U(400,000) = EU(Y + \tilde{Z})$$

or

$$(4.7)^{-2} \cong \frac{1}{2}(6)^{-2} + \frac{1}{2}(4)^{-2}$$

or

$$0.0452694 \cong 0.04513; \text{ confirmed}$$

Note also that for this preference class, the insurance premium is directly proportional to the parameter γ .

Can we convert these ideas into statements about rates of return? Let the equivalent risk-free return be defined by

$$U(Y(1 + r_f)) = U(Y + CE(Y, \tilde{Z}))$$

The random payoff \tilde{Z} can also be converted into a rate of return distribution via $\tilde{Z} = \tilde{r}Y$, or, $\tilde{r} = \tilde{Z}/Y$. Therefore, r_f is defined by the equation

$$U(Y(1 + r_f)) \equiv EU(Y(1 + \tilde{r}))$$

By risk aversion, $E\tilde{r} > r_f$. We thus define the utility-specific rate of return risk premium Π' as $\Pi' = E\tilde{r} - r_f$, or $E\tilde{r} = r_f + \Pi'$, where Π' depends on the degree of risk aversion of the agent in question. We conclude this section by computing the rate of return premium in a particular case. Suppose that $U(Y) = \ln Y$ and that the random payoff \tilde{Z} satisfies

$$\tilde{Z} = \begin{cases} \$100,000 & \text{with probability } \frac{1}{2} \\ -\$50,000 & \text{with probability } \frac{1}{2} \end{cases}$$

from a base of $Y = \$500,000$. The risky rate of return implied by these numbers is clearly

$$\tilde{r} = \begin{cases} 20\% & \text{with probability } \frac{1}{2} \\ -10\% & \text{with probability } \frac{1}{2} \end{cases}$$

with an expected return of 5%. The $CE(Y, \tilde{Z})$ must satisfy

$$\ln(500,000 + CE(Y, \tilde{Z})) = \frac{1}{2}\ln(600,000) + \frac{1}{2}\ln(450,000), \text{ or}$$

$$CE(Y, \tilde{Z}) = e^{\frac{1}{2}\ln(600,000) + \frac{1}{2}\ln(450,000)} - 500,000$$

$$CE(Y, \tilde{Z}) = 19,618, \text{ so that}$$

$$(1 + r_f) = \frac{519,618}{500,000} = 1.0392$$

The utility-specific rate of return risk premium is thus $5 - 3.92\% = 1.08\%$. Let us be clear: This rate of return risk premium does not represent a market or equilibrium premium.² Rather, it reflects personal preference characteristics and corresponds to the premium over the risk-free rate necessary to compensate, utility-wise, a specific individual, with the postulated preferences and initial wealth, for engaging in the risky investment.

4.5 Assessing the Degree of Relative Risk Aversion

Suppose that agents' utility functions are of the form $U(Y) = (Y^{1-\gamma})/(1-\gamma)$ class. As noted earlier, a quick calculation informs us that $R_R(Y) \equiv \gamma$, and we say that $U(\)$ is of the CRRA

² Accordingly, we use a different symbol, Π' , than the one used in Chapter 2 for the market risk premium (π).

class. To get a feeling as to what this measure means, consider the following uncertain payoff:

$$\begin{cases} \$50,000 \text{ with probability } \pi = 0.5 \\ \$100,000 \text{ with probability } \pi = 0.5 \end{cases}$$

Assuming your utility function is of the type just noted, what would you be willing to pay for such an opportunity (i.e., what is the CE for this uncertain prospect) if your current wealth were Y ? The interest in asking such a question resides in the fact that, given the amount you are willing to pay, it is possible to infer your coefficient of relative risk aversion $R_R(Y) = \gamma$, provided your preferences are adequately represented by the postulated functional form. This is achieved with the following calculation.

The CE, the maximum amount you are willing to pay for this prospect, is defined by the equation

$$\frac{(Y + \text{CE})^{1-\gamma}}{1 - \gamma} = \frac{\frac{1}{2}(Y + 50,000)^{1-\gamma}}{1 - \gamma} + \frac{\frac{1}{2}(Y + 100,000)^{1-\gamma}}{1 - \gamma}$$

Assuming zero initial wealth ($Y = 0$), we obtain the following sample results (clearly, $\text{CE} > 50,000$):

$\gamma = 0$	$\text{CE} = 75,000$ (risk neutrality)
$\gamma = 1$	$\text{CE} = 70,711$
$\gamma = 2$	$\text{CE} = 66,667$
$\gamma = 5$	$\text{CE} = 58,566$
$\gamma = 10$	$\text{CE} = 53,991$
$\gamma = 20$	$\text{CE} = 51,858$
$\gamma = 30$	$\text{CE} = 51,209$

Alternatively, if we suppose a current wealth of $Y = \$100,000$ and a degree of risk aversion of $\gamma = 5$, the equation results in a $\text{CE} = \$66,532$.

4.6 The Concept of Stochastic Dominance

In response to dissatisfaction with the standard ranking of risky prospects based on mean and variance, a theory of choice under uncertainty with general applicability has been developed. In this section, we show that the postulates of expected utility lead to a definition of two weaker alternative concepts of dominance with wider applicability than the concept of state-by-state dominance. These are of interest because they circumscribe the situations in which rankings among risky prospects are preference free, or can be defined independently of the specific trade-offs (among return, risk, and other characteristics of probability distributions) implicit in the form of an agent's utility function.

Table 4.1: Sample investment alternatives

Payoffs	10	100	2000
Probability Z_1	0.4	0.6	0
Probability Z_2	0.4	0.4	0.2
$EZ_1 = 64, \sigma_{z_1} = 44$ $EZ_2 = 444, \sigma_{z_2} = 779$			

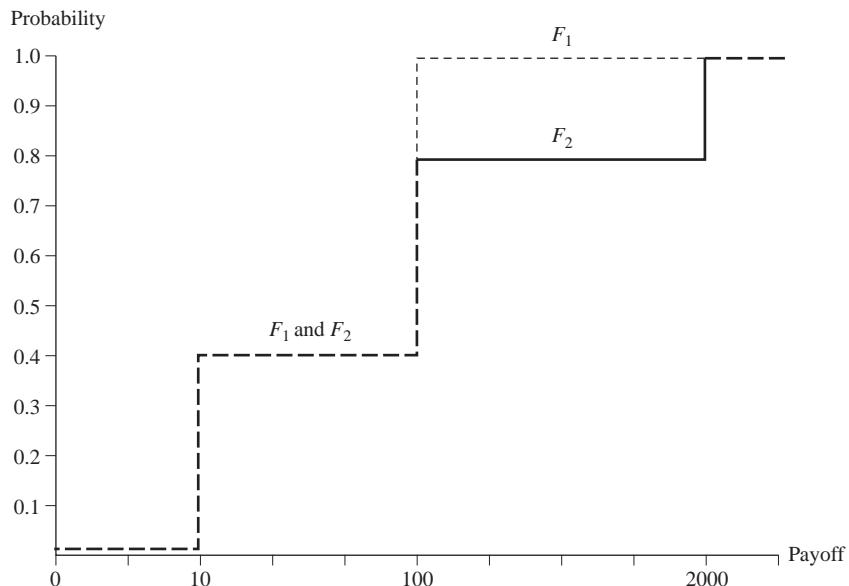


Figure 4.4
An example of FSD.

We start with an illustration. Consider two investment alternatives, \tilde{Z}_1 and \tilde{Z}_2 , with the characteristics outlined in [Table 4.1](#).

First, observe that under standard mean–variance analysis, these two investments cannot be ranked. Although investment \tilde{Z}_2 has the greater mean, it also has the greater variance. Yet, all of us would clearly prefer to own investment 2. It at least matches investment 1 and has a positive probability of exceeding it.

To formalize this intuition, let us examine the cumulative probability distributions associated with each investment, $F_1(\bar{Z})$ and $F_2(\bar{Z})$ where $F_i(\bar{Z}) = \text{Prob}(\tilde{Z}_i \leq \bar{Z})$.

In [Figure 4.4](#), we see that $F_1(\cdot)$ always lies above $F_2(\cdot)$. This observation leads to [Definition 4.1](#).

Definition 4.1 Let $F_A(\tilde{x})$ and $F_B(\tilde{x})$, respectively, represent the cumulative distribution functions of two random variables (cash payoffs) that, without loss of generality, assume

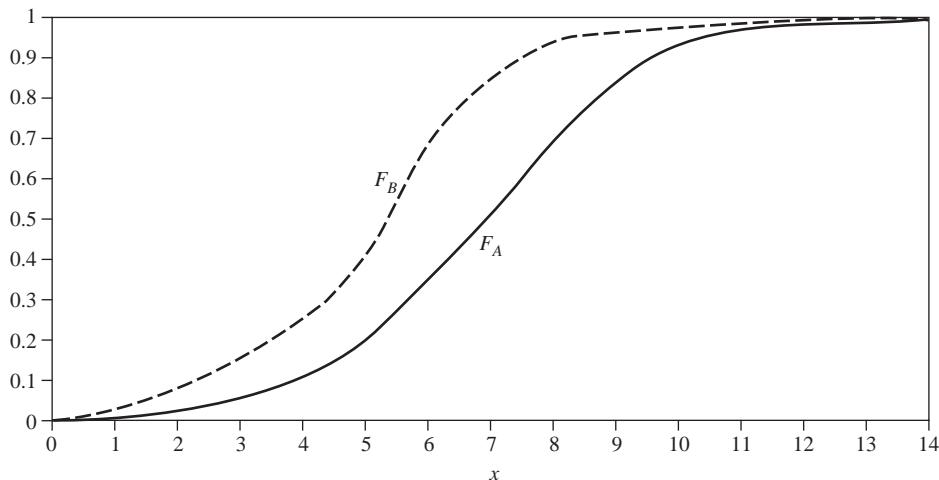


Figure 4.5
FSD: a more general representation.

values in the interval $[a,b]$. We say that $F_A(\tilde{x})$ first-order stochastically dominates (FSD) $F_B(\tilde{x})$ if and only if $F_A(x) \leq F_B(x)$ for all $x \in [a,b]$.

Distribution A in effect assigns more probability to higher values of x ; in other words, higher payoffs are more likely. Accordingly, the distribution functions of A and B generally conform to the following pattern: if F_A FSD F_B , then F_A is everywhere below and to the right of F_B as represented in Figure 4.5. By this criterion, investment 2 in Figure 4.4 first order stochastically dominates investment 1. It should, intuitively, be preferred. Theorem 4.2 summarizes our intuition in this latter regard.

Theorem 4.2 Let $F_A(\tilde{x})$ and $F_B(\tilde{x})$ be two cumulative probability distributions for random payoffs $\tilde{x} \in [a, b]$. Then $F_A(\tilde{x})$ FSD $F_B(\tilde{x})$ if and only if $E_A U(\tilde{x}) \geq E_B U(\tilde{x})$ for all nondecreasing utility functions $U(\cdot)$.

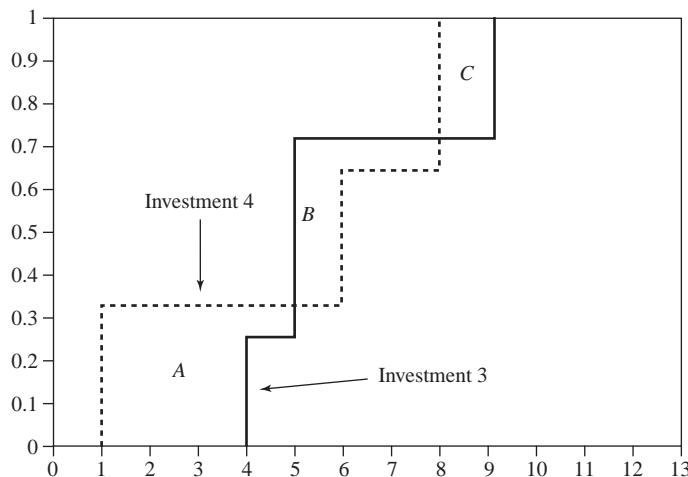
Proof See the Appendix.

Although it is not equivalent to state-by-state dominance (see Exercise 4.8), FSD is an extremely strong condition. As is the case with the former, state-by-state dominance, it is so strong a concept that it induces only a very incomplete ranking among uncertain prospects. Can we find a broader measure of comparison, for instance, which would make use of the hypothesis of risk aversion as well? Consider the two independent investments in Table 4.2.³

³ In this example, contrary to the previous one, the two investments considered are statistically independent.

Table 4.2: Two independent investments

Investment 3		Investment 4	
Payoff	Probability	Payoff	Probability
4	0.25	1	0.33
5	0.50	6	0.33
9	0.25	8	0.33

Figure 4.6
SSD illustrated.

Which of these investments is better? Clearly, neither investment (first order) stochastically dominates the other, as Figure 4.6 confirms. The probability distribution function corresponding to investment 3 is not everywhere below the distribution function of investment 4. Yet, we would probably prefer investment 3. Can we formalize this intuition (without resorting to the mean/variance criterion, which in this case accords with intuition: $ER_4 = 5$, $ER_3 = 5.75$; $\sigma_4 = 2.9$, and $\sigma_3 = 1.9$)? This question leads to a weaker notion of stochastic dominance that explicitly compares distribution functions.

Definition 4.2 Second-order stochastic dominance (SSD) Let $F_A(\tilde{x})$ and $F_B(\tilde{x})$ be two cumulative probability distributions for random payoffs in $[a,b]$. We say that $F_A(\tilde{x})$ SSD $F_B(\tilde{x})$ if and only if for any x :

$$\int_{-\infty}^x [F_B(t) - F_A(t)]dt \geq 0$$

(with strict inequality for some meaningful interval of values of t).

Table 4.3: Investment 3 second-order stochastically dominates investment 4

Values of x	$\int_0^x f_x(t)dt$	$\int_0^x F_3(t)dt$	$\int_0^x f_4(t)dt$	$\int_0^x F_4(t)dt$	$\int_0^x [F_4(t) - F_3(t)]dt$
0	0	0	0	0	0
1	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
2	0	0	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$
3	0	0	$\frac{1}{3}$	1	1
4	0.25	0.25	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{13}{12}$
5	0.75	1	$\frac{1}{3}$	$\frac{5}{3}$	$\frac{2}{3}$
6	0.75	1.75	$\frac{2}{3}$	$\frac{7}{3}$	$\frac{7}{12}$
7	0.75	2.5	$\frac{2}{3}$	3	$\frac{1}{2}$
8	0.75	3.25	1	4	$\frac{3}{4}$
9	1	4.25	1	5	0.75
10	1	5.25	1	6	0.75
11	1	6.25	1	7	0.75
12	1	7.25	1	8	0.75
13	1	8.25	1	9	0.75

The calculations in [Table 4.3](#) reveal that, in fact, investment 3 SSD investment 4 (let $f_i(x)$, $i = 3, 4$, denote the density functions corresponding to the cumulative distribution function $F_i(x)$). In geometric terms ([Figure 4.6](#)), this would be the case as long as area B is smaller than area A .

As [Theorem 4.3](#) shows, this notion makes sense, especially for risk-averse agents:

Theorem 4.3 Let $F_A(\tilde{x})$ and $F_B(\tilde{x})$ be two cumulative probability distributions for random payoffs \tilde{x} defined on $[a, b]$. Then, $F_A(\tilde{x})$ SSD $F_B(\tilde{x})$ if and only if $E_A U(\tilde{x}) \geq E_B U(\tilde{x})$ for all nondecreasing and concave U .

Proof See [Laffont \(1989\)](#), Section 2.5.

That is, all risk-averse agents will prefer the SSD asset. Of course, FSD implies SSD: if for two investments Z_1 and Z_2 , Z_1 FSD Z_2 , then it is also true that Z_1 SSD Z_2 . But the converse is not true.⁴

4.7 Mean Preserving Spreads

[Theorems 4.2 and 4.3](#) attempt to characterize a notion of “better/worse” that is relevant for probability distributions or random variables representing investment payoffs. But there are two aspects to such a comparison: the notion of “more or less risky” and the trade-off between risk and return. Let us now attempt to isolate the former effect by comparing only

⁴ It turns out that SSD is a concept fundamental to understanding the benefits to diversification (see Chapter 6).

those probability distributions with identical means. We will then review [Theorem 4.3](#) in the context of this latter requirement.

The concept of more or less risky is captured by the notion of a mean preserving spread. In our context, this notion can be informally stated as follows: Let $f_A(x)$ and $f_B(x)$ describe, respectively, the probability density functions on payoffs to assets A and B . If $f_B(x)$ can be obtained from $f_A(x)$ by removing some of the probability weight from the center of $f_A(x)$ and distributing it to the tails in such a way as to leave the mean unchanged, we say that $f_B(x)$ is related to $f_A(x)$ via a **mean preserving spread**. [Figure 4.7](#) suggests what this notion would mean in the case of normal-type distributions with identical mean, yet different variances.

How can this notion be made both more intuitive and more precise? Consider a set of possible payoffs \tilde{x}_A that are distributed according to $F_A(\cdot)$. We further randomize these payoffs to obtain a new random variable \tilde{x}_B according to

$$\tilde{x}_B = \tilde{x}_A + \tilde{z} \quad (4.10)$$

where, for any x_A value, $E(\tilde{z}) = \int z dH_{x_A}(\tilde{z}) = 0$; in other words, we add some pure randomness to \tilde{x}_A . Let $F_B(\cdot)$ be the distribution function associated with \tilde{x}_B . We say that $F_B(\cdot)$ is a mean preserving spread of $F_A(\cdot)$. A simple example of this is as follows. Let

$$\tilde{x}_A = \begin{cases} 5 & \text{with probability } \frac{1}{2} \\ 2 & \text{with probability } \frac{1}{2} \end{cases}$$

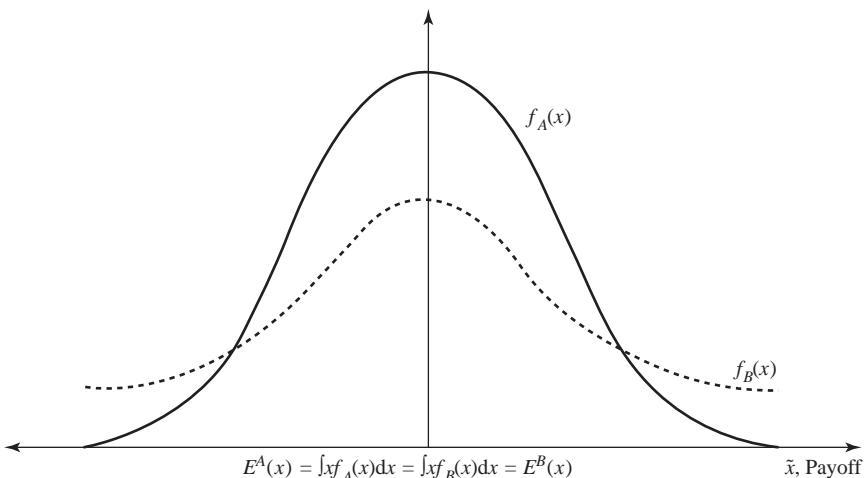


Figure 4.7
Mean preserving spread.

and suppose

$$\tilde{z} = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

Then,

$$\tilde{x}_B = \begin{cases} 6 & \text{with probability } \frac{1}{4} \\ 4 & \text{with probability } \frac{1}{4} \\ 3 & \text{with probability } \frac{1}{4} \\ 1 & \text{with probability } \frac{1}{4} \end{cases}$$

Clearly, $E\tilde{x}_A = E\tilde{x}_B = 3.5$; we would also all agree that $F_B(\cdot)$ is intuitively riskier.

Our final theorem ([Theorem 4.4](#)) relates the sense of a mean preserving spread, as captured by [Eq. \(4.10\)](#), to our earlier results.

Theorem 4.4 Let $F_A(\cdot)$ and $F_B(\cdot)$ be two distribution functions defined on the same state space with identical means. The following statements are equivalent:

- (i) $F_A(\tilde{x}) \leq F_B(\tilde{x})$
- (ii) $F_B(\tilde{x})$ is a mean preserving spread of $F_A(\tilde{x})$ in the sense of [Eq. \(4.10\)](#).

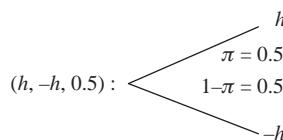
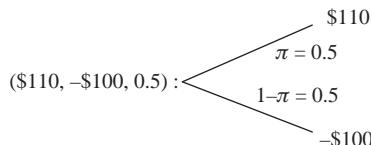
Proof See [Rothschild and Stiglitz \(1970\)](#).

But what about distributions that are not stochastically dominant under either definition and for which the mean–variance criterion does not give a relative ranking? For example, consider (independent) investments 5 and 6 in [Table 4.4](#).

In this case we are left to compare distributions by computing their respective expected utilities. That is to say, the ranking between these two investments is preference dependent. Some risk-averse individuals will prefer investment 5, while other risk-averse individuals will prefer investment 6. This is not bad. There remains a systematic basis of comparison. The task of the investment advisor is made more complex, however, as he will have to elicit more information on the preferences of his client if he wants to be in position to provide adequate advice.

Table 4.4: Two investments; No dominance

Investment 5		Investment 6	
Payoff	Probability	Payoff	Probability
1	0.25	3	0.33
7	0.5	5	0.33
12	0.25	8	0.34

**Figure 4.8**
A fair lottery.**Figure 4.9**
An often-refused lottery.

4.8 An Unsettling Observation About Expected Utility

Our definition of a risk-averse investor is one who is unwilling to accept the lottery depicted in [Figure 4.8](#) at any wealth level. As noted in the discussion of loss aversion, however, there is also substantial experimental evidence (see, for example, [Kahneman and Tversky, 1979](#)) that risk-averse investors also refuse lotteries of the form depicted in [Figure 4.9](#) at all wealth levels.

Building on this experimental observation, [Rabin \(2000\)](#) goes on to show analytically that the rejection of the lottery depicted in [Figure 4.9](#) at all wealth levels implies unrealistically extreme risk aversion over large gambles. To illustrate, he shows that an investor with concave utility who rejects $(\$110, -\$100, 0.5)$ for all wealth levels less than or equal to \$300,000 will *necessarily* reject $(\$850,000,000,000, -\$2000, 0.5)$ at a wealth level of \$290,000 ([Rabin, 2000](#), Table 2). On the face of it, such a rejection seems a bit absurd: even at a relatively modest wealth level of \$290,000, most of us would be willing to hazard

a loss of \$2000 against a 50% chance at an \$850 billion jackpot. Rabin's (2000) argument rests on the observation that if the investor rejects $(\$110, -\$100, 0.5)$ at all wealth levels, his marginal utility of wealth must be decreasing geometrically as his wealth level increases. A very large offsetting payment is thus required to induce the acceptance of a lottery with even modest losses as wealth increases.^{5,6}

Rabin's observations suggest that if the VNM-expected utility framework is to be appropriate for evaluating large "gambles," investors must be essentially risk neutral over small ones. Yet, they are not, at least in experimental contexts. Apparently, investors' preference for small gambles are guided by some other preference structure such as loss aversion. It is at present unclear what this inconsistency means for important economic behavior such as the consumption-savings decision.

4.9 Applications: Leverage and Risk

Here we make more precise the notion of leverage first introduced in Section 2.5.3.

The word leverage is both a noun and a verb. As a noun, it simply means "borrowed money," obtained in exchange for issuing a debt security. As a verb (as in "to leverage" an asset), it means to acquire an asset using both the investor's own money (his equity contribution E) and borrowed funds (the debt contribution D). The degree of leverage applied to an asset (the relative extent of debt financing) is variously measured by the (D/E) ratio or the debt/value ratio (D/V) , where V is the market value of all claims to the underlying asset's cash flows.

One of the hallmarks of the financial crisis was the enormous levels of indebtedness across many different institutions that it revealed. Whether on the part of financial institutions or individual households (and, in some cases, sovereign states), the years leading up to the financial crisis were often marked by dramatic increases in leverage. We want to understand why this was the case, and a natural way to initiate the exploration is first to consider the effects of leverage on an asset's risk and return.

Consider an investor with equity capital E , who is interested to acquire asset A , with intrinsic rate of return distribution $r_A(\tilde{\theta})$. An unlevered (all-equity) investment in A would naturally imply that the return on the investor's equity, $r_u^e(\tilde{\theta})$ would satisfy

$$r_u^e(\theta) = r_A(\theta) \text{ for every state } \theta$$

⁵ Note that Rabin's (2000) results require only concavity and are not specific to any functional form. Accordingly, we invite the reader to confirm the rejection of $(\$850 \text{ billion}, -\$2000, 0.5)$ at a wealth level of \$290,000 for any of the utility functions considered in this chapter.

⁶ See this chapter's web notes for the details of Rabin's (2000) argument.

If, however, the investor borrows an amount D at a constant rate r_d to help acquire the asset, then the return on his now “levered equity,” $r_L^e(\tilde{\theta})$, is given by⁷

$$r_L^e(\theta) = r_u^e(\theta) + (D/E)(r_u^e(\theta) - r_d) \quad (4.11)$$

It follows from Eq. (4.11) that

$$E\tilde{r}_L^e = E\tilde{r}_u^e + (D/E)(E\tilde{r}_u^e - r_d), \text{ and} \quad (4.12)$$

$$\sigma_{\tilde{r}_L^e} = (1 + D/E)\sigma_{\tilde{r}_u^e} \quad (4.13)$$

where we assume the debt is risk free. Equations (4.12) and (4.13) make clear that an increase in leverage, as measured by an increase in (D/E) will increase both the expected return and risk of leveraged equity.

Intuitively, what mechanism lies behind Eqs. (4.12) and (4.13)? Leverage clearly only makes sense if $E\tilde{r}_u^e > r_d$: on average, the underlying asset’s return must exceed the cost of borrowing. It follows that there must be a preponderance of states (or such states have a preponderance of the probability) for which $r_u^e(\theta) > r_d$. By Eq. (4.11), in those states $\tilde{\theta}$ where $r_u^e(\tilde{\theta}) > r_d$, the equity investor not only receives what the underlying asset itself pays, $r_u^e(\tilde{\theta})$, on *his* invested capital, E , but also the return surplus the asset earns above the financing cost, $(r_u^e(\tilde{\theta}) - r_d)$, on the debt financed portion of the asset’s cost, D , in those states. This surplus is the source of the increase in expected returns for levered versus unlevered equity. In the event, the asset pays a return below the borrowing cost; however, the investor’s return is lowered because the asset return shortfall relative to r_d must be made up from his own residual payment. Therein lies the source of the increase in risk that leverage promotes. Note that when we speak of the investor’s return, we mean only the return on his own capital invested in the project.

⁷ The logic behind Eq. (4.12) is as follows: Consider two distinct financings for the purchase of asset A , one all equity, the other a mixture of debt and (leveraged) equity. If an investor owns all the unlevered equity (the first case financing) or all the debt and all the equity (the second case financing), in either case the return to his portfolio securities must be the same and coincide with the return on the underlying asset itself, $r_A(\tilde{\theta})$. Why? In either case, the investor would, in the aggregate, receive all the asset’s free cash flows. Algebraically, this discussion can be summarized as:

$$r_A(\theta) = \underbrace{r_u^e(\theta)}_{\text{all equity financing's return}} = \underbrace{\left(\frac{E}{D+E} \right) r_L^e(\theta) + \left(\frac{D}{D+E} \right) r_d}_{\text{average return to levered equity and debt under partial debt financing}}$$

Rearranging the above equation to solve for $r_L^e(\theta)$ gives Eq. (4.11). (We have anticipated future chapters by relying on the fact that the return on a portfolio is the value-weighted average of the returns to its constituent assets.) Note that we first introduced relationship (4.11) in Chapter 2.

4.9.1 An Example

Let us imagine an investor who purchases a \$300,000 home in Phoenix, Arizona, with a \$20,000 down payment. The investor foresees $E\tilde{r}_A = 10\%$ (house price appreciation) with $\sigma_{\tilde{r}_A} = 5\%$. Net of assorted tax benefits, the mortgage cost, r_d , is 5%. In this case, $D/E = \$280,000/\$20,000 = 14$, and $E\tilde{r}_L^e = E\tilde{r}_u^e + D/E(E\tilde{r}_u^e - r_d) = 0.10 + 14(0.10 - 0.05) = 0.80$ or 80% relative to $E\tilde{r}_A$, a huge increase. Risk, however, also rises dramatically:

$$\sigma_{\tilde{r}_L^e} = \left(1 + \frac{D}{E}\right)\sigma_{\tilde{r}_u^e} = (1 + 14)(0.05) = 0.75 \text{ or } 75\%$$

Note that this investor has only a \$20,000 equity investment in the house with \tilde{r}_L^e representing the return on that amount alone. Relative to an all-equity purchase of the house, note also that expected returns rose by a factor of 8 while risk increased by a factor of 15.

How does this enormous risk manifest itself? Consider a decline in house prices of 10% (not a large decline; in the recent housing crisis, Phoenix area homes declined in value by nearly 50% on average), and let $\hat{\theta}$ denotes the specific state in which the 10% decline is experienced; then

$$\begin{aligned} r_L^e(\hat{\theta}) &= r_u^e(\hat{\theta}) + \left(\frac{D}{E}\right)(r_u^e(\hat{\theta}) - r_d) \\ &= -0.10 + (14)(-0.10 - 0.05) \\ &= -2.20 \end{aligned}$$

which means that the investor “lost” 220% of his initial wealth. While an all-equity investor can lose at most his entire investment ($\tilde{r}_u^e(\theta) \geq -1.00$), a leveraged investor can lose even more because of the debt he must personally repay.⁸ This occurs as follows:⁹

⁸ This statement is country specific as regards the responsibility of the mortgagee, and applies more to Europe than to the United States. In the USA most home mortgages are ‘non-recourse’, meaning that the mortgagee can walk away from his mortgage and surrender his house to the lender without the lender being able to seize his other assets to cover any discrepancy between the home’s market value and the loan’s outstanding balance. In this sense the purchase of a home is closely similar to an equity investment in much of the USA (particulars are governed by state law).

⁹ The situation may be analyzed alternatively by observing that the investor effectively owns the following portfolio P :

$$P = \{1 \text{ Phoenix house, 1 loan}\} \text{ where } w_{\text{house}} = \frac{\$300,000}{\$20,000} = 15 \text{ and } w_{\text{loan}} = \frac{-\$280,000}{\$20,000} = -14; \text{ thus}$$

$$\begin{aligned} r_P(\hat{\theta}) &= w_{\text{house}}r_{\text{house}}(\hat{\theta}) + w_{\text{loan}}r_d \\ &= 15(-.10) - 14(.05) \\ &= -2.20 = r_L^e(\hat{\theta}) \end{aligned}$$

If the event $\hat{\theta}$ is realized, this mortgage is said to go “underwater.”

$$\begin{aligned}\text{Loss on the value of the house: } & -0.10 (\$300,000) = -\$30,000 \\ \text{Mortgage interest payment: } & -0.05 (\$280,000) = \underline{-\$14,000} \\ \text{Total losses/outflows} & = -\$44,000\end{aligned}$$

$$1 + r_L^e(\hat{\theta}) = \frac{-\$44,000}{\$20,000} = -2.2, \text{ or } -220\%$$

4.9.2 Is Leverage a Good Thing?

From our discussion at the start of Chapter 3, we know it is generally impossible to rank two assets directly, one of which has both a higher mean and a higher standard deviation of returns. The leveraged asset return has this feature relative to its unleveraged counterpart. But what about the Sharpe ratio? We observe

$$\frac{E\tilde{r}_L^e - r_d}{\sigma_{\tilde{r}_L^e}} = \frac{E\tilde{r}_u^e + \frac{D}{E}(E\tilde{r}_u^e - r_d) - r_d}{(1 + \frac{D}{E})\sigma_{\tilde{r}_u^e}} \quad (4.14)$$

$$= \frac{(1 + \frac{D}{E})E\tilde{r}_u^e - (1 + \frac{D}{E})r_d}{(1 + \frac{D}{E})\sigma_{\tilde{r}_u^e}} = \frac{E\tilde{r}_u^e - r_d}{\sigma_{\tilde{r}_u^e}} \quad (4.15)$$

i.e., leverage has no effect on the Sharpe ratio.¹⁰

Since a higher Sharpe ratio is viewed as indicative of superior investment performance, the fact that an increase in debt leaves this fundamental measure unchanged must be viewed as a discouragement to its use. Yet, in the years immediately preceding the 2008 financial crisis, household, firm, and government debt levels all rose to record levels in both the United States and certain European countries (e.g., Ireland, Spain). At the time of its collapse (September 15, 2008) Lehman Brothers Corporation had a (D/E) of roughly 50. What was going on?

There are, of course, various tax incentives for debt; in particular, the home mortgage interest deduction in the US. Homeowners may also be willing to shoulder the risk of debt financing because they have a strong preference for home ownership, and debt financing makes this preference more immediately accessible. Firms also benefit from the tax deductability of interest payments. Prior to the financial crisis, home mortgages in the United States, however, usually began with a (D/E) = 4 (a 20% down payment). Yet in the last years of the housing boom, mortgages were being written with no down payment at all ($D/E = \infty$). Since tax incentives to debt financing have been in place for a long time, other incentives must have been at work.

¹⁰ The ratio $E\tilde{r}_L^e/\sigma_{\tilde{r}_L^e}$ actually declines as (D/E) increases.

To propose one such possible incentive, it is necessary to be more precise regarding the effects of leverage on equity return distributions. This is the focus of [Theorem 4.4](#). It basically asserts that in the favorable states of nature, $\{\theta : \tilde{r}_u^e(\theta) > r_d\}$, the leveraged return distribution FSD its unlevered counterpart (and thus also more highly leveraged return distributions will FSD less highly leveraged ones on the set of favorable states).

Theorem 4.5 Let Θ denote the underlying space of events, and consider the following partition of Θ :

$$A = \{\theta \in \Theta : r_u^e(\theta) > r_d\}, \quad \text{and} \quad \Theta - A = \{\theta \in \Theta : r_u^e(\theta) \leq r_d\}$$

Furthermore, let $F_{L/B}$ and $F_{u/B}$ denote the conditional distribution functions restricted to some set $B \subseteq \Theta$ for $\tilde{r}_L^e(\theta)$ and $\tilde{r}_u^e(\theta)$ respectively. Then,

- (i) $F_{L/A}(\tilde{r}_L^e(\theta))$ FSD $F_{u/A}(\tilde{r}_u^e(\theta))$
- (ii) $F_{u/\Theta-A}(\tilde{r}_u^e(\theta))$ FSD $F_{L/\Theta-A}(\tilde{r}_L^e(\theta))$

Proof We demonstrate only (i), as (ii) follows similarly.

$$\begin{aligned} F_{L/A}(\hat{r}) &= \frac{\text{Prob}(\tilde{r}_L^e(\theta) \leq \hat{r} : \theta \in A)}{\text{Prob}(\theta \in A)} \\ &= \frac{\text{Prob}\left(\tilde{r}_u^e(\theta) + \frac{D}{E}(\tilde{r}_u^e(\theta) - r_d) \leq \hat{r} : \theta \in A\right)}{\text{Prob}(\theta \in A)} \\ &= \frac{\text{Prob}\left(\tilde{r}_u^e(\theta) \leq \frac{\hat{r} + \frac{D}{E}r_d}{\left(1 + \frac{D}{E}\right)} : \theta \in A\right)}{\text{Prob}(\theta \in A)} \\ &< \frac{\text{Prob}\left(\tilde{r}_u^e(\theta) \leq \frac{\hat{r} + \frac{D}{E}\hat{r}}{\left(1 + \frac{D}{E}\right)} : \theta \in A\right)}{\text{Prob}(\theta \in A)} \\ &< \frac{\text{Prob}(\tilde{r}_u^e(\theta) \leq \hat{r} : \theta \in A)}{\text{Prob}(\theta \in A)} \\ &= F_{u/A}(\hat{r}) \end{aligned}$$

A simple paraphrase of [Theorem 4.5](#) is as follows: in high return states, $\{\theta : \tilde{r}_u^e(\theta) > r_d\}$, leverage makes them even higher; in low return states $\{\theta : \tilde{r}_u^e(\theta) < r_d\}$ it makes them even lower. Accordingly, agents will desire high leverage if they can somehow ignore the low return states ($\{\theta : \tilde{r}_u^e(\theta) < r_d\}$), or if a belief bias leads them to underestimate the low return state probabilities. Both investors and their managers may be guilty of overoptimistic beliefs, but limited liability managers, whose compensation is linked principally to payoffs only in the favorable states, are uniquely positioned to ignore the low return states altogether.

4.9.3 An Application to Executive Compensation

Certain private equity firms employ high leverage ratios in their managed portfolios, yet the portfolio managers of these firms are surely aware of high risk and unchanged Sharpe ratio that such leverage entails. Accordingly, the popular press has speculated that high leverage choices may simply represent a response of the managers of these firms to the way they are paid. For example, the “2 and 20” rule is pervasive: managers receive 2% of the amount invested as a one-time payment, plus 20% of net positive returns above some benchmark, computed annually, for the contracting period. If the investment returns fall short of the benchmark, the managers receive nothing. But, neither are they assessed any penalty payment: it is a limited liability contract.

In the simplified setting of the present discussion, the limited liability, incentive portion of the contract may be expressed as

$$\text{Incentive Comp}(\theta) = \begin{cases} \gamma_1 Y(\tilde{r}_A(\theta) - r_d); & r_A(\theta) > r_d \\ 0; & \text{otherwise} \end{cases} \quad (4.16)$$

where we choose, for simplicity, r_d as the benchmark. The quantity Y is the amount invested by the client, and $\tilde{r}_A(\theta)$ the uncertain return on some portfolio A of assets selected by the manager on the investor’s behalf with, say, $\gamma_1 = 0.20$.

Let $r_L^e(\theta)$ represent the return on a leveraged portfolio of these assets with the degree of leverage measured by the (D/E) ratio. In a like fashion we naturally identify $r_A(\theta) = r_u^e(\theta)$. Define the set A' by

$$\begin{aligned} A' &= \{\theta : \theta \in \Theta : r_A(\theta) > r_d\} \\ &= \{\theta : \theta \in \Theta : \text{Incentive Comp}(\theta) > 0\} \end{aligned}$$

The following corollaries to [Theorem 4.4](#) point in the direction of the likely behavior of a self-interested manager with compensation contract [Eq. \(4.16\)](#).

Corollary 4.5.1 Let $F_{L/A'}^C$ and $F_{u/A'}^C$ denote the distribution functions of the manager's incentive compensation restricted to the set A' when, respectively, he leverages the underlying portfolio of assets ($D/E > 0$) and when he does not ($D/E = 0$). Then,

- (i) $F_{L/A'}^C \text{ FSD } F_{u/A'}^C$ for any ($D/E > 0$).

Since the manager's incentive compensation is zero on $\Theta - A'$, we can also claim

- (ii) $F_{L/\Theta}^C \text{ FSD } F_{u/\Theta}^C$ for any ($D/E > 0$).¹¹

Proof Adaptation of the proof of [Theorem 4.5](#).

Corollary 4.5.2 Every manager with an increasing and concave VNM-expected utility representation and a limited liability incentive contract of the form [\(4.16\)](#) will employ leverage ($D/E > 0$).

Proof By [Theorem 4.3](#) and [Corollary 4.5.1](#) since FSD implies SSD.

The results in [Corollaries 4.5.1–4.5.2](#) are really nothing more than the colloquial observation that asset managers with no “skin in the game” have little regard for (downside) risk.¹² These observations likely explain many instances where leverage ratios grew rapidly in the precrisis years when credit was readily available to “grow” the leverage.

4.10 Conclusions

The main topic of this chapter was the VNM-expected utility representation specialized to admit risk aversion. Two measures of the degree of risk aversion were presented. Both are functions of an investor's current level of wealth, and, as such, we would expect them to change as wealth changes. Is there any systematic relationship between $R_A(Y)$, $R_R(Y)$, and Y which it is reasonable to assume?

In order to answer that question, we must move away from the somewhat artificial setting of this chapter. As we will see in Chapter 5, systematic relationships between wealth and the measures of absolute and relative risk aversion are closely related to investors' portfolio behavior.

¹¹ This argument is largely unchanged if the manager also receives an additional fixed salary payment.

¹² The theory of optimal contracting between investors (principals) and financial advisors (their agents) is not yet well understood and is not a general topic we consider. See [Stracca \(2006\)](#) for an excellent survey. It is known, however, that simple linear contracts such as [Eq. \(4.16\)](#) are not optimal for delegated portfolio managers. See [Admati and Pfleiderer \(1997\)](#).

References

- Admati, A., Pfeiferer, P., 1997. Does it all add up? Benchmarks and the compensation of active portfolio managers. *J. Bus.* 102, 323–350.
- Arrow, K.J., 1971. Essays in the Theory of Risk Bearing. Markham, Chicago, IL.
- De Groot, M., 1970. Optimal Statistical Decisions. McGraw-Hill, New York, NY.
- Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica*. 47, 263–291.
- Laffont, J.-J., 1989. The Economics of Uncertainty and Information. MIT Press, Cambridge, MA.
- Pratt, J., 1964. Risk aversion in the small and the large. *Econometrica*. 32, 122–136.
- Rabin, M., 2000. Risk aversion and expected utility theory: a calibration theorem. *Econometrica*. 68, 1281–1292.
- Rothschild, M., Stiglitz, J.E., 1970. Increasing risk: a definition. *J. Econ. Theory*. 2, 225–243.
- Stracca, L., 2006. Delegated portfolio management: a survey of the theoretical literatures. *J. Econ. Surv.* 20, 823–848.

Appendix: Proof of Theorem 4.2

⇒ There is no loss in generality in assuming $U(\cdot)$ is differentiable, with $U'(\cdot) > 0$.

Suppose $F_A(x)$ FSD $F_B(x)$, and let $U(\cdot)$ be a utility function defined on $[a,b]$ for which $U'(\cdot) > 0$. We need to show that

$$E_A U(\tilde{x}) = \int_a^b U(\tilde{x}) dF_A(\tilde{x}) > \int_a^b U(\tilde{x}) dF_B(\tilde{x}) = E_B U(\tilde{x})$$

This result follows from integration by parts (recall the relationship $\int_a^b u dv = uv \Big|_a^b - \int_a^b v du$).

$$\begin{aligned} & \int_a^b U(\tilde{x}) dF_A(\tilde{x}) - \int_a^b U(\tilde{x}) dF_B(\tilde{x}) \\ &= U(b)F_A(b) - U(a)F_A(a) - \int_a^b F_A(\tilde{x})U'(\tilde{x})d\tilde{x} \\ &\quad - \{U(b)F_B(b) - U(a)F_B(a) - \int_a^b F_B(\tilde{x})U'(\tilde{x})d\tilde{x}\} \\ &= - \int_a^b F_A(\tilde{x})U'(\tilde{x})d\tilde{x} + \int_a^b F_B(\tilde{x})U'(\tilde{x})d\tilde{x}, \\ & \text{(since } F_A(b) = F_B(b) = 1, \text{ and } F_A(a) = F_B(a) = 0\text{)} \\ &= \int_a^b [F_B(\tilde{x}) - F_A(\tilde{x})]U'(\tilde{x})d\tilde{x} \geq 0 \end{aligned}$$

The desired inequality follows since, by the definition of FSD and the assumption that the marginal utility is always positive, both terms within the integral are positive. If there is some subset $(c,a) \subset [a,b]$ on which $F_A(x) > F_B(x)$, the final inequality is strict.

\Leftarrow Proof by contradiction. If $F_A(\tilde{x}) \leq F_B(\tilde{x})$ is false, then there must exist an $\bar{x} \in [a, b]$ for which $F_A(\bar{x}) > F_B(\bar{x})$. Define the following nondecreasing function $\hat{U}(x)$ by

$$\hat{U}(x) = \begin{cases} 1 & \text{for } b \geq x > \tilde{x} \\ 0 & \text{for } a \leq x < \tilde{x} \end{cases}$$

We will use integration by parts again to obtain the required contradiction.

$$\begin{aligned} & \int_a^b \hat{U}(\tilde{x}) dF_A(\tilde{x}) - \int_a^b \hat{U}(\tilde{x}) dF_B(\tilde{x}) \\ &= \int_a^b \hat{U}(\tilde{x}) [dF_A(\tilde{x}) - dF_B(\tilde{x})] \\ &= \int_{\tilde{x}}^b 1 [dF_A(\tilde{x}) - dF_B(\tilde{x})] \\ &= F_A(b) - F_B(b) - [F_A(\bar{x}) - F_B(\bar{x})] - \int_{\bar{x}}^b [F_A(\tilde{x}) - F_B(\tilde{x})](0) d\tilde{x} \\ &= F_B(\bar{x}) - F_A(\bar{x}) < 0 \end{aligned}$$

Thus we have exhibited an increasing function $\hat{U}(x)$ for which $\int_a^b \hat{U}(\tilde{x}) dF_A(\tilde{x}) < \int_a^b \hat{U}(\tilde{x}) dF_B(\tilde{x})$, a contradiction.

Risk Aversion and Investment Decisions, Part 1

Chapter Outline

5.1 Introduction	115
5.2 Risk Aversion and Portfolio Allocation: Risk-Free Versus Risky Assets	116
5.2.1 The Canonical Portfolio Problem	116
5.2.2 Illustration and Examples	117
5.3 Portfolio Composition, Risk Aversion, and Wealth	118
5.4 Special Case of Risk-Neutral Investors	121
5.5 Risk Aversion and Risky Portfolio Composition	122
5.6 Risk Aversion and Savings Behavior	124
5.6.1 Savings and the Riskiness of Returns	124
5.6.2 Illustrating Prudence	128
5.6.3 The Joint Saving–Portfolio Problem	129
5.7 Generalizing the VNM-Expected Utility Representation	130
5.7.1 Preferences for the Timing of Uncertainty Resolution	131
5.7.2 Preferences That Guarantee Time-Consistent Planning	133
5.7.2.1 Quasi-Hyperbolic Discounting	135
5.7.3 Separating Risk and Time Preferences	137
5.8 Conclusions	139
References	140

5.1 Introduction

Chapters 3 and 4 provided a systematic procedure for assessing an investor's relative preference for various investment payoffs: rank them according to expected utility using a VNM-utility representation constructed to reflect the investor's preferences over random payments. The subsequent postulate of risk aversion further refined this idea: it is natural to hypothesize that the utility-of-money function entering the investor's VNM index is concave ($U''() < 0$). Two widely used measures were introduced and interpreted, each permitting us to assess an investor's degree of risk aversion. In the setting of a zero-cost investment paying either $(+h)$ or $(-h)$, these measures were shown to be linked with the

minimum probability of success above one-half necessary for a risk-averse investor to take on such a prospect willingly. They differ only as to whether (h) measures an absolute amount of money or a proportion of the investors' initial wealth.

In this chapter, we begin to use these ideas with a view toward understanding the two most important financial decisions an investor will make. The first is his portfolio composition decision: what are his relative demands for assets of different risk classes and, in particular, his demand for risk-free versus risky assets? This topic is dealt with in [Sections 5.2–5.5](#). The second concerns an investor's consumption/savings decision: how much of an investor's current income should he allocate to savings and how much to consumption? When an investor saves, he augments his wealth in the future which implies extending our modeling context to multiple periods. This is the focus of [Section 5.6](#).

A multiperiod setting naturally gives rise to preference phenomena not present in the atemporal context of earlier chapters. In [Section 5.7](#), we review a number of departures from the strict VNM-expected utility paradigm that allows these phenomena to be given formal representation.

5.2 Risk Aversion and Portfolio Allocation: Risk-Free Versus Risky Assets

5.2.1 The Canonical Portfolio Problem

Consider an investor with wealth level Y_0 , who is deciding what amount, a , to invest in a risky portfolio with uncertain rate of return \tilde{r} . We can think of the risky asset as being, in fact, the market portfolio under the “old” capital asset pricing model (CAPM), to be reviewed in Chapter 8. The alternative is to invest in a risk-free asset that pays a certain rate of return r_f . The time horizon is one period. The investor's wealth at the end of the period is given by

$$\tilde{Y}_1 = (1 + r_f)(Y_0 - a) + a(1 + \tilde{r}) = Y_0(1 + r_f) + a(\tilde{r} - r_f)$$

The choice problem he must solve can be expressed as

$$\max_a EU(\tilde{Y}_1) = \max_a EU(Y_0(1 + r_f) + a(\tilde{r} - r_f)) \quad (5.1)$$

where $U(\)$ is his utility-of-money function and E the expectations operator.

This formulation of the investor's problem is fully in accord with the lessons of the prior chapter. Each choice of a leads to a different uncertain payoff distribution, and we want to find the choice that corresponds to the most preferred such distribution. By construction of his VNM representation, this is the payoff pattern that maximizes his expected utility.

Under risk aversion ($U''() < 0$), the necessary and sufficient first-order condition (FOC) for problem (5.1) is given by

$$E[U'(Y_0(1 + r_f) + a(\tilde{r} - r_f))(\tilde{r} - r_f)] = 0 \quad (5.2)$$

Analyzing Eq. (5.2) allows us to describe the relationship between the investor's degree of risk aversion and his portfolio's composition as per the following theorem.

Theorem 5.1 Assume $U'(\) > 0$, and $U''(\) < 0$ and let \hat{a} denote the solution to problem (5.1). Then

$$\begin{aligned}\hat{a} > 0 &\Leftrightarrow E\tilde{r} > r_f \\ \hat{a} = 0 &\Leftrightarrow E\tilde{r} = r_f \\ \hat{a} < 0 &\Leftrightarrow E\tilde{r} < r_f\end{aligned}$$

Proof Since this is a fundamental result, it is worthwhile to make clear its (straightforward) justification. We follow the argument presented in Arrow (1971), Chapter 2.

Define $W(a) = E\{U(Y_0(1 + r_f) + a(\tilde{r} - r_f))\}$. The FOC (5.2) can then be written $W'(a) = E[U'(Y_0(1 + r_f) + a(\tilde{r} - r_f))(\tilde{r} - r_f)] = 0$. By risk aversion ($U'' < 0$), $W''(a) = E[U''(Y_0(1 + r_f) + a(\tilde{r} - r_f))(\tilde{r} - r_f)^2] < 0$, i.e., $W'(a)$ is everywhere decreasing. It follows that \hat{a} will be positive if and only if $W'(0) = U'(Y_0(1 + r_f))E(\tilde{r} - r_f) > 0$ (since then a will have to be increased from the value of 0 to achieve equality in the FOC). Since U' is always strictly positive, this implies $\hat{a} > 0$ if and only if $E(\tilde{r} - r_f) > 0$.

The other assertions follow similarly.

Theorem 5.1 asserts that a risk-averse agent will invest in the risky asset or portfolio only if the expected return on the risky asset exceeds the risk-free rate. From another perspective, a risk-averse agent will *always* participate (possibly via an arbitrarily small stake) in a risky investment when the odds are favorable.

5.2.2 Illustration and Examples

It is worth pursuing the above result to get a sense of how large a is relative to Y_0 . Our findings will, of course, be preference dependent. Let us begin with the fairly standard and highly tractable utility function $U(Y) = \ln Y$. For added simplicity, let us also assume that the risky asset is forecast to pay either of two returns (corresponding to an “up” or “down” stock market), $r_2 > r_1$, with probabilities π and $1 - \pi$, respectively. It makes sense (why?) to assume $r_2 > r_f > r_1$ and $E\tilde{r} = \pi r_2 + (1 - \pi)r_1 > r_f$.

Under this specification, the FOC (5.2) becomes

$$E \left\{ \frac{\tilde{r} - r_f}{Y_0(1 + r_f) + a(\tilde{r} - r_f)} \right\} = 0$$

Writing out the expectation explicitly yields

$$\frac{\pi(r_2 - r_f)}{Y_0(1 + r_f) + a(r_2 - r_f)} + \frac{(1 - \pi)(r_1 - r_f)}{Y_0(1 + r_f) + a(r_1 - r_f)} = 0$$

which, after some straightforward algebraic manipulation, gives

$$\frac{a}{Y_0} = \frac{-(1 + r_f)[E\tilde{r} - r_f]}{(r_1 - r_f)(r_2 - r_f)} > 0 \quad (5.3)$$

This is an intuitive sort of expression: the fraction of wealth invested in risky assets increases with the return premium paid by the risky asset ($E\tilde{r} - r_f$) and decreases with an increase in the return dispersion around r_f as measured by $(r_2 - r_f)$ ($r_f - r_1$).¹

Suppose $r_f = 0.05$, $r_2 = 0.40$, and $r_1 = -0.20$, and $\pi = \frac{1}{2}$ (the latter information guarantees $E\tilde{r} = 0.10$). In this case, $a/Y_0 = 0.6$: 60% of the investor's wealth turns out to be invested in the risky asset. Alternatively, suppose $r_2 = 0.30$ and $r_1 = -0.10$ (same r_f , π , and $E\tilde{r}$); here we find that $a/Y_0 = 1.4$. This latter result must be interpreted to mean that an investor would prefer to invest at least his full wealth in the risky portfolio. If possible, he would even want to borrow an additional amount, equal to 40% of his initial wealth, at the risk-free rate, and invest this amount in the risky portfolio as well. In comparing these two examples, we see that the return dispersion is much smaller in the second case (lower risk in a mean-variance sense) with an unchanged return premium. With less risk and unchanged mean returns, it is not surprising that the proportion invested in the risky asset increases very substantially. We will see, however, that, somewhat surprisingly, this result does not generalize without further assumption on the form of the investor's preferences.

5.3 Portfolio Composition, Risk Aversion, and Wealth

In this section, we consider how an investor's portfolio decision is affected by his degree of risk aversion and his wealth level. A natural first exercise is to compare the portfolio composition across individuals of differing risk aversion. The answer to this first question conforms with intuition: if John is more risk averse than Amos, he optimally invests a smaller fraction of his wealth in the risky asset. This is the essence of our next two theorems.

¹ That this fraction is independent of the wealth level is not a general result, as we shall find out in Section 5.3.

Theorem 5.2 (Arrow, 1971) Suppose, for all wealth levels Y , $R_A^1(Y) > R_A^2(Y)$, where $R_A^i(Y)$ is the measure of absolute risk aversion of investor $i, i = 1, 2$. Then $\hat{a}_1(Y) < \hat{a}_2(Y)$.

That is, the more risk-averse agent, as measured by his absolute risk-aversion measure, will always invest less in the risky asset, given the same level of wealth. This result does not depend on measuring risk aversion via the absolute Arrow–Pratt measure. Indeed, since $R_A^1(Y) > R_A^2(Y) \Leftrightarrow R_R^1(Y) > R_R^2(Y)$, [Theorem 5.2](#) can be restated as

Theorem 5.3 Suppose, for all wealth levels $Y > 0$, $R_R^1(Y) > R_R^2(Y)$ where $R_R^i(Y)$ is the measure of relative risk aversion of investor $i, i = 1, 2$. Then $\hat{a}_1(Y) < \hat{a}_2(Y)$.

Continuing with the example of [Section 5.2.2](#), suppose now that the investor's utility function has the form $U(Y) = (Y^{1-\gamma})/(1-\gamma)$, $\gamma > 1$. This utility function displays both greater absolute and greater relative risk aversion than $U(Y) = \ln Y$ (you are invited to prove this statement). From [Theorems 5.2 and 5.3](#), we would expect this greater risk aversion to manifest itself in a reduced willingness to invest in the risky portfolio. Let us see if this is the case.

For these preferences, the expression corresponding to [Eq. \(5.3\)](#) is

$$\frac{a}{Y_0} = \frac{(1+r_f)\left\{[(1-\pi)(r_f-r_1)]^{\frac{1}{\gamma}} - (\pi(r_2-r_f))^{\frac{1}{\gamma}}\right\}}{(r_1-r_f)\{\pi(r_2-r_f)\}^{\frac{1}{\gamma}} - (r_2-r_f)\{(1-\pi)(r_f-r_1)\}^{\frac{1}{\gamma}}} \quad (5.4)$$

In the case of our first example, but with $\gamma = 3$, we obtain, by simple direct substitution,

$$\frac{a}{Y_0} = 0.24$$

indeed, only 24% of the investor's assets are invested in the risky portfolio, down from 60% earlier.

The next logical question is to ask how the investment in the risky asset varies with the investor's total wealth as a function of his degree of risk aversion. Let us begin with statements appropriate to the absolute measure of risk aversion.

Theorem 5.4 (Arrow, 1971) Let $\hat{a} = \hat{a}(Y_0)$ be the solution to Problem [\(5.1\)](#), then:

- i. $R'_A(Y) < 0 \Leftrightarrow \hat{a}'(Y_0) > 0$
- ii. $R'_A(Y) = 0 \Leftrightarrow \hat{a}'(Y_0) = 0$
- iii. $R'_A(Y) > 0 \Leftrightarrow \hat{a}'(Y_0) < 0$

Case (i) is referred to as declining absolute risk aversion (DARA). Agents with this property become more willing to accept greater bets as they become wealthier. [Theorem 5.4](#) says that such agents will also increase the amount invested in the risky asset ($\hat{a}'(Y_0 > 0)$). To state matters slightly differently, an agent with the indicated DARA will, if he becomes wealthier, be willing to put some of that additional wealth at risk. Utility functions of this form are quite common: those considered in the example, $U(Y) = \ln Y$ and $U(Y) = (Y^{1-\gamma})/(1-\gamma)$, $\gamma > 0$, display this property. It also makes intuitive sense.

Under constant absolute risk aversion or CARA, case (ii), the amount invested in the risky asset is unaffected by the agent's wealth. This result is somewhat counterintuitive. One might have expected that a CARA decision maker, in particular one with little risk aversion, would invest some of his or her increase in initial wealth in the risky asset.

[Theorem 5.4](#) disproves this intuition. An example of a CARA utility function is $U(Y) = -e^{-vY}$. Indeed,

$$R_A(Y) = \frac{-U''(Y)}{U'(Y)} = \frac{-(-v^2)e^{-vY}}{ve^{-vY}} = v$$

Let's verify the claim of [Theorem 5.4](#) for this utility function. Consider

$$\max_a E(-e^{-v(Y_0(1+r_f)+a(\tilde{r}-r_f))})$$

The FOC is

$$E[v(\tilde{r} - r_f)e^{-v(Y_0(1+r_f)+a(\tilde{r}-r_f))}] = 0$$

Now compute da/dY_0 ; by differentiating the above equation, we obtain

$$E\left[v(\tilde{r} - r_f)e^{-v(Y_0(1+r_f)+a(\tilde{r}-r_f))}\left(1 + r_f + (\tilde{r} - r_f)\frac{da}{dY_0}\right)\right] = 0$$

$$(1 + r_f)E\underbrace{\left[v(\tilde{r} - r_f)e^{-v(Y_0(1+r_f)+a(\tilde{r}-r_f))}\right]}_{= 0 \text{ (by the FOC)}} + E\left[\underbrace{v(\tilde{r} - r_f)^2}_{> 0} \frac{da}{dY_0} \underbrace{e^{-v(Y_0(1+r_f)+a(\tilde{r}-r_f))}}_{> 0}\right] = 0$$

therefore, $da/dY_0 \equiv 0$.

For the above preference ordering, and our original two-state risky distribution,

$$\hat{a} = \frac{1}{v} \left(\frac{1}{r_1 - r_2} \right) \ln \left(\frac{(1-\pi)}{\pi} \left(\frac{r_f - r_1}{r_2 - r_f} \right) \right)$$

Note that in order for \hat{a} to be positive, it must be that

$$0 < \frac{(1 - \pi)}{\pi} \left(\frac{r_f - r_1}{r_2 - r_f} \right) < 1$$

A sufficient condition is that $\pi > 1/2$.

Case (iii) is one with increasing absolute risk aversion (IARA). It says that as an agent becomes wealthier, he reduces his investments in risky assets. This does not make much sense, and we will generally ignore this possibility. Note, however, that the quadratic utility function, which is of some significance as we will see later on, possesses this property.

Let us now think in terms of the relative risk-aversion measure. Since it is defined for bets expressed as *a proportion of wealth*, it is appropriate to think in terms of *elasticities*, or of how the fraction invested in the risky asset changes as wealth changes. Define $\eta(Y, \hat{a}) = (\hat{a}/\hat{a})/(dY/Y) = (Y/\hat{a})(\hat{a}/dY)$, i.e., the wealth elasticity of investment in the risky asset. For example, if $\eta(Y, \hat{a}) > 1$, as wealth Y increases, the **percentage** increase in the amount optimally invested in the risky portfolio exceeds the percentage increase in Y . Or as wealth increases, the **proportion** optimally invested in the risky asset increases. Analogous to [Theorem 5.4](#) is [Theorem 5.5](#).

Theorem 5.5 (Arrow, 1971) If, for all wealth levels Y ,

- i. $R'_R(Y) = 0$ (CRRA) then $\eta = 1$
- ii. $R'_R(Y) < 0$ (DRRA) then $\eta > 1$
- iii. $R'_R(Y) > 0$ (IRRA) then $\eta < 1$

In his article, Arrow gives support for the hypothesis that the rate of relative risk aversion should be constant with constant relative risk aversion (CRRA) ≈ 1 . In particular, it can be shown that if an investor's utility of wealth is to be bounded above as wealth tends to ∞ , then $\lim_{Y \rightarrow \infty} R'_R(Y) \geq 1$; similarly, if $U(Y)$ is to be bounded below as $Y \rightarrow 0$, then $\lim_{Y \rightarrow 0} R'_R(Y) \leq 1$. These results suggest that if we wish to assume CRRA, then $CRRA = 1$ is the appropriate value.² Utility functions of the CRRA class include $U(Y) = (Y^{1-\gamma})/(1-\gamma)$, where $R_R(Y) = \gamma$ and $R_A(Y) = \gamma/Y$.

5.4 Special Case of Risk-Neutral Investors

As noted in Chapter 4, a risk-neutral investor is one who does not care about risk; he ranks investments solely on the basis of their expected returns. The utility function of such an

² Note that the above comments also suggest the appropriateness of weakly increasing relative risk aversion as an alternative working assumption.

agent is necessarily of the form $U(Y) = c + dY$, c, d constants, $d > 0$. (Check that $U'' = 0$ in this case.)

What proportion of his wealth will such an agent invest in the risky asset? The answer is: provided $E\tilde{r} > r_f$ (as we have assumed), **all** of his wealth will be invested in the risky asset.³ This is clearly seen from the following. Consider the agent's portfolio problem:

$$\max_a E(c + d(Y_0(1 + r_f) + a(\tilde{r} - r_f))) = \max_a [c + d(Y_0(1 + r_f)) + da(E\tilde{r} - r_f)]$$

With $E\tilde{r} > r_f$ and, consequently, $d(E\tilde{r} - r_f) > 0$, this expression is increasing in a .

This means that if the risk-neutral investor is unconstrained, he will attempt to borrow as much as possible at r_f and reinvest the proceeds in the risky portfolio. He is willing, without bound, to exchange certain payments for uncertain claims of greater expected value. As such he stands willing to absorb all of the economy's financial risk. If we specify that the investor is prevented from borrowing, then the maximum will occur at $a = Y_0$.

5.5 Risk Aversion and Risky Portfolio Composition

So far we have considered the question of how an investor should allocate his wealth between a risk-free asset and a risky asset or portfolio. We now go one step further and ask the following question: when is the *composition* of the portfolio (i.e., the percentage of the portfolio's value invested in each of the J risky assets that compose it) independent of the agent's wealth level? This question is particularly relevant in light of current investment practices whereby portfolio decisions are usually taken in steps. Step 1, often associated with the label "asset allocation," is the choice of instruments: stocks, bonds, and riskless assets (possibly alternative investments as well, such as hedge funds, private equity, and real estate); Step 2 is the country or sector allocation decision: here the issue is to optimize not across asset class but across geographical regions or industrial sectors. Step 3 consists of the individual stock-picking decisions made on the basis of information provided by financial analysts. The issuing of asset and country/sector allocation "grids" by all major financial institutions, tailored to the risk profile of the different clients, but independent of their wealth levels (and of changes in their wealth), is predicated on the hypothesis that differences in wealth (across clients) and changes in their wealth do not require adjustments in portfolio composition provided risk tolerance is either unchanged or controlled for.

Let us illustrate the issue in more concrete terms; take the example of an investor with invested wealth equal to \$12,000 and optimal portfolio proportions of $a_1 = \frac{1}{2}$, $a_2 = \frac{1}{3}$, and $a_3 = \frac{1}{6}$ (only three assets are considered). In other words, this individual's portfolio holdings are \$6000 in asset 1, \$4000 in asset 2, and \$2000 in asset 3. The implicit assumption

³ One way to interpret [Theorem 5.1](#) in this light is to infer that it suggests investors are approximately risk neutral over small gambles ([Arrow, 1971](#)). Yet this is apparently not the case (c.f. (110, -100; %)).

behind the most common asset management practice is that, were the investor's wealth to double to \$24,000, the new optimal portfolio would naturally be

$$\text{Asset 1: } \frac{1}{2}(\$24,000) = \$12,000$$

$$\text{Asset 2: } \frac{1}{3}(\$24,000) = \$8,000$$

$$\text{Asset 3: } \frac{1}{6}(\$24,000) = \$4,000$$

The question we pose in the present section is: Is this hypothesis supported by theory? The answer is generally no, in the sense that it is only for very specific preferences (utility functions) that the asset allocation is optimally left unchanged in the face of changes in wealth levels. Fortunately, these specific preferences include some of the major utility representations. The principal result in this regard is as follows.

Theorem 5.6 (Cass and Stiglitz, 1970) Let the vector $\begin{pmatrix} \hat{a}_1(Y_0) \\ \vdots \\ \hat{a}_J(Y_0) \end{pmatrix}$ denote the amount

optimally invested in the J risky assets if the wealth level is Y_0 .

$$\text{Then } \begin{pmatrix} \hat{a}_1(Y_0) \\ \vdots \\ \hat{a}_J(Y_0) \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_J \end{pmatrix} f(Y_0)$$

(for some positive function $f(\cdot)$) if and only if either

- i. $U'(Y_0) = (\theta Y_0 + \kappa)^\Delta$ or
- ii. $U'(Y_0) = \xi e^{-vY_0}$

There are, of course, implicit restrictions on the choice of θ , κ , Δ , ξ , and v to ensure, in particular, that $U''(Y_0) < 0$.⁴

Integrating (i) and (ii), respectively, in order to recover the utility functions corresponding to these marginal utilities, one finds, significantly, that the first includes the CRRA class of functions:

$U(Y_0) = \frac{1}{1-\gamma} Y_0^{(1-\gamma)} \quad \gamma \neq 1$, and $U(Y_0) = \ln(Y_0)$, while the second corresponds to the CARA:

$$U(Y_0) = \frac{\xi}{-v} e^{-vY_0}$$

⁴ For (i), we must have either $\theta > 0$, $\Delta < 0$, and Y_0 such that $\theta Y_0 + \kappa \geq 0$ or $\theta < 0$, $\kappa < 0$, $\Delta > 0$, and $Y_0 \leq -\frac{\kappa}{\theta}$. For (ii), $\xi > 0$, $-v < 0$, and $Y_0 \geq 0$.

In essence, **Theorem 5.6** states that it is only in the case of utility functions satisfying constant absolute or constant relative risk-aversion preferences (and some generalization of these functions of minor interest) that the relative composition of the risky portion of an investor's optimal portfolio is invariant to changes in his wealth.⁵ Only in these cases should the investor's portfolio composition be left unchanged as invested wealth increases or decreases. It is only with such utility specifications that the standard "grid" approach to portfolio investing is formally justified.⁶

5.6 Risk Aversion and Savings Behavior

5.6.1 Savings and the Riskiness of Returns

We have thus far considered the relationship between an agent's degree of risk aversion and the composition of his portfolio. This was accomplished in an atemporal one-period context. A related, though significantly different, question is to ask how an agent's **savings rate** is affected by an increase in the degree of risk facing him. Note that the savings question will be the first occasion where we must explicitly introduce a time dimension into the analysis (i.e., where the important trade-off involves the present versus the future). It is to be expected that the answer to this question will be influenced, in a substantial way, by the agent's degree of risk aversion.

Consider first an agent solving the following two-period consumption–savings problem:

$$\begin{aligned} & \max_s E\{U(Y_0 - s) + \delta U(s\tilde{R})\} \\ \text{s.t. } & Y_0 \geq s \geq 0 \end{aligned} \tag{5.5}$$

where Y_0 is initial (period zero) wealth, s is the amount saved and entirely invested in a risky portfolio with uncertain gross risky return, $\tilde{R} = 1 + \tilde{r}$, $U(\cdot)$ is the agent's period utility-of-consumption function, and δ is his subjective discount factor.⁷ The subjective discount rate $\delta < 1$ captures the extent to which the investor values consumption in the future less than current consumption.

⁵ As noted earlier, the constant absolute risk-aversion class of preferences has the property that the total amount invested in risky assets is invariant to the level of wealth. It is not surprising therefore that the proportionate allocation among the available risky assets is similarly invariant as this theorem asserts.

⁶ **Theorem 5.6** does not mean, however, that the fraction of initial wealth invested in the risk-free asset versus the risky "mutual fund" is invariant to changes in Y_0 . The CARA class of preferences discussed in the previous footnote is a case in point.

⁷ Note that this $U(c) = U(Y_0 - s)$ is in principle no different from the (indirect) utility of *wealth* function considered earlier in Chapter 3: the income remaining after savings is the income available for consumption.

The FOC for this problem (assuming an interior solution) is given by

$$U'(Y_0 - s) = \delta E(U'(s\tilde{R})\tilde{R}) \quad (5.6)$$

It is clear from the above equation that the properties of the return distribution \tilde{R} will influence the optimal level of s . One is particularly interested to know how optimal savings is affected by the riskiness of returns.

To be concrete, let us think of two return distributions \tilde{R}_A, \tilde{R}_B such that \tilde{R}_B is riskier than \tilde{R}_A and $E\tilde{R}_A = E\tilde{R}_B$. From our previous work (Theorem 4.4), we know this can be made precise by stating that \tilde{R}_A SSD \tilde{R}_B or that \tilde{R}_B is a mean-preserving spread of \tilde{R}_A . In other words, one can write $\tilde{R}_B = \tilde{R}_A + \tilde{\varepsilon}$, where $\tilde{\varepsilon}$ is a zero mean random variable uncorrelated with \tilde{R}_A . Let s_A and s_B be, respectively, the savings out of Y_0 corresponding to the return distributions \tilde{R}_A and \tilde{R}_B . The issue is whether s_A is larger than s_B or if the converse is true. In other words, can one predict that a representative risk-averse agent *will save more or less* when confronted with riskier returns on his or her savings?

Let us try to think intuitively about this issue. On the one hand, one may expect that more risk will mean a decrease in savings because “a bird in the hand is worth two in the bush.” This can be viewed as a substitution effect: A riskier return can be likened to an increase in the cost of future consumption. A rational individual may then opt to consume more today. On the other hand, a risk-averse individual may want to increase savings in the face of uncertainty as a precautionary measure, in order to guarantee a minimum standard of living in the future. This reaction indeed is associated with the notion of “precautionary savings.” The reader is invited to verify that this ambiguity is resolved in a mean–variance world in favor of the first argument. In that context, riskier returns imply a decrease in the RHS of Eq. (5.6), or a decrease in the expected future marginal utility of consumption weighted by the gross return. For the equality to be restored, consumption today must increase, and, consequently, savings must decrease.

It is important to realize, however, that the mean–variance response is not representative of the reactions of all risk-averse agents. Indeed, observers seeking to explain the increase in the US personal savings rate post 2007 have regularly pointed to the rising uncertainties surrounding the macroeconomic situation in general and the pace of economic growth in particular.⁸ As our discussion suggests, the key technical issue is whether the RHS of Eq. (5.6) is increased or decreased by an increase in risk. Applying reasoning similar to that used when discussing risk aversion (see Section 5.2), it is easy to see that this issue, in fact, revolves around whether the RHS of Eq. (5.6) (i.e., $\delta U'(sR)R = sg(R)$), is convex (in which case it increases) or concave (in which case it decreases) in R .

⁸ Which, if they are right, would tend to suggest that “the world is not mean–variance.”

Suppose, to take an extreme case, that the latter is linear in R ; we know that linearity means that the RHS of Eq. (5.6) can be written as $\delta E(g(R)) = \delta g(ER)$. But since R_A and R_B have the same mean, this implies that the RHS of Eq. (5.6), and consequently optimal savings, are unaffected by an increase in risk. If, on the other hand, $g(R)$ is concave, then $E(g(R)) < g(E(R))$; this is the previous case with a concave $g(\cdot)$, upward deviations from the mean produce smaller changes in the values attained by the function than downward deviations. It is then the case that the RHS will be decreased as a result of the mean-preserving spread on R . The reverse is true if $g(\cdot)$ is convex.

Note that in the all important case where $U(c) = \ln(c)$, $g(R)$ is in fact a constant function of R , with the obvious result that the savings decision is not affected by the increase in the riskiness of the return distribution. This difference in results between two of the workhorses of finance (mean variance and log utility) is worth emphasizing.

Let us now compute the second derivative of $g(R)$:⁹

$$g''(R) = 2U''(sR)s + s^2RU'''(sR) \quad (5.7)$$

Using Eq. 5.7, one can express the sign of g'' in terms of the relative rate of risk aversion as in Theorem 5.7.

Theorem 5.7 (Rothschild and Stiglitz, 1971) Let \tilde{R}_A and \tilde{R}_B be two return distributions with identical means such that \tilde{R}_A SSD \tilde{R}_B , and let s_A and s_B be, respectively, the savings out of Y_0 corresponding to the return distributions \tilde{R}_A and \tilde{R}_B .

If $R'_R(Y) \leq 0$ and $R_R(Y) > 1$, then $s_A < s_B$

If $R'_R(Y) \geq 0$ and $R_R(Y) < 1$, then $s_A > s_B$

Proof To prove this assertion, we need the following Lemma 5.7.

Lemma 5.7 $R'_R(Y)$ has the same sign as $-[U'''(Y)Y + U''(Y)(1 + R_R(Y))]$.

Proof Since $R_R(Y) = \frac{-YU''(Y)}{U'(Y)}$

$$R'_R(Y) = \frac{[-U'''(Y)Y - U''(Y)]U'(Y) - [-U''(Y)Y]U''(Y)}{[U'(Y)]^2}$$

⁹ $g(R) = U'(sR)R \Rightarrow g'(R) = U''(sR)sR + U'(sR)$ and $g''(R)$ as in Eq. (5.6). In the log case, we get $g(R) = 1/s$ and thus $g'(R) = g''(R) = 0$.

Since $U'(Y) > 0$, $R'_R(Y)$ has the same sign as

$$\begin{aligned} & \frac{[-U'''(Y)Y - U''(Y)]U'(Y) - [-U''(Y)Y]U''(Y)}{U'(Y)} \\ &= -U'''(Y)Y - U''(Y) - \left[\frac{-U''(Y)Y}{U'(Y)} \right] U''(Y) \\ &= -\{U'''(Y)Y + U''(Y)[1 + R_R(Y)]\} \end{aligned}$$

Now we can proceed with the theorem. We will show only the first implication as the second follows similarly. By the lemma, since $R'_R(Y) < 0$,

$$\begin{aligned} & -\{U'''(Y)Y + U''(Y)[1 + R_R(Y)]\} < 0, \text{ or} \\ & \{U'''(Y)Y + U''(Y)[1 + R_R(Y)]\} > 0 \end{aligned}$$

In addition, since $U''(Y) < 0$ and $R_R(Y) > 1$,

$$U'''(Y)Y + U''(Y)(2) > \{U'''(Y)Y + U''(Y)[1 + R_R(Y)]\} > 0$$

This is true for all Y ; hence

$2U''(sR) + sRU'''(sR) > 0$. Multiplying left and right by $s > 0$, one gets
 $2U''(sR)s + s^2RU'''(sR) > 0$, which by Eq. (5.6) implies

$$g''(R) > 0$$

But by the earlier remarks, this means that $s_A < s_B$ as required.

Theorem 5.7 implies that for the class of constant relative risk-aversion utility functions, i.e., functions of the form

$$U(c) = (1-\gamma)^{-1}c^{1-\gamma}$$

($0 < \gamma \neq 1$), an increase in risk increases savings if $\gamma > 1$ and decreases it if $\gamma < 1$, with the $U(c) = \ln(c)$ case being the watershed for which savings is unaffected. For broader classes of utility functions, this theorem provides a partial characterization only, suggesting different investors react differently according to whether they display declining or increasing relative risk aversion.

A more complete characterization of the issue of interest is afforded if we introduce the concept of prudence, first proposed by [Kimball \(1990\)](#). Let

$P(c) = (-U'''(c))/(U''(c))$ be a measure of absolute prudence, while by analogy with risk aversion,

$cP(c) = (-cU'''(c))/(U''(c))$ then measures relative prudence. **Theorem 5.7** can now be restated as [Theorem 5.8](#).

Theorem 5.8 Let \tilde{R}_A and \tilde{R}_B be two return distributions such that $\tilde{R}_A \leq_{SSD} \tilde{R}_B$, and let s_A and s_B be, respectively, the savings out of Y_0 corresponding to the return distributions \tilde{R}_A and \tilde{R}_B . Then,

$$s_A > s_B \text{ iff } c\mathbf{P}(c) \leq 2$$

and conversely,

$$s_A < s_B \text{ iff } c\mathbf{P}(c) > 2$$

That is, risk-averse individuals with relative prudence lower than 2 decrease savings while those with relative prudence above 2 increase savings in the face of an increase in the riskiness of returns.

Proof We have seen that $s_A < s_B$ if and only if $g''(R) > 0$. From Eq. (5.6), this means

$$sRU'''(sR)/U''(sR) < s - 2, \text{ or}$$

$$c\mathbf{P}(c) = \frac{sRU'''(sR)}{-U''(sR)} > 2$$

as claimed. The other part of the proposition is proved similarly.

5.6.2 Illustrating Prudence

The relevance of the concept of prudence can be illustrated in the simplest way if we turn to a slightly different problem, where one ignores uncertainty in returns (assuming, in effect, that the net return is identically zero) while asking how savings in period zero is affected by uncertain labor income in period 1. Our remarks in this context are drawn from Kimball (1990).

Let us write the agent's second period labor income, Y , as $Y = \bar{Y} + \tilde{Y}$, where \bar{Y} is the mean labor income and \tilde{Y} measures deviations from the mean (of course, $E\tilde{Y} = 0$). The simplest form of the decision problem facing the agent is thus:

$$\max_s E[U(Y_0 - s) + \beta U(s + \bar{Y} + \tilde{Y})]$$

where $s = s_i$ satisfies the FOC.

i. $U'(Y_0 - s_i) = \beta E(U'(s_i + \bar{Y} + \tilde{Y})\}$

It will be of interest to compare the solution s_i to the above FOC with the solution to the analogous decision problem, denoted s_{ii} , in which the uncertain labor income component is absent. The latter FOC is simply

$$\text{ii. } U'(Y_0 - s_{ii}) = \beta U'(s_{ii} + \bar{Y})$$

The issue once again is whether and to what extent s_i differs from s_{ii} .

One approach to this question, which gives content to the concept of prudence, is to ask what the agent would need to be paid (what compensation is required in terms of period 2 income) to ignore labor income risk—in other words, for his first-period consumption and savings decision to be unaffected by uncertainty in labor income.

The answer to this question leads to the definition of the *compensating precautionary premium* $\psi = \psi(\bar{Y}, \tilde{Y}, s)$ as the amount of additional second period wealth (consumption) that must be given to the agent in order that the solution to (i) coincides with the solution to (ii). That is, the compensatory precautionary premium $\psi(\bar{Y}, \tilde{Y}, s)$ is defined as the solution of

$$U'(Y_0 - s_{ii}) = \beta E\{U'(s_{ii} + \bar{Y} + \tilde{Y} + \psi(\bar{Y}, \tilde{Y}, s))\}$$

[Kimball \(1990\)](#) proves the following two results.

Theorem 5.9 Let $U(\cdot)$ be three times continuously differentiable and $\mathbf{P}(s)$ be the index of absolute prudence. Then

$$\text{i. } \psi(\bar{Y}, \tilde{Y}, s) \approx 1/2\sigma_{\tilde{Y}}^2 \mathbf{P}(s + \bar{Y})$$

Furthermore, let $U_1(\cdot)$ and $U_2(\cdot)$ be two second-period utility functions for which

$$\mathbf{P}_1(s) = \frac{-U'''_1(s)}{U''_1(s)} < \frac{-U'''_2(s)}{U''_2(s)} = \mathbf{P}_2(s), \text{ for all } s$$

Then

$$\text{ii. } \psi_2(\bar{Y}, \tilde{Y}, s) > \psi_1(\bar{Y}, \tilde{Y}, s) \text{ for all } s, \bar{Y}, \tilde{Y}$$

Theorem 5.9 (i) shows that investors' precautionary premia are directly proportional to the product of their prudence index and the variance of their uncertain income component, a result analogous to the characterization of the measure of absolute risk aversion obtained in Section 4.3. The result of [Theorem 5.9](#) (ii) confirms the intuition that the more "prudent" the agent, the greater the compensating premium.

5.6.3 The Joint Saving–Portfolio Problem

Although for conceptual reasons we have so far distinguished the consumption–savings and the portfolio allocation decisions, it is obvious that the two decisions must be considered jointly. We now formalize the consumption/savings/portfolio allocation problem:

$$\max_{\{a, s\}} U(Y_0 - s) + \delta EU(s(1 + r_f) + a(\tilde{r} - r_f)) \quad (5.8)$$

where s denotes the total amount saved and a is the amount invested in the risky asset. Specializing the utility function to the form $U(Y) = (Y^{1-\gamma})/(1 - \gamma)$, the FOCs for this joint decision problem are

$$\begin{aligned}s &: (Y_0 - s)^{-\gamma}(-1) + \delta E([s(1 + r_f) + a(\tilde{r} - r_f)]^{-\gamma}(1 + r_f)) = 0 \\ a &: E[(s(1 + r_f) + a(\tilde{r} - r_f))^{-\gamma}(\tilde{r} - r_f)] = 0\end{aligned}$$

The first equation spells out the condition to be satisfied at the margin for the savings level—and by corollary, consumption—to be optimal. It involves comparing the marginal utility today with the expected marginal utility tomorrow, with the rate of transformation between consumption today and consumption tomorrow being the product of the discount factor and the gross risk-free return. This FOC need not occupy us any longer here.

The interesting element is the solution to the second FOC: it has the exact same form as Eq. (5.2) with the endogenous (optimal) s replacing the exogenous initial wealth level Y_0 . Let us rewrite this equation as

$$s^{-\gamma} E \left[\left((1 + r_f) + \frac{a}{s} (\tilde{r} - r_f) \right)^{-\gamma} (\tilde{r} - r_f) \right] = 0$$

which implies

$$E \left[\left((1 + r_f) + \frac{a}{s} (\tilde{r} - r_f) \right)^{-\gamma} (\tilde{r} - r_f) \right] = 0$$

This equation confirms the lessons of Eqs. (5.3) and (5.4): For the selected utility function, the proportion of savings invested in the risky asset is independent of s , the amount saved. This is an important result: while it does not generalize to other utility functions, it nevertheless opens up the possibility of a straightforward extension of the savings–portfolio problem to many periods. We pursue this important extension in Chapter 16.

5.7 Generalizing the VNM-Expected Utility Representation

Project evaluation exercises and the household’s consumption–savings problem are fundamentally intertemporal in nature. In an intertemporal context, however, various preference phenomena provide additional challenges to the VNM-expected utility framework. It is these extensions that we propose to address in the remaining section of this chapter.

While proving to be extremely useful in explaining real-world phenomena, the extensions we consider are not “radical.” In particular, preferences continue to be defined over money payoffs rather than gains or losses relative to some benchmark, and all proposed extensions reduce to our standard VNM-expected utility set up for specific choices of the relevant parameters.

5.7.1 Preferences for the Timing of Uncertainty Resolution

Under the VNM-expected utility representation, investors are assumed to be concerned only with actual final payoffs and the cumulative probabilities of attaining them. In particular, they are assumed to be indifferent to the timing of uncertainty resolution. To get a better idea of what this means in a choice-theoretic context, consider the two investment payoff trees depicted in [Figure 5.1](#). These investments are to be evaluated from the viewpoint of date 0 (today).

Under the expected utility postulates, these two payoff structures would be valued (in utility terms) identically as

$$EU(\tilde{P}) = U(100) + \delta[\pi U(150) + (1 - \pi)U(25)]$$

where δ is the investor's subjective discount factor.

This means that a VNM investor would not care if the uncertainty were resolved in period 1 or one period later.¹⁰ Yet, people are, in fact, very different in this regard. Some want to know the outcome of an uncertain event as soon as possible; others prefer to postpone it as long as possible.¹¹

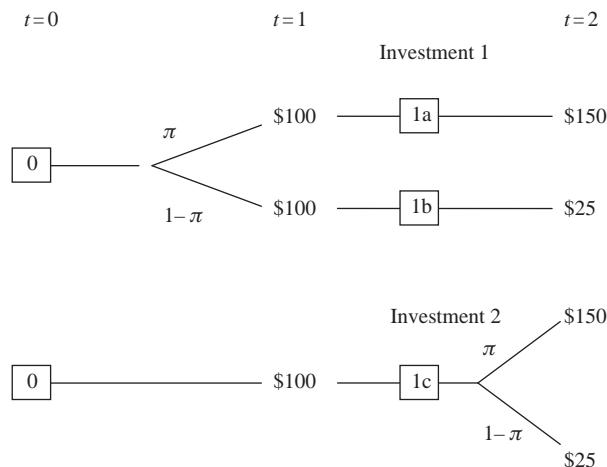


Figure 5.1
Two investments: payoff patterns for earlier versus later resolution of uncertainty.

¹⁰ In particular, preference for the timing of uncertainty resolution violates VNM-expected utility axiom C.1a (see Chapter 3).

¹¹ One commonplace illustration of these differences arises in persons' differing attitudes "to seeing the doctor." When confronted with an unforeseen pain, some individuals immediately visit their physician to have it precisely diagnosed. Others let the days pass in the hope the pain will "go away" of its own accord. Only if it persists will they (reluctantly) see their physician for a diagnosis. There are, of course, many other reasons for avoiding the doctor, such as the inability to pay his fee.

Kreps and Porteus (1978) were the first to develop a theory that allowed for preferences for the timing of uncertainty resolution. They showed that if investor preferences over uncertain sequential payoffs were of the form

$$U_0(P_1, P_2(\tilde{\theta})) = W(P_1, E(U_1(P_1, P_2(\tilde{\theta}))))$$

then investors would prefer early (late) resolution of uncertainty according to whether $W(P_1, \cdot)$ is convex (concave) (loosely, whether $W_{22} > 0$ or $W_{22} < 0$). In the above representation P_i is the payoff in period $i = 1, 2$. If $W(P_1, \cdot)$ were concave, for example, the expected utility of investment 1 would be lower than investment 2.

Let us illustrate this distinction using the choices in Figure 5.1. We assume functional forms similar to those used in an illustration of Kreps and Porteus (1978); in particular, assume $W(P_1, EU) = EU^{1.5}$, and $U_1(P_1, P_2(\tilde{\theta})) = (P_1 + P_2(\tilde{\theta}))^{1/2}$. Let $\pi = 0.5$ and note that the overall composite function $U_0(\cdot)$ is concave in all of its arguments. In computing utilities at the decision nodes [0], [1a], [1b], and [1c] (the latter decisions are trivial ones), we must be especially scrupulous to observe exactly the dates at which the uncertainty is resolved under the two alternatives:

$$[1a]: EU_1^{1a}(P_1, P_2(\theta)) = (100 + 150)^{1/2} = 15.811$$

$$[1b]: EU_1^{1b}(P_1, P_2(\theta)) = (100 + 25)^{1/2} = 11.18$$

$$[1c]: EU_1^{1c}(P_1, P_2(\theta)) = 0.5(100 + 150)^{1/2} + 0.5(100 + 25)^{1/2} = 13.4955$$

At $t = 0$, the expected utility on the upper tree is

$$\begin{aligned} EU_0^{1a,1b}(P_1, P_2(\tilde{\theta})) &= EW^{1a,1b}(P_1, P_2(\tilde{\theta})) \\ &= 0.5W(100, 15.811) + 0.5W(100, 11.18) \\ &= 0.5(15.811)^{1.5} + 0.5(11.18)^{1.5} \\ &= 50.13 \end{aligned}$$

while on the lower tree

$$EU_0^{1c}(P_1, P_2(\tilde{\theta})) = W(100, 13.4955) = (13.4955)^{1.5} = 49.57$$

This investor clearly prefers early resolution of uncertainty, which is consistent with the convexity of the $W(\cdot)$ function. Note that the result is simply an application of Jensen's inequality.¹² If $W(\cdot)$ had been concave, the ordering would be reversed.

Recent empirical evidence (e.g., Brown and Kim (2013)) suggests that a majority of individuals prefer early resolution of uncertainty, a fact that may have special significance

¹² Let $a = (100 + 150)^{1/2}$, $b = (100 + 25)^{1/2}$, $g(x) = x^{1.5}$ (convex), $EU_0^{1a,1b}(P_1, P_2(\tilde{\theta})) = Eg(x) > g(Ex) = EU_0^{1c}(P_1, P_2(\tilde{\theta}))$, where $x = a$ with prob = 0.5 and $x = b$ with prob = 0.5.

for the pricing of stocks.¹³ The intuition goes along the following lines: Bansal and Yaron (2004) demonstrate that the aggregate US consumption and dividend (for the CRSP index of US stocks) data series are consistent with both having a common, small, highly persistent mean growth component.¹⁴ This means (e.g., in the case of dividends) there will be long periods of high average dividend growth followed by long periods where average dividend growth is low, and vice versa. Because of the high persistence, these series reflect the late resolution of uncertainty. When combined with preferences for early resolution of uncertainty on the part of market participants, this persistence phenomenon leads to low equity prices in an equilibrium asset pricing context (see Chapter 10): investors will only hold late-resolution-of-uncertainty equity securities if their prices are very low. In fact, in the Bansal and Yaron (2004) equilibrium asset pricing context average equity returns are high enough (equity prices are low enough) to resolve the equity premium puzzle (see Chapter 2).¹⁵

The functional forms used in the prior illustrative example are not widely used. In Section 5.7.3, however, we introduce the popular Epstein-Zin (1991) utility specification which, depending on the choice of parameter values, can also display preference for the early or late uncertainty resolution.

5.7.2 Preferences That Guarantee Time-Consistent Planning

The notion of *time-consistent planning* is this: if, at each date, the agent could plan against any future contingency, what is the required relationship among the family of orderings $\{\succeq_t : t = 0, 1, 2, \dots, T\}$ that will cause plans that were optimal with respect to preferences \succeq_0 to remain optimal in all future time periods given all that may happen in the interim (i.e., intermediate consumption experiences and the specific way uncertainty has evolved)? In particular, what utility function representation will guarantee this property?

When considering decision problems over time, such as portfolio investments over a multiperiod horizon, time consistency would appear to be a desirable property. In its absence, one would observe portfolio rebalancing not motivated by any outside event or information flow, but simply resulting from the inconsistency of the date t preference ordering of the investor compared with the preferences on which her earlier portfolio positions were based. Asset trades would then be fully motivated by endogenous and unobservable preference issues and would thus be basically unexplainable.

¹³ For the experimental results presented in Brown and Kim (2013), “a majority (60.4%) of subjects demonstrate a preference for early resolution (of uncertainty), about one-third are indifferent (36.6%), and 3% exhibit a preference for late resolution” (Brown and Kim, 2013, p. 2).

¹⁴ The CRSP index comprises nearly all traded USA stocks and is compiled by the Center for Research in Security Prices at the University of Chicago.

¹⁵ We explore this topic in greater detail in Chapter 10.

To see what it takes for a utility function to be time consistent, let us consider two periods where at date 1 any one of $\theta \in S$ possible states of nature may be realized. Let c_0 denote a possible consumption level at date 0, and let $c_1(\theta)$ denote a possible consumption level in period 1 if state “ θ ” occurs. [Johnsen and Donaldson \(1985\)](#) demonstrate that if initial preferences \succeq_0 , with utility representation $U(\cdot)$, are to guarantee time-consistent planning, there must exist continuous and monotone increasing functions $f(\cdot)$ and $\{U_\theta(\cdot, \cdot) : \theta \in S\}$ such that

$$U(c_0, c_1(\theta)) : \theta \in S = f(c_0, U_\theta(c_0, c_1(\theta))) : \theta \in S \quad (5.9)$$

where $U_\theta(\cdot, \cdot)$ is the state θ contingent utility function.¹⁶

This result means the utility function must be of a form such that the utility representations in future states can be recursively nested as individual arguments of the overall utility function. This condition is satisfied by the VNM-expected utility form

$$U(c_0, c_1(\theta)) : \theta \in S = U_0(c_0) + \sum_{\theta} \pi_\theta U(c_1(\theta)) \quad (5.10)$$

which clearly display the structure of [Eq. \(5.9\)](#).¹⁷ The VNM-utility representation is thus time consistent, but the latter property can also be accommodated by more general utility functions which are not VNM. To see this, consider the following specialization of representation [\(5.9\)](#), where there are three possible states at $t = 1$:

$$U(c_0, c_1(1), c_1(2), c_1(3)) = \left\{ c_0 + \pi_1 U_1(c_0, c_1(1)) + [\pi_2 U_2(c_0, c_1(2))]^{\frac{1}{3}} \pi_3 U_3(c_0, c_1(3)) \right\}^{\frac{1}{2}} \quad (5.11)$$

where

$$\begin{aligned} U_1(c_0, c_1(1)) &= \log(c_0 + c_1(1)), \\ U_2(c_0, c_1(2)) &= (c_0)^{\frac{1}{2}} (c_1(2))^{\frac{1}{2}}, \text{ and} \\ U_3(c_0, c_1(3)) &= c_0 c_1(3) \end{aligned}$$

¹⁶ With the utility-of-money function defined here over state-contingent consumption, we are implicitly assuming one composite consumption good with a normalized price of one in every period. State-contingent consumption, $c_t(\theta)$, thus also represents state contingent income $Y_t(\theta)$ available for spending after savings have been set aside. The consumption-income identity legitimizes our use of the utility-of-money function $U(\cdot)$ as simultaneously representing a utility of consumption function.

¹⁷ The many period extension (slightly generalized) of [Eq. \(5.10\)](#) is to postulate that investor preferences over contingent consumption plans assume the form $E_0(\sum_{t=0}^T \delta^t u(c_t))$, $0 < \delta < 1$. At any future date $t > 0$, the operative ordering thus becomes $E_t(\sum_{j=0}^{T-t} \delta^{t+j} u(c_{t+j})) = \delta^t E_0(\sum_{j=0}^{T-t} \delta^j u(c_{t+j}))$ which, by the ordinal property of utility functions, is equivalent to [Eq. \(5.9\)](#). Accordingly, preferences of this form always admit time consistent planning.

In this example, preferences are clearly not linear in the probabilities and thus they are not of the VNM-expected utility type. Nevertheless, identification (5.11) is of the form of Eq. (5.9). It also has the feature that preferences in any future state are independent of irrelevant alternatives, where the irrelevant alternatives are those consumption plans for states that do not occur. As such, agents with these preferences will never experience regret, and the Allais paradox will not be operational.

Consistency of choice makes sense and turns out to be important for individual and national savings behavior, but is it borne out empirically? Unfortunately, the answer is: frequently not. A simple illustration of this is a typical pure-time-preference experiment from the psychology literature (uncertainty in future states is not even needed). Participants are asked to choose among the following monetary prizes:¹⁸

Question 1: Would you prefer \$100 today or \$200 in 2 years?

Question 2: Would you prefer \$100 in 6 years or \$200 in 8 years?

Respondents often prefer the \$100 in Question 1 and the \$200 in Question 2, not realizing that Question 2 involves the same choice as Question 1 but with a 6-year delay. If these people behave true to their answers, they will be time inconsistent: in the case of Question 2, although they state their preference now for the \$200 prize in 8 years, when year 6 arrives they will take the \$100 and run! Is there a way this choice reversal can be conveniently represented in a conventional, utility-based framework? Providing an answer to this question constitutes our next topic.

5.7.2.1 Quasi-Hyperbolic Discounting

The commonplace reversal cited above suggests very different attitudes toward *intended* saving in the future versus actual savings today. This is apparent from a slight reframing of the two questions posed above into a consumption–savings context.

Question 1': Would you prefer \$100 today or the opportunity to save and invest the \$100 for an assured payoff in 2 years of \$200?

Question 2': Would you prefer \$100 in 6 years or the opportunity to save and invest the \$100 for 2 more years for an assured payoff of \$200?

While Questions 1' and 2' are framed slightly differently from Questions 1 and 2, a study of savings patterns would reveal similar responses: \$100 today preferred to \$200 in 2 years, yet investing the \$100 in 6 years being preferred to receiving the \$100 at that time. In other words, investors apparently are willing to commit to savings in the future while being less willing to save today: there is a bias for “immediate gratification.” A convenient way to

¹⁸ See Ainslie and Haslan (1992) for details. Similar illustrations may be found in Thaler (1981).

summarize these choices formally within the general expected utility framework is to define a family of preference orderings $\{\mathbb{U}_t(\cdot): t = 0, 1, 2, \dots\}$ by

$$\mathbb{U}_t(\cdots) = E_t \left\{ u(c_t) + \beta \sum_{j=1}^T \delta^{t+j} u(c_{t+j}) \right\} \quad (5.12)$$

where $\beta < 1$.¹⁸ Note that for such an investor, his consumption/savings decision between periods $t+j$ and $t+j+1$, $j > 0$, as planned today (period t) is governed by the subjective discount factor δ . Yet when period t transpires, his consumption savings behavior will be governed by the subjective discount factor $\delta\beta < \delta$. It is as though the investor has a dual personality. Individuals of this persuasion will behave dynamically time inconsistently: they will plan to save a lot in the future, whereas “when the future arrives” they end up saving very much less. Following [Laibson \(1997\)](#) preferences of the form (5.12) are now referred to as quasi-hyperbolic.

In situations where preferences are of the form (5.12), the ability of the individual consumer–investor to achieve an appropriate life-cycle consumption/savings plan becomes an issue of the need for self-control in the form of a supporting strategy of “commitment.” By a commitment strategy, we mean one by which “self t ” binds his “future selves” to save or retain assets in excess of what they would otherwise voluntarily accomplish when they themselves become the decision makers. In practice, commitment investing frequently takes the form of investing in illiquid assets, assets for which the early sale is either forbidden, except in very narrowly defined circumstances of ill-health, or limited by costly penalties for early withdrawal. Savings in the form of contributions to employer-sponsored retirement accounts, IRAs, and even short-term Christmas-club accounts (recently reintroduced by Walmart) are commonplace illustrations. The acquisition of a home financed by a mortgage contract is the most widespread of commitment devices: not only is a home difficult to sell (liquidate) quickly but its associated mortgage payments become a form of required savings. [Laibson \(1997\)](#) illustrates the notion of commitment in the context of illiquid versus liquid asset acquisition in a formal game-theoretic (current self versus future selves) consumption–savings context.

[Phelps and Pollak \(1968\)](#) originally used preferences of the form (5.12) to explain historically low US savings rates in the United States. Using analysis based on preference ordering (5.12) [Laibson \(1997\)](#) similarly argues that the decline in the US savings rate beginning in the 1980s may have been due to the proliferation of unsecured debt in the

¹⁸ This family of ordering was first proposed by [Phelps and Pollak \(1968\)](#) and later resurrected and further developed by [Laibson \(1997\)](#).

form of easily accessible credit cards.¹⁹ Already low national savings rates can be lowered even further when financial liberalization allows the commitment devices noted above to be circumvented. The increased ease of obtaining home equity loans (which allow the partial liquidation of real estate wealth) in the years leading up to the financial crisis is a case in point. We are also reminded of the work of Jappelli and Pagano (1994) (Chapter 1), who demonstrated that financial deregulation in Italy led to a reduction in the Italian savings rate. Although the focus of this deregulation was the removal of financial constraints, in some respects commitment devices were thereby removed as well.

Sections 5.7.1 and 5.7.2 have focused largely on time preference related issues. In the final segment of our discussion, we bring back risk preferences and, in particular, consider the joint interaction of time and risk preferences.

5.7.3 Separating Risk and Time Preferences

Consider the standard consumption–savings problem (5.5), and suppose once again that the agent’s period utility function has been specialized to have the standard CRRA form,

$$U(c) = \frac{Y^{1-\gamma}}{1-\gamma}, \quad \gamma > 0$$

For this utility function, the single-parameter γ captures not only the agent’s sensitivity to atemporal risk, but also his sensitivity to consumption variation across time periods or, equivalently, his willingness to substitute consumption in one period for consumption in another. A high γ signals a strong desire to avoid atemporal consumption risk and, simultaneously, a strong reluctance to substitute consumption in one period for consumption in another. To see this more clearly, consider a deterministic version of Problem (5.5) where $\delta < 1$, $\tilde{R} \equiv 1$:

$$\max_{0 \leq s \leq Y_0} \{U(Y_0 - s) + \delta U(s)\}$$

The necessary and sufficient FOC

$$\begin{aligned} -(Y_0 - s)^{-\gamma} + \delta s^{-\gamma} &= 0 \text{ or} \\ \left(\frac{1}{\delta}\right)^{\frac{1}{\gamma}} &= \left(\frac{Y_0 - s}{s}\right) \end{aligned}$$

¹⁹ The national savings rate is important because it largely determines an economy’s long-run capital stock available to each of its workers, and thus per worker productivity and consumption. Economies that save more will, ceteris paribus, also enjoy lower interest rates. See Web Chapter A.

With $\delta < 1$, as the agent becomes more and more risk averse ($\gamma \mapsto \infty$), $((Y_0 - s)/s) = (c_0/c_1) \mapsto 1$; i.e., $c_0 \approx c_1$. For this preference structure, a highly risk-averse agent will also seek an intertemporal consumption profile that is very smooth.

We have stressed repeatedly the pervasiveness of the preference for smooth consumption whether across time or across states of nature, and its relationship with the notion of risk aversion. It is time to recognize that while in an atemporal setting a desire for smooth consumption across states of nature is the very definition of risk aversion, in a multiperiod environment, risk aversion and the desire for intertemporal consumption smoothing should not necessarily be equated. After all, one may speak of intertemporal consumption smoothing in a no-risk, deterministic setting, and one may speak of risk aversion in an uncertain, atemporal environment. The situation considered so far where the same parameter determines both is thus restrictive. Indeed, empirical studies tend to suggest that typical individuals are more averse to intertemporal substitution (they desire very smooth consumption intertemporally) than they are averse to risk *per se*. This latter fact cannot be captured in the aforementioned, single-parameter setting.

Is it possible to generalize the standard utility specification and break this coincidence of time and risk preferences? [Epstein and Zin \(1989, 1991\)](#) answer positively and propose a class of utility functions that allows each dimension to be parameterized separately while still preserving the time consistency property discussed in [Section 5.7.2](#). They provide, in particular, the axiomatic basis for preferences over lotteries leading to the [Kreps and Porteus \(1978\)](#)-like utility representation:

$$U_t = U(c_t, c_{t+1}, c_{t+2}, \dots) = W(c_t, \text{CE}(\tilde{U}_{t+1})) \quad (5.13)$$

where

- i. $U_t = U(c_t, \tilde{c}_{t+1}, \tilde{c}_{t+2}, \dots)$ describes the investor's period t future lifetime utility associated with consumption stream $(c_t, \tilde{c}_{t+1}, \tilde{c}_{t+2}, \dots)$
- ii. $W(,)$ is an aggregator function analogous to the $f()$ function in [Eq. \(5.9\)](#); and
- iii. $\text{CE}(\tilde{U}_{t+1})$ is the period t certainty equivalent of uncertain lifetime utility beginning at date $t + 1$, measured in terms of period $t + 1$ consumption units.

[Epstein and Zin \(1989\)](#) and [Weil \(1989\)](#) explore the following CES (constant elasticity of intertemporal substitution)-like specialized version of [Eq. \(5.13\)](#):

$$W(c_t, \text{CE}(\tilde{U}_{t+1})) = \left[(1-\delta)c_t^{1-\rho} + \delta(\text{CE}(\tilde{U}_{t+1}))^{1-\rho} \right]^{\frac{1}{1-\rho}}, \quad \rho \neq 1 \text{ or} \quad (5.14a)$$

$$W(c_t, \text{CE}(\tilde{U}_{t+1})) = (1 - \delta)\log c_t + \delta\log(\text{CE}(\tilde{U}_{t+1})), \quad \rho = 1 \quad (5.14b)$$

where $1/\rho$ is the elasticity of intertemporal substitution. Since a higher ρ means the investor prefers a smoother intertemporal, certain consumption stream, ρ is viewed as the investor's time preference parameter.

The certainty equivalent of future lifetime utility, $\text{CE}(\tilde{U}_{t+1})$, is computed in a manner directly analogous to what was done in Sections 4.4 and 4.5. It thus identifies γ as the risk preference parameter, where γ and ρ may be chosen independently of one another.

$$[\text{CE}(\tilde{U}_{t+1})] = \left(E_t \left(\tilde{U}_{t+1}^{1-\gamma} \right) \right)^{\frac{1}{1-\gamma}}, 1 \neq \gamma > 0, \text{ or} \quad (5.15a)$$

$$\log \text{CE}(\tilde{U}_{t+1}) = E_t(\log \tilde{U}_{t+1}), \gamma = 1 \quad (5.15b)$$

By Eq. (5.13), $U_t(\cdot)$ is recursively determined and, in general, has no time-separable representation. Accordingly, Epstein and Zin (1989) preferences are not easy to use. In a consumption savings context, for example, the lifetime utility function U_t , the savings decision s_t and the portfolio allocation decision (risk free versus risky assets) must all be determined simultaneously and endogenously.²⁰

As for preference for the timing of uncertainty resolution, if $\gamma > \rho$, it can be demonstrated that the investor prefers early resolution and vice versa if $\gamma < \rho$. If $\gamma = \rho$, recursive substitution of U_t yields

$$U_t = \left[(1-\delta) E_t \sum_{j=0}^{\infty} \delta^j c_{t+j}^{1-\gamma} \right]^{\frac{1}{1-\gamma}}$$

which represents the same preference as

$$E_t \sum_{j=0}^{\infty} \delta^j c_{t+j}^{1-\gamma}$$

and is thus equivalent to the usual time-separable expected utility case with CRRA utility. In general Eqs. (5.11)–(5.12) are not of the expected utility form as probabilities do not enter linearly.

As amply demonstrated in Weil (1989), Epstein-Zin (1989) preferences, in and of themselves, will not resolve the equity premium puzzle. As noted at the close of Section 5.7.1, further structure must be placed on the manner in which consumption uncertainty is modeled (see Chapter 10).

5.8 Conclusions

We have considered, in a very simple context, the relationship between an investor's degree of risk aversion, on the one hand, and his desire to save and the composition of his portfolio

²⁰ See van Binsbergen et al. (2008) for one numerical solution technique; also see the web notes to this chapter.

on the other. Most of the results were intuitively acceptable, and that, in itself, makes us more confident of the VNM representation.

Are there any lessons here for portfolio managers to learn? At least three lessons are suggested:

1. Regardless of the level of risk, some investment in risky assets is warranted, even for the most risk-averse clients (provided $E\tilde{r} > r_f$). This is the substance of [Theorem 5.1](#).
2. As the value of a portfolio changes significantly, the asset allocation (proportion of wealth invested in each asset class) and the risky portfolio composition should be reconsidered. How that should be done depends critically on the client's attitudes toward risk. This is the substance of [Theorems 5.4–5.6](#).
3. Investors are willing, in general, to pay to reduce income (consumption) risk and would like to enter into mutually advantages transactions with institutions less risk averse than themselves. The extreme case of this is illustrated in [Section 5.5](#).

We went on to consider how greater return uncertainty influences savings behavior. On this score and in some other instances, this chapter has illustrated the fact that, somewhat surprisingly, risk aversion is not always a sufficient hypothesis to recover intuitive behavior in the face of risk. The third derivative of the utility function often plays a role. The notion of prudence permits an elegant characterization in these situations.

We concluded the chapter by considering three plausible modifications of the general expected utility framework in order to capture phenomena not representable under the strict VNM-expected utility axioms. As we will see in future chapters, preference representations allowing for both the separation of time and risk parameters and a preference for the early timing of uncertainty resolution, will be fundamental to a resolution of the equity premium phenomenon.

In many ways, this chapter has aimed at providing a broad perspective allowing us to place modern portfolio theory and its underlying assumptions in their proper context. We are now prepared to revisit this pillar of modern finance.

References

- Ainslie, G., Haslan, N., 1992. Hyperbolic discounting. In: Lowenstein, G., Elster, J. (Eds.), *Choice Over Time*. Russell Sage Foundation, New York, NY.
- Arrow, K.J., 1971. *Essays in the Theory of Risk Bearing*. Markham, Chicago, IL.
- Bansal, R., Yanon, A., 2004. Risks for the long run: a potential resolution of asset pricing puzzles. *J. Finan.* 59, 1481–1509.
- Brown, A., Kim, H., 2013, Do individuals have preferences used in macro-finance models? An experimental investigation, Working Paper, Texas A&M University, Department of Economics.
- Cass, D., Stiglitz, J.E., 1970. The structure of investor preference and asset returns and separability in portfolio allocation: a contribution to the pure theory of mutual funds. *J. Econ. Theory*. 2, 122–160.

- Epstein, L.G., Zin, S.E., 1989. Substitution, risk aversion, and the temporal behavior of consumption growth and asset returns I: theoretical framework. *Econometrica*. 57, 937–969.
- Epstein, L.G., Zin, S.E., 1991. Substitution, risk aversion, and the temporal behavior of consumption growth and asset returns II: an empirical analysis. *J. Empir. Econ.* 99, 263–286.
- Jappelli, T., Pagano, M., 1994. Savings, growth, and liquidity constraints. *Q. J. Econ.* 109, 83–109.
- Johnsen, T., Donaldson, J.B., 1985. The structure of intertemporal preferences under uncertainty and time consistent plans. *Econometrica*. 53, 1451–1458.
- Kimball, M.S., 1990. Precautionary savings in the small and in the large. *Econometrica*. 58, 53–73.
- Kreps, D., Porteus, E., 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*. 46, 185–200.
- Laibson, D., 1997. Golden eggs and hyperbolic discounting. *Q. J. Econ.* 112, 443–477.
- Phelps, E.S., Pollak, R.A., 1968. On second best national savings and game-equilibrium growth. *Rev. Econ. Stud.* 35, 185–199.
- Rothschild, M., Stiglitz, J., 1971. Increasing risk II: its economic consequences. *J. Econ. Theory*. 3, 66–85.
- Thaler, R., 1981. Some empirical evidence on dynamic inconsistency. *Econ. Lett.* 8, 201–207.
- van Binsbergen, J., Fernandez-Villaverde, J., Koijen, R., Rubio-Ramirez, J., 2008. Working with Epstein–Zin preferences: computation and likelihood estimation of DSGE models with recursive preferences, Working Paper, Duke University.
- Weil, P.h., 1989. The equity premium puzzle and the risk free rate puzzle. *J. Monet. Econ.* 24, 401–421.

Risk Aversion and Investment Decisions, Part II: Modern Portfolio Theory

Chapter Outline

6.1 Introduction	144
6.2 More About Utility Functions and Return Distributions	144
6.3 Refining the Normality-of-Returns Assumption	149
6.4 Description of the Opportunity Set in the Mean–Variance Space: The Gains from Diversification and the Efficient Frontier	152
6.5 The Optimal Portfolio: A Separation Theorem	158
6.6 Stochastic Dominance and Diversification	159
6.7 Conclusions	165
References	166
Appendix 6.1: Indifference Curves Under Quadratic Utility or Normally Distributed Returns	166
Part I	166
Part II	167
<i>U Is Quadratic</i>	168
<i>The Distribution if R Is Normal</i>	168
Proof of the Convexity of Indifference Curves	170
Appendix 6.2: The Shape of the Efficient Frontier; Two Assets; Alternative Hypotheses	171
Perfect Positive Correlation (Figure 6.3)	171
Imperfectly Correlated Assets (Figure 6.4)	171
Perfect Negative Correlation (Figure 6.5)	172
One Riskless and One Risky Asset (Figure 6.6)	172
Appendix 6.3: Constructing the Efficient Frontier	173
The Basic Portfolio Problem	173
Generalizations	174
Nonnegativity Constraints	174
Composition Constraints	175
Adjusting the Data (Modifying the Means)	176
Constraints on the Number of Securities in the Portfolio	177

6.1 Introduction

In the context of the previous chapter, we encountered the following canonical portfolio problem:

$$\max_a EU(\tilde{Y}_1) = \max_a EU[Y_0(1 + r_f) + a(\tilde{r} - r_f)] \quad (6.1)$$

Here the portfolio choice is limited to allocating investable wealth, Y_0 , between a risk-free and a risky asset, a being the amount invested in the latter.

More generally, we can admit N risky assets, with returns $(\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_N)$, as in the Cass–Stiglitz theorem. The above problem in this case becomes

$$\max_{\{a_1, a_2, \dots, a_N\}} EU(Y_0(1 + r_f) + \sum_{i=1}^N a_i(\tilde{r}_i - r_f)) = \max_{\{w_1, w_2, \dots, w_N\}} EU(Y_0(1 + r_f) + \sum_{i=1}^N w_i Y_0(\tilde{r}_i - r_f)) \quad (6.2)$$

[Equation \(6.2\)](#) reexpresses the problem with $w_i = (a_i/Y_0)$, the proportion of wealth invested in the risky asset i , being the key decision variable rather than a_i , the amount of money invested.

The latter expression may further be written as

$$\max_{\{w_1, w_2, \dots, w_N\}} EU \left\{ Y_0 \left[(1 + r_f) + \sum_{i=1}^N w_i (\tilde{r}_i - r_f) \right] \right\} = EU\{Y_0[1 + \tilde{r}_P]\} = EU\{\tilde{Y}_1\} \quad (6.3)$$

where \tilde{Y}_1 denotes the end-of-period wealth and \tilde{r}_P the rate of return on the overall portfolio of assets held.

Modern portfolio theory (MPT) explores the details of portfolio choice such as Problem (6.3), (i) under the mean–variance utility hypothesis and (ii) for an arbitrary number of risky investments, with or without a risk-free asset.¹ The goal of this chapter is to review the fundamentals underlying this theory. We first draw the connection between the mean–variance utility hypothesis and our earlier utility development.

6.2 More About Utility Functions and Return Distributions

What provides utility? As noted in Chapter 3, financial **theory** assumes that the ultimate source of a consumer's satisfaction lies in consuming the goods and services he is able to

¹ [Markowitz \(1952\)](#) developed the theory for portfolios exclusively composed of risky assets. [Tobin \(1958\)](#) added a risk-free asset.

purchase.² Preference relations and utility functions are accordingly defined on bundles of consumption goods (recall Theorem 3.1):

$$u(c_1, c_2, \dots, c_M) \quad (6.4)$$

where the indexing $i = 1, \dots, M$ is across date-state (contingent) commodities: goods are characterized not only by their identity as a product or service but also by the time and state in which they may be consumed. States of nature, however, are mutually exclusive. For each date and state of nature (θ), there is a traditional budget constraint

$$p_{1\theta}c_{1\theta} + p_{2\theta}c_{2\theta} + \dots + p_{M\theta}c_{M\theta} \leq Y_\theta \quad (6.5)$$

where the indexing runs across goods for a given state θ ; in other words, the M quantities $c_{i\theta}$, $i = 1, \dots, M$, and the M prices $p_{i\theta}$, $i = 1, \dots, M$ correspond to the M goods available in state of nature θ , while Y_θ is the (“end-of-period”) wealth level available in that same state. We quite naturally assume that the number of goods available in each state is constant.³

In this context, and in some sense summarizing what we discussed in Chapter 5, it is quite natural to think of an individual’s decision problem as being undertaken sequentially, in three steps.

Step 1, The consumption–savings decision. Here, the issue is deciding how much to consume versus how much to save today: how to split period zero income Y_0 between current consumption spending C_0 and saving S_0 for consumption in the future where

$$C_0 + S_0 = Y_0$$

Step 2, The portfolio problem. At this second step, the problem is to choose assets in which to invest one’s savings so as to obtain the desired pattern of end-of-period wealth across the various states of nature. This means, in particular, allocating $(Y_0 - C_0)$ between the risk-free and the N risky assets with $(1 - \sum_{i=1}^N w_i)(Y_0 - C_0)$ representing the investment in the risk-free asset, and $(w_1(Y_0 - C_0), w_2(Y_0 - C_0), \dots, w_N(Y_0 - C_0))$, representing the vector of investments in the various risky assets.

Step 3, Tomorrow’s consumption choice. Given the realized state of nature and the level of wealth obtained, there remains the issue of choosing consumption bundles to maximize the utility function (Expression (6.4)) subject to Constraint (6.5) where

² Of course this does not mean that nothing else in life provides utility or satisfaction (!) but the economist’s inquiry is normally limited to the realm of market phenomena and economic choices.

³ This is purely formal: if a good is not available in a given state of nature, it is understood nevertheless to exist but with a total, tradable economy-wide endowment of the good being zero.

$$Y_\theta = (Y_0 - C_0) \left[(1 + r_f) + \sum_{i=1}^N w_i (r_{i\theta} - r_f) \right]$$

and $r_{i\theta}$ denotes the *ex post* return to asset i in state θ .

In such problems, it is fruitful to work via backward induction, starting from the end (step 3). Step 3 is a standard microeconomic problem, and for our purpose its solution can be summarized by a utility-of-money function $U(Y_\theta)$ representing the (maximum) level of utility that results from optimizing in step 3 given that the wealth available in state θ is Y_θ .

In other words, we define $U(Y_\theta)$ as

$$U(Y_\theta) \equiv \max_{(c_{1\theta}, \dots, c_{M\theta})} u(c_{1\theta}, \dots, c_{M\theta})$$

$$\text{s.t. } p_{1\theta}c_{1\theta} + \dots + p_{M\theta}c_{M\theta} \leq Y_\theta$$

Naturally enough, maximizing the expected utility of Y_θ across all states of nature becomes the objective of step 2:

$$\max_{\{w_1, w_2, \dots, w_N\}} EU(\tilde{Y}) = \sum_{\theta} \pi_\theta U(Y_\theta)$$

Here π_θ is the probability of state of nature θ . The end-of-period wealth (a random variable) can now be written as $\tilde{Y} = (Y_0 - C_0)(1 + \tilde{r}_P)$, with $(Y_0 - C_0)$ the initial wealth net of date 0 consumption and $\tilde{r}_P = r_f + \sum_{i=1}^N w_i (\tilde{r}_i - r_f)$ the rate of return on the portfolio of assets in which $(Y_0 - C_0)$ is invested. This brings us back to Eq. (6.3):

$$\max_{w_1, \dots, w_N} EU(\tilde{Y}) = \max_{w_1, \dots, w_N} EU((Y_0 - C_0)(1 + \tilde{r}_P))$$

In general, the optimal portfolio allocation decision and the optimal consumption savings decision must be considered jointly. In the case that $U(Y)$ has the constant relative risk aversion (CRRA) form (and, more generally, if $U(Y)$ is homogeneous⁴), however, these decisions can be separated and undertaken sequentially with the former preceding the latter.⁵ To confirm this assertion, observe that if

$$U(Y) = \frac{1}{1-\gamma} (Y)^{1-\gamma}$$

⁴ A utility function $U(Y)$ is said to be homogeneous if and only if it has the property that for any $\Delta > 0$, $U(\Delta Y) = \Delta^v U(Y)$ for some v , a real number. As such, $U(Y)$ is said to be homogeneous of degree v .

⁵ Footnote 1 of Chapter 4 details the form of $u(c_1, \dots, c_M)$ such that $U(Y)$ is CRRA.

$$\begin{aligned}
 \max_{w_1, \dots, w_N} EU(\tilde{Y}) &= \max_{w_1, \dots, w_N} E\left(\frac{1}{1-\gamma}((Y_0 - C_0)(1 + \tilde{r}_P))^{1-\gamma}\right) \\
 &= \max_{w_1, \dots, w_N} E\left((Y_0 - C_0)^{1-\gamma}\left(\frac{1}{1-\gamma}\right)(1 + \tilde{r}_P)^{1-\gamma}\right) \\
 &= \left\{(Y_0 - C_0)^{1-\gamma} \cdot \max_{w_1, \dots, w_N} E\hat{U}(1 + \tilde{r}_P)\right\}
 \end{aligned} \tag{6.6}$$

where $\hat{U}(1 + r_P) = ((1 + r_P)^{1-\gamma})/(1 - \gamma)$. Although seemingly defined over rates of return, $\hat{U}(\cdot)$ technically remains a utility-of-money function because its domain represents the wealth accruing to the investor in period 1 if he saves and invests \$1.00 in period zero at the rate r_P . Note also that decomposition (6.6) allows the investor to complete, first, step 2 of his overall decision problem and then step 1 as per backward induction. Step 1 requires that the investor choose his level of savings ($Y_0 - C_0$) so as optimally to trade off utility at $t = 0$ against expected utility at $t = 1$ given that each unit of savings is invested optimally, i.e., in a portfolio whose proportions are the solution to Problem (6.6).

Accordingly, the investor's comprehensive savings–portfolio composition problem becomes

$$\max_{C_0, w_1, \dots, w_N} \frac{(C_0)^{1-\gamma}}{1-\gamma} + \delta \left\{ (Y_0 - C_0)^{1-\gamma} \max_{w_1, \dots, w_N} E\hat{U}(1 + \tilde{r}_P) \right\}$$

as per our discussion in Section 5.6.3. In particular, the investor's optimal portfolio proportions will be the same irrespective of the amount he decides to invest.

The remainder of this chapter will focus exclusively on the determination of optimal portfolio proportions, i.e., on step 2 above (for step 1, see Chapter 5). Accordingly, we will assume $U(Y)$ is homogeneous, drop the $U(Y), \hat{U}(r_P)$ distinction, and consider utility functions $U(\cdot)$ defined interchangeably over r_P or Y . We next take advantage of a very useful mathematical approximation.

Using a simple Taylor series approximation, one can see that the mean and variance of an investor's wealth distribution are the critical elements to the determination of his expected utility *for any distribution* and any concave utility function. Let \tilde{Y} denote an investor's uncertain end period wealth, and $U(\cdot)$ as his utility-of-money function. The Taylor series approximation for the investor's utility of wealth $U(Y)$ around $E(\tilde{Y})$ yields

$$U(Y) = U[E(\tilde{Y})] + U'[E(\tilde{Y})][Y - E(\tilde{Y})] + \frac{1}{2}U''[E(\tilde{Y})][Y - E(\tilde{Y})]^2 + H_3 \tag{6.7}$$

where $H_3(Y) = \sum_{j=3}^{\infty} \frac{1}{j!} U^{(j)}[E(\tilde{Y})][Y - E(\tilde{Y})]^j$.

Now let us compute expected utility using this approximation:

$$\begin{aligned} EU(\tilde{Y}) &= U[E(\tilde{Y})] + U'[E(\tilde{Y})] \underbrace{[E(\tilde{Y}) - E(\tilde{Y})]}_{=0} \\ &\quad + \frac{1}{2} U''[E(\tilde{Y})] \underbrace{E[\tilde{Y} - E(\tilde{Y})]^2}_{=\sigma^2(\tilde{Y})} + E\tilde{H}_3 \\ &= U[E(\tilde{Y})] + \frac{1}{2} U''[E(\tilde{Y})]\sigma^2(\tilde{Y}) + E\tilde{H}_3 \end{aligned}$$

If $E\tilde{H}_3$ is small, $E(\tilde{Y})$ and $\sigma^2(\tilde{Y})$ become central to determining $EU(\tilde{Y})$, at least to a first approximation.⁶ Since \tilde{Y} inherits the form of the distribution on $\tilde{r}_P(\tilde{Y} = Y_0(1 + \tilde{r}_P))$, μ_P and σ_P are, by extension, central components of \tilde{r}_P .⁷ In other words, we could substitute r_P for Y in the discussion above and arrive at the analogous conclusion. It remains to dispense with the approximation error term $E\tilde{H}_3$ and in this regard there are two routes open to us.

First, if $U(Y)$ is a quadratic function, U'' is a constant and, as a result, $H_3 \equiv 0$, so $E(\tilde{Y})$ and $\sigma^2(\tilde{Y})$ are all that matter.⁸ It follows by equivalence that μ_P and σ_P alone determine $EU(\tilde{r}_P)$. Assuming the utility function is quadratic, however, is not fully satisfactory since the preference representation would then possess an attribute we earlier deemed fairly implausible: increasing absolute risk aversion (IARA) (see Chapter 4). On this ground, supposing all or most investors have a quadratic utility function is very restrictive.

Second, we may straightforwardly assume that individual asset returns are normally distributed: as a weighted average of normal random variables, \tilde{r}_P will also be normally distributed. It follows immediately that we may assert $\mathbb{U}(\tilde{r}_P) = EU(\tilde{r}_P) \equiv \mathbb{U}(\mu_P, \sigma_P)$

To confirm this assertion, observe that

$$\tilde{z} = \frac{\tilde{r}_P - \mu_P}{\sigma_P} \sim N(0, 1),$$

if \tilde{r}_P is distributed normally. As a result

$$EU(\tilde{r}_P) \equiv EU(\mu_P + \tilde{z}\sigma_P)$$

so that $EU(\tilde{r}_P)$ depends only on μ_P and σ_P , and we may write $\mathbb{U}(\tilde{r}_P)$ as $\mathbb{U}(\mu_P, \sigma_P)$. This latter formulation allows us to compute investor indifference curves in $\mu_P \times \sigma_P$ space easily.

⁶ Since $U''(E(\tilde{Y})) < 0$, by concavity, and H_3 is small, the above equation confirms Jensen's inequality.

⁷ If \tilde{Y} is normally distributed (and thus \tilde{r}_P), H_3 can be expressed in terms of $E(\tilde{Y})$ and $\sigma^2(\tilde{Y})$, so, again, $E(\tilde{Y})$ and $\sigma^2(\tilde{Y})$ alone determine $EU(\tilde{Y})$; similarly μ_P and σ_P alone determine \tilde{r}_P .

⁸ These well-known assertions are detailed in Appendix 6.1 where it is also shown that, under either of the above hypotheses, indifference curves in the mean-variance space are increasing and convex to the origin.

Accordingly, our operating assumptions will be expanded to include the assumed normality of asset return distributions.

The normality assumption on the rate of return processes for individual stocks and stock indices is fairly robust empirically but it is not satisfied exactly: there is typically too much probability in the tails of the return distributions (see [Section 6.3](#)). Option-based instruments, which are increasingly prevalent in investor portfolios, are also characterized by rate of return probability distributions that are far from normal. These remarks taken together suggest that the analysis to follow must be viewed as a useful and productive approximation, and not more.

Let us summarize our setting going forward. First, the investor has a one-period investment horizon and a homogeneous utility function. Second, we assume the investor knows and takes as exogenously given the vector of return random variables $(\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_N)$; in particular, he knows returns are normally distributed, and he knows each security's μ and σ . Together these assumptions allow us to infer his preferences have the VNM utility form $\mathbb{U}(\mu_P, \sigma_P)$, where $\mathbb{U}_1(\mu_P, \sigma_P) > 0$, and $\mathbb{U}_2(\mu_P, \sigma_P) < 0$ (see [Appendix 6.2](#) for details).

We have not yet discussed how these rate of return distributions arise. We are at the stage of identifying demand curves, and not yet attempting to describe how equilibrium prices or returns are determined.

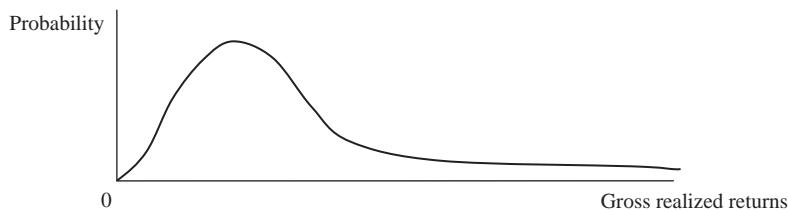
6.3 Refining the Normality-of-Returns Assumption

Before we embark upon a (μ_P, σ_P) -based theory of portfolio formation, we must first address two final objections to the normality-of-returns assumption. First, the assumption that period returns (e.g., daily, monthly, annual) are normally distributed is inconsistent with the limited liability feature of most financial instruments, i.e., $\tilde{r}_i \geq -1$ for most securities i and certainly for stocks: the worst an equity investor can do is to lose his initial investment. Normality presents a further problem for the discrete compounding of cumulative returns: the product of normally distributed random variables (returns) is not itself normally distributed.

As first suggested in Box 3.1, both of these objections are made moot if we assume that all returns are continuously compounded and define a stock's rate of return as

$$\tilde{r}_{i,t} = \ln((\tilde{q}_{i,t+1}^e + \widetilde{\text{div}}_{i,t+1})/(q_{i,t}^e)).$$

First and foremost, continuous compounding preserves limited liability since $Y_0 e^r \geq 0$ for any $r \in (-\infty, +\infty)$. It has the added feature that compounding preserves normality since the sum of normally distributed random variables is normally distributed (see again the comments in Box 3.1).

**Figure 6.1**

Lognormal probability density functions for gross realized returns.

Accordingly, the working assumption in empirical financial economics is that equity returns are continuously compounded and normally distributed in any period of time; in other words, for any stock i and any time t ,

$$\tilde{r}_{i,t} = \ln\left(\frac{\tilde{q}_{i,t+1}^e + \tilde{\text{div}}_{i,t+1}}{q_{i,t}^e}\right) \sim N(\mu_i, \sigma_i)$$

With $\tilde{r}_{i,t} \sim N(\mu_i, \sigma_i)$ the corresponding probability density function for gross realized returns, $(\tilde{q}_{i,t}^e + \tilde{\text{div}}_{i,t})/(q_{i,t-1}^e)$, has the form represented in Figure 6.1 with a mean of $e^{\mu_i + (1/2)\sigma_i^2}$ and a variance of $e^{(2\mu_i + \sigma_i^2)}(e^{\sigma_i^2} - 1)$.

Extending the idea a bit, we can say more generally that if an investor's wealth in period t is Y_t , then his wealth in period $t + 1$, $\tilde{Y}_{t+1} = e^{\tilde{r}_{t+1}^p} Y_t$, has a probability density function of the same form. By way of language, we say that realized gross equity returns are lognormally distributed because their logarithm is normally distributed.

There is substantial statistical evidence to support the normality assumption, subject to two very important qualifications:

1. While the normal distribution is perfectly symmetric about its mean, individual daily stock returns are frequently skewed to the right. Conversely, the returns to certain stock indices appear skewed to the left.⁹
2. Sample daily return distributions for most individual stocks exhibit “excess kurtosis” or “fat tails,” i.e., there is more probability in the tails than would be justified by the normal distribution.¹⁰ In plain language this fact means that extreme events—very positive or very negative returns—are observed with greater frequency than the normal distribution

⁹ Skewness: The extent to which a probability density is “pushed to the left or right” is measured by the skewness statistic $S(\tilde{r}_{it})$, defined by $S(\tilde{r}_{it}) = E((\tilde{r}_{it} - \mu_i)^3 / \sigma_1^3)$. $S(\tilde{r}_{it}) \equiv 0$ if \tilde{r}_{it} is normally distributed. $S(\tilde{r}_{it}) > 0$ suggests a rightward bias, and conversely if $S(\tilde{r}_{it}) < 0$.

¹⁰ Kurtosis is measured as the normalized fourth moment: $K(\tilde{r}_{it}) = E((\tilde{r}_{it} - \mu_i)^4 / \sigma_1^4)$. If \tilde{r}_{it} is normal, then $K(\tilde{r}_{it}) = 3$, but fat-tailed distributions with extra probability weight in the tail areas have higher kurtosis measures.

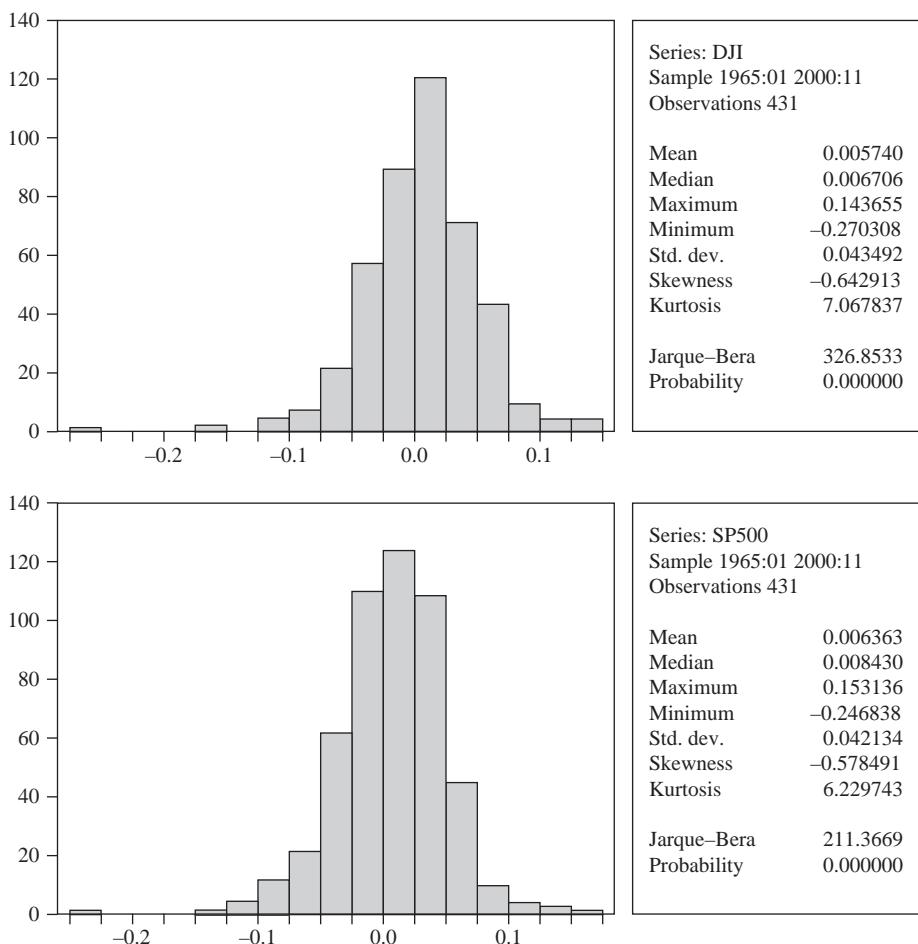


Figure 6.2
Empirical return distributions: Dow Jones and S&P₅₀₀.

would predict. The same is true of stock indices. The extent of this excess kurtosis diminishes substantially, however, when monthly data is used, but it is not eliminated.

Figure 6.2 illustrates the returns on the Dow Jones and the S&P 500. Both indices display negative skewness and a significant degree of kurtosis (Box 6.1).¹¹

¹¹ Note on the Jarque–Bera statistic: This is a test statistic for normality. It is based on the skewness and kurtosis of the sample data. For a large sample, say $n \geq 100$, with skewness S and kurtosis K , the Jarque–Bera statistic, $(n/6)(S^2 + ((K-3)^2)/4)$, follows a Chi Square distribution with 2 degrees of freedom if the underlying distribution is normal. For each of these samples, n is 431 and the reported measures of skewness and kurtosis imply that the likelihood of the distribution being strictly normal is very low (≈ 0.00), although the deviation from normality does not appear to be too extreme (the Jarque–Bera values of truly nonnormal distributions may exceed 10,000).

BOX 6.1 Connection to Factor Models

Our analysis so far has characterized the benefits of diversification in terms of the mean–variance efficient frontier of portfolio returns. Factors, however, underlie returns. Recall in Chapter 2 where we hypothesized the existence of multiple factors $\tilde{F}^1, \tilde{F}^2, \dots, \tilde{F}^J$ for which, for any asset i ,

$$\tilde{r}_i = \alpha_i + \beta_i^1 \tilde{F}^1 + \beta_i^2 \tilde{F}^2 + \dots + \beta_i^J \tilde{F}^J + \tilde{\varepsilon}_i$$

with different stocks, via the β_i^j s, displaying different factor sensitivities, and $\text{cov}(\tilde{\varepsilon}_i, F^j) = 0$ for all j . In this context, risk reduction via diversification is concerned with the $\tilde{\varepsilon}_i$ terms, their correlations across different securities, etc. The risks associated with the factors themselves cannot be diversified away as these risks affect all stock returns similarly. Accordingly, for any asset i ,

$$\mu_i = \alpha_i + \beta_i^1 E\tilde{F}^1 + \dots + \beta_i^J E\tilde{F}^J$$

and $\sigma_i^2 = \sigma_{\tilde{F}^1, \dots, \tilde{F}^J}^2 + \sigma_{\tilde{\varepsilon}_i}^2$, where $\sigma_{\tilde{F}^1, \dots, \tilde{F}^J}^2$ denotes aggregate factor risk.

There is one further complication. Even if individual stock returns are lognormally distributed, the returns to a portfolio of such stocks need not be lognormal because the log of a sum is not equal to the sum of the logs:

$$\ln(1 + \tilde{r}_P) = \ln(1 + w_1 \tilde{r}_1 + \dots + w_N \tilde{r}_N) \neq w_1 \ln(1 + \tilde{r}_1) + \dots + w_N \ln(1 + \tilde{r}_N)$$

The extent of the error introduced by assuming lognormal portfolio returns is usually quite small, however, if the return period is short (e.g., daily) so that $\ln(1 + \tilde{r}_i) \approx \tilde{r}_i$.

Let us pause at this point and review the sum total of all our underlying assumptions:

1. For all risky assets under consideration for portfolio inclusion, $\tilde{r}_{i,t} \sim N(\mu_i, \sigma_i)$. This is a reasonable first approximation.
2. The investor's utility-of-money function $U(\cdot)$ is homogeneous so that the same optimal portfolio proportions, once determined, apply for all investable wealth levels.
3. Individual asset returns and portfolio returns are continuously compounded.
4. As a result of Assumptions 1–3, the investor's VNM utility function $\mathbb{U}(\cdot)$, defined over portfolio return distributions \tilde{r}_P , can be expressed in the form $\mathbb{U}(\mu_P, \sigma_P)$.

With these four requirements in mind, let us proceed to the central concept of this chapter.

6.4 Description of the Opportunity Set in the Mean–Variance Space: The Gains from Diversification and the Efficient Frontier

The main idea of this chapter is as follows: The expected return to a portfolio is the *weighted* average of the expected returns of the assets composing the portfolio. The same

result is not generally true for the variance. The variance of a portfolio is generally *smaller* than the weighted average of the variances of individual asset returns corresponding to this portfolio. Therein lie the gains from diversification.

Let us illustrate this assertion, starting with the case of a portfolio of two assets only. The typical investor's objective is to maximize a function $U(\mu_R, \sigma_P)$, where $U_1 > 0$ and $U_2 < 0$: the investor likes expected return (μ_P) and dislikes standard deviation (σ_P). In this context, one recalls that an asset (or portfolio) A is said to **mean–variance dominate** an asset (or portfolio) B if $\mu_A \geq \mu_B$ and simultaneously $\sigma_A < \sigma_B$, or if $\mu_A > \mu_B$ while $\sigma_A \leq \sigma_B$. We can then define the **efficient frontier** as the locus of all nondominated portfolios in the mean–standard deviation space. By definition, no (“rational”) mean–variance investor would choose to hold a portfolio not located on the efficient frontier. The shape of the efficient frontier is thus of primary interest.

Next consider the efficient frontier in the two-asset case for a variety of possible asset return correlations. The basis for the results of this section is the formula for the variance of a portfolio of two assets, 1 and 2, defined by their respective expected returns, μ_1, μ_2 , standard deviations, σ_1 and σ_2 , and their correlation $\rho_{1,2}$:

$$\sigma_P^2 = w_1^2 \sigma_1^2 + (1 - w_1)^2 \sigma_2^2 + 2w_1(1 - w_1)\sigma_1\sigma_2\rho_{1,2}$$

where w_i is the proportion of the portfolio allocated to asset i . The following results, detailed in [Appendix 6.2](#), are of special importance.

Case 1 (Reference): In the case of two risky assets with perfectly positively correlated returns, the efficient frontier is linear. In that extreme case, the two assets are essentially identical, there is no gain from diversification, and the portfolio's standard deviation is nothing other than the average of the standard deviations of the component assets:

$$\sigma_R = w_1\sigma_1 + (1 - w_1)\sigma_2$$

As a result, the equation of the efficient frontier is

$$\mu_R = \mu_1 + \frac{\mu_2 - \mu_1}{\sigma_2 - \sigma_1}(\sigma_R - \sigma_1)$$

as depicted in [Figure 6.3](#). It assumes that positive amounts of both assets are held.

Case 2: In the case of two risky assets with imperfectly correlated returns, the standard deviation of the portfolio is necessarily smaller than it would be if the two component assets were perfectly correlated. By the previous result, one must have

$\sigma_P < w_1\sigma_1 + (1 - w_1)\sigma_2$, provided the proportions are not 0 or 1. Thus, the efficient frontier must stand left of the straight line in [Figure 6.3](#). This is illustrated in [Figure 6.4](#) for different values of $\rho_{1,2}$.

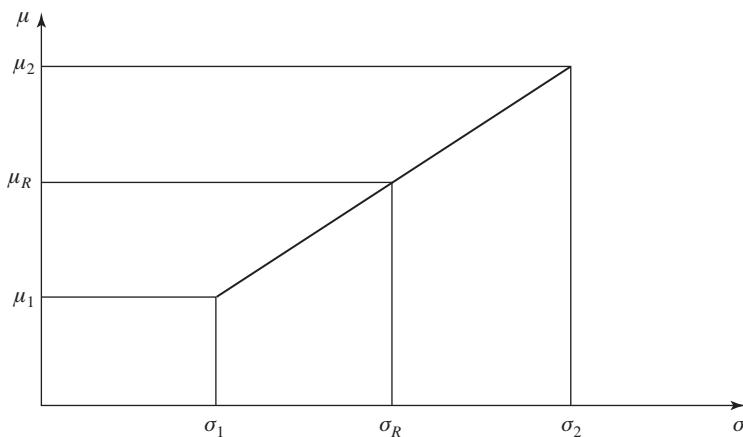


Figure 6.3
The efficient frontier: two perfectly correlated risky assets.

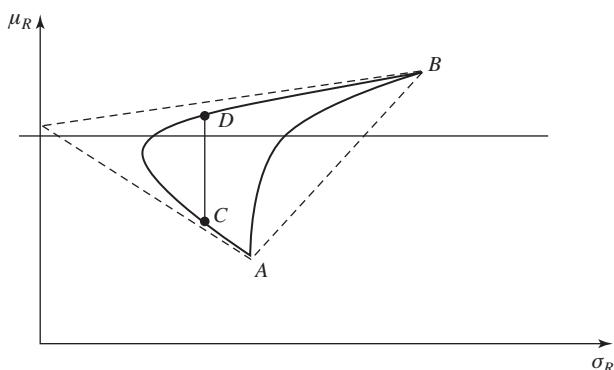


Figure 6.4
The efficient frontier: two imperfectly correlated risky assets.

The smaller the correlation (the further away from +1), the more to the left is the efficient frontier as demonstrated formally in [Appendix 6.2](#). Note that the diagram makes clear that in this case, some portfolios made up of assets 1 and 2 are, in fact, dominated by other portfolios. Unlike in Case 1, not all portfolios are efficient. In view of future developments, it is useful to distinguish the **minimum variance frontier** from the efficient frontier. In the present case, all portfolios between A and B belong to the minimum variance frontier, i.e., they correspond to the combination of assets with minimum variance for all arbitrary levels of expected returns. However, certain levels of expected returns are not efficient targets since higher levels of returns can be obtained for identical levels of risk. Thus, portfolio C has minimum variance, but it is

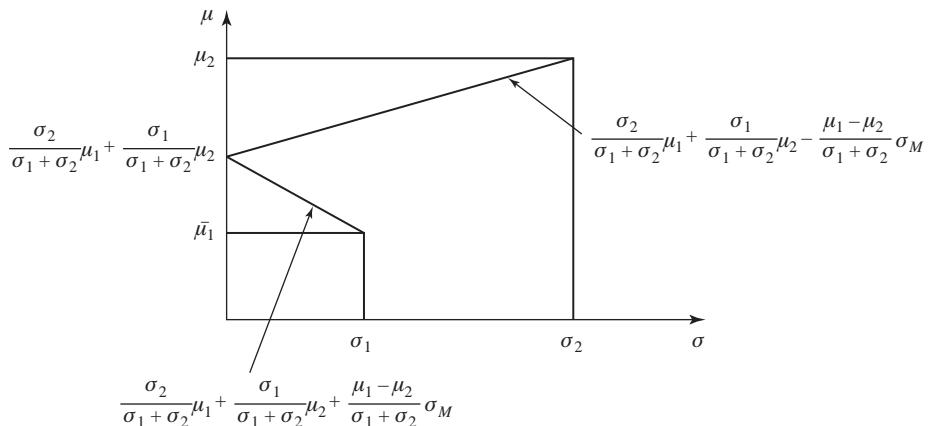


Figure 6.5

The efficient frontier: two perfectly negatively correlated risky assets.

not efficient, being dominated, for instance, by portfolio *D*. Figure 6.4 again assumes positive amounts of both assets (*A* and *B*) are held.

Case 3: If the two risky assets have returns that are perfectly negatively correlated, one can show that the minimum variance portfolio is risk free while the frontier is once again linear. Its graphical representation in that case is in Figure 6.5, with the corresponding demonstration found in Appendix 6.2.

Case 4: If one of the two assets is risk free, then the efficient frontier is a straight line originating on the vertical axis at the level of the risk-free return. In the absence of a short sales restriction, i.e., if it is possible to borrow at the risk-free rate to leverage one's holdings of the risky asset, then, intuitively enough, the overall portfolio can be made riskier than the riskiest among the existing assets. In other words, it can be made riskier than the one risky asset, and it must be that the efficient frontier is projected to the right of the (μ_2, σ_2) point (defining asset 1 as the risk-free asset). This situation is depicted in Figure 6.6, with the corresponding results demonstrated in Appendix 6.2.

Case 5 (*n* risky assets): It is important to realize that a portfolio is also an asset, fully defined by its expected return, its standard deviation, and its correlation with other existing assets or portfolios. Thus, the previous analysis with two assets is more general than it appears: It can easily be repeated, with one of the two assets being a portfolio. In that way, one can extend the analysis from two to three assets, from three to four, and so on. If there are *n* risky, imperfectly correlated assets, then the efficient frontier will have the bullet shape of Figure 6.7. Adding an extra asset to the two-asset framework implies that the diversification possibilities are improved and that, in principle, the efficient frontier is displaced to the left.

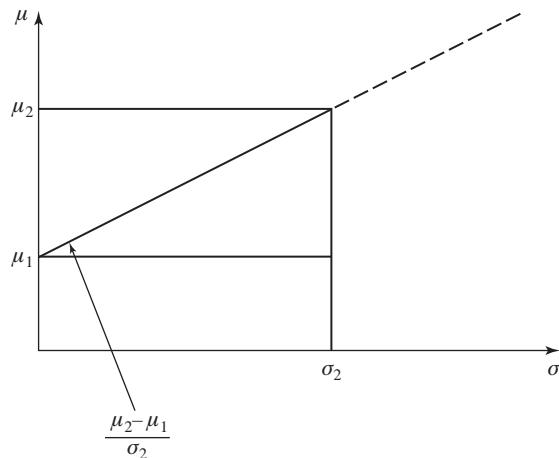


Figure 6.6
The efficient frontier: one risky and one risk-free asset.

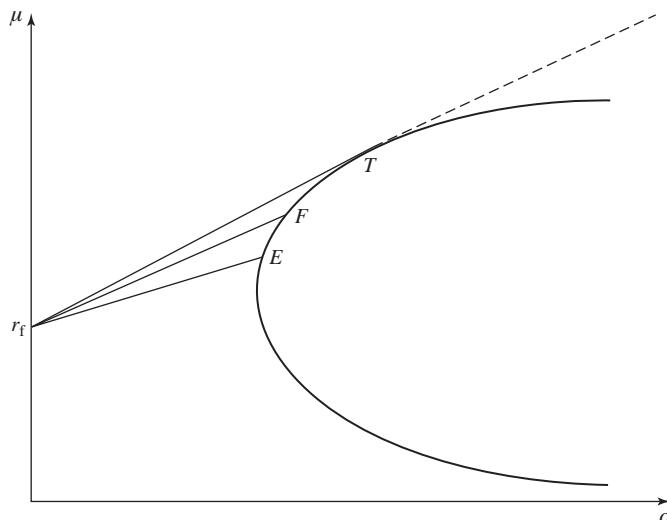


Figure 6.7
The efficient frontier: one risk-free and n risky assets.

Case 6: If there are n risky assets and a risk-free one, the efficient frontier is a straight line once again. To arrive at this conclusion, let us arbitrarily pick one portfolio on the efficient frontier when there are n risky assets only, say portfolio E in [Figure 6.7](#), and make up all possible portfolios combining E and the risk-free asset.

What we learned above tells us that the set of such portfolios is the straight line joining the point $(0, r_f)$ to E . Now we can quickly check that all portfolios on this line are dominated by those we can create by combining the risk-free asset with portfolio F . Continuing our reasoning in this way and searching for the highest similar line joining $(0, r_f)$ with the risky asset bullet-shaped frontier, we obtain, as the truly efficient frontier, the straight line originating from $(0, r_f)$ that is tangent to the risky asset frontier. Let T be the tangency portfolio. As before, if we allow a short position in the risk-free asset, the efficient frontier extends beyond T ; it is represented by the broken line in [Figure 6.7](#).

Formally, with n assets (possibly one of them risk free), the efficient frontier is obtained as the relevant (nondominated) portion of the minimum variance frontier, the latter being the solution, for all possible expected returns μ , to the following quadratic program (QP):

$$(QP) \quad \begin{aligned} & \min_{w_i's} \sum_i \sum_j w_i w_j \sigma_{ij} \\ & \text{s.t. } \sum_i w_i \mu_i = \mu \\ & \qquad \sum_i w_i = 1 \end{aligned}$$

In (QP) we search for the vector of weights that minimizes the variance of the portfolio (verify that you understand the writing of the portfolio variance in the case of n assets) under the constraint that the expected return on the portfolio must be μ . This defines one point on the minimum variance frontier. One can then change the fixed value of μ , equating it successively to all plausible levels of portfolio expected return; in this way one effectively draws the minimum variance frontier.¹² Program (QP) is the simplest version of a family of similar quadratic programs used in practice. This is because (QP) includes the minimal set of constraints. The first is only an artifice in that it defines the expected return to be reached in a context where μ is a parameter; the second constraint is simply the assertion that the vector of w_i 's defines a portfolio (and thus that they add up to one).

Many other constraints can be added to customize the portfolio selection process without altering the basic structure of problem (QP). Probably the most common implicit or explicit constraint for an investor involves limiting her investment universe. As mentioned earlier (Chapter 3, Box 3.2), the well-known *home bias puzzle* reflects the difficulty in explaining, from the MPT viewpoint, why investors do not invest a larger fraction of their portfolios in stocks quoted “away from home,” i.e., in international, or emerging markets. This can be viewed as the result of an unconscious limitation of the investment universe considered by the investor. Self-limitation may also be fully conscious and explicit as in the case of

¹² While in principle one could as well maximize the portfolio's expected return for given levels of standard deviation, it turns out to be more efficient computationally to do the reverse.

“ethical” mutual funds that exclude arms manufacturers or companies with a tarnished ecological record from their investment universe. These constraints are easily accommodated in our setup, as they simply appear or do not appear in the list of the N assets under consideration.

Other common constraints are nonnegativity constraints ($w_i \geq 0$), indicating the impossibility of short selling some or all assets under consideration. Without nonnegativity constraints, mean–variance analysis often leads to efficient portfolios with large short positions in some assets.¹³ Short selling may be impossible for feasibility reasons (exchanges or brokers may not allow it for certain instruments) or, more frequently, for regulatory reasons applying to specific types of investors, e.g., pension funds.

An investor may also wish to construct an efficient portfolio subject to the constraint that his holdings of some stocks should not, in value terms, fall below a certain level (perhaps because of potential tax liabilities or because ownership of a large block of this stock affords some degree of managerial control). This requires a constraint of the form

$$w_j \geq \frac{V_j^*}{V_P}$$

where V_j^* is the postulated lower bound on the value of his holdings of stock j and V_P is the overall value of his portfolio.

Other investors may wish to obtain the lowest risk subject to a required expected return constraint and/or be subject to a constraint that limits the number of stocks in their portfolio (in order, possibly, to economize on transaction costs). An investor may, for example, wish to hold at most 3 out of a possible 10 stocks, yet to hold those 3 that give the minimum risk subject to a required return constraint. With certain modifications, this possibility can be accommodated into (QP) as well. [Appendix 6.3](#) details how Microsoft Excel® can be used to construct the portfolio efficient frontier under these and other constraints.

6.5 The Optimal Portfolio: A Separation Theorem

The optimal portfolio is naturally defined as that portfolio maximizing the investor’s (mean–variance) utility; in other words, that portfolio for which the investor is able to reach the highest indifference curve, which we know to be increasing and convex to the origin. If the efficient frontier has the shape described in [Figure 6.6](#), i.e., if there is a risk-free asset, then all tangency points must lie on the same efficient frontier, regardless of the rate of risk aversion of the investor. Let there be two investors sharing the same perceptions as to expected returns, variances, and return correlations but differing in their willingness to

¹³ An illustration of this particular phenomenon is presented in Chapter 7.

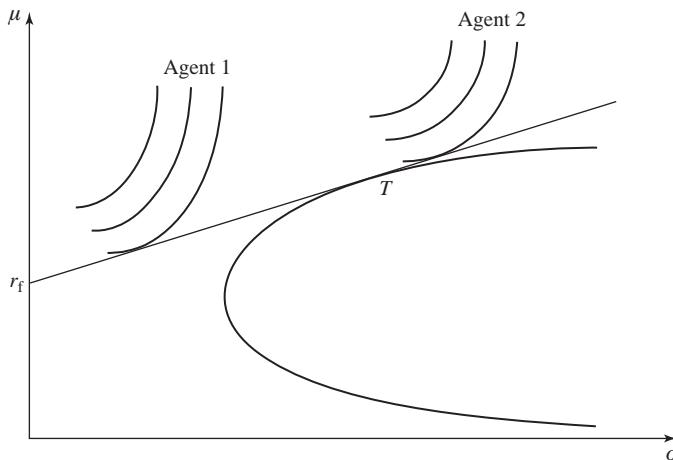


Figure 6.8
The optimal portfolios of two differently risk-averse investors.

take risks. The relevant efficient frontier will be identical for these two investors, although their optimal portfolios will be represented by different points on the same line: with differently shaped indifference curves the tangency points must differ. See [Figure 6.8](#).

Nevertheless, it is a fact that our two investors will invest in the same two “funds,” the risk-free asset on the one hand, and the risky portfolio (T) identified by the tangency point between the straight line originating from the vertical axis and the bullet-shaped frontier of risky assets, on the other. This is the **two-fund theorem**, also known as the **separation theorem**, because it implies that the optimal portfolio of risky assets can be identified separately from an investor’s knowledge of his risk preference. This result will play a significant role in Chapter 8 when we construct the capital asset pricing model.

6.6 Stochastic Dominance and Diversification

In this concluding section, we propose to deal with a typical question facing an investor: should a particular asset be added to the investor’s portfolio? More precisely, if the particular asset in question is added, what characteristics must it have in order that the investor can be assured that his expected utility of wealth will increase?

To be more precise, suppose the investor’s current portfolio P is characterized by (μ_P, σ_P) . He is considering adding some of asset A , characterized by $(\mu_A, \sigma_A, \rho_{AP})$ to his portfolio. Although the strict attainment of mean–variance efficiency may well require it, he does not plan at this juncture to alter the relative composition of the assets he already has. Indeed, to implement substantial changes in his portfolio’s existing composition is likely to be a costly

proposition, and let us assume also that the avoidance of this cost is presently dominant in his decision making. These considerations mean that if the investor decides to allocate some of his wealth to A , his new portfolio will then be composed of two assets, some units of the original portfolio P and some amount of money invested in asset A . We denote this new portfolio P' .¹⁴

Let us go on and further assume that by undertaking this investment, the investor hopes to improve the overall return on his wealth. Since the expected return on a portfolio is simply the weighted average of the expected returns of the assets contained in the portfolio, the only way for the investor to increase his original portfolio's expected returns is to add assets for which $\mu_A > \mu_P$. Accordingly, we will assume this is true for the asset in question: $\mu_A > \mu_P$ (e.g., we may think of A as itself a portfolio of emerging markets stocks which had high returns in the 1980–2000 period). Few benefits in life are totally free, however, which means that $\sigma_A^2 > \sigma_P^2$ as well (this fact also characterized emerging markets stocks relative to the S&P₅₀₀ in the 1980s and 1990s). If $\rho_{AP} \approx 1$, so that asset A and the investor's existing portfolio P have identical return patterns, the investor will be confronted by a difficult choice: by allocating some of his cash to A , his portfolio's μ will certainly rise, but its risk will almost surely increase as well. If $\rho_{AP} < 1$, however, it is possible that the investor's overall portfolio risk would fall with the inclusion of A , i.e., A and P 's return patterns could be sufficiently different so as to allow risk to decline. How different would the return patterns have to be for this “diversification” effect to dominate? Would a VNM-expected utility maximize necessarily agree to include asset A in that case? These are the questions and context under consideration in this final chapter topic. We formalize them as follows:

Question 1: Under what conditions will the addition of an asset to a portfolio (in the aforementioned sense) allow the portfolio's expected return to increase and its standard deviation to decline? This is one characterization of the gains to diversification in a mean/variance environment.

Question 2: Will a VNM-expected utility maximizing investor also prefer to be more highly diversified in the sense of Question 1? To say it differently, although the inclusion of a higher expected return, higher standard deviation asset may make sense from a mean–variance perspective, would a typical VNM-expected utility investor always agree?

¹⁴ There are a number of scenarios to which this problem description will apply. For one such scenario, suppose an investor is confronted with an investment opportunity (A) when his savings are invested in an ETF. If he sells some units of his ETF to finance the purchase of A he is not altering the proportions of the other securities he implicitly owns by virtue of owning units of the ETF. A second scenario is as follows: An investor owns a portfolio (P) with a number N of specific assets. He has accumulated some cash to invest, either by increasing the assets in (P) in a way that leaves the proportions unchanged (given the data available to him it is on the efficient frontier for those assets), or he can invest the same cash in a new asset (A) which has presented itself. Under either scenario, the analysis will apply.

If their respective returns patterns differ, we know that some of A 's return variation (risk) can be canceled out by variation in the return of P , and that the extent of this “cancelation effect” increases the lower the ρ_{AP} . Accordingly, Question 1 can be forthrightly restated as follows: How low must ρ_{AP} be in order that the inclusion of A into portfolio P will reduce overall portfolio risk?¹⁵ The answer is provided in [Theorem 6.1](#).

Theorem 6.1 Consider an N -asset portfolio P characterized by (μ_P, σ_P) to which an investor is considering adding an asset A where (i) $\mu_A > \mu_P$ and (ii) $\rho_{AP} < (\sigma_P/\sigma_A)$. Then there exists a portfolio P^* such that $\mu_{P^*} > \mu_P$ and $\sigma_{P^*} < \sigma_P$ where P^* contains A in positive amount.

Proof Consider a portfolio P' defined by proportions $w_A \geq 0$ in A and $1 - w_A$ in P . Then,

$$\begin{aligned}\sigma_{P'}^2 &= (1 - w_A)^2 \sigma_P^2 + w_A^2 \sigma_A^2 + 2 w_A(1 - w_A) \text{ cov}(\tilde{r}_A, \tilde{r}_P) \\ \frac{\partial \sigma_{P'}^2}{\partial w_A} &= -2(1 - w_A)\sigma_P^2 + 2 w_A \sigma_A^2 + 2 \text{ cov}(\tilde{r}_A, \tilde{r}_P) - 4 w_A \text{ cov}(\tilde{r}_A, \tilde{r}_P)\end{aligned}$$

Thus, $(\partial \sigma_{P'}^2 / \partial w_A) = 2 \text{ cov}(\tilde{r}_A, \tilde{r}_P) - 2\sigma_P^2$ for $w_A = 1$. Thus

$(\partial \sigma_{P'}^2 / \partial w_A) < 0$ for $w_A = 1$ if and only if $\text{cov}(\tilde{r}_A, \tilde{r}_P) < \sigma_P^2$, which is true if and only if $\rho_{AP} < (\sigma_P/\sigma_A)$, as assumed.

By the continuity of $(\partial \sigma_{P'}^2 / \partial w_A)$ as a function of w_A , there exists a $\hat{w}_A > 0$ such that for all $0 \leq w_A < \hat{w}_A$

$\mu_{P'} = \pi_A \mu_A - (1 - \pi_A) \mu_P > \mu_P$ (by assumption (i)), and
 $\sigma_{P'} < \sigma_P$ (by the above computation).

Pick one $\hat{w}_A > 0$ with corresponding portfolio P^* . Then

$$\mu_{P^*} > \mu_P \tag{6.8a}$$

$$\sigma_{P^*} < \sigma_P \tag{6.8b}$$

Portfolio P^* thus strictly dominates P in a mean–variance framework.

Portfolio P^* as constructed in the Lemma may not be efficient: further gains may be obtainable by altering the composition of the fraction $(1 - w_A)$ of P^* that is composed of the original P . Our investor accepts this possibility, however, because of his interest in minimizing his transactions costs.

¹⁵ To clarify this language, we will take “inclusion” to mean that the investor’s original portfolio with $(w_P = 1, w_A = 0)$ becomes one where $(w_P < 1, w_A > 0)$.

It is also worth noting that condition (ii) of the Lemma is always satisfied provided $\rho_{AP} = 0$, i.e., when the return on the asset under consideration for portfolio inclusion is statistically independent of the returns to the portfolio to which it may be “joined.” It follows that if enough independent (of one another) assets are included in a portfolio (at least with equal proportions), its risk can be made arbitrarily small. This assertion encapsulates the insurance principle.¹⁶ For equity investors, it is essentially impossible to find numerous assets with this property, however, as our discussion in Chapter 7 will confirm.

At this juncture, we have given an answer to Question 1: if conditions (i) and (ii) of Proposition 6.1 are satisfied, then the investor can increase his portfolio’s overall expected return and reduce its risk by adding some of asset A. To answer Question 2, however, additional results are needed. They are as follows.

Theorem 6.2 Consider two return distributions \tilde{r}_1 and \tilde{r}_2 with, respectively, cumulative distribution functions $F_{\tilde{r}_1}$ and $F_{\tilde{r}_2}$, where $\tilde{r}_1 \sim N(\mu_1, \sigma_1)$ and $\tilde{r}_2 \sim N(\mu_2, \sigma_2)$. Then $F_{\tilde{r}_1}$ FSD $F_{\tilde{r}_2}$ if and only if the following hold:

- a. $\mu_1 > \mu_2$
- b. $\sigma_1 = \sigma_2$.

Proof See [Levy \(2006\)](#), Theorem 6.1

Theorem 6.3 Consider two return distributions \tilde{r}_1 and \tilde{r}_2 with respective cumulative distribution functions $F_{\tilde{r}_1}$ and $F_{\tilde{r}_2}$, where $\tilde{r}_1 \sim N(\mu_1, \sigma_1)$ and $\tilde{r}_2 \sim N(\mu_2, \sigma_2)$. Then $F_{\tilde{r}_1}$ SSD $F_{\tilde{r}_2}$ if and only if the following hold:

- a. $\mu_1 \geq \mu_2$
- b. $\sigma_1 \leq \sigma_2$.

with at least one strong inequality.

Proof See [Levy \(2006\)](#), Theorem 6.2.

Corollary 6.1 applies these results (Theorem 6.3, in particular) to our investor’s situation.

Corollary 6.1 Consider portfolios P and P^* of Proposition 6.1, which satisfy [Eqs. \(6.8a\)](#) and [\(6.8b\)](#). Let their respective cumulative distribution functions be denoted $F_{\tilde{r}_{P^*}}$ and $F_{\tilde{r}_P}$. Then $F_{\tilde{r}_{P^*}}$ SSD $F_{\tilde{r}_P}$.

¹⁶ Insurance companies write policies to many insurers (car, fire, etc.) whose risks to the company are independent, e.g., the risks to a 64-year-old driver in New York are independent of the risks of a similarly aged driver in Illinois.

Proof Application of [Theorem 6.3](#).

We are getting closer to answering our investor's second question but we are not yet quite there. It would be convenient to answer this question with an immediate appeal to [Theorem 4.3](#), but that theorem is expressed in terms of payoffs while [Corollary 6.1](#) is expressed in terms of returns.¹⁷ This (small) inconsistency, however, can be easily overcome as follows.

Let $F_{\tilde{y}}$ denote the cumulative distribution function associated with some random variable \tilde{y} . By [Corollary 6.1](#), we know that $F_{\tilde{r}_{P^*}} \text{ SSD } F_{\tilde{r}_P}$. It follows immediately that $F_{1+\tilde{r}_{P^*}} \text{ SSD } F_{1+\tilde{r}_P}$ and thus, by [Theorem 4.2](#),

$$EU(1 + \tilde{r}_{P^*}) > EU(1 + \tilde{r}_P)$$

for any increasing, concave utility function $U(\cdot)$, where these expectations are “linear in the probabilities” as per VNM-expected utility theory.

If we further assume $U(\cdot)$ is homogeneous of degree ν and that the start-of-period wealth is Y_0 , then

$$\begin{aligned} Y_0^\nu EU(1 + \tilde{r}_{P^*}) &> Y_0^\nu EU(1 + \tilde{r}_P) \\ EU(Y_0(1 + \tilde{r}_{P^*})) &> EU(Y_0(1 + \tilde{r}_P)) \end{aligned} \quad (6.9)$$

for any increasing, concave utility function $U(\cdot)$. If our investor's utility function is increasing, concave, and homogeneous, his expected utility will thus increase with the inclusion of A .

Note that expectations in [Eq. \(6.9\)](#) are taken with respect to the investor's actual future wealth payoffs. We summarize these thoughts in [Theorem 6.4](#).

Theorem 6.4 Consider a universe of assets whose returns are normally distributed and consider greater diversification in the sense of [Theorem 6.1](#) as reflected in portfolios P^* and P . Then for any VNM-expected utility investor with an increasing, concave, and homogeneous utility function, it will be welfare improving to include some positive amount of asset A .

Proof The discussion above.

As a conclusion to our discussion, [Theorem 6.4](#) is a bit unsatisfying, however, because it assumes that the investor's next-period wealth distribution is normally distributed, which is certainly not the case if the assets in the portfolio are equities, as we have assumed. Rather, the investor's wealth distribution will, in fact, be lognormal. To modify our results to accommodate this additional fact, let us first denote \tilde{Y}_1^P and $\tilde{Y}_1^{P^*}$ as the next-period random

¹⁷ These quantities are of course equivalent if our investor's wealth is \$1 (1 CHF, etc.), but we need greater generality.

wealth levels if Y_0 , the investor's present wealth, is invested, respectively, in P or P^* . Under continuous compounding,

$$\begin{aligned} E\tilde{Y}_1^P &= EY_0 e^{\tilde{r}_P} = Y_0 e^{\mu_P + 1/2\sigma_P^2} \\ \sigma_{\tilde{Y}_1^P}^2 &= Y_0^2 e^{2\mu_P + \sigma_P^2} (e^{\sigma_P^2} - 1) \\ E\tilde{Y}_1^{P^*} &= EY_0 e^{\tilde{r}_{P^*}} = Y_0 e^{\mu_{P^*} + 1/2\sigma_{P^*}^2} \\ \sigma_{\tilde{Y}_1^{P^*}}^2 &= Y_0^2 e^{2\mu_{P^*} + \sigma_{P^*}^2} (e^{\sigma_{P^*}^2} - 1) \end{aligned}$$

For the relationships above, since $\mu_{P^*} > \mu_P$ and $\sigma_{P^*} < \sigma_P$ it is not guaranteed that $E\tilde{Y}_1^{P^*} > E\tilde{Y}_1^P$ because $\sigma_{P^*}^2 < \sigma_P^2$! The dependence of $E\tilde{Y}_1^P$ on σ_{P^*} (also for P) follows from the asymmetric nature of the lognormal distribution (which governs both \tilde{Y}_1^P and $\tilde{Y}_1^{P^*}$): a decrease in the variance (which applies to \tilde{r}_{P^*} relative to \tilde{r}_P) must be accompanied by a shift in probability toward the left tail which inevitably diminishes the mean as well. To obtain results analogous to [Theorem 6.5](#), the following result is needed.

Theorem 6.5 Let \tilde{Y}_1 and \tilde{Y}_2 be two lognormally distributed random variables with cumulative distribution functions $F_{\tilde{Y}_1}$ and $F_{\tilde{Y}_2}$, respectively. Then, $F_{\tilde{Y}_1}$ SSD $F_{\tilde{Y}_2}$ if and only if

- a. $E\tilde{Y}_1 \geq E\tilde{Y}_2$
- b. $\sigma_{\tilde{Y}_1}/E\tilde{Y}_1 \leq \sigma_{\tilde{Y}_2}/E\tilde{Y}_2$,

with at least one strict inequality.

Proof [Levy \(2006\)](#), Theorem 6.5

Our final result is a direct application of [Theorem 6.5](#).

Theorem 6.6 Let \tilde{Y}_1^P and $\tilde{Y}_1^{P^*}$ be as in the discussion above. Assume that Eqs. (6.8a), (6.8b), and (6.10)

$$\mu_{P^*} + (1/2)\sigma_{P^*}^2 > \mu_P + (1/2)\sigma_P^2 \quad (6.10)$$

hold. Then $F_{Y_1^{P^*}}$ SSD $F_{Y_1^P}$, and a VNM-expected utility investor with an increasing and concave utility-of-money function prefers greater diversification in the sense of adding asset A to his portfolio.

Proof A direct application of [Theorem 6.5](#) is as follows. By assumption (6.10), $EY_1^{P^*} \geq EY_1^P$ which satisfies condition (a) of [Theorem 6.5](#). Furthermore,

$$\begin{aligned} \left(\frac{\sigma_{\tilde{Y}_1^{P^*}}}{EY_1^{P^*}} \right)^2 &= \left(\frac{Y_0^2 e^{2\mu_{P^*} + \sigma_{P^*}^2}}{[Y_0 e^{\mu_{P^*} + 1/2\sigma_{P^*}}]^2} \right) (e^{\sigma_{P^*}^2} - 1) \\ &= (e^{\sigma_{P^*}^2} - 1) < (e^{\sigma_P^2} - 1) = \left(\frac{\sigma_{\tilde{Y}_1^P}}{EY_1^P} \right)^2 \end{aligned}$$

by Eqs. (6.8a) and (6.8b).

Thus, $(\sigma_{\tilde{Y}_1^{P^*}}/EY_1^{P^*}) < (\sigma_{\tilde{Y}_1^P}/EY_1^P)$ and (b) of Theorem 6.5 is satisfied.

What is to be learned from this discussion? First and foremost, we learned that adding an asset A to a portfolio P for which $\mu_A > \mu_P$ and $\rho_{AP} < (\sigma_P/\sigma_A)$ will create a second-order stochastically dominating distribution relative to that of the original portfolio P . In general, however, this is not enough to guarantee that the investor's expected utility of wealth will be enhanced by its inclusion. That requires the new portfolio (P^* , the portfolio with some A included) to satisfy $\mu_{P^*} + (1/2)\sigma_{\tilde{r}_{P^*}} > \mu_P + (1/2)\sigma_{\tilde{r}_P}$. This amounts to requiring that the asset up for inclusion should have a significantly higher expected return than the portfolio into which it may be placed. This is an illustration where "means matter more than variances" (and, as we shall see, means are also more difficult to estimate precisely).

We now conclude this chapter.

6.7 Conclusions

First, it is important to keep in mind that everything said so far in this chapter applies to the *subjective* expectations regarding future return distributions that the investor may hold. There is no requirement that these distributions be objective or "rational."

Second, although initially conceived in the context of descriptive economic theories, the success of portfolio theory arose primarily from the possibility of giving it a normative interpretation, i.e., of seeing the theory as providing a guide on how to proceed to identify a potential investor's optimal portfolio. In particular, it points to the information requirements to be fulfilled (ideally). Under the formal restrictions spelled out in this chapter, one cannot identify an optimal portfolio without estimates of mean returns, standard deviations of returns, and correlations among returns. As Chapter 7 will make clear, this is no easy task empirically. Investors may thus simply fall back on simpler allocation strategies such as an equally weighted portfolio. As we will also see in Chapter 7, there is a substantial body of empirical evidence that argues for this alternative.

One can view the role of the financial analyst as providing plausible figures for the relevant statistics or offering alternative scenarios for consideration to the would-be investor. This is the first absolutely critical step in the search for an optimal portfolio.

The computation of the (subjective) efficient frontier is the second step, and it essentially involves solving the QP problem possibly in conjunction with constraints specific to the investor. The third and final step consists of defining, at a more or less formal level, the investor's risk tolerance and, on that basis, identifying his optimal allocation between risk free and risky assets.

References

- Levy, H., 2006. Stochastic Dominance: Investment Decision Making Under Uncertainty. Springer Science and Business Media, New York.
Markowitz, H.M., 1952. Portfolio selection. *J. Finan.* 7, 77–91.
Tobin, J., 1958. Liquidity preference as behavior towards risk. *Rev. Econ. Stud.* 26, 65–86.

Appendix 6.1: Indifference Curves Under Quadratic Utility or Normally Distributed Returns

In this appendix, we demonstrate more rigorously that if an investor's utility function is quadratic or if returns are normally distributed, the investor's expected utility of the portfolio's rate of return is a function of the portfolio's mean return and standard deviation only (Part I). We subsequently show that in either case, investor indifference curves are convex to the origin (Part II).

Part I

If the utility function is **quadratic**, it can be written as

$$U(r_P) = a + b r_P + c r_P^2$$

where r_P denotes a portfolio's rate of return. Let the constant $a = 0$ in what follows since it does not play any role. For this function to make sense, we must have $b > 0$ and $c < 0$. The first and second derivatives are, respectively,

$$U'(r_P) = b + 2c r_P \text{ and } U''(r_P) = 2c < 0$$

Expected utility is then of the following form:

$$E(U(\tilde{r}_P)) = bE(\tilde{r}_P) + c(E(\tilde{r}_P^2)) = b\mu_P + c\mu_P^2 + c\sigma_P^2$$

that is of the form $g(\sigma_P, \mu_P)$. As shown in [Figure A6.1](#), this function is strictly concave. But it must be restricted to ensure positive marginal utility: $\tilde{r}_P < -b/2c$. Moreover, the coefficient of absolute risk aversion is increasing ($R'_A > 0$). These two characteristics are unpleasant, and they prevent a more systematic use of the quadratic utility function.

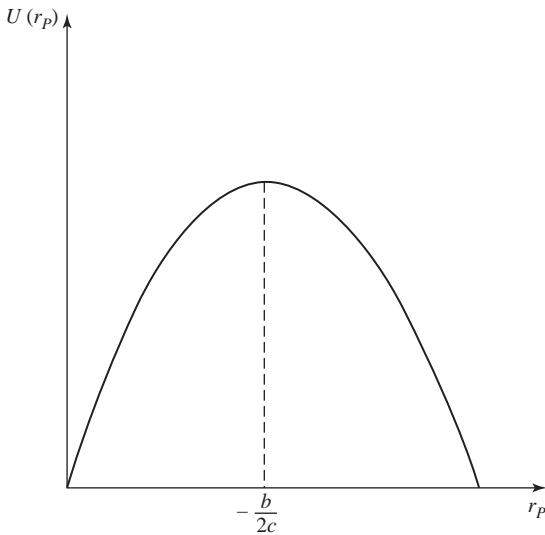


Figure A6.1
The graph of a quadratic utility function.

Alternatively, if the individual asset returns \tilde{r}_i are normally distributed, $\tilde{r}_P = \sum_i w_i \tilde{r}_i$ is normally distributed as well. Let \tilde{r}_P have density $f(\tilde{r}_P)$, where

$$f(\tilde{r}_P) = N(\tilde{r}_P; \mu_P, \sigma_P)$$

The standard normal variate \tilde{Z} is defined by

$$\tilde{Z} = \frac{\tilde{r}_P - \mu_P}{\sigma_P} \sim N(\tilde{Z}; 0, 1)$$

$$\text{Thus, } \tilde{r}_P = \sigma_P \tilde{Z} + \mu_P \quad (6.11)$$

$$\mathbb{E}(\tilde{r}_P) = E(U(r_P)) = \int_{-\infty}^{+\infty} U(r_P) f(r_P) dr_P = \int_{-\infty}^{+\infty} U(\sigma_P Z + \mu_P) N(Z; 0, 1) dZ \quad (6.12)$$

The quantity $E(U(r_P))$ is again a function of σ_P and μ_P only. Maximizing $E(U(\tilde{r}_P))$ amounts to choosing w_i so that the corresponding σ_P and μ_P maximize the integral (6.12).

Part II

Construction of indifference curves in the mean–variance space. There are again two cases.

U Is Quadratic

An indifference curve in the mean–variance space is defined as the set: $\{(\sigma_P, \mu_P) | E(U(\tilde{r}_P)) = b\mu_P + c\mu_P^2 + c\sigma_P^2 = k\}$, for some utility level k .

This can be rewritten as

$$\sigma_P^2 + \mu_P^2 + \frac{b}{c}\mu_P + \frac{b^2}{4c^2} = \frac{k}{c} + \frac{b^2}{4c^2}$$

$$\sigma_P^2 + \left(\mu_P + \frac{b}{2c}\right)^2 = \frac{k}{c} + \frac{b^2}{4c^2}$$

This equation defines the set of points (σ_P, μ_P) located in the circle of radius $\sqrt{\frac{k}{c} + \frac{b^2}{4c^2}}$ and of center $(0, -(b/2c))$ as in [Figure A6.2](#).

In the relevant portion of the (σ_P, μ_P) space, indifference curves thus have positive slope and are convex to the origin.

The Distribution if R Is Normal

One wants to describe

$$\left[(\sigma_P, \mu_P) \left| \int_{-\infty}^{+\infty} U(\sigma_P Z + \mu_P) N(Z; 0, 1) dZ = \bar{U} \right. \right]$$

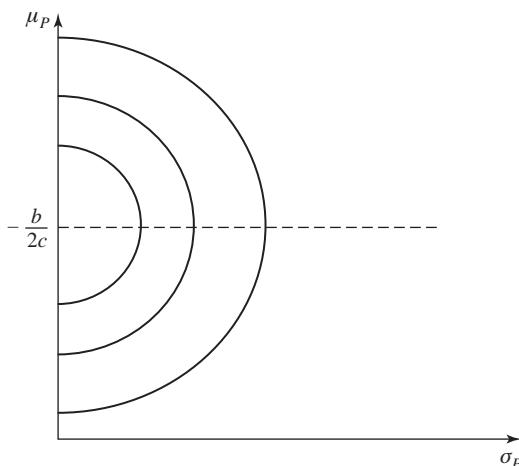


Figure A6.2
The indifference curves of a quadratic utility agent.

Differentiating totally yields:

$$0 = \int_{-\infty}^{+\infty} U'(\sigma_P Z + \mu_P)(Z d\sigma_P + d\mu_P)N(Z; 0, 1) dZ, \text{ or}$$

$$\frac{d\mu_P}{d\sigma_P} = - \frac{\int_{-\infty}^{+\infty} U'(\sigma_P Z + \mu_P)ZN(Z; 0, 1) dZ}{\int_{-\infty}^{+\infty} U'(\sigma_P Z + \mu_P)N(Z; 0, 1) dZ}$$

If $\sigma_P = 0$ (at the origin),

$$\frac{d\mu_P}{d\sigma_P} = - \frac{\int_{-\infty}^{+\infty} ZN(Z; 0, 1) dZ}{\int_{-\infty}^{+\infty} N(Z; 0, 1) dZ} = 0$$

If $\sigma_P > 0$, $d\mu_P/d\sigma_P > 0$.

Indeed, the denominator is positive since $U'(\cdot)$ is positive by assumption, and $Z \sim N(0, 1)$ is a probability density function, hence it is always positive.

The expression $\int_{-\infty}^{+\infty} U'(\sigma_P Z + \mu_P)ZN(Z; 0, 1) dZ$ is negative under the hypothesis that the investor is risk averse—in other words, that $U(\cdot)$ is strictly concave. If this hypothesis is verified, the marginal utility associated with each negative value of Z is larger than the marginal utility associated with positive values. Since this is true for all pairs of $\pm Z$, the integral on the numerator is negative. See [Figure A6.3](#) for an illustration.

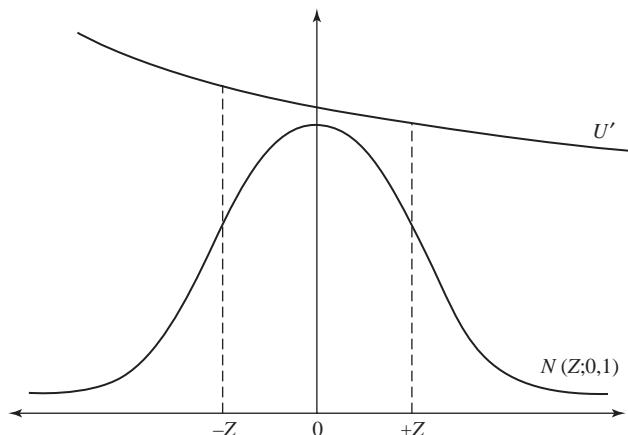


Figure A6.3

The marginal utility for negative values of Z higher than for positive ones.

Proof of the Convexity of Indifference Curves

Let two points (σ_P, μ_P) and $(\sigma_{P'}, \mu_{P'})$ lie on the same indifference curve offering the same level of expected utility \bar{U} . Let us consider the point $(\sigma_{P''}, \mu_{P''})$, where $\sigma_{P''} = \alpha\sigma_P + (1 - \alpha)\sigma_{P'}$ and $\mu_{P''} = \alpha\mu_P + (1 - \alpha)\mu_{P'}$.

One would like to prove that

$$E(U(\sigma_{P''}Z + \mu_{P''})) > \alpha E(U(\sigma_P Z + \mu_P)) + (1 - \alpha)E(U(\sigma_{P'} Z + \mu_{P'})) = \bar{U}.$$

By the strict concavity of U , the inequality

$$U(\sigma_{P''}Z + \mu_{P''}) > \alpha U(\sigma_P Z + \mu_P) + (1 - \alpha)U(\sigma_{P'} Z + \mu_{P'})$$

is verified for all (σ_P, μ_P) and $(\sigma_{P'}, \mu_{P'})$ and any Z value.

One may thus write

$$\begin{aligned} & \int_{-\infty}^{+\infty} U(\sigma_{P''}Z + \mu_{P''})N(Z; 0, 1)dZ > \\ & \alpha \int_{-\infty}^{+\infty} U(\sigma_P Z + \mu_P)N(Z; 0, 1)dZ + (1 - \alpha) \int_{-\infty}^{+\infty} U(\sigma_{P'} Z + \mu_{P'})N(Z; 0, 1)dZ, \text{ or} \\ & E(U(\sigma_{P''}Z + \mu_{P''})) > \alpha E(U(\sigma_P Z + \mu_P)) + (1 - \alpha)E(U(\sigma_{P'} Z + \mu_{P'})), \text{ or} \\ & E(U(\sigma_{P''}Z + \mu_{P''})) > \alpha \bar{U} + (1 - \alpha) \bar{U} = \bar{U} \end{aligned}$$

See [Figure A6.4](#) for an illustration.

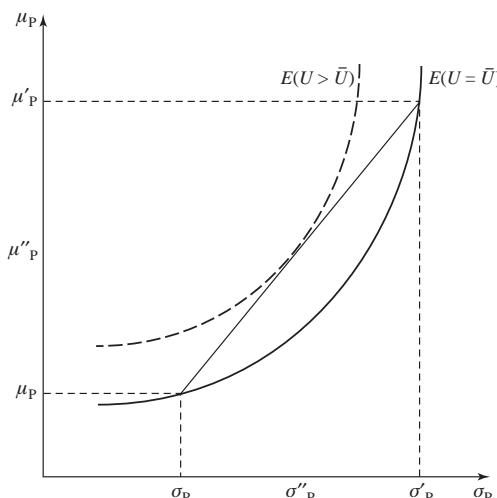


Figure A6.4
The indifference curves are convex-shaped.

Appendix 6.2: The Shape of the Efficient Frontier; Two Assets; Alternative Hypotheses

Perfect Positive Correlation (Figure 6.3)

$$\rho_{12} = 1$$

$\sigma_P = w_1\sigma_1 + (1 - w_1)\sigma_2$, the weighted average of the standard deviations of individual asset returns

$$\rho_{1,2} = 1$$

$$\mu_P = w_1\mu_1 + (1 - w_1)\mu_2 = \mu_1 + (1 - w_1)(\mu_2 - \mu_1)$$

$$\sigma_P^2 = w_1^2\sigma_1^2 + (1 - w_1)^2\sigma_2^2 + 2w_1w_2\sigma_1\sigma_2\rho_{1,2}$$

$$= w_1^2\sigma_1^2 + (1 - w_1)^2\sigma_2^2 + 2w_1w_2\sigma_1\sigma_2$$

$$= (w_1\sigma_1 + (1 - w_1)\sigma_2)^2 \text{ [perfect square]}$$

$$\sigma_P = \pm (w_1\sigma_1 + (1 - w_1)\sigma_2) \Rightarrow w_1 = \frac{\sigma_P - \sigma_2}{\sigma_1 - \sigma_2}; 1 - w_1 = \frac{\sigma_1 - \sigma_P}{\sigma_1 - \sigma_2}$$

$$\mu_P = \mu_1 + \frac{\sigma_1 - \sigma_P}{\sigma_1 - \sigma_2}(\mu_2 - \mu_1) = \mu_1 + \frac{\mu_2 - \mu_1}{\sigma_2 - \sigma_1}(\sigma_P - \sigma_1)$$

Imperfectly Correlated Assets (Figure 6.4)

$$-1 < \rho_{12} < 1$$

Reminder: $\mu_P = w_1\mu_1 + (1 - w_1)\mu_2$

$$\sigma_P^2 = w_1^2\sigma_1^2 + (1 - w_1)^2\sigma_2^2 + 2w_1(1 - w_1)\sigma_1\sigma_2\rho_{1,2}$$

Thus,

$$\frac{\partial \sigma_P^2}{\partial \rho_{1,2}} = 2w_1(1 - w_1)\sigma_1\sigma_2 > 0$$

which implies $\sigma_P < w_1\sigma_1 + (1 - w_1)\sigma_2$. σ_P is smaller than the weighted average of the σ 's, there are gains from diversifying.

Fix μ_P , hence w_1 , and observe: as one decreases $\rho_{1,2}$ (from +1 to -1), σ_P^2 diminishes (and thus also σ_P). Hence, the opportunity set for $\rho = \bar{\rho} < 1$ must be to the left of the line AB ($\rho_{1,2} = 1$) except for the extremes.

$$\begin{aligned} w_1 = 0 \Rightarrow \mu_P &= \mu_2 \text{ and } \sigma_P^2 = \sigma_2^2 \\ w_1 = 1 \Rightarrow \mu_P &= \mu_1 \text{ and } \sigma_P^2 = \sigma_1^2 \end{aligned}$$

Perfect Negative Correlation (Figure 6.5)

$$\rho_{1,2} = -1$$

$$\sigma_P^2 = w_1^2 \sigma_1^2 + (1-w_1)^2 \sigma_2^2 - 2w_1(1-w_1)\sigma_1\sigma_2 \text{ with } (w_2 = (1-w_1))$$

$$= (w_1\sigma_1 - (1-w_1)\sigma_2)^2 \text{ [perfect square again]}$$

$$\sigma_P = w \pm [w_1\sigma_1 - (1-w_1)\sigma_2] = \pm [w_1(\sigma_1 + \sigma_2) - \sigma_2]$$

$$w_1 = \frac{\pm \sigma_P + \sigma_2}{\sigma_1 + \sigma_2}$$

$$\sigma_P = 0 \Leftrightarrow w_1 = \frac{\sigma_2}{\sigma_1 + \sigma_2}$$

$$\begin{aligned} \mu_P &= \frac{\pm \sigma_P + \sigma_2}{\sigma_1 + \sigma_2} \mu_1 + \left(1 - \frac{\pm \sigma_P + \sigma_2}{\sigma_1 + \sigma_2}\right) \mu_2 \\ &= \frac{\pm \sigma_P + \sigma_2}{\sigma_1 + \sigma_2} \mu_1 + \frac{\sigma_1 \pm \sigma_P}{\sigma_1 + \sigma_2} \mu_2 \\ &= \frac{\sigma_2}{\sigma_1 + \sigma_2} \mu_1 + \frac{\sigma_1}{\sigma_1 + \sigma_2} \mu_2 \pm \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \sigma_P \end{aligned}$$

One Riskless and One Risky Asset (Figure 6.6)

Asset 1: $\mu_1, \sigma_1 = 0$

Asset 2: μ_2, σ_2

$\mu_1 < \mu_2$

$$\mu_P = w_1\mu_1 + (1-w_1)\mu_2$$

$$\sigma_P^2 = w_1^2 \sigma_1^2 + (1-w_1)^2 \sigma_2^2 + 2w_1(1-w_1)\text{cov}_{1,2}$$

$= (1-w_1)^2 \sigma_2^2$ since $\sigma_1^2 = 0$ and $\text{cov}_{1,2} = \rho_{1,2}\sigma_1\sigma_2 = 0$; thus,

$\sigma_P = (1-w_1)\sigma_2$, and

$$w_1 = 1 - \frac{\sigma_P}{\sigma_2}$$

Appendix 6.3: Constructing the Efficient Frontier

In this appendix, we outline how Excel's SOLVER program may be used to construct an efficient frontier using historical data on returns. Our method does not require the explicit computation of means, standard deviations, and return correlations for the various securities under consideration; they are implicitly obtained from the data directly.

The Basic Portfolio Problem

Let us, for purposes of illustration, assume that we have assembled a time series of four data points (monthly returns) for each of three stocks, and let us further assume that these four realizations fully describe the relevant return distributions. We also assign equal probability to the states underlying these realizations.

Following our customary notation, let w_i represent the fraction of wealth invested in asset i , $i = 1, 2, 3$, and let r_{P,θ_j} represent the return for a portfolio of these assets in the case of event θ_j , $j = 1, 2, 3, 4$. The Excel formulation analogous to problem (QP) of the text is found in [Table A6.1](#), where (A1) through (A4) define the portfolio's return in each of the four states; (A5) defines the portfolio's average return; (A6) places a bound on the expected return; by varying μ , it is possible to trace out the efficient frontier; (A7) defines the standard deviation when each state is equally probable; and (A8) is the *budget constraint*.

The Excel-based solution to this problem is

$$w_1 = 0.353$$

$$w_2 = 0.535$$

$$w_3 = 0.111$$

Table A6.1: The excel formulation of the (QP) problem

$\min_{\{w_1, w_2, w_3, w_4\}} \text{SD}$
 (Minimize portfolio standard deviation)
 Subject to:
 (A1) $r_{P,\theta_1} = 6.23w_1 + 5.10w_2 + 7.02w_3$
 (A2) $r_{P,\theta_2} = -0.68w_1 + 4.31w_2 + 0.79w_3$
 (A3) $r_{P,\theta_3} = 5.55w_1 - 1.27w_2 - 0.21w_3$
 (A4) $r_{P,\theta_4} = -1.96w_1 + 4.52w_2 + 10.30w_3$
 (A5) $\mu_P = 0.25r_1^P + 0.25r_2^P + 0.25r_3^P + 0.25r_4^P$
 (A6) $\mu_P \geq \mu = 3$
 (A7) $\text{SD} = \text{SQRT}(\text{SUMPRODUCT}(r_{P,\theta_1}, r_{P,\theta_2}, r_{P,\theta_3}, r_{P,\theta_4}))$
 (A8) $w_1 + w_2 + w_3 = 1$

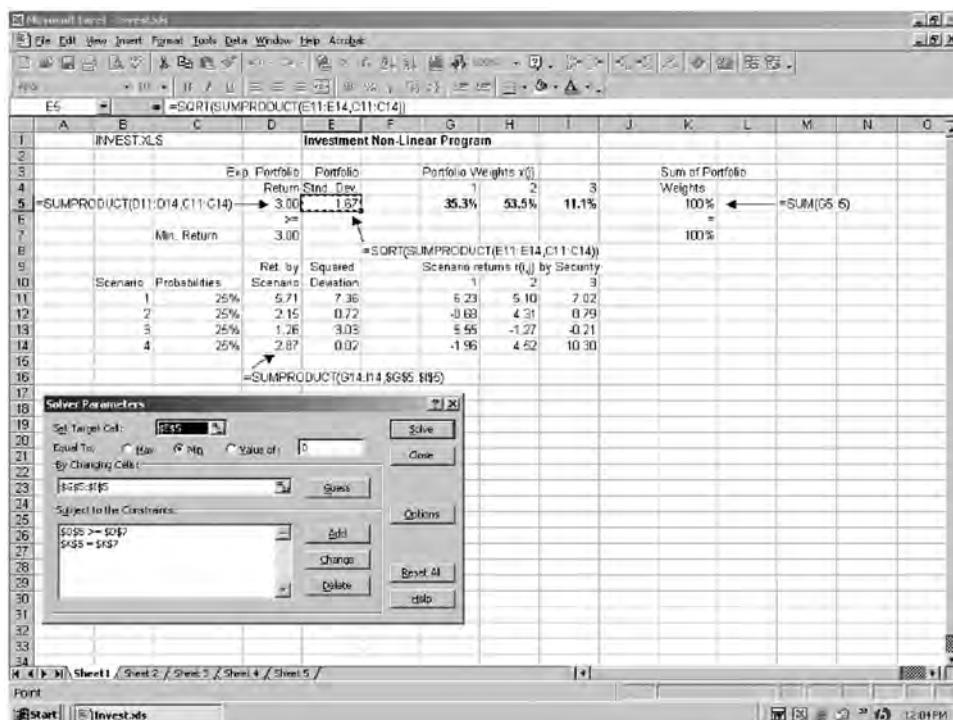


Figure A6.5
Excel Screen 1.

when μ is fixed at $\mu = 3.0\%$. The corresponding portfolio mean and standard deviation are $\mu_P = 3.00$, and $\sigma_P = 1.67$. Screen 1 (Figure A6.5) describes the Excel setup for this case.

Note that this approach does not require the computation of individual security expected returns, variances, or correlations, but it is fundamentally no different from Problem (QP) in the text, which does require them. Note also that by recomputing “min SD” for a number of different values of μ , the efficient frontier can be well approximated.

Generalizations

The approach described above is very flexible and accommodates a number of variations, all of which amount to specifying further constraints.

Nonnegativity Constraints

These amount to restrictions on short selling. It is sufficient to specify the additional constraints

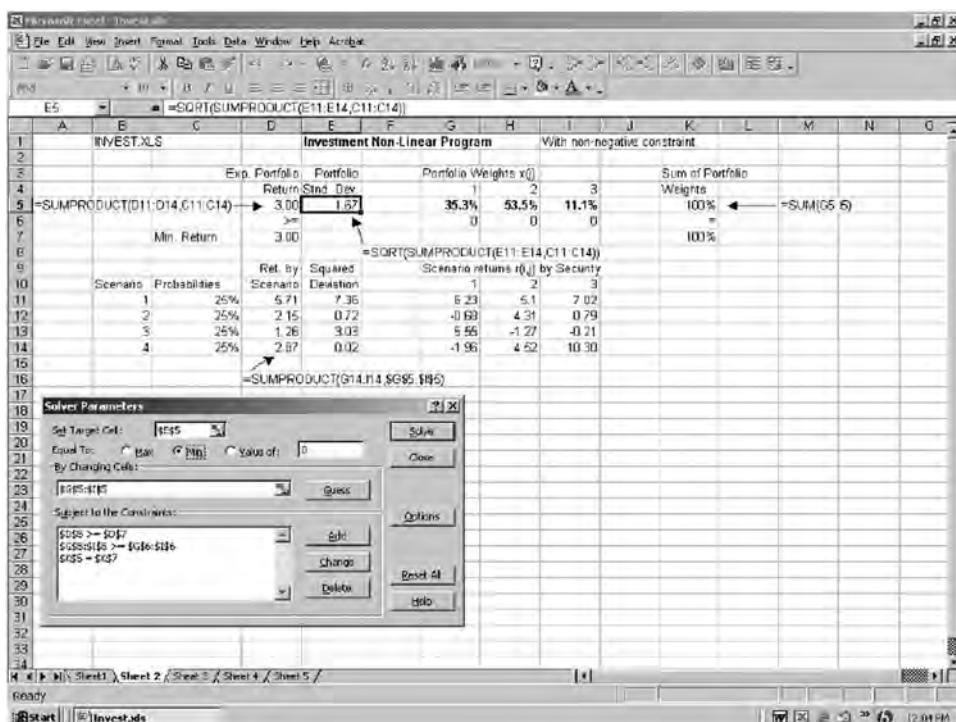


Figure A6.6
Excel Screen 2.

$$w_1 \geq 0$$

$$w_2 \geq 0$$

$$w_3 \geq 0$$

The functioning of SOLVER is unaffected by these added restrictions (although more constraints must be added), and for the example above the solution remains unchanged. (This is intuitive since the solutions were all positive.) See Excel Screen 2 (Figure A6.6).

Composition Constraints

Let us enrich the scenario. Assume that the market prices of stocks 1, 2, and 3 are, respectively, \$25, \$32, and \$17, and that the current composition of the portfolio consists of 10,000 shares of stock 1, 10,000 shares of stock 2, and 30,000 shares of stock 3, with an aggregate market value of \$1,080,000. You wish to obtain the lowest standard deviation for

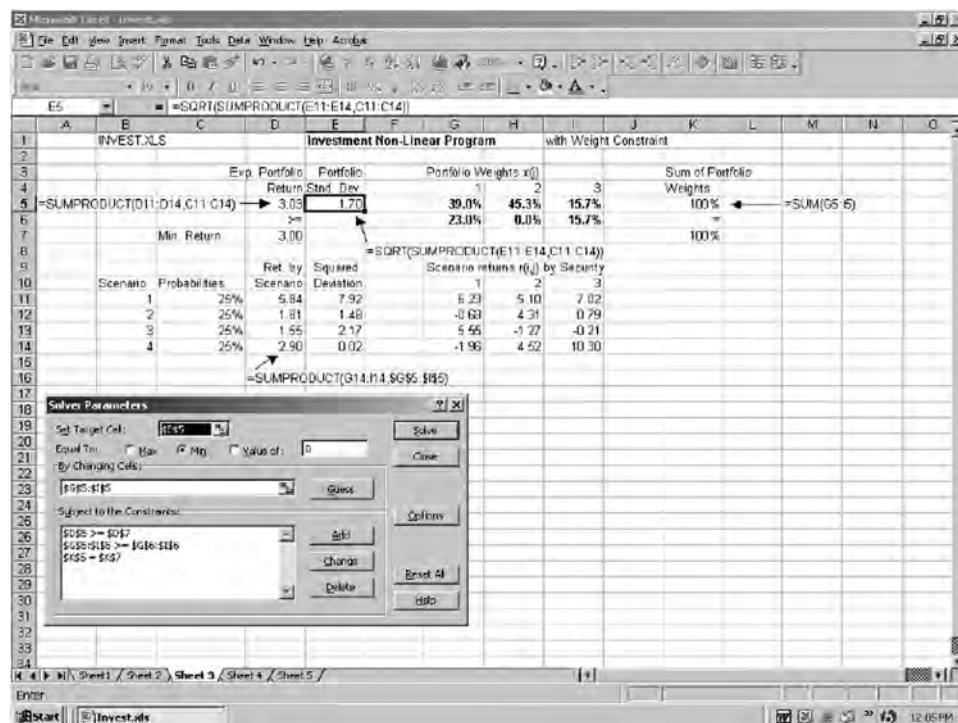


Figure A6.7
Excel Screen 3.

a given expected return subject to the constraints that you retain 10,000 shares of stock 1 and 10,000 shares of stock 3. Equivalently, you wish to constrain portfolio proportions as follows:

$$w_1 \geq \frac{10,000 \times \$25}{\$1,080,000} = 0.23$$

$$w_3 = \frac{10,000 \times \$17}{\$1,080,000} = 0.157$$

while w_2 is free to vary. Again SOLVER easily accommodates this. We find $w_1 = 0.39$, $w_2 = 0.453$, and $w_3 = 0.157$, yielding $\mu_P = 3.03\%$ and $\sigma_P = 1.70\%$. Both constraints are binding. See Excel Screen 3 (Figure A6.7).

Adjusting the Data (Modifying the Means)

On the basis of the information in Table A6.2,

Table A6.2: Hypothetical return data

	Probability	Stock 1 (%)	Stock 2 (%)	Stock 3 (%)
State 1	0.25	6.23	5.10	7.02
State 2	0.25	-0.68	4.31	0.79
State 3	0.25	5.55	-1.27	-0.21
State 4	0.25	-1.96	4.52	10.30

Table A6.3: Modified return data

	Probability	Stock 1 (%)	Stock 2 (%)	Stock 3 (%)
Event 1	0.25	7.23	6.10	6.02
Event 2	0.25	0.32	5.31	-0.21
Event 3	0.25	6.55	-0.27	-1.21
Event 4	0.25	-0.96	5.52	9.30

$$\mu_1 = 2.3\%$$

$$\mu_2 = 3.165\%$$

$$\mu_3 = 4.47\%$$

Suppose other information becomes available suggesting that, over the next portfolio holding period, the returns on stocks 1 and 2 would be 1% higher than their historical mean and the return on stock 3 would be 1% lower. This supplementary information can be incorporated into min SD by modifying [Table A6.2](#). In particular, each return entry for stocks 1 and 2 must be increased by 1% while each entry of stock 3 must be decreased by 1%. Such changes do not in any way alter the standard deviations or correlations implicit in the data. The new input table for SOLVER is found in [Table A6.3](#).

Solving the same problem, min SD without additional constraints yields $w_1 = 0.381$, $w_2 = 0.633$, and $w_3 = -0.013$, yielding $\mu_P = 3.84$ and $\sigma_P = 1.61$. See Excel Screen 4 ([Figure A6.8](#)).

Constraints on the Number of Securities in the Portfolio

Transactions costs may be substantial. In order to economize on these costs, suppose an investor wished to solve min SD subject to the constraint that his portfolio would contain at most two of the three securities. To accommodate this change, it is necessary to introduce three new binary variables that we will denote x_1, x_2, x_3 , corresponding to stocks 1, 2, and 3, respectively. For all $x_i, i = 1, 2, 3, x_i \in \{0, 1\}$. The desired result is obtained by adding the following constraints to the problem min SD:

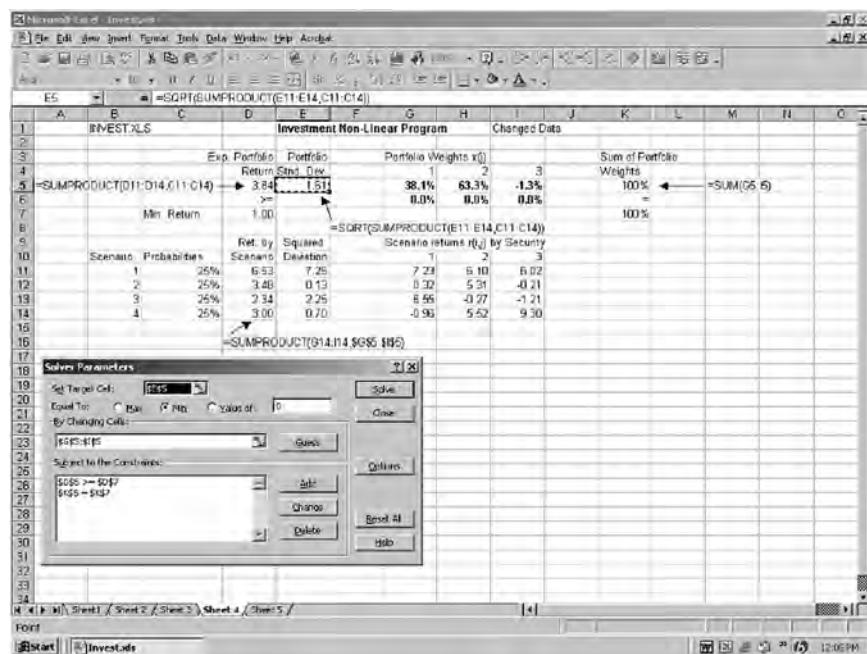


Figure A6.8
Excel Screen 4.

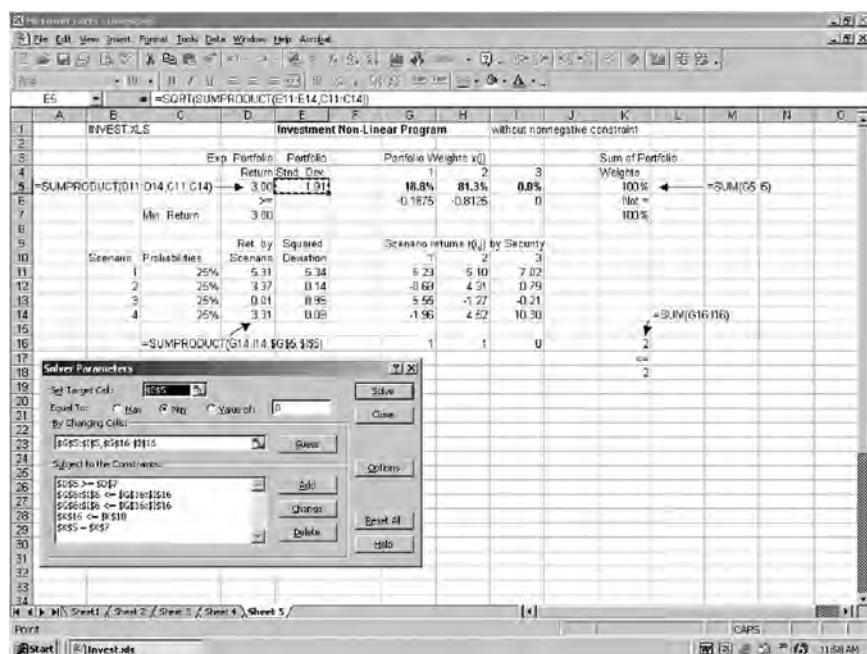


Figure A6.9
Excel Screen 5.

$$\begin{aligned}w_1 &\leq x_1 \\w_2 &\leq x_2 \\w_3 &\leq x_3 \\x_1 + x_2 + x_3 &\leq 2, \\x_1, x_2, x_3 &\text{ are binary}\end{aligned}$$

In the previous example, the solution is to include only securities one and two with proportions $w_1 = 0.188$, and $w_2 = 0.812$. See Excel Screen 5 ([Figure A6.9](#)).

Risk Aversion and Investment Decisions, Part III: Challenges to Implementation

Chapter Outline

7.1 Introduction 181

7.2 The Consequences of Parameter Uncertainty 183

7.3 Trends and Cycles in Stock Market Return Data 187

 7.3.1 Trends in International Stock Market Cross-Correlations 188

 7.3.2 Asset Correlations in Cyclical Periods of High Volatility 190

 7.3.3 The Financial Crisis 191

7.4 Equally Weighted Portfolios 193

7.5 Are Stocks Less Risky for Long Investment Horizons? 195

 7.5.1 Long- and Short-Run Equity Riskiness: Historical Patterns 195

 7.5.2 Intertemporal Stock Return Behavior Through Time: The Random Walk Model 197

 7.5.3 Are Stocks Less Risky in the Long Run? A Predictive Perspective 201

7.6 Conclusions 203

References 204

Appendix 7.1 205

7.1 Introduction

Prior to the advent of modern portfolio theory (MPT), financial researchers (and practitioners) had no comprehensive mental or analytical structure for understanding either how asset “risk” might be measured or how the return patterns of a portfolio’s constituent assets interact to determine its overall risk. Although practitioners were certainly aware of the qualitative benefits of portfolio diversification, their understanding often took the form of simply noting that the act of spreading one’s wealth across many risky assets (or investment projects) increased the likelihood that some, at least, would pay positive returns.¹ MPT dramatically extended this qualitative understanding by making

¹ The notion of diversification to mitigate investment risk is very old. See Ecclesiastes 11: 1–2; New Revised Standard Version, Harper Collins Study Bible: “Send out your bread upon the waters, for after many days you will get it back. Divide your means seven ways, or even eight, for you do not know what disaster may happen on earth.”

precise the measurement of risk overall and the measurement and source of the “gains to diversification.”

It is not a large step from descriptive to normative applications of MPT. As a result, MPT has become a fundamental tool for assisting investors in making rational portfolio allocation decisions. The message is clear: there are large gains to diversification and these are achieved by allocating one’s wealth between risk-free assets and the available tangency portfolio.

Normative applications of the theory, however, require estimates of the fundamental statistical quantities upon which it is built: the (μ_i, σ_i) for every candidate asset i , and the cross-correlations ρ_{ij} for all asset return pairs. These numbers are necessary to characterize the investor’s future risk/return possibilities. How are they to be obtained?

As noted in Box 3.1, it is natural to look to historical return data as the source of these estimates and, more precisely, to use historical average returns, historical standard deviations, and historical correlations as the estimates, respectively, for the required μ_i, σ_i , and ρ_{ij} s. If the returns and return interactions are statistically stationary with a unique ergodic set and if the historical data series is sufficiently long, the aforementioned quantities can be estimated with a high degree of precision.² In many applied applications, however, it is customary to base these estimates only on the prior 5 years of monthly return data. This choice follows from the fact that many investors have 1-year (or less) investment horizons and are therefore more concerned with conditional estimates rather than the unconditional ones that long data series are designed to provide. Equity return patterns over the coming year, for example, may be highly dependent on the macroeconomy’s present state: is it in recession or expansion?

Furthermore, the present state can change fairly rapidly as we have seen recently. The financial crisis was preceded by a long period of gradual macroeconomic expansion often referred to as the “Great Moderation.” It ended abruptly with the bankruptcies of Lehman Brothers, Bear Stearns, and AIG (all within a period of a few months), and the onset

² Our notion of stationarity for a discrete time (Markov) process $\{x_t\}$ is as follows; Let s, t be arbitrary time indices and X the state space with $\hat{x} \in X$, and $B \subseteq X$, B a subset. Define

$$P(s, \hat{x}, t, B) = \text{Prob}(x_t \in B; x_s = \hat{x})$$

Then for any integer u , if

$$P(s + u, \hat{x}, t + u, B) = \text{Prob}(x_{t+u} \in B; x_{s+u} = \hat{x}) = P(s, \hat{x}, t, B)$$

the (Markov) process is said to be stationary.

The same (Markov) process possesses an invariant distribution $\hat{G}(\cdot)$ on X if and only if for any $B \subseteq X$, $\text{Prob}(x_{t+1} \in B) = \int_{X \in X} P(x_{t+1} \in B; x_t = x) \hat{G}(dx)$. The process is ergodic if the invariant distribution describes the long-run average pattern in the data.

of the “Great Recession.” Statistical estimates drawn from “Great Moderation” data were not very informative vis-à-vis stock market behavior in the first year of the financial crisis.

As these thoughts suggest, while investors should be guided by the implications of MPT, they must also be aware of its practical limitations. We elaborate upon these issues in the remainder of the chapter.

7.2 The Consequences of Parameter Uncertainty

In order to construct the efficient frontier of risky assets, an investor or his agent must obtain estimates $\hat{\mu}_i$, $\hat{\sigma}_i$, and $\hat{\rho}_{ij}$, $\forall i \neq j$ for all risky assets (individual shares, mutual funds, exchange traded funds (ETFs), etc.) under consideration for portfolio inclusion. These estimates, obtained from historical data are imperfect “stand-ins” for the true μ_i , σ_i , and ρ_{ij} $\forall i \neq j$ which are themselves presumed to be stable inclusive of the historical period of estimation and the investor’s future investment horizon. As computed in Chapter 3, the $\hat{\mu}_i$, $\hat{\sigma}_i$, and the $\hat{\rho}_{ij}$ s represent unbiased estimates of their true counterparts, but they may not be precise estimates if the samples from which they are estimated are “too small.”

Unfortunately, errors in these estimates interact with the technique of mean–variance portfolio optimization to trace out an “estimated efficient frontier” which may be radically distorted relative to its true counterpart.

Most of the distortion has its origin in the misestimation of mean returns (see [Box 7.1](#)). To illustrate how mean misestimation can distort an investor’s perception of his risk/return tradeoffs, consider the construction of an efficient frontier of six assets with true monthly mean returns of $(\mu_1, \mu_2, \dots, \mu_6)$ and estimated mean returns of $(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_6)$. It is the latter quantities that influence actual efficient frontier construction. Roughly speaking, one would expect $\hat{\mu}_j > \mu_j$ for three of the assets and $\hat{\mu}_j < \mu_j$ for the remaining three. When computing the efficient frontier, these errors do not “average out,” however, because mean–variance optimization will, by design, overweight (relative to what would be the case using the true values) the former group of assets, and underweight the latter group. Accordingly, the estimated frontier will lie above the true efficient frontier, with the distortion being the least in the region of the minimum variance portfolio where mean return values are downplayed.

BOX 7.1 On the Precision of Mean Estimates

Straightforward estimations typically rely on monthly data going back at most 5 years yielding 60 data points. Suppose that $\hat{\mu}_i = 0.8\%$ and $\hat{\sigma}_i = 6\%$ at monthly frequencies. The standard error of the estimate of the mean is $(\hat{\sigma}_i/\sqrt{J}) = (0.06/\sqrt{60}) = 0.008$, yielding a confidence interval around the true mean of

(Continued)

BOX 7.1 On the Precision of Mean Estimates (Continued)

$$\begin{aligned} \text{Prob}\left(\hat{\mu}_i - 2\left(\frac{\hat{\sigma}_i}{\sqrt{J}}\right) \leq \mu_i \leq \hat{\mu}_i + 2\left(\frac{\hat{\sigma}_i}{\sqrt{J}}\right)\right) &= 0.95 \\ &= \text{Prob}(0.008 - 2(0.008) \leq \mu_i \leq 0.008 + 2(0.008)) = 0.95 \\ &= \text{Prob}(-0.008 \leq \mu_i \leq 0.024) = 0.95 \end{aligned}$$

By implication, we cannot even conclude to a high degree of confidence that μ_i is positive. Variances computed on the basis of 60 data points, however, are more precisely estimated than means. The standard error in this case is $(\hat{\sigma}_i/\sqrt{2J})$.

Neither of these frontiers, however, describes the operative frontier confronting the investor. That frontier must be the one which uses portfolio weights obtained via MPT analysis using the historical estimates in conjunction with the true means, variances, and return correlations. By construction, the operative frontier must lie below the true efficient frontier and in fact need not even be concave. Figure 7.1 illustrates these relationships qualitatively. While the position of the efficient frontier constructed using the true μ_i , σ_i , and ρ_{ij} s is not known, it must lie between its two “half-brother” frontiers, a fact of little use since the position of the operative frontier is similarly unknown. Note that the investor ends up effectively selecting portfolios with the worst risk-return characteristics.

To get some feel for the actual magnitudes involved, we compute the above three frontiers using six assets whose plausible monthly return statistics are described in Table 7.1. The results of this exercise are presented in Figure 7.2.³

Note that at a 3% portfolio SD, the difference between the estimated and operative expected returns is roughly .65% per month ($0.019 - 0.0125$), or 7.8% on an annualized basis. In this case, the distortions introduced by estimation errors lead to a substantially exaggerated sense of expected portfolio returns.⁴ In general mean-variance optimization tends to overweight securities with large estimated returns, small variances, and negative correlations relative to other candidate assets, and vice versa.

³ To generate the Operative Frontier in Figure 7.2, we first generated a random time series of 24 return entries for each of the securities in Table 7.1 in a manner that respected the stated moments and cross-correlations. From this artificial data, sample means, variances, and cross-correlations were then computed. It is these latter quantities that formed the basis for constructing both Estimated and Operative Frontiers in the following manner: For a given σ , the Estimated Frontier uses the $\hat{\mu}_i$, $\hat{\sigma}_i$, and $\hat{\rho}_{ij}$ obtained from the generated data to get proportions $(\hat{w}_1, \dots, \hat{w}_6)$. The Operative Frontier uses these $(\hat{w}_1, \dots, \hat{w}_6)$ values in conjunction with the true μ_i , σ_i , and ρ_{ij} .

⁴ The magnitude of the discrepancy between estimated and operative returns generally declines, however, as the number of available securities increases.

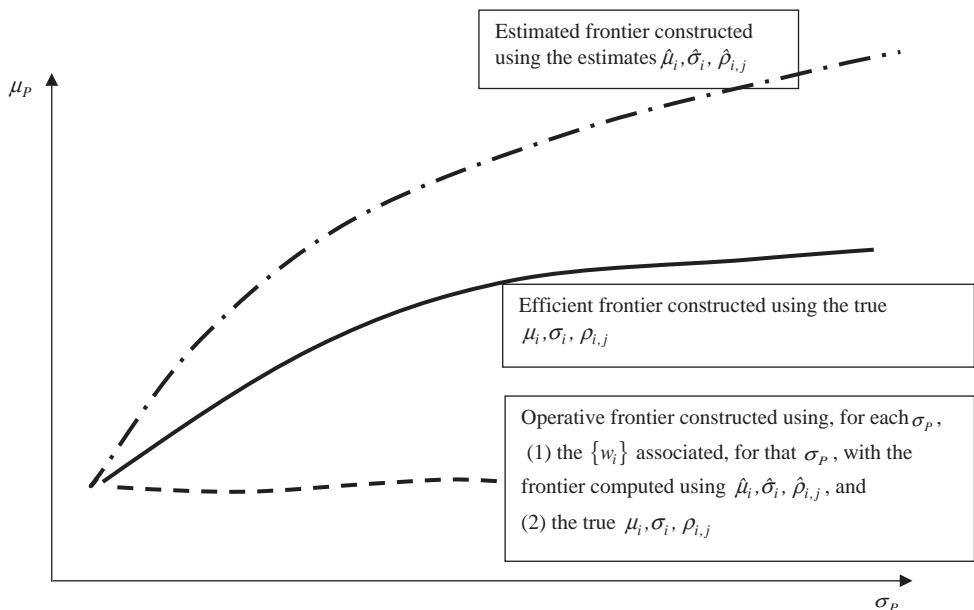


Figure 7.1
True, actual, and operative frontiers: general pattern.

Table 7.1: Underlying moments for the six securities

Security	1	2	3	4	5	6
True mean	0.016	0.014	0.013	0.011	0.01	0.012
True SD	0.06	0.05	0.055	0.04	0.035	0.051
Correlation matrix: for all $i \neq j$, $\rho_{ij} = 0.10$						

The portfolios we have been discussing are collections of risky assets alone. When risk-free assets are included as an investment option in conjunction with eligible portfolios on the estimated frontier, their inclusion may lead to the construction of portfolios with excessive allocations to risky assets, at least relative to what would result using candidate portfolios on the “true” or “operative” frontiers. In a mean–variance investment environment, “overoptimistic” parameter estimates may thus contribute to excessive risk taking. This assertion follows from the fact that the efficient frontier of risk-free and risky assets using the estimated risky asset frontier is steeper than the one based on the operative frontier of risky assets, i.e., its Sharpe ratio is higher. As a result, the investor will tend to allocate a larger fraction (given his $U(\mu_i, \sigma_i)$) of his investible wealth to risky assets (Figure 7.3). If the equity portfolio ends up being very highly leveraged (as in a hedge-fund context) and actual *ex post* returns turn out to be low (as the position of the operative frontier suggests), the result could be cumulatively devastating. Investors, of course, do not know the operative frontiers, but they should be aware that it lies below their estimated one.

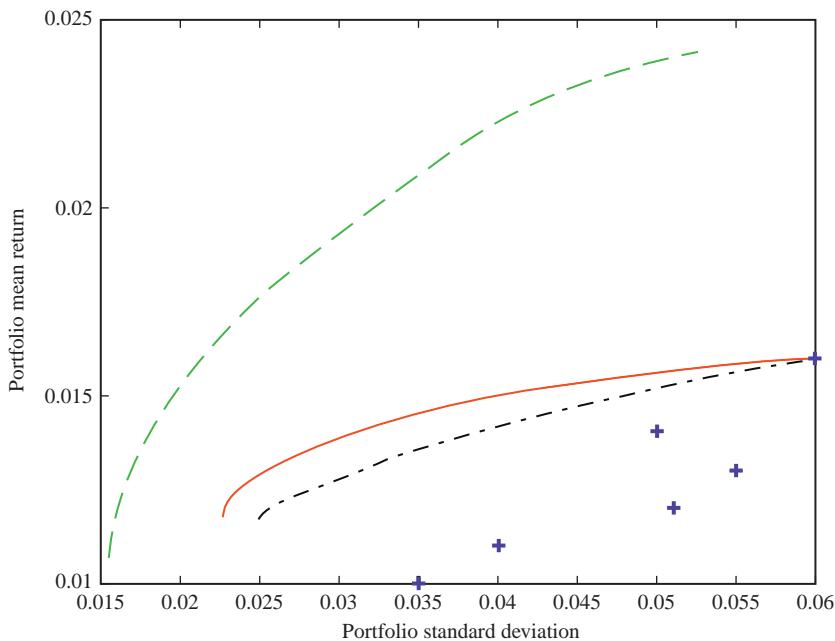


Figure 7.2: True, estimated, and operative frontiers⁽ⁱ⁾

(i) The three frontiers are pictured in red (true), green (estimated), and black (operative) respectively; the + signs denote the individual securities based on simulated return data. Computations based on 24 monthly return data points.

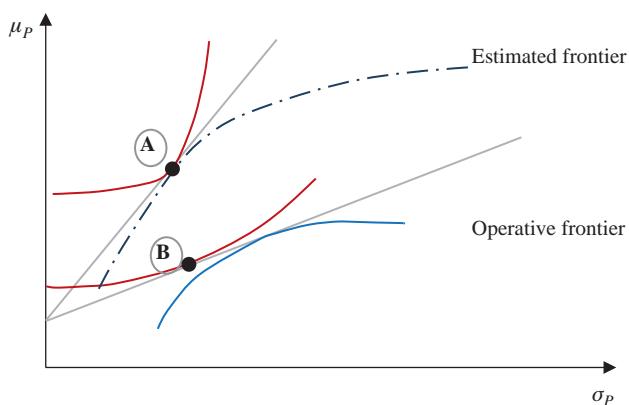


Figure 7.3: Tangency portfolios, estimated and operative frontiers⁽ⁱ⁾

(i) Consider optimal allocations (A) using estimated and (B) using operative frontiers. The set of mean-variance indifference curves is the same for all frontiers. Note that in this example, the estimated frontier suggests an all stock portfolio while the operative frontier (were it actually to be identified) suggests equal proportions in stocks and risk-free assets.

These observations do not imply that MPT is “wrong” in any way. Due to data limitations, however, it may give rise to misleading conclusions.

There is another aspect of MPT which arises because of its emphasis on mean portfolio returns, but that has nothing to do with estimation issues. This is the tendency of MPT to construct efficient portfolios with large long and short positions when short selling constraints are not imposed. Many investors are prohibited from short selling, or from undertaking short positions as large as unconstrained MPT would recommend. We defer this discussion to Appendix 7.1. The appendix relies on data presented in the next section and thus should be read subsequent to it.

7.3 Trends and Cycles in Stock Market Return Data

When the input data to an efficient frontier calculation is based on historical return series, it is implicitly assumed that the stochastic return relationships on which the data is based will be preserved for the future investment horizon. This is a simple characterization of stationarity. A related issue is the precision of the estimates: precise estimates require lots of data, which means relying on data series into the more distant past. This requirement, however, makes the stationarity assumption more difficult to accept: economic relationships are not static.

In recent years, changing economic relationships have been especially manifest in the international arena. The globalization of industry and finance, in particular, has brought the economies of the world into closer alignment and interconnectedness. There is increasing evidence of a world business cycle, a phenomenon we would expect to translate into increased cross-country stock market return correlations. We also observe increased world financial market integration, with the accompanying increase in cross-border capital flows.⁵ Especially in times of financial stress (high return volatility), the ease of capital flows allows the consequences of destructive economic events in one country, most especially, to be transmitted almost instantaneously across all of the world’s principal financial markets. The recent financial crisis is a major case in point: while originating in the United States, it quickly spread to Europe and then to Asia.

There are both trend and cyclical aspects to these considerations which we document shortly. In either case, they represent challenges for investors because they suggest that the potential for risk reduction via international diversification may in general be diminishing (trend) and that this phenomenon may be most acutely experienced in cyclical periods of

⁵ Global capital flows as a percentage of world GDP rose from 4% in 1994 to 20% in 2007. By 2009, these flows had collapsed to 2.5% of world GDP. See Milesa-Ferretti and Tille (2010).

extreme financial uncertainty (“tail events”). In essence, “when investors need most to be well diversified, the possibilities for diversification disappear.” In the remainder of this section, we illustrate these considerations.

7.3.1 Trends in International Stock Market Cross-Correlations

To get some flavor for trends in the inputs to the efficient frontier, let’s explore the gains to international diversification across national stock markets. We consider estimates based on historical monthly return data for the period 1.1.1996 through 12.31.2006. These estimates are found in [Tables 7.2 and 7.3](#) for, respectively, the full sample period and the most recent half sample.

First, note the enormous variation in average returns across the two samples. In passing from 10 years of data to the more recent 5 years, average monthly French returns go from

Table 7.2: Major stock markets summary risk/return statistics 1.1.1996–12.31.2006^a

	FRA	GER	Japan	UK	US	Brazil	Korea
Average monthly return	0.71%	0.65%	− 0.21%	0.34%	0.57%	1.57%	0.63%
Standard deviation	5.9%	7.2%	5.6%	4.1%	4.5%	10.4%	11.0%
Correlation table							
France	1						
Germany	0.90	1					
Japan	0.45	0.43	1				
UK	0.79	0.79	0.49	1			
US	0.74	0.77	0.37	0.79	1		
Brazil	0.58	0.60	0.27	0.61	0.65	1	
Korea	0.40	0.39	0.45	0.51	0.48	0.39	1

^aAuthors’ calculations.

Table 7.3: Major stock markets summary risk/return statistics 1.1.2002–12.31.2006^a

	FRA	GER	Japan	UK	US	Brazil	Korea
Average monthly return	− 0.40%	− 0.38%	0.25%	− 0.19%	− 0.15%	1.08%	1.36%
Standard deviation	5.9%	7.9%	5.2%	4.1%	4.3%	8.2%	7.4%
Correlation table							
France	1						
Germany	0.95	1					
Japan	0.53	0.49	1				
UK	0.90	0.88	0.51	1			
US	0.89	0.89	0.49	0.87	1		
Brazil	0.67	0.65	0.43	0.67	0.72	1	
Korea	0.74	0.74	0.56	0.71	0.72	0.54	1

^aAuthors’ calculations.

0.71% to, more recently, -0.40% , average returns to the Brazilian market index jump by 50% and Korean returns double. At least for shorter sample periods, these numbers illustrate the instability of the mean estimates. Much less variation across the data samples is observed for estimates of σ_j ; changes are comparatively small and in two cases, France and the United Kingdom, the estimate is unchanged. Note also that within this sample of countries, stock market returns become uniformly more highly correlated with the passage of time: the correlation matrix in [Table 7.3](#) dominates its counterpart in [Table 7.2](#) entry by entry. While not exhausted, the gains to international diversification across these countries over a purely domestic portfolio appear to be diminishing with time. This fact in itself is not surprising: as national economies become more highly integrated into the world economy and, as such, become more influenced by the world business cycle, we would expect stock returns to become more highly correlated. For countries with long-term historical trading relationships, stock market returns are already highly correlated; see, for example, France and Germany. Note also that there is generally less cross-country variation in standard deviations than in the means for either data set.

Confirming our earlier discussion, the locations of the efficient frontier based on information in, respectively, [Tables 7.2 and 7.3](#) for the G5 subsample (France, Germany, Japan, the UK and the USA) differ substantially. This is captured in [Figure 7.4](#).

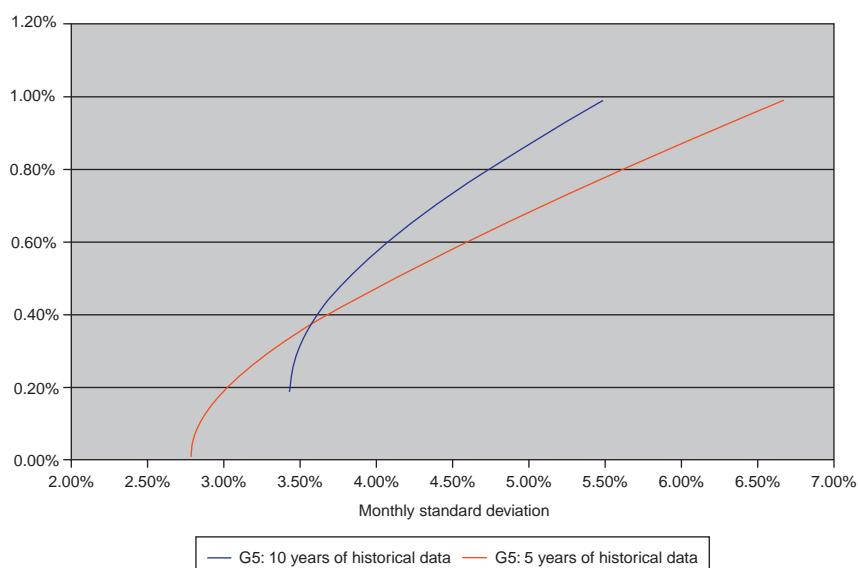


Figure 7.4: Effect of updating the period used to calculate risk/return estimates (most recent 5 years versus 10 years of historical data) for the G5 countries.

Source: Based on data, respectively, in [Tables 7.2 and 7.3](#).

7.3.2 Asset Correlations in Cyclical Periods of High Volatility

Another aspect of parameter instability is directly associated with periods of high stock market return volatility. In particular, cross-country market returns for many of the world's major stock markets tend to become more highly correlated during periods of general market decline, e.g., large losses in the US market tend to coincide with large losses across other major markets. The classic reference in this regard is [Longin and Solnik \(2001\)](#), who measure the extent of these effects across the aggregate stock markets of the United States, Germany, France, the United Kingdom, and Japan using monthly return data for the period 1959–1996. [Figure 7.5](#) captures the essence of their results for US–UK return relationship.

Some interpretation is required. First, the ‘black dot’ on the vertical axis describes the correlation of US and UK aggregate stock market returns using the full data sample: 0.519. The continuous line represents the conditional correlation of US/UK stock market returns in the subsamples where the returns in either country exceeded the indicated levels (numbers on the horizontal axis). For example, in the periods for which either return was less than –10%, the aggregate US–UK return correlation was 0.676. Generally, similar patterns are observed across all the sample stock markets, i.e., the pattern is similar for the US–Germany, US–France, etc.

The important qualitative implication to note is that the greater the severity of the market decline in either country, the greater the cross-correlation of returns—and the less effective cross-country wealth diversification would be in reducing risk. Note also that this pattern is

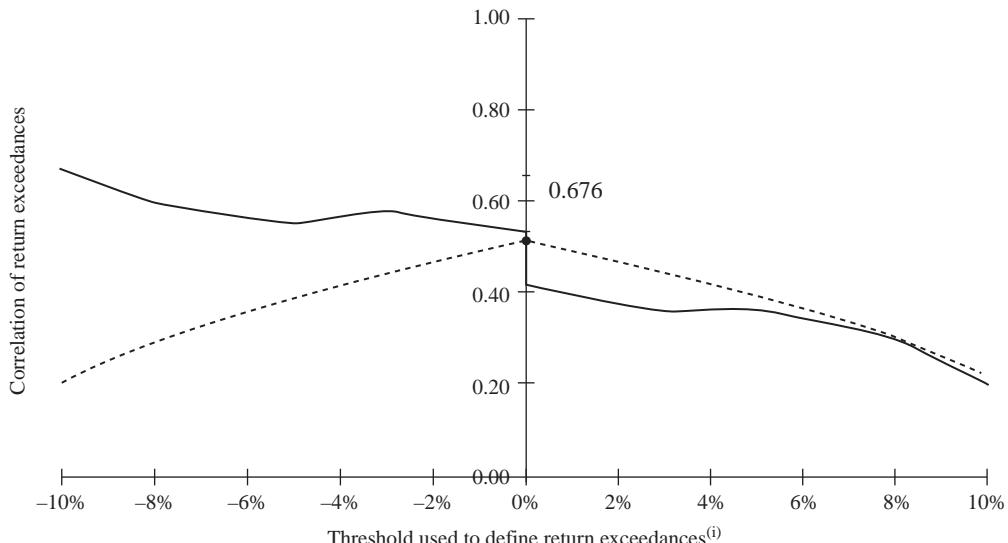


Figure 7.5: Correlation between US and UK aggregate stock market return exceedances

⁽ⁱ⁾ Exceedances signify returns above the corresponding r_f . Source: [Longin and Solnik \(2001\)](#).

reversed in the case of positive returns. While Longin and Solnik's data set does not include the period immediately preceding the financial crisis, there is no reason to believe the pattern conveyed in [Figure 7.5](#) has otherwise changed. While these results again do not refute the principle of diversification, they do remind us that the benefits to diversification will likely diminish in periods of severe market stress. The dashed line gives the predictive correlations for the same benchmarks under return normality assumptions where the UK–US return cross-correlation is taken as its full sample value of 0.519.

Similar results were found by [Riberio and Veronesi \(2002\)](#) using a different data set. They report average stock market and industrial production cross correlations across the United Kingdom, Japan, Germany, Canada, France, Switzerland, and the United States exclusively during periods of recession and boom as defined by the National Bureau of Economic Research (NBER). These are reported in [Table 7.4](#).

Focusing on quarterly data, [Table 7.4](#) makes us aware not only of increased stock market return correlations in “bad times” (recessions) but also the underlying increased business cycle correlations that these return correlations reflect. In traditional factor models of security returns (e.g., [Chen, Roll and Ross, 1986](#)), industrial production is often selected as an underlying macroeconomic factor. Note also that both correlations are less strong in periods of boom relative to periods of recession which is in line with the [Longin and Solnik \(2001\)](#) results, suggesting that “booms” are essentially more idiosyncratic than recessions.

7.3.3 The Financial Crisis

National stock markets whose returns were already highly correlated generally became even more so during the period of the financial crisis. [Table 7.5](#) provides cross-correlations for the world's major stock markets (by aggregate market value) for the period

Table 7.4: Average correlations for aggregate stock market returns and industrial production in recessions and booms^a

		Average Correlation Based on Monthly Data	
Recession	Stock Returns		Industrial Production
	0.7847		0.5295
Average Correlation Based on Quarterly Data			
Boom	0.7229		0.5207
	0.8546		0.6439
Recession	0.7263		0.5558

^aBased on data from 1970–2000; sample average correlation across all country pairs for United Kingdom, United States, Japan, Germany, Canada, France, and Switzerland.

Source: [Riberio and Veronesi \(2002\)](#).

Table 7.5: Summary risk/return statistics for major stock markets during the Financial Crisis, 1.1.2008–12.31.2012^a

		FRA	GER	Japan	UK	US	Italy	Canada
Average monthly return		− 0.09%	0.079%	− 0.21%	0.05%	0.30%	− 0.72%	0.31%
Monthly standard deviation		8.28%	8.87%	5.26%	6.55%	5.47%	9.40%	7.57%
Correlation table								
	FRA	1						
	GER	0.95	1					
	Japan	0.76	0.78	1				
	UK	0.92	0.89	0.79	1			
	US	0.89	0.90	0.77	0.91	1		
	Italy	0.97	0.92	0.75	0.85	0.85	1	
	Canada	0.82	0.81	0.73	0.78	0.86	0.78	1

^aWe thank G. Bekaert for providing these statistics to us.

January 2008–December 2012 which more than encompasses the worst of the crisis. The Lehman Brothers bankruptcy filing, which is sometimes cited as the start of the “true crisis phase,” occurred early in the morning of September 15, 2008.

In comparing [Table 7.5](#) with [Table 7.3](#), much of the message of [Section 7.3.2](#) is confirmed. Where comparable, all the cross-correlations in [Table 7.5](#) exceed or are equal to their counterparts in [Table 7.3](#) and strictly exceed their counterparts in [Table 7.2](#). For countries whose stock market returns were already highly correlated such as Germany and France, there is little change ([Table 7.5](#) versus [Table 7.3](#)). Of greater interest is the case of Japan: its stock market returns became much more highly correlated with the rest of the world: on an entry-by-entry basis, the column in [Table 7.5](#) for Japan exceeds its counterpart in [Table 7.3](#), reflecting the worldwide nature of the crisis.

Note also that the return standard deviations are uniformly higher during the crisis than in either earlier period. It is of interest that the highest return standard deviations are observed among the European countries, with Germany’s returns the most volatile despite the fact that its macroeconomy was less affected by the crisis than those of France, the United States, and the United Kingdom. This may reflect the lingering sovereign debt crisis in Europe, and Germany’s central role in any plan for its alleviation. The European sovereign debt crisis had no counterpart in the United States or the United Kingdom.

Mean monthly returns are, in general, not disastrously low (except in the case of Italy), reflecting the stock market recoveries that prevailed during the latter half of the sample period. Returns were uniformly negative for all stock markets in the year following the Lehman Brothers bankruptcy. As of this writing (May 2014), the macroeconomic consequences of the crisis (e.g., the loss of employment) are still manifest.

7.4 Equally Weighted Portfolios

Our discussion has thus far highlighted some of the difficulties encountered in the actual practice of portfolio formation. We emphasize that these considerations do not represent the failure of MPT, but rather its quantitative limitations due to data availability. In this section, we present one potential manifestation of these data restrictions.

There is substantial evidence that many investors follow a simple, equal-weight allocation rule for the assets they have decided to hold ([Benartzi and Thaler, 2001](#); [Huberman and Jiang, 2006](#)). Most mean–variance-based portfolio selection rules, even those with constraints (e.g., no short sales) arrive at portfolio weights, however, that are far from equal. While the origins of this practice may be behavioral, we note that an equally weighted portfolio does not require parameter (means, variances) estimation and thus does not suffer from the consequences of misestimating these quantities as discussed in [Section 7.2](#). Could this fact alone justify the widespread adoption of the equal weighting rule?

Various authors, in particular [DeMiguel et al. \(2009\)](#) and [Plyakha et al. \(2012\)](#), have recently compared the return characteristics of naïve, equally weighted portfolios with monthly rebalancing over and against those arising from various mean–variance motivated sophisticated selection rules. In nearly all cases and using a variety of performance measures (e.g., the classic Sharpe ratio), the equally weighted portfolio strongly outperforms those organized around the mean-variance-efficiency principle.

How is their performance evaluation undertaken? [DeMiguel et al. \(2009\)](#) first select a particular initial historical time interval and assemble the actual return data during this period for all assets under consideration for portfolio inclusion. This data set constitutes the “data sample” and it is used to estimate, for every asset, its mean returns, return standard deviations, and return correlations with other assets. Using these estimates and the portfolio selection technique being evaluated, the resulting portfolio returns are constructed using the actual constituent asset returns in the subsequent time period (so called out of sample portfolio returns) on a monthly updating basis.⁶ The various performance measures are then applied to this newly constructed portfolio return data. In particular, [DeMiguel et al. \(2009\)](#)

⁶ This “rolling-sample” approach to model evaluation is quite clearly described in the [DeMiguel et al. \(2009, p. 1927\)](#) paper; we paraphrase them as follows: given a T period-long data set, they choose an estimation window of length $M = 60$ or $M = 120$ months, $M < T$. In each month t , starting from $t = M + 1$, they use the data in the previous M months to estimate the parameters (means, variances, etc.) needed to implement a particular strategy: they use these estimated parameters to determine the relative portfolio weights in the portfolio of risky assets. These weights determine the composition of the risky portfolio that is used to determine the actual portfolio return in period $t + 1$ under the given strategy. The process continues by adding the actual return realizations for the various underlying assets next period to the data set while dropping the earliest return until the end of the data set is reached. The result of this procedure is a series of $T - M$ monthly “out of sample” returns generated by the particular portfolio strategy under consideration. It is these returns that are subject to the various measures of performance evaluations.

emphasize three performance measures: the Sharpe ratio, the certainty equivalent (CEQ) return, and a measure of portfolio turnover.⁷ The latter measure is suggestive of the transactions costs associated with a particular strategy. The methodology is then applied to the evaluation of 14 strategies using 7 different sets of underlying assets. They find that the $1/N$ strategy delivers a statistically superior Sharpe ratio across all strategies for all but one of the 7 data sets. It is similarly dominant across the board for all strategies and data sets under CEQ evaluation, and is dead last with regard to the turnover measure. These results are due to the parameter misestimation arising from the limited size of the data sets. Comparing the results of standard mean–variance portfolio construction with the $(1/N)$ strategy based, in particular, on portfolios of 50 US-based stocks, the authors demonstrate that 6000 monthly data entries would be required for mean–variance allocation to be the dominant strategy. Typical estimations in the finance industry use time series for at most 120 quarters. In general DeMiguel et al. (2009) find that the naïve $(1/N)$ strategy dominates all others provided (1) the number of available assets N is large (allowing substantial risk reduction even under naïve diversification) and (2) the constituent assets do not have available sufficiently long term data series for precise parameter estimation.

At the start of this section, we mentioned that naïve diversification (the $1/N$ strategy) is often the strategy of choice for investors. In retirement portfolios, it often assumes the form of equal allocations to several distinct mutual funds. Perhaps this phenomenon can be attributed in part to portfolio managers and financial advisors being aware of the sophisticated analysis characteristic of DeMiguel et al. (2009) and the associated literature.⁸ It may also be due to investors noticing that this strategy on average “does relatively well” for them. DeMiguel et al. (2009) suggest one possible underlying reason for this relative success.

⁷ To be precise, for each strategy k , denote the mean and SD of the sample excess (above r^f) portfolio return series generated as per footnote (8) by, respectively, $\hat{\mu}_k$ and $\hat{\sigma}_k$. The three performance measures are:

- i. the Sharpe ratio, $\hat{\mu}_k/\hat{\sigma}_k$;
- ii. the CE return is defined as the risk free rate r_k^f that an investor is willing to accept rather than adopting a particular risky portfolio strategy k . In a mean–variance utility context it is defined by $r_k^f \equiv \hat{\mu}_k - (1/2)\sigma_k^2$;
- iii. the turnover measure attempts to capture the amount of trading required to implement a particular strategy k . It is defined as the average sum of trades across the N risky assets in the portfolio under a particular strategy implementation: turnover $k = (1/T - M) \sum_{t=1}^{T-M} \sum_{j=1}^N |\hat{w}_{k,j,t+1} - \hat{w}_{k,j,t+}|$ where $\hat{w}_{k,j,t+}$ is the portfolio weight under strategy k , just before rebalancing in $t + 1$ (after period t price changes) and $\hat{w}_{k,j,t+1}$ is the desired portfolio weights for period $t + 1$ before prices change in that period.

⁸ See, for example, Best and Grauer (1991), Bloomfield et al. (1977), Jorion (1991), Michaud (1989), and Plyakha et al. (2012).

7.5 Are Stocks Less Risky for Long Investment Horizons?

Our discussion of the consequences of uncertainty in mean return estimation for the position of the *ex ante* efficient frontier was thus far cast in the familiar one-period setting. As we will see now, these same considerations also have implications for whether stocks are more or less risky for long in contrast to short investment horizons.

7.5.1 Long- and Short-Run Equity Riskiness: Historical Patterns

Are stocks less risky at longer time horizons? How is the question to be addressed? Popular wisdom in this regard is based primarily on [Siegel \(2008\)](#) who argues that, by certain measures, the riskiness of US stocks (we are speaking here of the value-weighted portfolio of all publicly traded US equities) declines for longer investment horizons (30 years), even to the extent of being less risky than US default-free treasury securities on the basis of comparing inflation-adjusted real returns. [Siegel's \(2008\)](#) main assertion is forcefully summarized in [Figure 7.6](#).

[Figure 7.6](#) chooses to measure relative riskiness by focusing on worst- and best-case scenarios. By these measures, stocks have been less risky than bonds at long investment horizons: at 30-year horizons, for example, the worst annualized real return on the US stock market averaged 2.5%, while for long-term bonds and T-bills, the average was, respectively, -2.0% and -1.0%. The dramatic decline, for increasing time horizons, in the range of annualized returns for all three security types suggests strong mean reversion in the return patterns of each.⁹ Using standard measures, [Campbell and Viceira \(2002, 2005\)](#) also report conditional variance estimates that decline with the investment horizon.

Given stationarity, ergodicity, mean reversion, and any other statistical property which may enhance the power of historical return data to characterize future return distributions, the evidence presented in [Figure 7.6](#) might be sufficient to sway investors in favor of all stock portfolios, but only for those who can commit to long investment horizons of 20 years or more. Consider a 10-year horizon; the worst stocks did was to lose 4.1% compounded for 10 years which was less than for bonds and T-bills.¹⁰ Suppose this same investor was required, however, to liquidate after only 3 years. Could the investor possibly have lost 38% (the maximum 1-year loss) in each of these periods? In this case, the investor would have been wiped out. The point being made is simply to say that knowing the best and worst average scenarios over long time periods tells us little about the investor's wealth

⁹ Mean reversion is the idea that periods of high security returns will be followed by low return periods and vice versa. Formally, a rate of return series (r_t) is said to be mean-reverting if and only if for any integers $0 \leq s < t < v < u$, $\text{cov}(r_t - r_s, r_v - r_u) < 0$. (This is not the only definition: see the Web notes to the present chapter.)

¹⁰ A 4.1% annual loss for 10 years is actually quite bad; an investor's capital would have been depleted by roughly one half under that scenario.

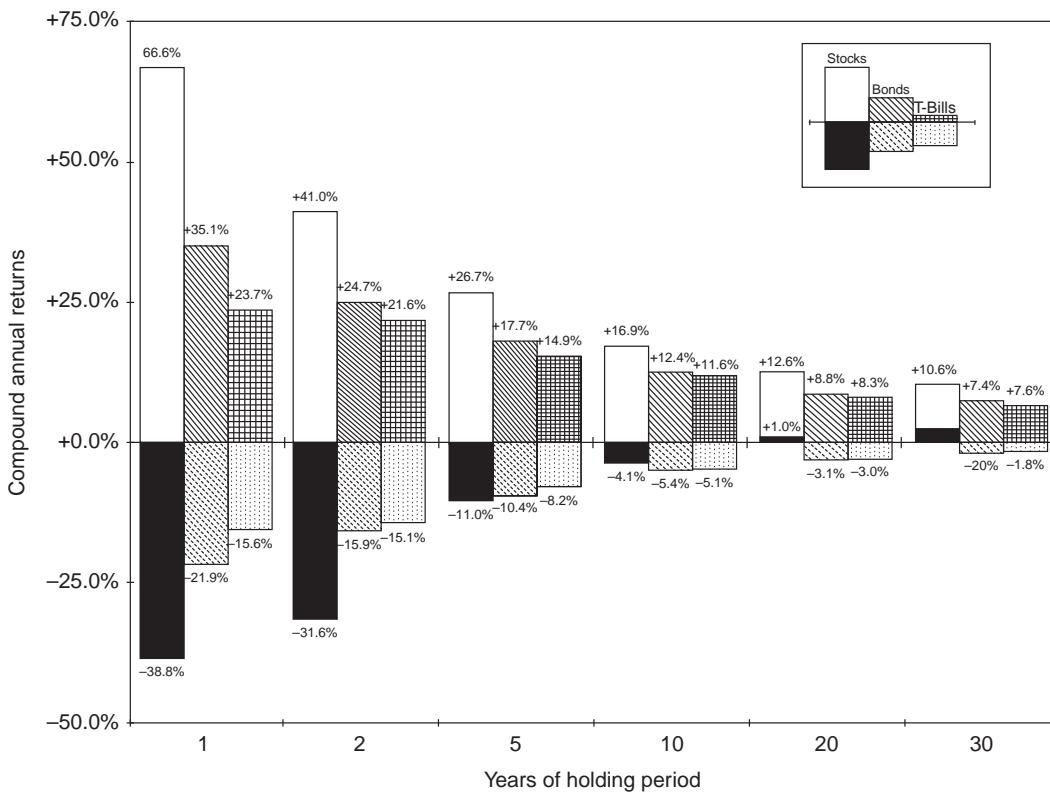


Figure 7.6: Maximum and minimum real holding period returns (1802–1997)⁽ⁱ⁾

⁽ⁱ⁾The maximum and minimum holding period returns, annualized, as reported in the figure, reflect the maximum and minimum returns over all contiguous time horizons of the indicated length during the 1802–1997 period. For example, the minimum (annualized) 30-year compounded equity return during 1802–1997 period was 2.6%. Note that the sample period does not include the Internet Bubble and subsequent collapses of the late 1990s, early 2000s, or the period of the Great Recession. Low negative returns on T-bills signifies periods of high inflation. Source: [Siegel \(2008\)](#), Figure 2.1.

volatility within the investment period, something that could be of vital significance in the event of an unexpected health emergency, job loss, etc. In particular, knowledge of the variance of returns would provide information on intermediate wealth volatility, something that is absent from a max–min discussion.¹¹

Accordingly, it would be useful to explore “[Siegel \(2008\)](#)-like” results in a model which explicitly accounts for means and variances in an *ex ante* setting. After all, we are

¹¹ For the United States, the σ of annualized real returns is generally accepted to be around 17%. A 1-year positive return of 66.6% (see [Figure 7.6](#) for a 1-year horizon) is more than three standard deviations from the mean, surely a tail event.

concerned about predictive returns—what future returns will be. Historical returns are useful principally as a guide to estimating moments. We thus need a model of equity return evolution with which to guide our thinking. The random walk model is the natural context in which to first explore these questions, and we introduce this model in [Section 7.5.2](#). In [Section 7.5.3](#), we then modify the random walk model and then revisit the original question: Are stocks more or less risky at longer time horizons?

7.5.2 Intertemporal Stock Return Behavior Through Time: The Random Walk Model

While our analysis so far has been quite detailed as regards to asset return distributions within a period (normality), we still have no working hypothesis as to how asset returns evolve intertemporally. Such a framework is especially important for long-term investors.

The benchmark model of equity return evolution is the random walk: returns are i.i.d. normally distributed across time periods.¹² It may seem surprising to start off with a statement concerning stock returns. We do this because stock returns can reasonably be thought of as stationary stochastic processes. Stock prices are not stationary because there is no upper bound to the values they may achieve and this unboundedness reflects nonstationary behavior. Statistical tests repeatedly confirm, however, that *stock returns* do represent stationary time series, and it is returns that are of interest to investors.¹³ Let us agree to measure returns and return statistics (μ, σ) on an annual basis, and let Δt represent an interval of time measured in years. The random walk hypothesis on a stock's rate of return over the time interval $[t, t + \Delta t]$ asserts that

$$\tilde{r}_{t,t+\Delta t} \sim N(\mu\Delta t, \sigma^2\Delta t) \quad (7.1)$$

where μ and σ^2 denote, respectively, the stock's known annualized mean return and the variance of its annual rate of return.¹⁴ It is frequently convenient to express [Eq. \(7.1\)](#) equivalently as

$$\tilde{r}_{t,t+\Delta t} = \mu\Delta t + \sigma\sqrt{\Delta t}\tilde{\varepsilon}_t \quad (7.2)$$

¹² For equity securities, the random walk model is an expression of the notion of “market efficiency.” The idea is this: by definition, new information flows to stock market participants in a statistically independent fashion (no predictability). If traders act on this new information to alter their positions immediately, the resulting price and return variation should be statistically independent as well.

¹³ The rate of return on wealth essentially captures its growth rate. By analogy to a stock's price, an economy's GDP is potentially without bound and nonstationary. The growth rate of an economy's GDP, however, does represent a stationary time series.

¹⁴ Expression [\(7.1\)](#) means that 1-year stock returns are distributed $N(\mu, \sigma^2)$; for 6-month intervals it is $N(.5\mu, .5\sigma^2)$; for monthly time intervals it is $N(1/12\mu, 1/12\sigma^2)$. The mean and variance are simply scaled up or down by the length of the time intervals.

where $\tilde{\varepsilon}_t \sim N(0, 1)$. To see this equivalence, note that it follows from Eq. (7.2) that

$$\begin{aligned} E\tilde{r}_{t,t+\Delta t} &= \mu\Delta t, \text{ and} \\ \sigma_{\tilde{r}_{t,t+\Delta t}}^2 &= \sigma^2\Delta t, \text{ as proposed in Eq. (7.1)} \end{aligned}$$

The benchmark random walk model (i.i.d. normal returns cum constant μ and σ) tends to be a very good model of security return evolution when the length of the time periods is short (e.g., a day where $\Delta t = (1/250)$, reflecting the fact of there being roughly 250 trading days per year), and the cumulative time horizon is less than 6 months. At longer horizons, return persistence may creep in: changes in μ and σ must be modeled systematically. In the case of μ , for example, it is reasonable to propose

$$\tilde{\mu}_{t+\Delta t} = (1 - \delta)\mu + \delta\mu_t + \tilde{\varepsilon}_{t+\Delta t}^\mu \quad (7.3)$$

where μ is the long-run mean (so that μ_t , the period t mean return, fluctuates about μ) and $\text{cov}(\tilde{\varepsilon}_t, \tilde{\varepsilon}_t^\mu) = 0$.¹⁵

While Eq. (7.2) describes the evolution of the period by period rate of return, we have to do a bit more work to understand how the asset's corresponding price evolves.¹⁶ If q_t^e is the price of a stock in period t , then $\tilde{q}_{t,t+\Delta t}^e = q_t^e(1 + \mu\Delta t + \sigma\sqrt{\Delta t}\tilde{\varepsilon}_t)$, assuming no dividend payments. In a world of continuous compounding

$$\begin{aligned} \ln\left(\frac{\tilde{q}_{t+\Delta t}^e}{q_t^e}\right) &\sim N(\hat{\mu}\Delta t, \hat{\sigma}^2\Delta t), \text{ or} \\ \ln \tilde{q}_{t+\Delta t}^e - \ln q_t^e &\sim N(\hat{\mu}\Delta t, \hat{\sigma}^2\Delta t) \text{ or} \\ \ln \tilde{q}_{t+\Delta t}^e &= \ln q_t^e + \hat{\mu}\Delta t + \hat{\sigma}\sqrt{\Delta t}\tilde{\varepsilon}_t, \text{ and} \end{aligned} \quad (7.5a)$$

$$\tilde{q}_{t+\Delta t}^e = q_t^e e^{\hat{\mu}\Delta t + \hat{\sigma}\sqrt{\Delta t}\tilde{\varepsilon}_t} \quad (7.5b)$$

¹⁵ Heston (1993) has also proposed a model that allows for the stochastic evolution of the return variance:

$$\tilde{\sigma}_{t+\Delta t}^2 = \sigma_t^2 + \kappa(\sigma^2 - \sigma_t^2)\Delta t + \sigma_v^2 \sqrt{\sigma_t^2} \sqrt{\Delta t} \tilde{\varepsilon}_{t+\Delta t}^\sigma \quad (7.4)$$

where σ^2 is the long run variance, σ_t^2 the period t variance, κ a positive constant capturing the speed of convergence of σ_t^2 back to σ^2 , σ_v^2 the volatility of the variance, and $\text{cov}(\tilde{\varepsilon}_t, \tilde{\varepsilon}_t^\sigma) = \rho$, which may differ from zero. The presence of the $\sqrt{\sigma_t^2}$ term guarantees that the variance remains positive as $\Delta t \rightarrow 0$. For the moment we will set aside generalizations (7.3) and (7.4) and focus on the benchmark formulation (7.2). Note that either generalization (7.3) or (7.4) introduces persistence into the model, and, in the case of Eq. (7.3), mean reversion.

¹⁶ The discussion above applies to a portfolio of assets as well.

with $\tilde{\varepsilon}_t \sim N(0, 1)$ and for any time interval Δt . We then say that the price of the asset is lognormally distributed because the \ln of its price at time $t + \Delta t$ is normally distributed (recall Section 6.3). It follows that

- i. $E\tilde{q}_{t,t+\Delta t}^e = q_t^e e^{(\hat{\mu} + \frac{\hat{\sigma}^2}{2})\Delta t}$ (7.6)

- ii. $\text{var}\tilde{q}_{t,t+\Delta t}^e = (q_t^e)^2 e^{(2\hat{\mu} + \hat{\sigma}^2)\Delta t} (e^{\hat{\sigma}^2\Delta t} - 1)$ (7.7)

Note that Eqs. (7.6) and (7.7) are the exact counterparts of the expressions in Section 6.6. This is no surprise: the one-period model of Section 6.6 is a special case of what we are considering here (Figure 7.7).

(Expressions (7.5a, b), (7.6), and (7.7) are really very general since we may substitute aggregate investor wealth Y_t invested in a portfolio p for q_t^e , let $t = 0$, $\hat{\mu} = \hat{\mu}_P$, $\hat{\sigma} = \hat{\sigma}_P$, and select Δt equal to our investment time horizon.)

When a stock's price evolves according to Eqs. (7.5a, b), the result is a familiar sawtooth pattern. In Figures 7.8 and 7.9, we present two potential price paths (there are an uncountably infinite number of possible paths) for the case where $q_t^e = \$120$, $\hat{\mu} = 0.20$, $\hat{\sigma} = 0.30$ and $\Delta t = (1/250)$. These price paths are generated by feeding into Eq. (7.5b) a sequence of independent draws from $N(0, 1)$ and updating the price accordingly.

If the reader sees a pattern in these figures, we gently remind him that he is mistaken: they are the result of purely i.i.d. draws from $N(0, 1)$.

With this perspective in mind, we return to the question of whether stocks—more precisely, very well diversified stock portfolios such as the S&P₅₀₀ portfolio—are less risky at longer time horizons. If this were to be the case, what additional assumption on the return processes, if any, would have to be true?

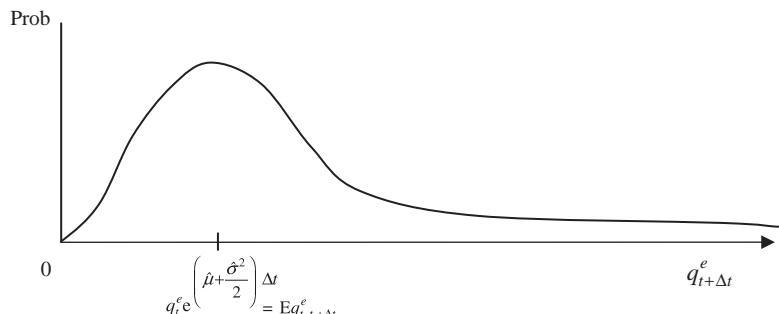
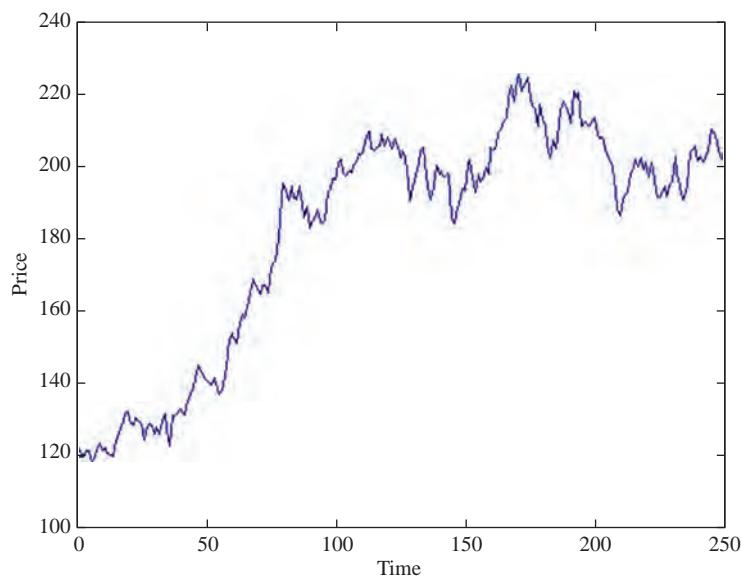
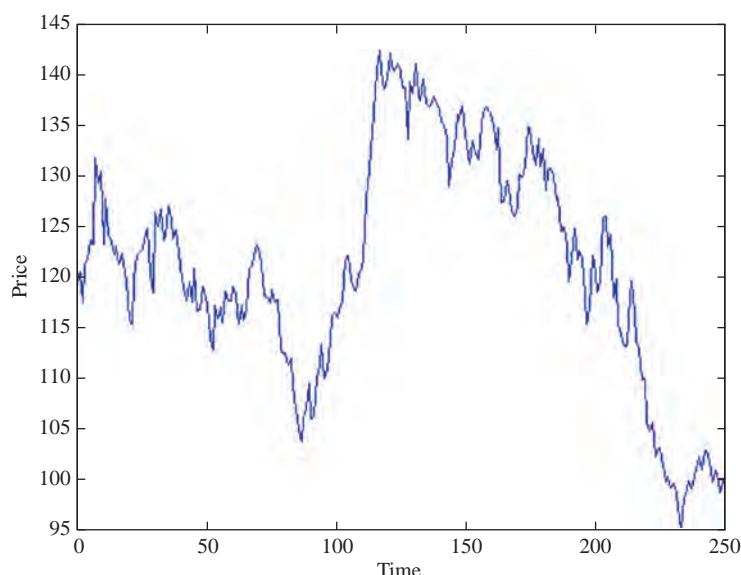


Figure 7.7
Asset price probability density function.

**Figure 7.8**

Sample price path of 250 periods; $q_t^e = \$120$, $\hat{\mu} = 0.20$, $\hat{\sigma} = 0.30$. Source: authors' simulations.

**Figure 7.9**

Sample price path of 250 periods; $P_0 = \$120$, $\hat{\mu} = 0.20$, $\hat{\sigma} = 0.30$. Source: authors' simulations.

7.5.3 Are Stocks Less Risky in the Long Run? A Predictive Perspective¹⁷

This is a substantive question from the perspective of an investor looking forward; let us first see what the random walk model has to say. To simplify notation, let $\Delta t = 1$ year. Equation (7.2) then reduces to

$$\tilde{r}_{t+1} = \mu + \sigma^2 \tilde{\varepsilon}_t$$

Investors are interested to know the variance of

$$\tilde{r}_{t,t+J} = \tilde{r}_{t+1} + \tilde{r}_{t+2} + \cdots + \tilde{r}_{t+J}$$

where J denotes their investment horizon. We refer to it as the predictive variance.

The question at hand is how the predictive variance changes with J . Assuming μ and σ^2 are precisely estimated, by the i.i.d. assumption underlying a random walk,

$$\text{var}(\tilde{r}_{t,t+J}) = J\sigma^2 \text{ or}$$

$(1/J)\text{var}(\tilde{r}_{t,t+J}) = \sigma^2$: the annualized risk is the same irrespective of the investment horizon. In keeping with the difficulty in estimating means precisely (recall Box 7.1), however, let us revisit this calculation under circumstances where σ^2 is known but μ is not. Let I_t denote all information pertinent to the estimation of μ . In this case

$$\begin{aligned} \text{var}(\tilde{r}_{t,t+J}) &= J\sigma^2 + \text{var}(J\mu|I_t) \\ &= J\sigma^2 + J^2 \text{var}(\mu|I_t) \text{ and} \\ \frac{1}{J} \text{var}(\tilde{r}_{t,t+J}) &= \sigma^2 + J \text{var}(\mu|I_t) \mapsto \infty \text{ as } J \mapsto \infty \end{aligned} \tag{7.8}$$

An investor who holds to the random walk model but who is uncertain as to the true mean return should thus generally view stocks as increasingly risky in the long run. This is a very different perspective than Siegel's (2008). If the uncertainty surrounding μ is measured conventionally by the standard error of the estimate of the mean, (σ/\sqrt{T}) (see Box 7.1), then expression (7.8) reduces to

$$\frac{1}{J} \text{var}(\tilde{r}_{t,t+J}) = \sigma^2 \left(1 + \frac{J}{T} \right)$$

Pastor and Stambaugh (2012) compute this quantity for $T = 206$ (the same historical data series as Siegel (2008)) and $J = 50$ (a 50-year time horizon looking forward); in this case:

$$\frac{1}{J} \text{var}(\tilde{r}_{t,t+J}|I_t) = \sigma^2 \left(1 + \frac{50}{206} \right) = 1.243\sigma^2$$

¹⁷ The remarks in this section are taken directly from Pastor and Stambaugh (2012).

the predictive return variance grows about 25% at long relative to short (one-year) horizons. Note that if the data set is much shorter, $T = 40$ years, for example, the long-run predictive variance is more than twice σ^2 .

Other sources of uncertainty can enter the long-/short-run variance comparison besides one-period return volatility. Suppose a security's return evolution conforms to the following system:

$$\tilde{r}_{t+1} = \mu_t + \tilde{\varepsilon}_{t+1} \quad (7.9a)$$

$$\tilde{\mu}_{t+1} = (1 - \beta)E\tilde{r} + \beta\mu_t + \tilde{w}_{t+1} \quad (7.9b)$$

where $\beta < 1$, $\rho_{\varepsilon,w} < 0$ and $\tilde{\varepsilon}_t, \tilde{w}_t$ are mean zero random components. Under specification (7.9a, b), there are five sources of potential uncertainty: (i) uncertainty in future expected returns as per Eq. (7.9b), (ii) i.i.d. uncertainty in $\tilde{\varepsilon}_t$, (iii) uncertainty in the current expected return ($E\tilde{r}$), and (iv) potential parameter uncertainty as regards $(\sigma_\varepsilon, \sigma_w, \rho_{\varepsilon,w}, \beta, E\tilde{r})$ (referred to as estimation risk), and the presence of (v) mean reversion which, in this model, is captured by $\rho_{\varepsilon,w} < 0$.¹⁸ In and of itself, mean reversion leads to a reduction in long-run uncertainty relative to the i.i.d. case. Under this specification, whether stocks are less or more risky in the long-run boils down to whether mean reversion overwhelms or is overwhelmed by the consequences of uncertainty in the other elements.

To gain an understanding of how these five sources of uncertainty interact to determine the predictive variance, [Pastor and Stambaugh \(2012\)](#) estimate the entire system (equations 7.9a,b and associated parameters) using annual USA return data for the period 1802-2007, the same data set as used by [Siegel \(2008\)](#). While we leave the myriad details of this estimation to the interested reader, the relative contributions of the different sources of uncertainty, and how they change as the investor's time horizon increases, are easily seen in [Figure 7.10](#). Note that the vertical axis measures annualized variance and that Panel A represents the vertical sum of the components in Panel B.

As expected, it is only the mean reversion in returns ($\rho_{\tilde{\varepsilon},\tilde{w}} < 0$) that tends to lower the predictive variance as the investor's time horizon increases. While its effect is substantial, it is overwhelmed by the increases in the other sources of uncertainty. In particular, the contribution of future mean uncertainty $\{\tilde{\mu}_{t+j}\}$ increases quite robustly and soon becomes dominant; the contribution from uncertainty as regards the current mean (μ_t) and the model parameters ($\sigma_{\tilde{w}}, \sigma_{\tilde{\varepsilon}}$, etc.) increase with the time horizon as well. By construction, the i.i.d. component's contribution ($\tilde{\varepsilon}_t$) is constant on an annualized basis.

What do we learn from the [Pastor and Stambaugh \(2012\)](#) exercise as regards our original question: "Are stocks more risky in the long run?" The answer appears to be model specific

¹⁸ [Pastor and Stambaugh \(2012\)](#) analyze this model in considerable detail, to which we refer the interested reader.

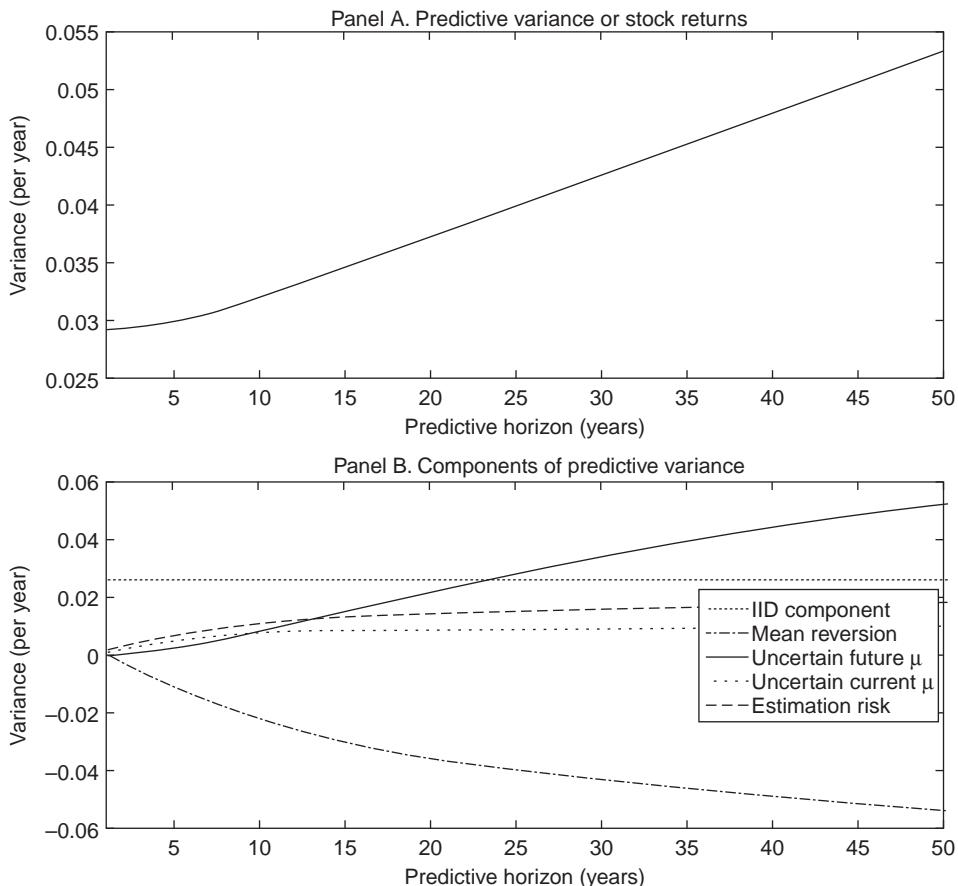


Figure 7.10: Predictive variance of multiperiod return and its components.

Panel A plots the variance of the predictive distribution of long-horizon returns, $\text{var}(r_{t,t+j} | I_t)$. Panel B plots the five components of the predictive variance. All quantities are divided by k , the number of periods in the return horizon. The results are obtained by estimating the predictive system on annual real U.S. stock market returns in 1802 to 2007. Source: Figure 6 in [Pastor and Stambaugh \(2012\)](#).

and largely depends on what the investor “knows — or believes he knows.” In particular, uncertainty as regards the initial period’s mean return and uncertainty regarding future expected returns can conspire such that long term investors in practice face substantially more return volatility than short-horizon investors.

7.6 Conclusions

First, it is important to keep in mind that everything said so far in this chapter applies regardless if the probability distributions (possibly normal) on returns represent the *subjective* expectations of the particular investor upon whom we are focusing or “objective” market forecasts.

Second, although initially conceived in the context of descriptive economic theories, the success of portfolio theory arose primarily from the possibility of giving it a normative interpretation, i.e., of seeing the theory as providing a guide on how to proceed to identify a potential investor's optimal portfolio. In particular, it points to the information requirements to be fulfilled (ideally). Even if we accept the formal restrictions implied by mean–variance analysis, one cannot identify an optimal portfolio without spelling out expectations on mean returns, standard deviations of returns, and correlations among returns. As this chapter has demonstrated, this is no easy task empirically. Investors may simply fall back on equally weighted portfolios. There is much to commend this alternative. One can view the role of the financial analyst as providing plausible figures for the relevant statistics or offering alternative scenarios for consideration to the would-be investor. This is the first absolutely critical step in the search for an optimal portfolio and one that may have statistical as well as intuitive aspects. The numbers proposed are forecasts and they may not be very precise ones at that. The computation of the (subjective) efficient frontier is the second step, and it essentially involves solving the quadratic programming problem (QP) possibly in conjunction with constraints specific to the investor. The third and final step consists of defining, at a more or less formal level, the investor's risk tolerance and, on that basis, identifying his optimal allocation to risk-free versus risky assets.

Market equilibrium considerations are next.

References

- Benartzi, S., Thaler, R., 2001. Naive diversification strategies in defined contribution savings plans. *Am. Econ. Rev.* 91, 79–98.
- Best, M., Grauer, R.R., 1991. On the sensitivity of mean–variance–efficient portfolios to changes in asset means: some analytical and computational results. *Rev. Finan. Stud.* 4, 315–342.
- Bloomfield, T., Leftwich, R., Long, J., 1977. Portfolio strategies and performance. *J. Finan. Econ.* 5, 201–218.
- Campbell, J., Viceira, L., 2002. Strategic asset allocation: Portfolio choice for long term investors. Oxford Univ. Press, Oxford, UK.
- Campbell, J., Viceira, L., 2005. The term structure of the risk-return tradeoff. *Finan. Analysts J.* 61, 34–44.
- Chen, N., Roll, R., Ross, S., 1986. Economic forces and the stock market. *J. Bus.* 59, 383–403.
- DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naïve diversification: how inefficient is the (1/N) portfolio strategy? *Rev. Econ. Stud.* 22, 1915–1953.
- Heston, S., 1993. A closed form solution for options with stochastic volatility. *Rev. Finan. Stud.* 6, 327–343.
- Huberman, G., Jiang, W., 2006. Offering vs. choice in 401(k) plans: equity exposure and the number of funds. *J. Finan.* 61, 763–801.
- Jorion, P., 1991. Bayesian and CAPM estimates of the means: implications for portfolio selection. *J. Bank. Finan.* 15, 717–727.
- Longin, F., Solnik, B., 2001. Extreme correlation of international equity markets. *J. Finan.* 56, 649–676.
- Michaud, R., 1989. The Markowitz optimization enigma: is “optimized” optimal? *Finan. Analysts J.* 45, 31–42.
- Milesa-Ferretti, G., Tille, C., 2010. The great retrenchment: international capital flows during the global financial crisis. CEPR, Economic Policy Working Paper.
- Pastor, L., Stambaugh, R., 2012. Are stocks really less volatile in the long run? *J. Finan.* 67, 431–477.

- Plyakha, Y., Uppal, R., Vilkov, G., 2012. Why does an equal-weighted portfolio outperform value and price-weighted portfolios? Working Paper, EDHEC Business School.
- Riberio, R., Veronesi, P., 2002. The excess co-movement of international stock markets in bad times: a rational expectations equilibrium model, mimeo, University of Chicago, Booth School.
- Siegel, J., 2008. Stocks for the Long Run. fifth ed. McGraw Hill, New York, NY.

Appendix 7.1

We can also use the data in [Table 7.2](#) to get an idea of the enormous impact that short selling constraints (see Section 6.4) can have on portfolio composition and the associated portfolio risk and return when the techniques of MPT are applied. Let us suppose that the entries in [Table 7.2](#) represent ex ante estimates in 1.1.1996. By comparing [Tables A.7.1 and A.7.2](#) (both based on [Table 7.2](#)), we see that for all cases in which four or more securities (national market portfolios) are eligible for portfolio admission, the lifting of the “no-short-sales” constraint leads to leverage levels exceeding 80% of the initial investment. Mean monthly returns in these cases are roughly 0.15% higher (1.8% annualized). In all cases, the portfolio is configured to have a monthly standard deviation of 5%.

Table A.7.1: Portfolio proportions when $\sigma_p = 5\%$ and short sales are not allowed

Short-Sale Constraints?	No	No	No	No	No
Horizon	120	120	120	120	120
France	122.28%	99.24%	89.59%	79.41%	79.79%
Germany	– 54.85%	– 56.23%	– 67.67%	– 71.09%	– 71.43%
Japan		– 31.24%	– 31.25%	– 21.86%	– 21.22%
UK	32.57%	88.23%	15.69%	20.93%	21.05%
US			93.64%	78.10%	79.06%
Brazil				14.51%	15.29%
Korea					– 2.55%
Mean monthly return	0.62%	0.70%	0.85%	0.89%	0.89%
SD of monthly rate	5.00%	5.00%	5.00%	5.00%	5.00%

Table A.7.2: Portfolio proportions when $\sigma_p = 5\%$ and short sales are permitted.

Short-Sale Constraints?	Yes	Yes	Yes	Yes	Yes
Horizon	120	120	120	120	120
France	62.99%	62.99%	59.09%	17.04%	17.04%
Germany	0.00%	0.00%	0.00%	0.00%	0.00%
Japan	0.00%	0.00%	0.00%	0.00%	0.00%
UK	37.01%	37.01%	0.00%	0.00%	0.00%
US			40.91%	68.82%	68.82%
Brazil				14.14%	14.14%
Korea					0.00%
Mean monthly return	0.57%	0.57%	0.65%	0.74%	0.74%
SD of monthly rate	5.00%	5.00%	5.00%	5.00%	5.00%

The Capital Asset Pricing Model

Chapter Outline

8.1 Introduction	209
8.2 The Traditional Approach to the CAPM	210
8.3 Valuing Risky Cash Flows with the CAPM	214
8.4 The Mathematics of the Portfolio Frontier: Many Risky Assets and No Risk-Free Asset	217
8.5 Characterizing Efficient Portfolios (No Risk-Free Assets)	222
8.6 Background for Deriving the Zero-Beta CAPM: Notion of a Zero-Covariance Portfolio	224
8.7 The Zero-Beta CAPM	227
8.8 The Standard CAPM	229
8.9 An Empirical Assessment of the CAPM	231
8.9.1 Fama and MacBeth (1973)	232
8.9.2 Banz (1981) and the “Size Effect”	234
8.9.3 Fama and French (1992)	234
8.9.4 Volatility Anomalies	235
8.10 Conclusions	239
References	240
Appendix 8.1: Proof of the CAPM Relationship	241
Appendix 8.2: The Mathematics of the Portfolio Frontier: An Example	242
Appendix 8.3: Diagrammatic Representation of the Fama—MacBeth Two-Step Procedure	245

8.1 Introduction

The capital asset pricing model (CAPM) is an equilibrium theory built on the foundation of modern portfolio theory. It is, however, an equilibrium theory with a somewhat peculiar structure. This is true for a number of reasons:

1. First, the CAPM is a theory of financial equilibrium only. Investors take the various statistical quantities—means, variances, covariances—that characterize a security’s return process as given. There is no attempt within the theory to link the return processes with events in the *real* side of the economy. In future model contexts, we shall generalize this feature (Chapter 11).

2. Second, as a theory of financial equilibrium, it makes the assumption that the supply of existing assets is equal to the demand for existing assets and, as such, that the currently observed asset prices are equilibrium ones. There is no attempt, however, to compute asset supply and demand functions explicitly. Only the equilibrium price vector is characterized. Let us elaborate on this point.

Under the CAPM, portfolio theory informs us about the demand side. If individual i invests a fraction w_{ij} of his initial wealth Y_{0i} in asset j , the value of his asset j holding is $w_{ij}Y_{0i}$. Absent any information that he wishes to alter these holdings, we may interpret the quantity $w_{ij}Y_{0i}$ as his demand for asset j at the prevailing price vector. If there are I individuals in the economy, the total value of all holdings of asset j is $\sum_i^I w_{ij}Y_{0i}$; by the same remark we may interpret this quantity as aggregate demand. At equilibrium, one must have $\sum_i^I w_{ij}Y_{0i} = P_j Q_j$, where p_j is the prevailing equilibrium price per share of asset j , Q_j is the total number of shares outstanding and, consequently, $p_j Q_j$ is the market capitalization of asset j . The CAPM derives its implications for prices by assuming that the actual economy-wide asset holdings are investors' aggregate optimal asset holdings.

3. Third, the CAPM expresses equilibrium in terms of relationships between the return distributions of individual assets and the return characteristics of the portfolio of all assets. We may view the CAPM as informing us, via modern portfolio theory, as to what asset return interrelationships must be in order for equilibrium asset prices to coincide with the observed asset prices.

In what follows we first present an overview of the traditional approach to the CAPM. This is followed by a more general presentation that permits at once a more complete and more general characterization. In both presentations we depart from earlier notation and denote a security's mean return not by μ , but rather by $E(\tilde{r})$. This notation is more typical of the empirical literature that we review at the end of the chapter.

8.2 The Traditional Approach to the CAPM

To get useful results in this complex world of many assets, we have to make simplifying assumptions. The CAPM approach essentially hypothesizes (1) that all agents have the *same beliefs* about future returns (i.e., homogeneous expectations), and, in its simplest form, (2) that there is a risk-free asset, paying a safe return r_f at which investors can borrow or lend as much as they wish. These assumptions guarantee (Chapter 6) that the mean-variance efficient frontier is the same for every investor, and furthermore, by the separation theorem, that all investors' optimal portfolios have an identical structure: a fraction of initial wealth is invested in the risk-free asset, the rest in the tangency portfolio (two-fund separation). It is then possible to derive a few key characteristics of equilibrium asset and portfolio returns without detailing the underlying equilibrium structure, i.e., the demand for and supply of assets, or discussing their prices.

Because all investors acquire shares in the same risky tangency portfolio T and make no other risky investments, all existing risky assets must belong to T by the definition of an equilibrium. Indeed, if some asset k were not found in T , there would be no demand for it; yet, it is assumed to exist in positive supply. Supply would then exceed demand, which is inconsistent with assumed financial market equilibrium. The same reasoning implies that the share of any asset j in portfolio T must correspond to the ratio of the market value of that asset $p_j Q_j$ to the market value of all assets $\sum_{j=1}^J p_j Q_j$. This, in turn, guarantees that tangency portfolio T must be nothing other than the market portfolio M , the portfolio of all existing assets where each asset appears in a proportion equal to the ratio of its market value to the total market capitalization.

This simple reasoning leads to a number of useful conclusions:

1. The market portfolio is efficient because it is on the efficient frontier.
2. All individual optimal portfolios are located on the line originating at point $(0, r_f)$ and going through $(E\tilde{r}_M, \sigma_M)$, which is also the locus of all efficient portfolios (Figure 8.1). This locus is usually called the **capital market line** or CML.
3. The slope of the CML is $(E\tilde{r}_M - r_f)/\sigma_M$. It tells us that an investor considering a marginally riskier efficient portfolio would obtain, in exchange, an increase in expected return of $(E\tilde{r}_M - r_f)/\sigma_M$. This is the price of, or reward for, risk taking—the price of risk as applicable to efficient portfolios. In other words, for efficient portfolios, we have the simple linear relationship in Eq. (8.1).

$$E\tilde{r}_p = r_f + \frac{E\tilde{r}_M - r_f}{\sigma_M} \sigma_p \quad (8.1)$$

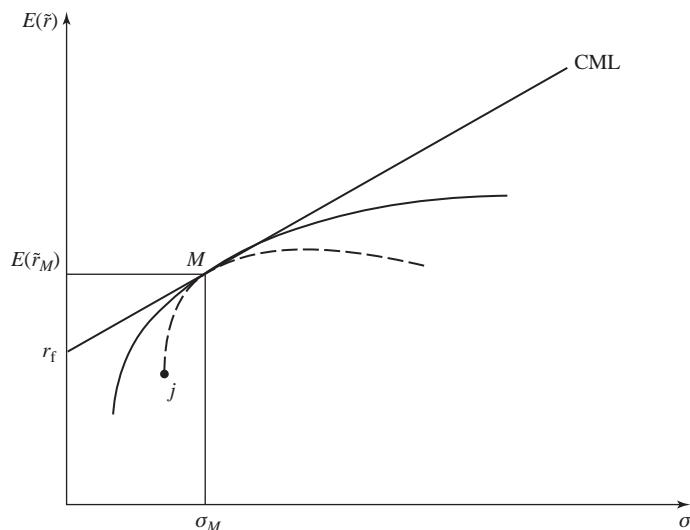


Figure 8.1
The CML.

The CML applies only to efficient portfolios. What can be said of an arbitrary asset j that does not belong to the efficient frontier? To discuss this essential part of the CAPM, we first rely on Eq. (8.2), formally derived in Appendix 8.1, and limit our discussion to its intuitive implications:

$$E\tilde{r}_j = r_f + (E\tilde{r}_M - r_f) \frac{\sigma_{jM}}{\sigma_M^2} \quad (8.2)$$

Let us define $\beta_j = \sigma_{jM}/\sigma_M^2$, i.e., the ratio of the covariance between the returns on asset j and the returns on the market portfolio divided by the variance of the market returns.

We can thus rewrite Eq. (8.2) as Eq. (8.3).

$$E\tilde{r}_j = r_f + \left(\frac{E\tilde{r}_M - r_f}{\sigma_M} \right) \beta_j \sigma_M = r_f + \left(\frac{E\tilde{r}_M - r_f}{\sigma_M} \right) \rho_{jM} \sigma_j \quad (8.3)$$

Comparing Eqs. (8.1) and (8.3), we obtain one of the major lessons of the CAPM: only the fraction ρ_{jM} of the total risk of an asset j , σ_j , is remunerated by the market. The remaining fraction $(1 - \rho_{jM})\sigma_j$ is not: it is “diversified away” when asset j is placed in the market portfolio. By “diversified away,” we mean that some of j ’s return variation is offset or canceled out by variation in the returns to other assets in the market portfolio.

To see this intuitively, let us first be reminded that under the CAPM, every investor holds only the market portfolio ($T = M$), and thus the relevant risk for an investor can only be the market’s standard deviation σ_M . As a consequence, what is important to the investor is the risk contribution of asset j to the risk of the market portfolio, i.e., the extent to which the inclusion of asset j into the overall portfolio M increases the latter’s standard deviation. This marginal contribution of j to the overall portfolio risk σ_M is appropriately measured by $\rho_{jM}\sigma_j$ ($= \beta_j\sigma_M$) as the following equivalence demonstrates:

$$\begin{aligned} \sigma_M^2 &= \sum_{j=1}^J w_j \text{cov}(\tilde{r}_j, \tilde{r}_M) = \sum_{j=1}^J w_j \rho_{jM} \sigma_j \sigma_M, \text{ and thus} \\ \sigma_M &= \sqrt{\sum_{j=1}^J w_j (\rho_{jM} \sigma_M)} \end{aligned} \quad (8.4)$$

Accordingly, the risk premium on a given asset j is the market price of risk, $(E\tilde{r}_M - r_f)/\sigma_M$, multiplied by the measure of the quantity of j ’s relevant risk: $\rho_{jM}\sigma_j$ (or $\beta_j\sigma_M$) $\leq \sigma_j$. As such, $\rho_{jM}\sigma_j$ measures the **systematic risk** of asset j , systematic in the sense that it is the portion of j ’s risk contributing to variation in the market portfolio’s return.¹ While the β_j of asset j

¹ Alternatively, and perhaps more informatively, systematic risk is also referred to as the “undiversifiable risk” or “market risk” of asset j .

is more typically referred to as asset j 's systematic risk measure (we will also use this language), it more precisely measures ($\beta_j = \sigma_{jM}/\sigma_M^2 = (\rho_{jM}\sigma_j)/\sigma_M$) the systematic risk of asset j *relative* to the systematic risk of the overall market (loosely, the (value-weighted) average systematic risk of the assets in M).

A comparison of [Eqs. \(8.1\) and \(8.3\)](#) reaffirms that an efficient portfolio is one for which all diversifiable risks have been eliminated, i.e., where $\rho_{jM} = 1$. For an efficient portfolio, total risk and systematic risk are one and the same: within the available universe of assets, there are no further opportunities for diversification.

These ideas can also be developed by writing, without loss of generality, the excess return on asset j as a linear function of the excess return on the market portfolio plus a random term that is independent of the market return:

$$\tilde{r}_j - r_f = \alpha_j + \beta_j(\tilde{r}_M - r_f) + \tilde{\varepsilon}_j \quad (8.5)$$

If we go one step further and regard [Eq. \(8.5\)](#) as an OLS (ordinary least squares) regression equation, we obtain the standard regression estimate of the coefficient on the market, $\hat{\beta}_j$, where

$$\hat{\beta}_j = \frac{\hat{\sigma}_{jM}}{\hat{\sigma}_M^2}$$

an identification identical in form to our theoretical beta, thus accounting for the “beta label.” Going on two steps further, the standard OLS variance decomposition yields

$$\hat{\sigma}_j^2 = \hat{\beta}_j^2 \hat{\sigma}_M^2 + \hat{\sigma}_{\varepsilon_j}^2 \quad (8.6)$$

Since $\hat{\beta}_j^2 \hat{\sigma}_M^2 = (\hat{\rho}_{jM} \hat{\sigma}_j)^2$ we may, accordingly, identify the first term in the variance decomposition as the systematic risk of j and identify the second as its diversifiable risk (relative to M). It is this latter term that disappears when j is placed in M .

Finally, [Eq. \(8.3\)](#) can equivalently be rewritten as

$$E\tilde{r}_j - r_f = (E\tilde{r}_M - r_f)\beta_j \quad (8.7)$$

which says that the expected excess return or the risk premium on an asset j is proportional to its β_j . [Equation \(8.7\)](#) defines the **security market line** or SML. It is depicted in [Figure 8.2](#).

We conclude the overview of the CAPM by drawing attention to the fact that [Eq. \(8.5\)](#), when viewed as an OLS regression, may also be interpreted as a single factor model of security return generation. As such the excess return on the market portfolio constitutes the single explanatory factor with the “CAPM beta” serving as its factor sensitivity.

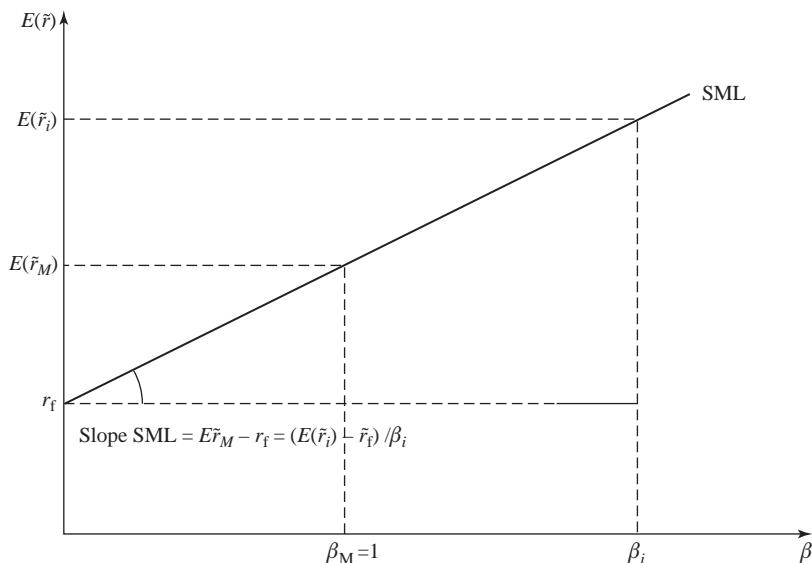


Figure 8.2
The SML.

8.3 Valuing Risky Cash Flows with the CAPM

We are now in position to make use of the CAPM not only to price assets but also to value nontraded risky cash flows such as those arising from an investment project. The traditional approach to this problem proposes to value an investment project at its present value price (i.e., at the appropriately discounted sum of the expected future cash flows). The logic is straightforward: to value a project equal to the present value of the expected future cash flows discounted at a particular rate is to price the project in a manner such that, at its present value price, it is expected to earn that discount rate. The appropriate rate, in turn, must be the analyst's estimate of the rate of return on *other* financial assets that represent title to cash flows *similar* in risk and timing to that of the project in question. This strategy has the consequence of pricing the project to pay the prevailing competitive rate for its risk class.

Enter the CAPM, which makes a definite statement regarding the appropriate discount factor to be used or, equivalently, the risk premium that should be applied to discount expected future cash flows. Strictly speaking, the CAPM is a one-period model; it is thus formally appropriate to use it only for one-period cash flows or projects.² In practice,

² Merton (1973) derives a multiperiod continuous-time version of the CAPM, provided cash flows follow a Markov process from period to period.

its use is more general and a multiperiod cash flow is typically viewed as the sum of one-period cash flows, each of which can be evaluated with the approach we now describe.

Consider some project j with cash-flow pattern

$$\begin{array}{c|c} t & t+1 \\ \hline -p_{j,t} & | \tilde{CF}_{j,t+1} \end{array}$$

The link with the CAPM is immediate once we define the rate of return on project j . For a financial asset, we would naturally write $\tilde{r}_{j,t+1} = (\tilde{p}_{j,t+1} + \tilde{d}_{j,t+1} - p_{j,t})/p_{j,t}$, where $\tilde{d}_{j,t}$ is the dividend or any flow payment associated with the asset between date t and $t+1$. Similarly, if the initial value of the project with cash flow $\tilde{CF}_{j,t+1}$ is $p_{j,t}$, the return on the project is $\bar{r}_{j,t+1} = (\tilde{CF}_{j,t+1} - p_{j,t})/p_{j,t}$.

One thus has

$$\begin{aligned} 1 + E(\tilde{r}_j) &= E\left(\frac{\tilde{CF}_{j,t+1}}{p_{j,t}}\right) = \frac{E(\tilde{CF}_{j,t+1})}{p_{j,t}}, \text{ and by the CAPM,} \\ E(\tilde{r}_j) &= r_f + \beta_j(E(\tilde{r}_M) - r_f), \text{ or} \\ 1 + E(\tilde{r}_j) &= 1 + r_f + \beta_j(E(\tilde{r}_M) - r_f), \text{ or} \\ \frac{E(\tilde{CF}_{j,t+1})}{p_{j,t}} &= 1 + r_f + \beta_j(E(\tilde{r}_M) - r_f). \text{ Thus,} \\ p_{j,t} &= \frac{E(\tilde{CF}_{j,t+1})}{1 + r_f + \beta_j(E(\tilde{r}_M) - r_f)} \end{aligned}$$

According to the CAPM, the project is thus priced at the present value of its expected cash flows discounted at the risk-adjusted rate appropriate to its risk class as identified by its β_j .

As discussed in Chapter 2, there is another potential approach to the pricing problem. It consists in altering the *numerator* of the pricing equations (the sum of expected cash flows) so that it is permissible to discount at the risk-free rate. This approach is based on the concept of certainty equivalent, which we discussed in Chapter 3. The idea is simple: If we replace each element of the future cash flow by its *CE*, it is clearly permissible to discount at the risk-free rate. Since we are interested in equilibrium valuations, however, we need a market certainty equivalent rather than an individual investor one. It turns out that this approach raises exactly the same set of issues as the more common one just considered: an equilibrium asset pricing model is required to tell us what market risk premium is

appropriate to deduct from the expected cash flow to obtain its *CE*. Again the CAPM helps solve this problem.³

In the case of a one-period cash flow, transforming period-by-period cash flows into their market certainty equivalents can be accomplished in a straightforward fashion by applying the CAPM equation to the rate of return expected on the project. With $\tilde{r}_j = (\tilde{C}F_{j,t+1}/p_{j,t}) - 1$, the CAPM implies

$$E\left(\frac{\tilde{C}F_{j,t+1}}{p_{j,t}} - 1\right) = r_f + \beta_j(E(\tilde{r}_M) - r_f) = r_f + \frac{\text{cov}\left(\frac{\tilde{C}F_{j,t+1}}{p_{j,t}} - 1, \tilde{r}_M\right)}{\sigma_M^2}(E(\tilde{r}_M) - r_f)$$

or

$$E\left(\frac{\tilde{C}F_{j,t+1}}{p_{j,t}} - 1\right) = r_f + \frac{1}{p_{j,t}} \text{cov}(\tilde{C}F_{j,t+1}, \tilde{r}_M) \left[\frac{E(\tilde{r}_M) - r_f}{\sigma_M^2} \right]$$

Solving for $p_{j,t}$ yields

$$p_{j,t} = \frac{E(\tilde{C}F_{j,t+1}) - \text{cov}(\tilde{C}F_{j,t+1}, \tilde{r}_M) \left[\frac{E(\tilde{r}_M) - r_f}{\sigma_M^2} \right]}{1 + r_f}$$

which one may also write

$$p_{j,t} = \frac{E(\tilde{C}F_{j,t+1}) - p_{j,t} \beta_j [E(\tilde{r}_M) - r_f]}{1 + r_f}$$

By appropriately transforming the expected cash flows, i.e., by subtracting what we have called an insurance premium (in Chapter 3), one can thus discount at the risk-free rate.

The equilibrium certainty equivalent can thus be defined using the CAPM relationship. Note the information requirements in the procedure: if what we are valuing is indeed a one-off, nontraded cash flow, the estimation of the β_j , or of $\text{cov}(\tilde{C}F_{j,t+1}, \tilde{r}_M)$, is far from straightforward; in particular, it cannot be based on historical data since there are none for the project at hand. It is here that the standard prescription calls for identifying a traded asset that can be viewed as similar in the sense of belonging to the same risk class. The estimated β for that traded asset is then to be used as an approximation in the above valuation formulas.

In the sections that follow, we first generalize the analysis of the efficient frontier presented in Chapter 6 to the $N \geq 2$ asset case. Such a generalization will require the use of elementary matrix algebra and is one of those rare situations in economic science where a

³ As does the arbitrage pricing theory (APT); see Chapter 14.

more general approach yields a greater specificity of results. We will, for instance, be able to detail a version of the CAPM without a risk-free asset. This is then followed by the derivation of the standard CAPM where a risk-free asset is present.

As noted in the introduction, the CAPM is essentially an interpretation that we are able to apply to the efficient frontier. Not surprisingly, therefore, we begin this task with a return to characterizing that frontier.

8.4 The Mathematics of the Portfolio Frontier: Many Risky Assets and No Risk-Free Asset

Notation. Assume $N \geq 2$ risky assets; assume further that no asset has a return that can be expressed as a linear combination of the returns to a subset of the other assets (the returns are linearly independent). Let V denote the variance–covariance matrix, in other words, $V_{ij} = \text{cov}(r_i, r_j)$; by construction V is symmetric. Linear independence in the above sense implies that V^{-1} exists. Let w represent a column vector of portfolio weights for the N assets. The expression $w^T V w$ then represents the portfolio's return variance: $w^T V w$ is always positive (i.e., V is positive definite).

Let us illustrate this latter assertion in the two-asset case

$$\begin{aligned} w^T V w &= (w_1 \ w_2) \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = (w_1 \sigma_1^2 + w_2 \sigma_{21} \quad w_1 \sigma_{12} + w_2 \sigma_2^2) \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \\ &= w_1^2 \sigma_1^2 + w_1 w_2 \sigma_{21} + w_1 w_2 \sigma_{12} + w_2^2 \sigma_2^2 \\ &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2w_1 w_2 \sigma_{12} \geq 0 \end{aligned}$$

since $\sigma_{12} = \rho_{12} \sigma_1 \sigma_2 \geq -\sigma_1 \sigma_2$.

Definition 8.1 formalizes the notion of a portfolio lying on the efficient frontier. Note that every portfolio is ultimately defined by the weights that determine its composition.

Definition 8.1 A frontier portfolio is one that displays minimum variance among all feasible portfolios with the same $E(\tilde{r}_p)$.

A portfolio p , characterized by w_p , is a frontier portfolio, if and only if w_p solves.⁴

⁴ The problem below in vector notation is problem (*QP*) of Chapter 6.

$$\begin{aligned} & \min_w \frac{1}{2} w^T V w \\ (\lambda) \text{ s.t. } & w^T e = E \left(\sum_{i=1}^N w_i E(\tilde{r}_i) = E(\tilde{r}_p) = E \right) \\ (\gamma) \text{ } & w^T \mathbf{1} = 1 \quad \left(\sum_{i=1}^N w_i = 1 \right) \end{aligned}$$

where the superscript T stands for *transposed* (i.e., transforms a column vector into a row vector) and reciprocally, e denotes the column vector of expected returns to the N assets, $\mathbf{1}$ represents the column vector of ones, and λ, γ are Lagrange multipliers. Short sales are permitted (no nonnegativity constraints are present). The solution to this problem can be characterized as the solution to $\min_{\{w, \lambda, \gamma\}} L$, where L is the Lagrangian:

$$L = \frac{1}{2} w^T V w + \lambda(E - w^T e) + \gamma(1 - w^T \mathbf{1}) \quad (8.8)$$

Under these assumptions, the optimal w_p , λ and γ must satisfy Eqs. (8.9) through (8.11), which are the necessary and sufficient first-order conditions (FOCs):

$$\frac{\partial L}{\partial w} = Vw - \lambda e - \gamma \mathbf{1} = 0 \quad (8.9)$$

$$\frac{\partial L}{\partial \lambda} = E - w^T e = 0 \quad (8.10)$$

$$\frac{\partial L}{\partial \gamma} = 1 - w^T \mathbf{1} = 0 \quad (8.11)$$

In the lines that follow, we manipulate these equations to provide an intuitive characterization of the optimal portfolio proportions (8.17). From Eq. (8.9), $Vw_p = \lambda e + \gamma \mathbf{1}$, or

$$w_p = \lambda V^{-1} e + \gamma V^{-1} \mathbf{1}, \text{ and} \quad (8.12)$$

$$e^T w_p = \lambda(e^T V^{-1} e) + \gamma(e^T V^{-1} \mathbf{1}) \quad (8.13)$$

Since $e^T w_p = w_p^T e$, we also have, from Eq. (8.10), that

$$E(\tilde{r}_p) = \lambda(e^T V^{-1} e) + \gamma(e^T V^{-1} \mathbf{1}) \quad (8.14)$$

From Eq. (8.12), we have

$$\begin{aligned} \mathbf{1}^T w_p &= w_p^T \mathbf{1} = \lambda(\mathbf{1}^T V^{-1} e) + \gamma(\mathbf{1}^T V^{-1} \mathbf{1}) \\ &= 1 \text{ (by Eq. (8.10))} \\ 1 &= \lambda(\mathbf{1}^T V^{-1} e) + \gamma(\mathbf{1}^T V^{-1} \mathbf{1}) \end{aligned} \quad (8.15)$$

Notice that Eqs. (8.14) and (8.15) are two *scalar* equations in the unknowns λ and γ (since such terms as $e^T V^{-1} e$ are pure numbers!). Solving this system of two equations in two unknowns, we obtain

$$\lambda = \frac{CE - A}{D} \quad \text{and} \quad \gamma = \frac{B - AE}{D} \quad (8.16)$$

where

$$\begin{aligned} A &= \mathbf{1}^T V^{-1} e = e^T V^{-1} \mathbf{1} \\ B &= e^T V^{-1} e > 0 \\ C &= \mathbf{1}^T V^{-1} \mathbf{1} \\ D &= BC - A^2 \end{aligned}$$

Here we have used the fact that the inverse of a positive definite matrix is itself positive definite. It can be shown that D is also strictly positive. Substituting Eqs. (8.16) into Eq. (8.12), we obtain

$$\begin{aligned} w_p &= \underbrace{\frac{CE - A}{D} V^{-1} e}_{\substack{\text{vector} \\ \lambda \\ \text{scalar}}} + \underbrace{\frac{B - AE}{D} V^{-1} \mathbf{1}}_{\substack{\text{vector} \\ \gamma \\ \text{scalar}}} \\ &= \frac{1}{D} [B(V^{-1} \mathbf{1}) - A(V^{-1} e)] + \frac{1}{D} [C(V^{-1} e) - A(V^{-1} \mathbf{1})]E \\ w_p &= \underbrace{g}_{\substack{\text{vector}}} + \underbrace{h}_{\substack{\text{vector}}} \underbrace{E}_{\text{scalar}} \end{aligned} \quad (8.17)$$

Since the FOCs (Eqs. (8.9) through (8.11)) are a necessary and sufficient characterization for w_p to represent a frontier portfolio with expected return equal to E , any frontier portfolio can be represented by Eq. (8.17). This is a very nice expression; pick the desired expected return E and it straightforwardly gives the weights of the corresponding frontier portfolio with E as its expected return. The portfolio's variance follows as $\sigma_p^2 = w_p^T V w_p$, which is also straightforward. Efficient portfolios are those for which E exceeds the expected return on

the minimum risk, risky portfolio. Our characterization thus applies to efficient portfolios as well: Pick an efficient E and Eq. (8.17) gives its exact composition. See Appendix 8.2 for an example.

Can we further identify the vectors g and h in Eq. (8.17)? In particular, do they somehow correspond to the weights of easily recognizable portfolios? The answer is positive. Since, if $E = 0$, $g = w_p$, g then represents the weights that define the frontier portfolio with $E(\tilde{r}_p) = 0$. Similarly, $g + h$ corresponds to the weights of the frontier portfolio with $E(\tilde{r}_p) = 1$, since $w_p = g + hE(\tilde{r}_p) = g + h\mathbf{1} = g + h$.

The simplicity of the relationship in Eq. (8.17) allows us to make two claims.

Proposition 8.1 The entire set of frontier portfolios can be generated by (are affine combinations of) g and $g + h$.

Proof To see this, let q be an arbitrary frontier portfolio with $E(\tilde{r}_q)$ as its expected return. Consider portfolio weights (proportions) $\pi_g = 1 - E(\tilde{r}_q)$ and $\pi_{g+h} = E(\tilde{r}_q)$; then, as asserted,

$$[1 - E(\tilde{r}_q)]g + E(\tilde{r}_q)(g + h) = g + hE(\tilde{r}_q) = w_q$$

The prior remark is generalized in Proposition 8.2.

Proposition 8.2 The portfolio frontier can be described as affine combinations of any two frontier portfolios, not just the frontier portfolios g and $g + h$.

Proof To confirm this assertion, let p_1 and p_2 be any two distinct frontier portfolios; since the frontier portfolios are different, $E(\tilde{r}_{p_1}) \neq E(\tilde{r}_{p_2})$. Let q be an arbitrary frontier portfolio, with expected return equal to $E(\tilde{r}_q)$. Since $E(\tilde{r}_{p_1}) \neq E(\tilde{r}_{p_2})$, there must exist a unique number α such that

$$E(\tilde{r}_q) = \alpha E(\tilde{r}_{p_1}) + (1 - \alpha)E(\tilde{r}_{p_2}) \quad (8.18)$$

Now consider a portfolio of p_1 and p_2 with weights α , $1 - \alpha$, respectively, as determined by Eq. (8.18). We must show that $w_q = \alpha w_{p_1} + (1 - \alpha)w_{p_2}$.

$$\begin{aligned} \alpha w_{p_1} + (1 - \alpha)w_{p_2} &= \alpha[g + hE(\tilde{r}_{p_1})] + (1 - \alpha)[g + hE(\tilde{r}_{p_2})] \\ &= g + h[\alpha E(\tilde{r}_{p_1}) + (1 - \alpha)E(\tilde{r}_{p_2})] \\ &= g + hE(\tilde{r}_q) \\ &= w_q, \text{ since } q \text{ is a frontier portfolio} \end{aligned}$$

What does the set of frontier portfolios, which we have characterized so conveniently, look like? Can we identify, in particular, the minimum variance portfolio? Locating that

portfolio is surely key to a description of the set of all frontier portfolios. Fortunately, given our results thus far, the task is straightforward.

For any portfolio on the frontier,

$$\sigma^2(\tilde{r}_p) = [g + hE(\tilde{r}_p)]^T V[g + hE(\tilde{r}_p)], \text{ with } g \text{ and } h \text{ as defined earlier.}$$

Multiplying all this out (very messy) yields:

$$\sigma^2(\tilde{r}_p) = \frac{C}{D} \left(E(\tilde{r}_p) - \frac{A}{C} \right)^2 + \frac{1}{C} \quad (8.19)$$

where A , C , and D are the constants defined earlier. We can immediately identify the following: since $C > 0$, $D > 0$,

- i. the expected return of the minimum variance portfolio is A/C ;
- ii. the variance of the minimum variance portfolio is given by $1/C$;
- iii. [Equation \(8.19\)](#) is the equation of a parabola with vertex $(1/C, A/C)$ in the expected return/variance space and of a hyperbola in the expected return/standard deviation space. See [Figures 8.3 and 8.4](#).

The extended shape of this set of frontier portfolios is due to the allowance for short sales as underlined in [Figure 8.5](#).

What has been accomplished thus far? First and foremost, we have a much richer knowledge of the set of frontier portfolios: given a level of desired expected return, we can easily identify the relative proportions of the constituent assets that must be combined to

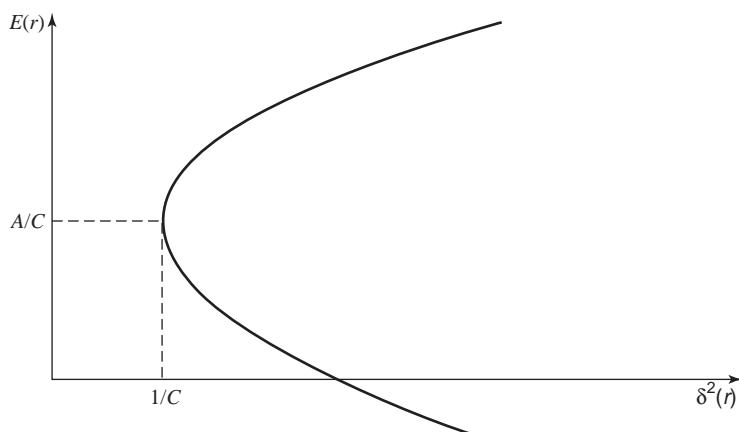
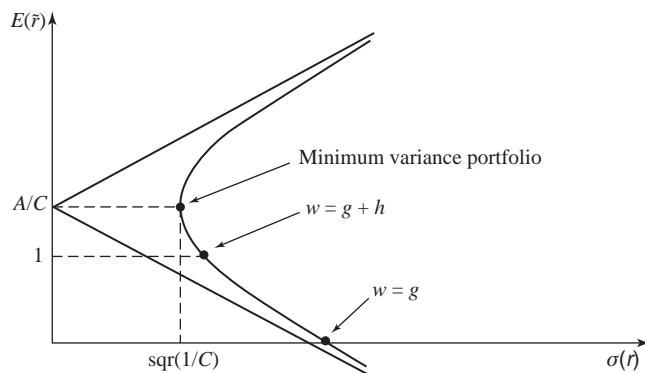
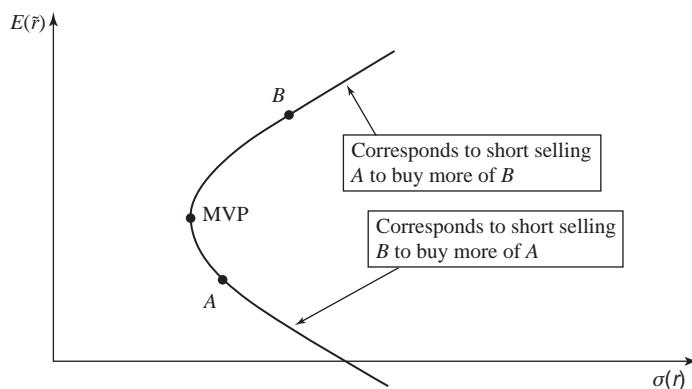


Figure 8.3

The set of frontier portfolios: Mean—variance space.

**Figure 8.4**

The set of frontier portfolios: Mean—standard deviation space.

**Figure 8.5**

The set of frontier portfolios: Short selling allowed.

create a portfolio with that expected return. This was illustrated in Eq. (8.17), and it is key. We then used it to identify the minimum risk portfolio and to describe the graph of all frontier portfolios.

All of these results apply to portfolios of any arbitrary collection of assets. So far, nothing has been said about financial market equilibrium. As a next step toward that goal, however, we need to identify the set of frontier portfolios that is efficient. Given Eq. (8.17), this is a straightforward task.

8.5 Characterizing Efficient Portfolios (No Risk-Free Assets)

Our first order of business is a definition.

Definition 8.2 Efficient portfolios are those frontier portfolios for which the expected return exceeds A/C , the expected return of the minimum variance portfolio.

Since Eq. (8.17) applies to all frontier portfolios, it applies to efficient ones as well. Fortunately, we also know the expected return on the minimum variance portfolio. As a first step, let us prove the converse of Proposition 8.2.

Proposition 8.3 Any convex combination of frontier portfolios is also a frontier portfolio.

Proof Let $(\bar{w}_1 \dots \bar{w}_N)$ define N frontier portfolios (\bar{w}_i represents the vector defining the composition of the i th portfolio) and α_i , $i = 1, \dots, N$ be real numbers such that $\sum_{i=1}^N \alpha_i = 1$. Lastly, let $E(\tilde{r}_i)$ denote the expected return of the portfolio with weights \bar{w}_i .

We want to show that $\sum_{i=1}^N \alpha_i \bar{w}_i$ is a frontier portfolio with $E(\tilde{r}) = \sum_{i=1}^N \alpha_i E(\tilde{r}_i)$.

The weights corresponding to a linear combination of the above N portfolios are:

$$\begin{aligned}\sum_{i=1}^N \alpha_i \bar{w}_i &= \sum_{i=1}^N \alpha_i (g + hE(\tilde{r}_i)) \\ &= \sum_{i=1}^N \alpha_i g + h \sum_{i=1}^N \alpha_i E(\tilde{r}_i) \\ &= g + h \left[\sum_{i=1}^N \alpha_i E(\tilde{r}_i) \right]\end{aligned}$$

Thus $\sum_{i=1}^N \alpha_i \bar{w}_i$ is a frontier portfolio with $E(\tilde{r}) = \sum_{i=1}^N \alpha_i E(\tilde{r}_i)$.

A corollary to the previous result is the following preposition.

Proposition 8.4 The set of efficient portfolios is a convex set.⁵

Proof Suppose each of the N portfolios under consideration was efficient; then $E(\tilde{r}_i) \geq A/C$, for every portfolio i . However, $\sum_{i=1}^N \alpha_i E(\tilde{r}_i) \geq \sum_{i=1}^N \alpha_i A/C = A/C$; thus, the convex combination is efficient as well. So the set of efficient portfolios, *as characterized by their portfolio weights*, is a convex set.

⁵ This does not mean, however, that the frontier of this set is convex-shaped in the risk-return space.

It follows from [Proposition 8.4](#) that if every investor holds an efficient portfolio, the market portfolio, being a weighted average of all individual portfolios, is also efficient. This is a key result.

The next section further refines our understanding of the set of frontier portfolios and, more especially, the subset of them that is efficient. Observe, however, that as yet we have said nothing about equilibrium.

8.6 Background for Deriving the Zero-Beta CAPM: Notion of a Zero-Covariance Portfolio

Proposition 8.5 For any frontier portfolio p , except the minimum variance portfolio, there exists a unique frontier portfolio with which p has zero covariance.

We will call this portfolio the *zero-covariance portfolio relative to p* and denote its vector of portfolio weights by $ZC(p)$.

Proof To prove this claim it will be sufficient to exhibit the (unique) portfolio that has this property. As we shall demonstrate shortly (see [Eq. \(8.25\)](#) and the discussion following it), the covariance of any two frontier portfolios p and q is given by the following general formula:

$$\text{cov}(\tilde{r}_p, \tilde{r}_q) = \frac{C}{D} \left[E(\tilde{r}_p) - \frac{A}{C} \right] \left[E(\tilde{r}_q) - \frac{A}{C} \right] + \frac{1}{C} \quad (8.20)$$

where A , C , and D are uniquely defined by e , the vector of expected returns, and V , the matrix of variances and covariances for portfolio p . These are, in fact, the same quantities A , C , and D defined earlier. If it exists, $ZC(p)$ must therefore satisfy

$$\text{cov}(\tilde{r}_p, \tilde{r}_{ZC(p)}) = \frac{C}{D} \left[E(\tilde{r}_p) - \frac{A}{C} \right] \left[E(\tilde{r}_{ZC(p)}) - \frac{A}{C} \right] + \frac{1}{C} = 0 \quad (8.21)$$

Since A , C , and D are all numbers, we can solve for $E(\tilde{r}_{ZC(p)})$

$$E(\tilde{r}_{ZC(p)}) = \frac{A}{C} - \frac{\frac{D}{C^2}}{E(\tilde{r}_p) - \frac{A}{C}} \quad (8.22)$$

Given $E(\tilde{r}_{ZC(p)})$, we can use [Eq. \(8.17\)](#) to uniquely define the portfolio weights corresponding to it.

From Eq. (8.22), since $A > 0$, $C > 0$, $D > 0$, if $E(\tilde{r}_p) > A/C$ (i.e., is efficient), then $E(r_{ZC(p)}) < A/C$ (i.e., is inefficient), and vice versa. The portfolio $ZC(p)$ will turn out to be crucial to what follows. It is possible to give a more complete geometric identification to the zero-covariance portfolio if we express the frontier portfolios in the context of the $E(\tilde{r}) - \sigma^2(\tilde{r})$ space (Figure 8.6).

The equation of the line through the chosen portfolio p and the minimum variance portfolio can be shown to be the following (it has the form $(y = b + mx)$):

$$E(\tilde{r}) = \frac{A}{C} - \frac{\frac{D}{C^2}}{E(\tilde{r}_p) - \frac{A}{C}} + \frac{E(\tilde{r}_p) - \frac{A}{C}}{\sigma^2(\tilde{r}_p) - \frac{1}{C}} \sigma^2(\tilde{r})$$

If $\sigma^2(\tilde{r}) = 0$, then

$$E(\tilde{r}) = \frac{A}{C} - \frac{\frac{D}{C^2}}{E(\tilde{r}_p) - \frac{A}{C}} = E(\tilde{r}_{ZC(p)})$$

(by Eq. (8.22)).

That is, the intercept of the line joining p and the minimum variance portfolio is the expected return on the zero-covariance portfolio. This identifies the zero-covariance portfolio to p geometrically. We already know how to determine its precise composition.

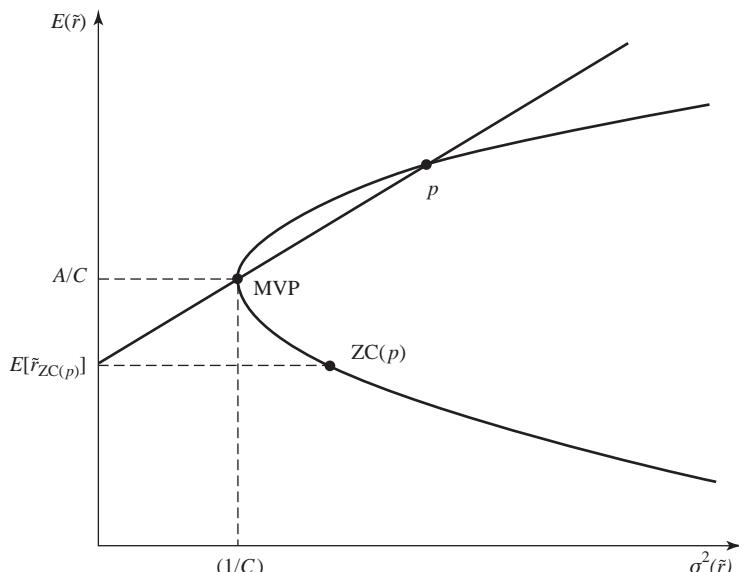


Figure 8.6

The set of frontier portfolios: Location of the zero-covariance portfolio.

Our next step is to describe the expected return on any portfolio in terms of frontier portfolios. After some manipulations, this will yield Eq. (8.29). The specialization of this relationship will give the zero-beta CAPM, which is a version of the CAPM when there is no risk-free asset. Recall that thus far we have not included a risk-free asset in our collection of assets from which we construct portfolios. Let q be any portfolio (which might not be on the portfolio frontier) and let p be any frontier portfolio.

$$\begin{aligned} \text{cov}(\tilde{r}_p, \tilde{r}_q) &= \underbrace{w_p^T V w_q}_{\text{by definition}} \\ &= [\lambda V^{-1} e + \gamma V^{-1} \mathbf{1}]^T V w_q \\ &= \lambda e^T V^{-1} V w_q + \gamma \mathbf{1}^T V^{-1} V w_q \\ &= \lambda e^T w_q + \gamma \left(\text{since } \mathbf{1}^T w_q = \sum_{i=1}^N w_q^i \equiv 1 \right) \end{aligned} \quad (8.23)$$

$$= \lambda E(\tilde{r}_q) + \gamma \left(\text{since } e^T w_q = \sum_{i=1}^N E(\tilde{r}_i) w_q^i \equiv E(\tilde{r}_q) \right) \quad (8.24)$$

where $\lambda = (CE(\tilde{r}_p) - A)/D$ and $\gamma = (B - AE(\tilde{r}_p))/D$, as per earlier definitions.

Substituting these expressions into Eq. (8.24) gives

$$\text{cov}(\tilde{r}_p, \tilde{r}_q) = \frac{CE(\tilde{r}_p) - A}{D} E(\tilde{r}_q) + \frac{B - AE(\tilde{r}_p)}{D} \quad (8.25)$$

Equation (8.25) is a short step from Eq. (8.20): Collect all terms involving expected returns, add and subtract $A^2 C / D C^2$ to get the first term in Eq. (8.20) with a remaining term equal to $(1/C)(BC/D - A^2/D)$. But the latter is simply $1/C$ since $D = BC - A^2$.

Let us go back to Eq. (8.24) and apply it to the case where q is $ZC(p)$; one gets

$$0 = \text{cov}(\tilde{r}_p, \tilde{r}_{ZC(p)}) = \lambda E(\tilde{r}_{ZC(p)}) + \gamma \text{ or } \gamma = -\lambda E(\tilde{r}_{ZC(p)}) \quad (8.26)$$

hence Eq. (8.24) becomes

$$\text{cov}(\tilde{r}_p, \tilde{r}_q) = \lambda [E(\tilde{r}_q) - E(\tilde{r}_{ZC(p)})] \quad (8.27)$$

Apply the latter to the case $p = q$ to get

$$\sigma_p^2 = \text{cov}(\tilde{r}_p, \tilde{r}_p) = \lambda [E(\tilde{r}_p) - E(\tilde{r}_{ZC(p)})] \quad (8.28)$$

and divide Eq. (8.27) by Eq. (8.28) and rearrange to obtain

$$E(\tilde{r}_q) = E(\tilde{r}_{ZC(p)}) + \beta_{pq} [E(\tilde{r}_p) - E(\tilde{r}_{ZC(p)})] \quad (8.29)$$

This equation bears more than a passing resemblance to the SML implication of the CAPM. But as yet it is simply a statement about the various portfolios that can be created from arbitrary collections of assets: (1) pick any frontier portfolio p ; (2) this defines an associated zero-covariance portfolio $ZC(p)$; and (3) any other portfolio q 's expected return can be expressed in terms of the returns to those portfolios and the covariance of q with the arbitrarily chosen frontier portfolio. [Equation \(8.29\)](#) would *very closely* resemble the SML if, in particular, we could choose $p = M$, the market portfolio of existing assets. The circumstances under which it is possible to do this form the subject to which we now turn.

8.7 The Zero-Beta CAPM

We would like to explain asset expected returns *in equilibrium*. The relationship in [Eq. \(8.29\)](#), however, is not the consequence of an equilibrium theory because it was derived for a *given* particular vector of expected asset returns, e , and a given covariance–variance matrix, V . In fact, it is the vector of returns e that we would like, in equilibrium, to understand. We need to identify a particular portfolio as being a frontier portfolio without specifying *a priori* the (expected) return vector and variance–covariance matrix of its constituent assets. The zero-beta CAPM tells us that under certain assumptions, this desired portfolio can be identified as the market portfolio M .

We may assume one of the following:

- i. agents maximize expected utility with increasing and strictly concave utility of money functions, and asset returns are multivariate normally distributed, or
- ii. each agent chooses a portfolio with the objective of maximizing a derived utility function of the form $W(e, \sigma^2)$, $W_1 > 0$, $W_2 < 0$, W concave.⁶

In addition, we assume that all investors have a common time horizon and homogeneous beliefs about e and V .

Under either set of assumptions, investors will only hold mean–variance efficient frontier portfolios.⁷ But this implies that, **in equilibrium**, the market portfolio, which is a convex combination of individual portfolios, is also on the efficient frontier.⁸

⁶ As noted in Chapter 6, the maintained perspective of this text is alternative i. In this sense, the zero-beta version of the CAPM can be viewed as an implication of expected utility theory.

⁷ Recall the demonstration in Section 6.3.

⁸ Note that, in the standard version of the CAPM, the analogous claim crucially depended on the existence of a risk-free asset.

Therefore, in Eq. (8.23), p can be chosen to be M , the portfolio of all risky assets, and Eq. (8.29) can, therefore, be expressed as

$$E(\tilde{r}_q) = E(\tilde{r}_{ZC(M)}) + \beta_{Mq}[E(\tilde{r}_M) - E(\tilde{r}_{ZC(M)})] \quad (8.30)$$

The relationship in Eq. (8.30) holds for any portfolio q , whether or not it is a frontier portfolio. This is the zero-beta CAPM.

An individual asset j is also a portfolio, so Eq. (8.30) applies to it as well:

$$E(\tilde{r}_j) = E(\tilde{r}_{ZC(M)}) + \beta_{Mj}[E(\tilde{r}_M) - E(\tilde{r}_{ZC(M)})] \quad (8.31)$$

The zero-beta CAPM (and the more familiar *Sharpe–Lintner–Mossin* CAPM) is an equilibrium theory: the relationships in Eqs. (8.30) and (8.31) hold in equilibrium.⁹

In equilibrium, investors will not be maximizing utility unless they hold efficient portfolios. Therefore, the market portfolio is efficient; we have identified one efficient frontier portfolio, and we can apply Eq. (8.30). By contrast, Eq. (8.29) is a pure mathematical relationship with no economic content; it simply describes relationships between frontier portfolio returns and the returns from any other portfolio of the same assets.

As noted in the introduction, the zero-beta CAPM does not, however, describe the process to or by which equilibrium is achieved. In other words, the process by which agents buy and sell securities in their desire to hold efficient portfolios, thereby altering security prices and thus expected returns, and requiring further changes in portfolio composition is not present in the model. When this process ceases and all agents are optimizing given the prevailing prices, then all will be holding efficient portfolios given the equilibrium expected returns e and covariance–variance matrix V . Thus M is also an efficient portfolio.

The efficiency of M is a principal implication of the CAPM.

Since, in equilibrium, agents' desired holdings of securities coincide with their actual holdings, we can identify M as the actual portfolio of securities held in the marketplace. There are many convenient approximations to M —the S&P 500 index of stocks being the most popular in the United States. The usefulness of these approximations, which are needed to give empirical content to the CAPM, is, however, debatable, as discussed later in the chapter.

As a final remark, let us note that the name “zero-beta CAPM” comes from the fact that $\beta_{ZC(M),M} = \text{cov}(\tilde{r}_M, \tilde{r}_{ZC(M)}) / \sigma_{ZC(M)}^2 = 0$, by construction of $ZC(M)$; in other words, the beta of $ZC(M)$ is zero.

⁹ Sharpe (1964), Lintner (1965), and Mossin (1966).

8.8 The Standard CAPM

Our development thus far did not admit the option of a risk-free asset. We need to add this if we are to achieve the standard form CAPM. On a purely formal basis, of course, a risk-free asset has zero covariance with M and thus $r_f = E(\tilde{r}_{ZC(M)})$. Hence we could replace $E(\tilde{r}_{ZC(M)})$ with r_f in Eq. (8.31) to obtain the standard representation of the CAPM, the SML. But this approach is not entirely appropriate since the derivation of Eq. (8.31) presumed the absence of any such risk-free asset.

More formally, the addition of a risk-free asset substantially alters the shape of the set of frontier portfolios in the $[E(\tilde{r}), \sigma(\tilde{r})]$ space. Let us briefly outline the development here, which closely resembles what was done above. Consider N risky assets with expected return vector e , and one risk-free asset, with *expected* return $= r_f$. Let p be a frontier portfolio, and let w_p denote the N vector of portfolio weights on the risky assets of p ; w_p in this case is the solution to

$$\begin{aligned} & \min_w \frac{1}{2} w^T V w \\ \text{s.t. } & w^T e + (1 - w^T \mathbf{1}) r_f = E \end{aligned}$$

Solving this problem gives

$$w_p = V^{-1}(e - r_f \mathbf{1}) \frac{E - r_f}{H}$$

where $H = B - 2Ar_f + Cr_f^2$ and A, B, C are defined as before.

Let us examine this expression for w_p more carefully:

$$w_p = \underbrace{V^{-1}}_{n \times n} \underbrace{(e - r_f \mathbf{1})}_{n \times 1} \underbrace{\frac{E(\tilde{r}_p) - r_f}{H}}_{a \text{ number}} \quad (8.32)$$

This expression tells us that if we wish to have a higher expected return, we should invest proportionally the same amount more in each risky asset so that the relative proportions of the risky assets remain unchanged. These proportions are defined by the $V^{-1}(e - r_f \mathbf{1})$ term. This is exactly the result we were intuitively expecting: Graphically, we are back to the linear frontier represented in Figure 8.1.

The weights w_p uniquely identify the tangency portfolio T . Also,

$$\sigma^2(\tilde{r}_p) = w_p^T V w_p = \frac{[E(\tilde{r}_p) - r_f]^2}{H}, \text{ and} \quad (8.33)$$

$$\text{cov}(\tilde{r}_q, \tilde{r}_p) = w_q^T V w_p = \frac{[E(\tilde{r}_q) - r_f][E(\tilde{r}_p) - r_f]}{H} \quad (8.34)$$

for any portfolio q and any frontier portfolio p . Note how all this parallels what we did before. Solving Eq. (8.34) for $E(\tilde{r}_q)$ gives

$$E(\tilde{r}_q) - r_f = \frac{H \text{cov}(\tilde{r}_q, \tilde{r}_p)}{E(\tilde{r}_p) - r_f} \quad (8.35)$$

Substituting for H via Eq. (8.33) yields

$$E(\tilde{r}_q) - r_f = \frac{\text{cov}(\tilde{r}_q, \tilde{r}_p)}{E(\tilde{r}_p) - r_f} \frac{[E(\tilde{r}_p) - r_f]^2}{\sigma^2(\tilde{r}_p)}$$

or

$$E(\tilde{r}_q) - r_f = \frac{\text{cov}(\tilde{r}_q, \tilde{r}_p)}{\sigma^2(\tilde{r}_p)} [E(\tilde{r}_p) - r_f] \quad (8.36)$$

Again, since T is a frontier portfolio, we can choose $p \equiv T$. But in equilibrium $T = M$; in this case, Eq. (8.36) gives

$$E(\tilde{r}_q) - r_f = \frac{\text{cov}(\tilde{r}_q, \tilde{r}_M)}{\sigma^2(\tilde{r}_M)} [E(\tilde{r}_M) - r_f]$$

or

$$E(\tilde{r}_q) = r_f + \beta_{qM}[E(\tilde{r}_M) - r_f] \quad (8.37)$$

for any asset (or portfolio) q . This is the standard CAPM.

Again, let us review the flow of logic that led to this conclusion. First, we identified the efficient frontier of risk-free and risky assets. This efficient frontier is fully characterized by the risk-free asset and a specific tangency frontier portfolio. The latter is identified in Eq. (8.32). We then observed that all investors, in equilibrium under homogeneous expectations, would hold combinations of the risk-free asset and that portfolio. Thus it must constitute the *market*—the portfolio of all risky assets. It is these latter observations that begin to give the CAPM empirical content.

8.9 An Empirical Assessment of the CAPM

Do observed patterns in financial data illustrate the conclusions of the CAPM? To answer this question, let us review its major assertions.¹⁰ They are as follows:

1. Investors are well diversified; more precisely, investors hold the risky portion of their portfolio wealth in the form of units of the market portfolio M . Empirically, investors holding risky assets should be found to hold them in the form of some market index fund or otherwise very well-diversified portfolio. This means idiosyncratic risks are irrelevant.
2. A security's equilibrium expected return is a linear function of its beta, with the slope coefficient approximately equal to $[E\tilde{r}_M - r_f]$, and the intercept approximately equal to r_f . The empirical confirmation of this statement would be the observation that overextended periods of time, securities' average returns are linearly related to their respective betas with slope equal to the observed market risk premium, and intercept equal to the average risk-free rate (all measurements for the same data set). This is a statement about the cross section of security returns.
3. No other characteristics of a security matter for the determination of its average returns beyond the level of its systematic risk as measured by beta. Empirically this means that it should not be possible to find any other security characteristic (in addition to beta) that is useful in explaining the cross section of observed average security returns.

If we summarize the CAPM by the above three implications, the answer to the question posed at the start of this section is “no”: the CAPM is an empirical failure.

In many respects, this failure is not at all surprising given the powerful assumptions underpinning the CAPM and their lack of empirical fulfillment. First and foremost, investors do not have identical expectations (the same forecasts of μ_i , σ_i , ρ_{ij} s for all risky assets under consideration). Second, the CAPM presumes that an investor's total wealth is derived only from his investments in the market portfolio and risk-free assets with the implication that *the market portfolio must contain every risky asset*. In reality, investors typically have wage income streams which are not perfectly positively correlated with \tilde{r}_M . These wage income streams are not tradable and thus are not part of the market portfolio. Furthermore, any convenient stock market proxy for the market portfolio, such as the S&P₅₀₀ portfolio or the Wiltshire 5000 portfolio, inevitably omits important asset classes, most especially residential real estate and risky corporate debt. Forty percent of private non-human wealth in the United States takes the form of residential real estate. Taken together, these considerations imply that any empirical test of the CAPM will, in essence, be a joint test of the theory itself and the “efficiency” of the chosen market proxy, an observation

¹⁰ What follows in this section is not a comprehensive review of the CAPM testing literature. Recent detailed surveys include [Ferson and Jagannathan \(1996\)](#) and [Shanken \(1996\)](#).

made forcefully by [Roll \(1977\)](#).¹¹ Lastly, many investors are constrained in their ability to borrow, which may lead them to form high-risk–high-expected return portfolios inside the efficient frontier.

With these qualifications in mind, let us now review the three CAPM summary conclusions in light of “the data.”

The failure of the first assertion is immediate: many investors are not well diversified domestically or internationally, many hold large positions in their own firm’s stock, and many own no stock at all (also a violation of the conclusion to Theorem 5.1), or only small amounts in the form of isolated individual shares.

To explore assertions (2) and (3) above, we first observe that they are basically statements about the form of the SML. Accordingly, it is this relationship that has been subject to formal statistical tests. We limit our review of these tests to three prominent papers and eschew any claim to being comprehensive. We also ignore the numerous details of data assembly and test design. These details are best addressed in a financial econometrics course. All three studies focus on the stock market where detailed data is readily available. The flow of the discussion is chronological.

8.9.1 Fama and MacBeth (1973)

These authors propose a two-step regression procedure, which has been used in numerous subsequent tests of the CAPM.¹² A simplified version of the essentials follows.

- i. First-pass regression: Select a portfolio of assets to serve as the proxy for M . For each of the risky assets in the portfolio and the portfolio itself (value weighted), assemble its time series return data for some representative historical period; for discussion purposes here let this be quarterly data from 1966.1 through 2000.4.¹³ For each asset i , estimate its historical beta via regression equation:

$$\tilde{r}_{i,s} - \tilde{r}_{f,s} = \hat{\alpha}_i + \hat{\beta}_i(\tilde{r}_{M,s} - \tilde{r}_{f,s}) + \tilde{\varepsilon}_{i,s} \quad (8.38)$$

[Equation \(8.38\)](#) is simply the linear representation of the CAPM (cf. [Eq. \(8.5\)](#)).

To emphasize the time series nature of this regression, we have assigned a subscript s

¹¹ Furthermore, modestly different proxies for M can sometimes lead to very different beta estimates for assets common to both.

¹² [Fama and MacBeth \(1973\)](#) were not the only authors to subject the CAPM to rigorous statistical tests. In particular, [Friend and Blume \(1970\)](#) for the 1960–1968 data period and [Haugen and Heins \(1975\)](#) using data from 1926 to 1971 provided robust evidence earlier on the limited ability of the CAPM to explain the data.

¹³ In the case of [Fama and MacBeth \(1973\)](#), the proxy for the market portfolio is the set of all NYSE stocks traded sometime in the period 1926.1 through 1968.6, with returns computed monthly. The risk free rate $r_{f,t}$ is the monthly T-bill rate.

to all the variables. Since individual stock betas are often imprecisely estimated, and in order to minimize the effects of these estimation errors in the regressions to follow, Fama and MacBeth (1973), for example, partition (“sort”) their universe of stocks into 20 portfolios based on their first-pass regression $\hat{\beta}_i$ s, ranking them from lowest to highest.¹⁴ Denote these portfolio betas as $\hat{\beta}_{j,t-1}^P$ where the added $t-1$ subscript is to be identified with the entire data period 1966.1 through 1970.4, and j indexes the portfolios.

- ii. Second-pass regression: For each set of portfolios constructed sequentially in step 1, compute its average return $\bar{r}_{j,t}^P, j = 1, 2, \dots, N = 20$ over a subsequent historical period, say the period 1971.1 through 1975.4. We index all quantities obtained from this second block of time series data with the subscript “ t .” The data points $(\bar{r}_{j,t}^P, \hat{\beta}_{j,t-1}^P)$ then allow the estimation of the following second-pass, cross-sectional (the variation is indexed by j) regression:

$$\bar{r}_{j,t}^P = \hat{\gamma}_{0,t} + \hat{\gamma}_{1,t} \hat{\beta}_{j,t-1}^P + \hat{\gamma}_{2,t} (\hat{\beta}_{j,t-1}^P)^2 + \hat{\gamma}_{3,t} s(\varepsilon_{j,t-1}^P) + \tilde{u}_{j,t}^P \quad (8.39)$$

The presence of the third term in the regression constitutes a simple test for nonlinearities in the return–systematic risk relationship across the collection of portfolios. The fourth term permits testing for evidence of an unsystematic risk-return relationship (thereby contradicting the first implication of the CAPM). In particular, $s(\varepsilon_{j,t-1}^P)$ is the average of the $\sigma_{e_{i,t-1}}$ for all stocks i in the j th portfolio.

This two step process is then repeated over and over again using progressively updated blocks of data. To illustrate this updating, consider the next first-pass regression to be based on the period 1967.1 through 1972.4 with the second-pass regression based on 1972.1 through 1976.4. The overall result is to generate a time series of estimates $\{\hat{\gamma}_{0,t}, \hat{\gamma}_{1,t}, \hat{\gamma}_{2,t}, \hat{\gamma}_{3,t}\}$ corresponding to the succession of data periods.¹⁵ Their procedure basically tests whether current betas have anything to say about future average returns.

The results of this large and carefully detailed study are as follows: First $\bar{\gamma}_2$ and $\bar{\gamma}_3$ (time averages, e.g., $\hat{\gamma}_2 = 1/T \sum_{t=1}^T \hat{\gamma}_{2,t}$) are not statistically different from zero which is support for both linearity and the basic form of the SML. While $\bar{\gamma}_1$ is positive, it is, however, too small relative to the average risk premium on the market. Essentially the average return/systematic risk relationship across the 20 portfolios is “too flat”: the returns to low beta portfolios are too high relative to what the CAPM would predict while high beta portfolios’ average returns are too low. Lastly, the $\bar{\gamma}_0$ estimate is positive but

¹⁴ In other words, the lowest beta portfolio is composed of the bottom 5% of stocks based on their beta rankings; the highest beta portfolio is composed of the highest 5% of stocks as measured by their estimated betas, etc. For each portfolio so composed, its portfolio beta is calculated as per $\beta_P = \sum_{k=1}^{N_p} w_k \hat{\beta}_k$, for all stocks k in the portfolio. In this way, the estimation errors are “averaged out.”

¹⁵ To make clear the timing of these regressions, see the diagrammatic representation in Appendix 8.3.

substantially in excess of the prevailing r_f . The [Fama and MacBeth \(1973\)](#) results have been replicated using other data sets covering different data periods. At best, their analysis represents a “lukewarm” confirmation of the CAPM.

8.9.2 Banz (1981) and the “Size Effect”¹⁶

The contribution of [Banz \(1981\)](#) is to initiate sorting on and to use explanatory variables based on characteristics unrelated to a stock’s market beta. In particular, [Banz \(1981\)](#) double sorts all NYSE (New York Stock Exchange) traded stocks into 25 portfolios, first sorting on increasing “size”—defined to be the aggregate market value of a firm’s equity (5 portfolio sort). In each size-ranked portfolio, the stocks are then “second-sorted” on the basis of increasing beta as per [Fama and MacBeth \(1973\)](#). In response to the criticism in [Roll \(1977\)](#), [Banz \(1981\)](#) also expands the market portfolio to include all publicly traded corporate and government debt. With these innovations, the [Banz \(1981\)](#) procedure amounts to the same as in [Fama and MacBeth \(1973\)](#). His “second-pass” regression takes the form

$$\bar{r}_{j,t}^P = \hat{\gamma}_{0,t} + \hat{\gamma}_{1,t} \hat{\beta}_{j,t-1}^P + \hat{\gamma}_{2,t} \left[\frac{ME_{j,t-1}^P - ME_{M,t-1}^P}{ME_{M,t-1}^P} \right] + \tilde{u}_{j,t} \quad (8.40)$$

In regression (8.40), $ME_{j,t}^P$ signifies the average market value of the stocks in portfolio j , while $ME_{M,t}^P$ is the average market value of all stocks in the market portfolio. [Banz \(1981\)](#) finds that the coefficients $\{\hat{\gamma}_{2,t}\}$ are on average negative and statistically significant, a finding in direct contradiction to the third assertion of the CAPM. While the “small firm effect” appears to have diminished in recent years, it nevertheless lives on in a somewhat transformed state (see the remarks on Fama and French (1993) in Chapter 14).

8.9.3 Fama and French (1992)

These authors cast a wider net for financial quantities with the power to explain the cross section of average equity returns. In particular, they evaluate BE/ME (see Section 2.5.6), E/P (the earnings to price ratio) and leverage (measured as A/ME , the book value of assets to market equity) in addition to the by-then-standard “size” (ME) and market-beta variables.

This selection has its basis in a number of earlier works. [Stattman \(1980\)](#) and [Rosenberg et al. \(1985\)](#) had found that firms’ average equity returns were positively related to their BE/ME ratio. [Chan et al. \(1991\)](#) confirmed that BE/ME was also significant in explaining the cross section of average Japanese equity returns, while [Basu \(1983\)](#) had demonstrated that E/P provided added explanatory power for US average stock returns when included in tests involving both size and market-beta variables.

¹⁶ See also [Reinganum \(1981\)](#) who simultaneously and independently discovered the “size effect.”

In Section 2 of their paper, [Fama and French \(1992\)](#) undertake second-pass regressions using the explanatory variables under study both individually and in combinations thereof. In the case of *BE/ME*, they investigate the cross-sectional regression

$$\tilde{r}_{i,t} = \hat{\gamma}_{0,t} + \hat{\gamma}_{1,t} \ln \left(\left(\frac{\widehat{BE}}{ME} \right)_{i,t-1} \right) + \tilde{u}_{i,t} \quad (8.41)$$

using monthly data for nearly all NYSE stocks for the period 1963.1 through 1990.4.¹⁷ Since $BE_{i,t}$ and $ME_{i,t}$ are measured precisely, there is no need to sort individual stocks into portfolios, and the entire panel of eligible stocks can be used with monthly updating. For regression (8.41), the time series average of the regression slopes is (0.50) with t -statistic (5.71). Clearly, the slope coefficient $\hat{\gamma}_1 = 1/T \sum_{t=1}^T \hat{\gamma}_{1,t}$ is both large and highly statistically different from zero.

This result is to be contrasted with the corresponding one for the market beta (with appropriate presorting into beta-ranked portfolios, etc.):

$$\bar{r}_{i,t} = \hat{\gamma}_{0,t} + \hat{\gamma}_{1,t} \hat{\beta}_{i,t-1} + \tilde{\varepsilon}_{i,t} \quad (8.42)$$

The average coefficient on beta is small (0.15) and its t -statistic (0.46) indicates that there is no statistical basis for concluding it is different from zero.

[Fama and French \(1992\)](#) offer three general conclusions to their study:

For the period 1963–1990:

1. “The market β does not seem to help to explain the cross section of average stock returns.”
2. “The combination of size and book-to-market equity seems to absorb the roles of leverage and E/P in average stock returns.”
3. The ratio of book-to-market equity is the most significant quantity for explaining the cross section of average security returns.¹⁸

While these observations leave us unsatisfied intellectually, it is clearly evident that the CAPM, however elegant and logical as a theory, fails to find empirical support.

8.9.4 Volatility Anomalies

CAPM theory, at its most basic, claims an *ex ante* positive relationship between an asset’s undiversifiable risk and its expected returns: assets which have more undiversifiable risk

¹⁷ Financial stocks are excluded from the sample. For each stock, $\bar{r}_{i,t}$ is the average monthly return for the 12 months following the end of period t .

¹⁸ Both quotes from [Fama and French \(1992\)](#), p. 428.

(higher β s) are less desirable and should sell for lower prices and pay higher expected returns. In the cross section of security returns, the systematic risk (β) and total risk (σ) measures are statistically very highly correlated. The same risk–return relationship should thus hold whether risk is measured as total risk σ or as (relative) systematic risk β .¹⁹ It is this claim that we presently consider.

Greenwood et al. (2010) (see also Ang et al. (2006, 2009)) explore the assertion using as their universe of stocks the Russell 1000, which they sort (with quarterly resorting) into quintiles based on total risk. Given this sorting, five equally weighted portfolios are formed, and their risks, returns, and betas measured using data for the period 1980–2009. The results are presented in Table 8.1 and graphed in Figure 8.7.

Table 8.1: Risk and return (annualized) quintiles of Russell 1000 stocks^a

	Return	σ (Total Risk)	β_{CAPM}
Quintile 1 (low risk)	13.84%	12.15	0.62
2	14.58%	15.1%	0.88
3	15.16%	17.3%	1.02
4	12.89%	20.9%	1.25
5	8.06%	32.6%	1.74
Russell 1000 (M)	11.50%	15.6%	1.00

Source: Table 8.1 and the subsequent Figure 8.9 are taken from Greenwood et al. (2010).

^aReturns and risks have been annualized from quarterly calculations; data period: (1980)–(2009); $E(\bar{r})_{\text{annual}} = 12E(\bar{r})_{\text{monthly}}$; $\sigma_{\text{annual}} = \sqrt{12}\sigma_{\text{monthly}}$. Recall our continuous compounding discussion in Chapter 3.

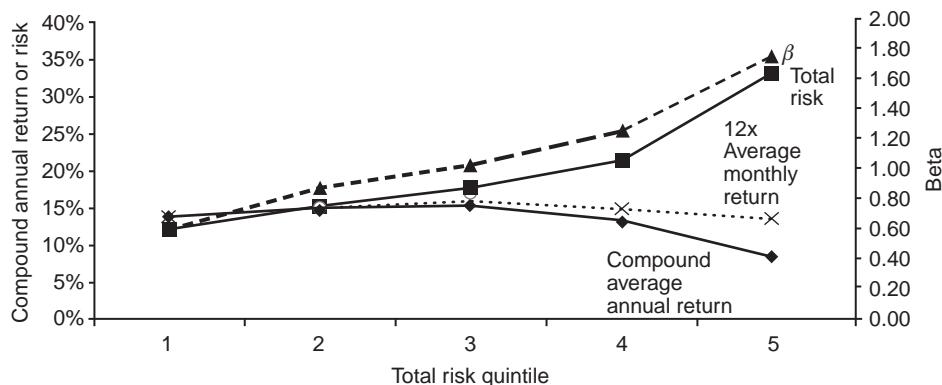


Figure 8.7
Graphical representation of Table 8.1.

¹⁹ This result is not theoretically guaranteed. It is easy to imagine two stocks, 1 and 2, where $\beta_1 < \beta_2$ yet $\sigma_1 > \sigma_2$. See the definition of the CAPM β .

Table 8.1 merits a number of observations. First, we confirm that higher average β portfolios also display higher total risk: systematic risk and total risk go together at this quintile level of aggregation. In **Table 8.1**, note that $\sigma_M = 15.6\%$ is the risk of the market proxy, in this case the Russell 1000 stock portfolio.

Second, we note that up to and including the third quintile, the CAPM risk/return relationship is observed. For quintile portfolios four and five, however, this “intuitive” ordering is reversed: increasingly risky (by either risk measure) assets pay, on average, lower returns. One way of summarizing the pattern observed in **Figure 8.7** is to say that the historical risk/return relationship is “too flat relative to the SML” (see also [Fama and MacBeth, 1973](#)). Not only is this phenomenon a violation of CAPM predictions, it appears also to conflict with the fundamental assumption of investor risk aversion.

There are several proposed explanations in the literature. [Frazzini and Pederson \(2013\)](#), for example, propose that the pattern in **Figure 8.8** may reflect the fact that high-risk tolerant investors are subject to borrowing constraints (a violation of the CAPM assumptions): these investors would prefer to increase their expected returns in an efficient manner by leveraging the market portfolio proxy, but such leverage is unavailable to them. [Frazzini and Pederson \(2013\)](#) propose that this same investor class, as a second best alternative, will attempt to assemble portfolios composed exclusively of high beta, high expected return stocks. If a significant fraction of stock market investors act in this way, the equilibrium effect will be to increase the relative prices and decrease the average returns of high market-beta stocks relative to the CAPM prediction, as suggested by **Figure 8.7**. We emphasize that the absence of borrowing constraints is a fundamental assumption underlying the CAPM.

It has also been suggested that the risk and return pattern in **Figure 8.7** may reflect high demand for what are called “lottery stocks.” These are stocks of near-to-bankruptcy firms that have characteristics (low prices, tiny probabilities of gigantic upward price increases) similar to a state lottery ticket. [Kumar \(2009\)](#) classifies a stock as a lottery stock on the basis of low price, high idiosyncratic risk, and high idiosyncratic skewness. These are stocks that typically underperform given their betas, possibly contributing to the pattern of **Figure 8.7**. [Kumar \(2009\)](#) introduces evidence that the socioeconomic characteristics associated with those who purchase lottery tickets are also associated with the class of investors who purchase “lottery stocks,” suggesting the possibility of the existence of a significant clientele that outweighs this type of security.

To characterize the purchasers of lottery tickets/lottery stocks simply as “risk lovers” would be, however, inaccurate. Rather, these persons identify themselves as individuals for whom the only way to accumulate significant wealth is to “hit the jackpot.” They eschew, for whatever reasons or circumstances, the standard notion of wealth accumulation via savings and reinvestment. As such, their behavior is beyond the scope of expected utility theory cum risk aversion which underlies the CAPM analysis. We also note that lottery stocks

have return distributions which are distinctly non-normal, and thus are formally excluded from consideration under the CAPM assumptions.

We recall from Chapter 2 that the idiosyncratic risk of most stocks represents 80–90% of their total return risk. Under the CAPM, idiosyncratic risks should not matter for asset pricing and average return determination: these are the risks that are diversified away. Emphasizing another departure from the assumptions of the CAPM, [Merton \(1987\)](#) and [Hirshleifer \(1988\)](#) show, via different mechanisms, that if investors face sizable trading frictions (and thus are not well diversified) and face incomplete markets, then idiosyncratic volatility should be positively linked with subsequent average returns.

In a pair of companion papers [Ang et al. \(2006, 2009\)](#) point out that for both US and international stocks, it is the reverse pattern that is actually observed. Measuring idiosyncratic volatility using the Fama–French three factor model (see Chapter 14), and sorting stocks into five idiosyncratic volatility portfolios, with quarterly rebalancing, [Ang et al. \(2006\)](#) uncover the reverse relationship most profoundly for high idiosyncratic volatility stocks. A similar relationship is observed if subsequent returns are compared across current period idiosyncratic risk rankings or if the returns and idiosyncratic risks are

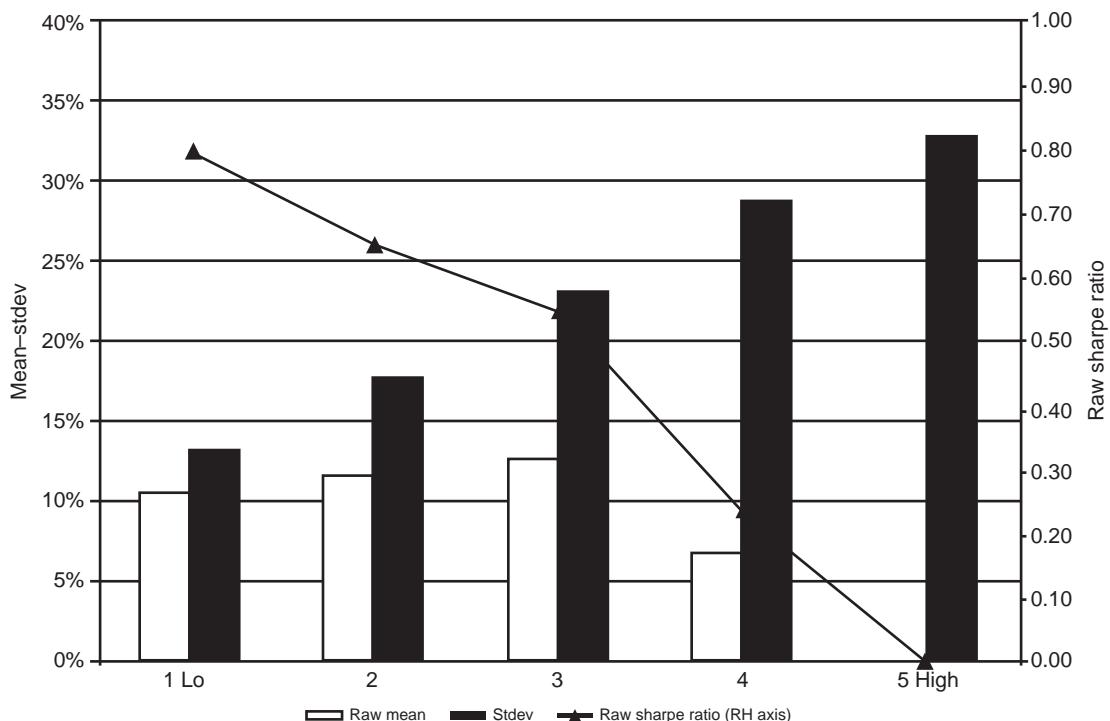


Figure 8.8

Average returns and idiosyncratic volatility: idiosyncratic volatility and subsequent returns.

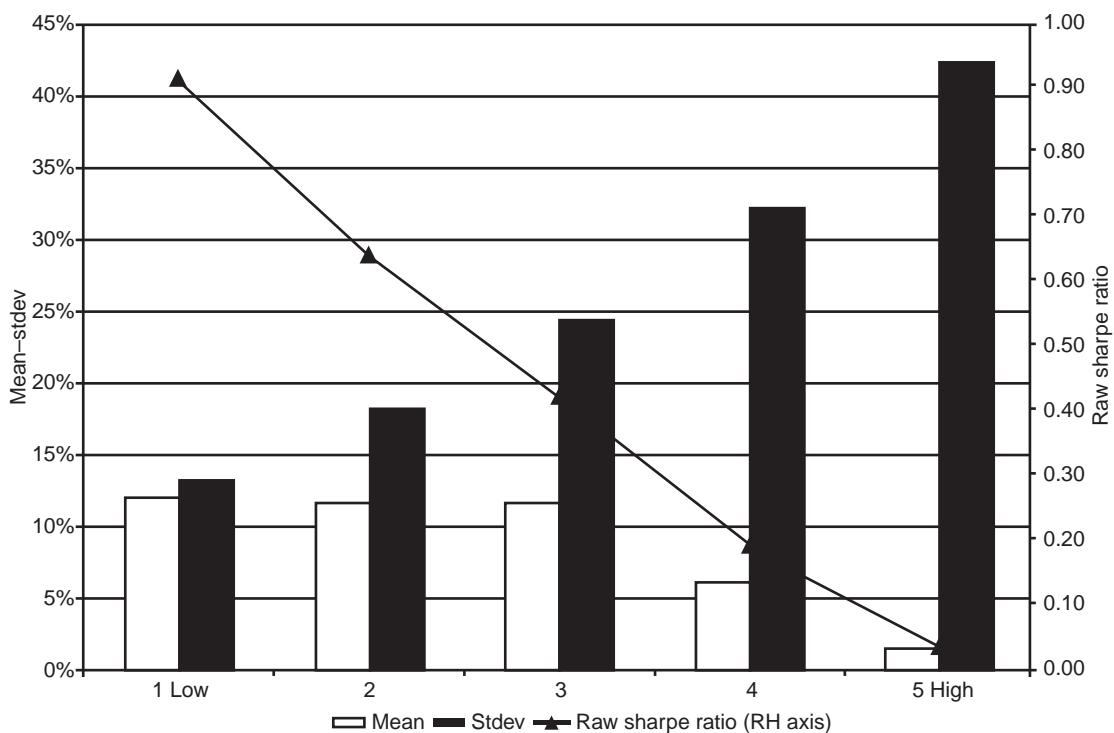


Figure 8.9
Average returns and idiosyncratic volatility contemporaneously measured.

measured contemporaneously. The results are presented, respectively, in Figures 8.8 and 8.9.

At the moment, there is no generally accepted explanation for these results, but they may also be related to the “borrowing constrained–lottery stock” nexus. The resulting mispricing should, however, be arbitrated away by short sellers, yet it is not.

As emphasized most recently in [Stambaugh et al. \(2012\)](#), this failure must be laid at the feet of the various forms of short-selling constraints. Borrowing and short selling constraints thus appear to complement one another in allowing mispricing (at least as defined by the CAPM) to persist for extended periods.²⁰

8.10 Conclusions

The CAPM was the first model to allow financial economists to organize their thoughts on the risk/return trade-off in a systematic way. While originally developed as a descriptive

²⁰ See [Lamont \(2004\)](#), [D’Avolio \(2002\)](#), and [Hong and Sraer \(2012\)](#) for various perspectives on “short selling constraints.”

theory, it has normative implications as well: at a minimum individual investors should be well diversified. This means investing that portion of their wealth designated for risky assets in the best available approximation to the market portfolio. Only in this way can an investor hope to approximate the best risk/return trade-off available to him. This assertion represents solid intuition and there is really no questioning of its legitimacy as a general principle.

The principal conclusion of the CAPM, the SML, is not borne out in direct tests of the data, however. Given the severity of its assumptions and the difficulty in assembling a verifiably accurate approximation to the true M , this outcome is not totally unexpected. The CAPM is also challenged by various anomalies which take the form of other factors that marginalize the market excess return factor in explaining the pattern of excess security returns.

Nevertheless, we are reminded that these anomalies have no meaning apart from the CAPM which provides the benchmark against which they are identified. The same must be said of the research accomplishments that have followed from them. As such the CAPM remains a fundamental device for organizing our thoughts about the relationship of risk and return.

References

- Ang, A., Hodrick, R., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. *J. Finan.* 61, 259–299.
- Ang, A., Hodrick, R., Xing, Y., Zhang, X., 2009. High idiosyncratic volatility and low returns: international and further U.S. evidence. *J. Finan. Econ.* 91, 1–23.
- Banz, R., 1981. The relationship between return and market value of common stocks. *J. Finan. Econ.* 9, 3–18.
- Basu, S., 1983. The relationship between earnings yield, market value, and the return for NYSE common stocks: further evidence. *J. Finan.* 43, 507–528.
- Chan, L., Hamao, Y., Lakonishok, J., 1991. Fundamentals and stock returns in Japan. *J. Finan.* 46, 1739–1789.
- D'Avolio, G., 2002. The market for borrowing stocks. *J. Finan. Econ.* 66, 271–306.
- Fama, E., French, K., 1992. The cross section of expected stock returns. *J. Finan.* 47, 427–465.
- Fama, E., MacBeth, J., 1973. Risk, return and equilibrium empirical tests. *J. Polit. Econ.* 81, 607–636.
- Ferson, W.E., Jagannathan, R., 1996. Econometric evaluation of asset pricing models. In: Maddala, G.S., Rao, C.R. (Eds.), *Statistical Methods in Finance, Handbook of Statistics*. Amsterdam, North Holland, p. 14.
- Frazzini, A., Pederson, L.H., 2013. Betting Against Beta, Working Paper, New York University.
- Friend, I., Blume, M., 1970. The measure of portfolio performance under uncertainty. *Am. Econ. Rev.* 60, 561–575.
- Greenwood, L., Viciera, L., Ang, A., Eysenbach, M., Jacques, B., 2010. Report on the Risk Anomaly, Martingale Asset Management.
- Haugen, R.A., Heins, J., 1975. The risk and rate of return on financial assets: some old wine in new bottles. *J. Finan. Quant. Anal.* 10, 775–784.
- Hirschleifer, D., 1988. Residual risk, trading costs and commodity futures risk premia. *Rev. Finan. Stud.* 1, 173–193.
- Hong, H., Sraer, D. 2012. Speculative Betas, Working Paper, Princeton University.
- Kumar, A., 2009. Who gambles in the stock market. *J. Finan.* 64, 1889–1933.
- Lamont, O., 2004. Going Down Fighting: Short Sellers vs. Firms, Working Paper, Yale University.

- Lintner, J., 1965. The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. *Rev. Econ. Stat.* 47 (1), 13–37.
- Merton, R., 1973. An intertemporal capital asset pricing model. *Econometrica*. 41, 867–887.
- Merton, R., 1987. A simple model of capital market equilibrium with incomplete information. *J. Finan.* 2, 483–510.
- Mossin, J., 1966. Equilibrium in a capital asset market. *Econometrica*. 34 (4), 768–783.
- Reinganum, M., 1981. A new empirical perspective on the CAPM. *J. Finan. Quant. Anal.* 16, 439–462.
- Roll, R., 1977. A critique of the asset pricing theory's test—Part I: on past and potential testability of the theory. *J. Finan. Econ.* 4, 129–176.
- Rosenberg, B., Reid, K., Lanstein, R., 1985. Persuasive evidence of market inefficiency. *J. Portf. Manag.* 11, 9–17.
- Shanken, J., 1996. Statistical methods in tests of portfolio efficiency: a synthesis". In: Maddala, G.S., Rao, C.R. (Eds.), *Statistical Methods in Finance, Handbook of Statistics*. Amsterdam, North Holland, p. 14.
- Sharpe, W.F., 1964. Capital asset prices: a theory of market equilibrium under conditions of risk. *J. Finan.* 19 (3), 425–442.
- Stambaugh, R.F., Yu, J., Yuan, Y., 2012. Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle, NBER Working Paper #18560.
- Stattman, D., 1980. Book values and stock returns, *The Chicago MBA: A Journal of Selected Papers*, 4, 25–45.

Appendix 8.1: Proof of the CAPM Relationship

Refer to [Figure 8.1](#). Consider a portfolio with a fraction $1-\alpha$ of wealth invested in an arbitrary security j and a fraction α in the market portfolio.

$$\begin{aligned} E(\tilde{r}_p) &= \alpha E(\tilde{r}_M) + (1 - \alpha)E(\tilde{r}_j) \\ \sigma_p^2 &= \alpha^2 \sigma_M^2 + (1 - \alpha)^2 \sigma_j^2 + 2\alpha(1 - \alpha)\sigma_{jM} \end{aligned}$$

As α varies we trace a locus that

- passes through M
(and through j)
- cannot cross the CML (why?)
- hence must be tangent to the CML at M

$$\text{Tangency} = \frac{dE(\tilde{r}_p)}{d\sigma_p} |_{\alpha=1} = \text{slope of the locus at } M = \text{slope of CML} = \frac{E(\tilde{r}_M) - r_f}{\sigma_M}$$

$$\frac{dE(\tilde{r}_p)}{d\sigma_p} = \frac{dE(\tilde{r}_p)/d\alpha}{d\sigma_p/d\alpha}$$

$$\frac{dE(\tilde{r}_p)}{d\alpha} = \bar{r}_M - \bar{r}_j$$

$$2\sigma_p \frac{d\sigma_p}{d\alpha} = 2\alpha\sigma_M^2 - 2(1 - \alpha)\sigma_j^2 + 2(1 - 2\alpha)\sigma_{jM}$$

$$\frac{dE(\tilde{r}_p)}{d\sigma_p} = \frac{(E(\tilde{r}_M) - E(\tilde{r}_j))\sigma_p}{\alpha\sigma_M^2 - (1 - \alpha)\sigma_j^2 + (1 - 2\alpha)\sigma_{jM}}$$

$$\frac{dE(\tilde{r}_p)}{d\sigma_p} \Big|_{\alpha=1} = \frac{(E(\tilde{r}_M) - E(\tilde{r}_j))\sigma_M}{\sigma_M^2 - \sigma_{jM}} = \frac{E(\tilde{r}_M) - \bar{r}_f}{\sigma_M}$$

$$(E(\tilde{r}_M) - E(\tilde{r}_j)) = \frac{(E(\tilde{r}_M) - \bar{r}_f)(\sigma_M^2 - \sigma_{jM})}{\sigma_M^2}$$

$$(E(\tilde{r}_M) - E(\tilde{r}_j)) = (E(\tilde{r}_M) - r_f) \left(1 - \frac{\sigma_{jM}}{\sigma_M^2} \right)$$

$$E(\tilde{r}_j) = r_f + (E(\tilde{r}_M) - r_f) \frac{\sigma_{jM}}{\sigma_M^2}$$

Appendix 8.2: The Mathematics of the Portfolio Frontier: An Example

Assume $e = \begin{pmatrix} E(\tilde{r}_1) \\ E(\tilde{r}_2) \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$; $V = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$, i.e., $\rho_{12} = \rho_{21} = -\frac{1}{2}$

Therefore,

$$V^{-1} = \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

check:

$$\begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{4}{3} - \frac{1}{3} & \frac{1}{3} - \frac{1}{3} \\ -\frac{4}{3} + \frac{4}{3} & -\frac{1}{3} + \frac{4}{3} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{aligned} A = 1^T V^{-1} e &= (1 \quad 1) \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \left(\underbrace{\frac{4}{3} + \frac{1}{3}}_{\frac{5}{3}}, \underbrace{\frac{1}{3} + \frac{1}{3}}_{\frac{2}{3}} \right) \begin{pmatrix} 1 \\ 2 \end{pmatrix} \\ &= \frac{5}{3} + 2 \left(\frac{2}{3} \right) = 3 \end{aligned}$$

$$B = e^T V^{-1} e = (1 \quad 2) \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \left(\frac{4}{3} + \frac{2}{3}, \frac{1}{3} + \frac{2}{3} \right) \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 4$$

$$C = 1^T V^{-1} 1 = (1 \quad 1) \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \left(\frac{5}{3}, \frac{2}{3} \right) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{7}{3}$$

$$D = BC - A^2 = 4 \left(\frac{7}{3} \right) - 9 = \frac{28}{3} - \frac{27}{3} = \frac{1}{3}$$

Now we can compute g and h :

$$\begin{aligned} 1. \quad g &= \frac{1}{D} [B(V^{-1} 1) - A(V^{-1} e)] \\ &= \frac{1}{\frac{1}{3}} \left[4 \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - 3 \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right] \\ &= 3 \left[4 \begin{pmatrix} \frac{5}{3} \\ \frac{2}{3} \end{pmatrix} - 3 \begin{pmatrix} \frac{6}{3} \\ \frac{3}{3} \end{pmatrix} \right] = 3 \left[\begin{pmatrix} \frac{20}{3} \\ \frac{8}{3} \end{pmatrix} - \begin{pmatrix} \frac{18}{3} \\ \frac{9}{3} \end{pmatrix} \right] \\ &= \left[\begin{pmatrix} 20 \\ 8 \end{pmatrix} - \begin{pmatrix} 18 \\ 9 \end{pmatrix} \right] = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} 2. \quad h &= \frac{1}{D} [C(V^{-1} e) - A(V^{-1} 1)] \\ &= \frac{1}{\frac{1}{3}} \left[7 \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} - 3 \begin{pmatrix} \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right] \end{aligned}$$

$$= 3 \left[\frac{7}{3} \begin{pmatrix} 2 \\ 1 \end{pmatrix} - 3 \begin{pmatrix} \frac{5}{3} \\ \frac{2}{3} \end{pmatrix} \right] = 7 \begin{pmatrix} 2 \\ 1 \end{pmatrix} - 3 \begin{pmatrix} \frac{5}{3} \\ \frac{5}{3} \end{pmatrix} = \begin{pmatrix} 14 \\ 7 \end{pmatrix} - \begin{pmatrix} 15 \\ 6 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Check by recovering the two initial assets; suppose $E(\tilde{r}_p) = 1$:

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} E(\tilde{r}_p) = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \Rightarrow \text{OK}$$

suppose $E(\tilde{r}_p) = 2$:

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} 2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -2 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \Rightarrow \text{OK}$$

The equation corresponding to Eq. (7.16) thus reads:

$$\begin{pmatrix} w_1^p \\ w_2^p \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} E(\tilde{r}_p)$$

Let us compute the minimum variance portfolio for these assets.

$$E(\tilde{r}_{p,\min \text{ var}}) = \frac{A}{C} = \frac{9}{7}$$

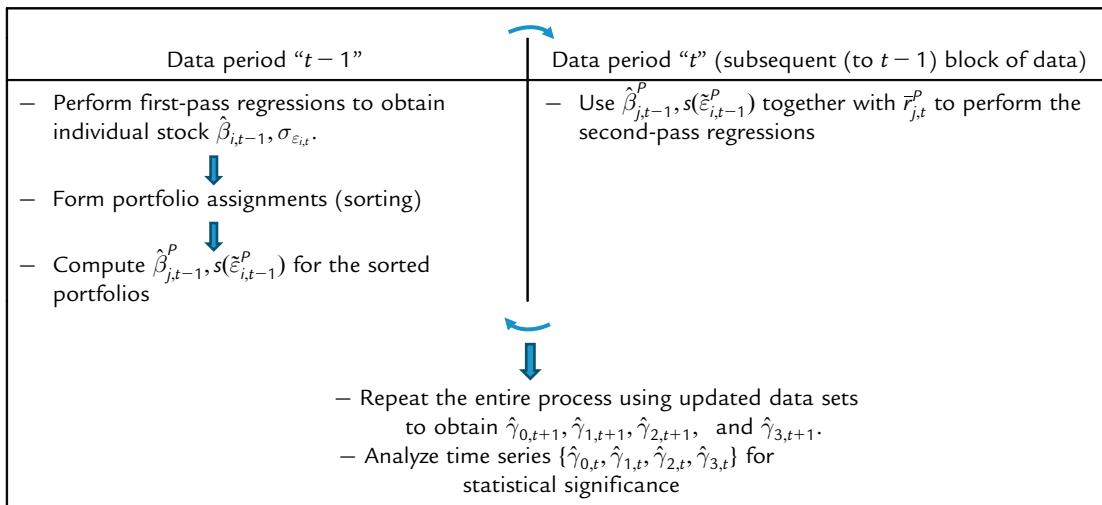
$$\sigma^2(\tilde{r}_{p,\min \text{ var}}) = \frac{1}{C} = \frac{3}{7} < \min\{1, 4\}$$

$$w^p = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \end{pmatrix} \frac{9}{7} = \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \begin{pmatrix} -\frac{9}{7} \\ \frac{9}{7} \end{pmatrix} = \begin{pmatrix} \frac{14}{7} \\ \frac{7}{7} \end{pmatrix} + \begin{pmatrix} -\frac{9}{7} \\ \frac{9}{7} \end{pmatrix} = \begin{pmatrix} \frac{5}{7} \\ \frac{2}{7} \end{pmatrix}$$

Let's check $\sigma^2(\tilde{r}_p)$ by computing it another way:

$$\sigma_p^2 = \begin{pmatrix} \frac{5}{7} & \frac{2}{7} \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix} \begin{pmatrix} \frac{5}{7} \\ \frac{2}{7} \end{pmatrix} = \begin{pmatrix} \frac{3}{7} & \frac{3}{7} \end{pmatrix} \begin{pmatrix} \frac{5}{7} \\ \frac{2}{7} \end{pmatrix} = \frac{3}{7} \Rightarrow \text{OK}$$

Appendix 8.3: Diagrammatic Representation of the Fama–MacBeth Two-Step Procedure



Arrow–Debreu Pricing, Part I

Chapter Outline

- 9.1 Introduction 247**
- 9.2 Setting: An Arrow–Debreu Economy 248**
- 9.3 Competitive Equilibrium and Pareto Optimality Illustrated 250**
- 9.4 Pareto Optimality and Risk Sharing 257**
- 9.5 Implementing Pareto Optimal Allocations: On the Possibility of Market Failure 260**
- 9.6 Risk-Neutral Valuations 263**
- 9.7 Conclusions 266**
- References 267**

9.1 Introduction

As interesting and popular as it is, the CAPM is a very limited theory of equilibrium pricing. We will devote the next chapters to reviewing alternative theories, each of which goes beyond the CAPM in one direction or another. The Arrow–Debreu pricing theory discussed in this chapter is a full general equilibrium theory as opposed to the partial equilibrium static view of the CAPM. Although also static in nature, it is applicable to a multiperiod setup and can be generalized to a broad set of situations. In particular, it is free of any preference restrictions and any distributional assumptions on returns. The Consumption CAPM considered subsequently (Chapter 10) is a fully dynamic construct. It is also an equilibrium theory, though of a somewhat specialized nature. With the Risk-Neutral Valuation Model and the Arbitrage Pricing Theory (APT), taken up in Chapters 12–14, we will move into the domain of arbitrage-based theories, after observing, however, that the Arrow–Debreu pricing theory itself may also be interpreted from the arbitrage perspective (Chapter 11).

The Arrow–Debreu model takes a more standard equilibrium view than the CAPM: it is explicit in stating that equilibrium means supply equals demand in every market. It is a very general theory accommodating production and, as already stated, very broad hypotheses on preferences. Moreover, no restriction on the distribution of returns is necessary. We will not, however, fully exploit the generality of the theory: In keeping with the objective of this text, we shall often limit ourselves to illustrating the theory with examples.

Arrow–Debreu modeling will be especially useful for equilibrium security pricing, especially the pricing of complex securities that pay returns in many different time periods and states of nature, such as common stocks or 30-year government coupon bonds. The theory will also enrich our understanding of project valuation because of the formal equivalence, underlined in Chapter 2, between a project and an asset. In so doing we will move beyond a pure equilibrium analysis and start using the concept of arbitrage. It is in the light of a set of no-arbitrage relationships that the Arrow–Debreu pricing takes its full force. As noted earlier, the arbitrage perspective on the Arrow–Debreu theory will be developed in Chapter 11.

9.2 Setting: An Arrow–Debreu Economy

In the basic setting that we shall use, the following apply:

1. There are two dates: 0, 1. This setup, however, is fully generalizable to multiple periods; see discussion that follows.
2. There are N possible states of nature at date 1, which we index by $\theta = 1, 2, \dots, N$ with probabilities π_θ .¹
3. There is one perishable (nonstorable) consumption good.
4. There are K agents, indexed by $k = 1, \dots, K$, with preferences:

$$U_0^k(c_0^k) + \delta^k \sum_{\theta=1}^N \pi_\theta U^\theta(c_\theta^k);$$

5. Agent k 's endowment is described by the vector $\{e_0^k, (e_\theta^k)_{\theta=1,2,\dots,N}\}$.

In this description, c_θ^k denotes agent k 's consumption of the sole consumption good in state θ , U is the real-valued utility representation of agent k 's period preferences, and δ^k is the agent's time discount factor. In fact, the theory allows for more general preferences than the time-additive expected utility form. Specifically, we could adopt the following representation of preferences:

$$U^k(c_0^k, c_{\theta_1}^k, c_{\theta_2}^k, \dots, c_{\theta_N}^k).$$

This formulation allows not only for a different way of discounting the future (implicit in the relative taste for present consumption relative to all future consumption), but it also permits heterogeneous, subjective views on the state probabilities (again implicit in the

¹ In the present chapter and in most of this text going forward, we assume that all agents hold the same (objective) probability beliefs. Such an assumption is most appropriate to a context where the economy's probabilistic structure over endowments and states does not change, allowing agents to learn the true structure, revising their own beliefs accordingly. In Chapter 18, the topic of heterogeneous beliefs is considered.

representation of relative preference for, say, $c_{\theta_2}^k$ vs $\cdot c_{\theta_3}^k$). In addition, it assumes neither time-additivity nor an expected utility representation. Since our main objective is not generality, we choose to work with the more restrictive but easier to manipulate time-additive expected utility form.

In this economy, the only traded securities are of the following type: One unit of security θ , with price q_θ , pays one unit of consumption if state θ occurs and nothing otherwise.

Its payout can thus be summarized by a vector with all entries equal to zero except for column θ where the entry is 1: $(0, \dots, 0, 1, 0, \dots, 0)$. These primitive securities are called **Arrow–Debreu securities**,² or *state-contingent claims* or simply *state claims*. Of course, the consumption of any individual k if state θ occurs equals the number of units of security θ that he holds. This follows from the fact that buying the relevant contingent claim is the only way for a consumer to secure purchasing power at a future date-state (recall that the good is perishable). An agent's decision problem can then be characterized by:

$$(P) \quad \begin{aligned} & \max_{(c_0^k, c_1^k, \dots, c_N^k)} U_0^k(c_0^k) + \delta^k \sum_{\theta=1}^N \pi_\theta U_\theta^k(c_\theta^k) \\ \text{s.t. } & c_0^k + \sum_{\theta=1}^N q_\theta c_\theta^k \leq e_0^k + \sum_{\theta=1}^N q_\theta e_\theta^k \\ & c_0^k, c_1^k, \dots, c_N^k \geq 0 \end{aligned}$$

The first inequality constraint will typically hold with equality in a world of nonsatiation. That is, the total value of goods and security purchases made by the agent (the left-hand side of the inequality) will exhaust the total value of his endowments (the right-hand side).

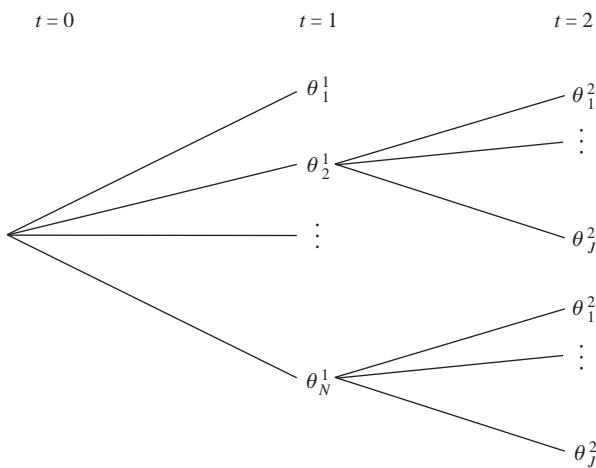
Equilibrium for this economy is a set of contingent claim prices (q_1, q_2, \dots, q_N) such that

1. at those prices (c_0^k, \dots, c_N^k) solve problem (P), for all k , and
2. $\sum_{k=1}^K c_0^k = \sum_{k=1}^K e_0^k$, $\sum_{k=1}^K c_\theta^k = \sum_{k=1}^K e_\theta^k$, for every θ .

Note that here the agents are solving for desired future and present consumption holdings rather than holdings of Arrow–Debreu securities. This is justified because, as just noted, there is a one-to-one relationship between the amount consumed by an individual in a given state θ and his holdings of the Arrow–Debreu security corresponding to that particular state θ , the latter being a promise to deliver one unit of the consumption good if that state occurs.

Note also that nothing in this formulation inherently restricts matters to two periods, if we define our notion of a state somewhat more richly, as a *date-state* pair. Consider three

² So named after the originators of modern equilibrium theory: see Arrow (1951) and Debreu (1959).

**Figure 9.1**

The structure of an economy with two dates and $(1 + NJ)$ states.

periods, for example. There are N possible states in date 1 and J possible states in date 2, regardless of the state achieved in date 1. Define $\hat{\theta}$ new states to be of the form $\hat{\theta}_s = (k, \theta_k^j)$, where k denotes the state in date 1 and θ_k^j denotes the state j in date 2, conditional that state k was observed in date 1 (refer to Figure 9.1). So $(1, \theta_1^5)$ would be a state and $(2, \theta_2^3)$ another state. Under this interpretation, the number of *states* expands to $1 + NJ$, with:

- 1: the date 0 state
- N : the number of date 1 states
- J : the number of date 2 states

With minor modifications, we can thus accommodate many periods and states. In this sense, our model is fully general and can represent as complex an environment as we might desire. In this model, the real productive side of the economy is in the background. We are, in effect, viewing that part of the economy as invariant to securities trading. The unusual and unrealistic aspect of this economy is that all trades occur at $t = 0$.³ We will relax this assumption in Chapter 10.

9.3 Competitive Equilibrium and Pareto Optimality Illustrated

Let us now develop an example. The essentials are found in Table 9.1.

There are two dates and, at the future date, two possible states of nature with probabilities $\frac{1}{3}$ and $\frac{2}{3}$. It is an exchange economy, and the issue is to share the existing endowments

³ Interestingly, this is less of a restriction for project valuation than for asset pricing.

Table 9.1: Endowments and preferences in our reference example

Agents	Endowments			Preferences	
	t = 0	t = 1			
		θ_1	θ_2		
Agent 1	10	1	2	$\frac{1}{2}c_0^1 + 0.9\left(\frac{1}{3}\ln(c_1^1) + \frac{2}{3}\ln(c_2^1)\right)$	
Agent 2	5	4	6	$\frac{1}{2}c_0^2 + 0.9\left(\frac{1}{3}\ln(c_1^2) + \frac{2}{3}\ln(c_2^2)\right)$	

between two individuals. Their (identical) preferences are linear in date 0 consumption, with constant marginal utility equal to $\frac{1}{2}$. This choice is made for ease of computation, but great care must be exercised in interpreting the results obtained in such a simplified framework. Date 1 preferences are concave and identical. The discount factor is 0.9.

Let q_1 be the price of a unit of consumption in date 1 state 1, and q_2 the price of one unit of the consumption good in date 1 state 2. We will solve for optimal consumption directly, knowing that this will define the equilibrium holdings of the securities. The prices of these consumption goods coincide with the prices of the corresponding state-contingent claims; period 0 consumption is taken as the numeraire, and its price is normalized to 1. This means that all prices are expressed in units of period 0 consumption: q_1, q_2 are prices for the consumption good at date 1, in states 1 and 2, respectively, measured in units of date 0 consumption. They can thus be used to add up or compare units of consumption at different dates and in different states, making it possible to add different date cash flows, with the q_i being the appropriate weights. This, in turn, permits computing an individual's wealth. Thus, in the previous problem, agent 1's wealth, which equals the present value of his current and future endowments, is $10 + 1q_1 + 2q_2$, while agent 2's wealth is $5 + 4q_1 + 6q_2$.

The respective agent problems are:

Agent 1.

$$\begin{aligned} & \max \frac{1}{2}(10 + 1q_1 + 2q_2 - c_1^1 q_1 - c_2^1 q_2) + 0.9\left(\frac{1}{3}\ln(c_1^1) + \frac{2}{3}\ln(c_2^1)\right) \\ & \text{s.t. } c_1^1 q_1 + c_2^1 q_2 \leq 10 + q_1 + 2q_2, \text{ and } c_1^1, c_2^1 \geq 0 \end{aligned}$$

Agent 2.

$$\begin{aligned} & \max \frac{1}{2}(5 + 4q_1 + 6q_2 - c_1^2 q_1 - c_2^2 q_2) + 0.9\left(\frac{1}{3}\ln(c_1^2) + \frac{2}{3}\ln(c_2^2)\right) \\ & \text{s.t. } c_1^2 q_1 + c_2^2 q_2 \leq 5 + 4q_1 + 6q_2 \text{ and } c_1^2, c_2^2 \geq 0 \end{aligned}$$

Note that in this formation, we have substituted out for the date 0 consumption; in other words, the first term in the max expression stands for $\frac{1}{2}(c_0)$, where we have substituted

for c_0 its value obtained from the constraint: $c_0 + c_1^1 q_1 + c_2^1 q_2 = 10 + 1q_1 + 2q_2$. With this trick, the only constraints remaining are the nonnegativity constraints requiring consumption to be nonnegative in all date-states.

The FOCs state that the intertemporal rate of substitution between future (in either state) and present consumption (i.e., the ratio of the relevant marginal utilities) should equal the price ratio. The latter is effectively measured by the price of the Arrow–Debreu security, the date 0 price of consumption being the numeraire. These first order conditions (FOCs) are (assuming interior solutions)

$$\text{Agent 1: } \begin{cases} c_1^1: \frac{q_1}{2} = 0.9 \left(\frac{1}{3} \right) \frac{1}{c_1^1} \\ c_2^1: \frac{q_2}{2} = 0.9 \left(\frac{2}{3} \right) \frac{1}{c_2^1} \end{cases} \quad \text{Agent 2: } \begin{cases} c_1^2: \frac{q_1}{2} = 0.9 \left(\frac{1}{3} \right) \frac{1}{c_1^2} \\ c_2^2: \frac{q_2}{2} = 0.9 \left(\frac{2}{3} \right) \frac{1}{c_2^2} \end{cases}$$

while the market-clearing conditions read: $c_1^1 + c_1^2 = 5$ and $c_2^1 + c_2^2 = 8$. Each of the FOCs is of the form $(q_\theta/1) = ((0.9)\pi_\theta(1/c_\theta^k))/(1/2)$, $k, \theta = 1, 2$, or

$$q_\theta = \frac{\delta\pi_\theta \frac{\partial U^k}{\partial c_\theta^k}}{\frac{\partial U_0^k}{\partial c_0^k}}, \quad k, \theta = 1, 2. \quad (9.1)$$

Together with the market-clearing conditions, Eq. (9.1) reveals the determinants of the equilibrium Arrow–Debreu security prices. It is of the form:

$$\frac{\text{Price of the good if state } \theta \text{ is realized}}{\text{Price of the good today}} = \frac{MU_\theta^k}{MU_0^k}.$$

In other words, the ratio of the price of the Arrow–Debreu security to the price of the date 0 consumption good must equal (at an interior solution; see Box 9.1) the ratio of the marginal utility of consumption tomorrow if state θ is realized to the marginal utility of today's consumption (the latter being constant at $\frac{1}{2}$). This is the *marginal rate of substitution* (MRS) between the contingent consumption in state θ and today's consumption. From this system of equations, one clearly obtains $c_1^1 = c_1^2 = 2.5$ and $c_2^1 = c_2^2 = 4$ from which one, in turn, derives:

$$q_1 = \frac{1}{\frac{1}{2}} (0.9) \left(\frac{1}{3} \right) \left(\frac{1}{c_1^1} \right) = 2(0.9) \left(\frac{1}{3} \right) \left(\frac{1}{2.5} \right) = (0.9) \left(\frac{1}{3} \right) \left(\frac{4}{5} \right) = 0.24$$

$$q_2 = \frac{1}{\frac{1}{2}} (0.9) \left(\frac{2}{3} \right) \left(\frac{1}{c_2^1} \right) = 2(0.9) \left(\frac{2}{3} \right) \left(\frac{1}{4} \right) = (0.9) \left(\frac{2}{3} \right) \left(\frac{4}{8} \right) = 0.3$$

BOX 9.1 Interior Versus Corner Solutions

We have described the *interior* solution to the maximization problem. By that restriction we generally mean the following: the problem under maximization is constrained by the condition that consumption at all dates should be nonnegative. There is no interpretation given to a negative level of consumption, and, generally, even a zero consumption level is precluded. Indeed, when we make the assumption of a log utility function, the marginal utility at zero is infinity, meaning that by construction the agent will do all that is in his power to avoid that situation. Effectively, an equation such as Eq. (9.1) will never be satisfied for finite and nonzero prices with log utility and period one consumption level equal to zero; that is, it will never be optimal to select a zero consumption level. Such is not the case with the linear utility function assumed to prevail at date 0. Here it is conceivable that, no matter what, the marginal utility in either state at date 1 (the numerator in the RHS of Eq. (9.1)) will be larger than $\frac{1}{2}$ times the Arrow–Debreu price (the denominator of the RHS in Eq. (9.1) multiplied by the state price). Intuitively, this would be a situation where the agent derives more utility from the good tomorrow than from consuming today, even when his consumption level today is zero. Fundamentally, the interior optimum is one where he would like to consume less than zero today to increase even further consumption tomorrow, something that is impossible. Thus the only solution is at a corner, that is, at the boundary of the feasible set, with $c_0^k = 0$ and the condition in Eq. (9.1) taking the form of an inequality.

In the present case, we can argue that corner solutions cannot occur with regard to future consumption (because of the log utility assumption). The full and complete description of the FOCs for problem (P) spelled out in Section 9.2 is then

$$q_\theta \frac{\partial U_0^k}{\partial c_0^k} \leq \delta \pi_\theta \frac{\partial U^k}{\partial c_\theta^k}, \text{ and if } c_0^k > 0, \text{ and } k, \theta = 1, 2. \quad (9.2)$$

In line with our goal of being as transparent as possible, we will often, in the sequel, satisfy ourselves with a description of interior solutions to optimizing problems, taking care to ascertain, ex-post, that the solutions do indeed occur at the interior of the choice set. This can be done in the present case by verifying that the optimal c_0^k is strictly positive for both agents at the interior solutions, so that Eq. (9.1) must indeed apply.

Notice how the Arrow–Debreu state-contingent prices reflect probabilities, on the one hand, and marginal rates of substitution (taking the time discount factor into account and computed at consumption levels compatible with market clearing) and thus relative scarcities, on the other. The prices computed above differ in that they take account both of the different state probabilities ($\frac{1}{3}$ for state 1, $\frac{2}{3}$ for state 2) and the differing marginal utilities as a result of the differing total quantities of the consumption good available in state 1 (5 units) and in state 2 (8 units). In our particular formulation, the total amount of goods available at date 0 is made irrelevant by the fact that date 0 marginal utility is constant. Note that if the date 1 marginal utilities were constant, as would be the case with

Table 9.2: Post-trade equilibrium consumptions

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	9.04	2.5	4
Agent 2	5.96	2.5	4
Total	15.00	5.0	8

linear (risk-neutral) utility functions, the goods endowments would not influence the Arrow–Debreu prices, which would then be exactly proportional to the state probabilities.

The date 0 consumptions, at the equilibrium prices, are given by

$$\begin{aligned} c_0^1 &= 10 + 1(0.24) + 2(0.3) - 2.5(0.24) - 4(0.3) = 9.04 \\ c_0^2 &= 5 + 4(0.24) + 6(0.3) - 2.5(0.24) - 4(0.3) = 5.96 \end{aligned}$$

The post-trade equilibrium consumptions are found in [Table 9.2](#).

This allocation is the best each agent can achieve at the equilibrium prices $q_1 = 0.24$ and $q_2 = 0.3$. Furthermore, at those prices, supply equals demand in each market, in every state and time period. These are the characteristics of a (general) competitive equilibrium.

In light of this example, it is interesting to return to some of the concepts discussed in our introductory chapter. In particular, let us confirm the (Pareto) optimality of the allocation emerging from the competitive equilibrium. Indeed, we have assumed as many markets as there are states of nature, so assumption H1 is satisfied. We have *de facto* assumed competitive behavior on the part of our two consumers (they have taken prices as given when solving their optimization problems), so H2 is satisfied. (Of course, in reality such behavior would not be privately optimal if indeed there were only two agents. Our example would not have changed materially had we assumed a large number of agents, but the notation would have become much more cumbersome.)

In order to guarantee the existence of an equilibrium, we need hypotheses H3 and H4 as well. H3 is satisfied in a weak form (no curvature in date 0 utility). Finally, ours is an exchange economy where H4 does not apply (or, if one prefers, it is trivially satisfied). Once the equilibrium is known to exist, as is the case here, H1 and H2 are sufficient to guarantee the optimality of the resulting allocation of resources. Thus, we expect to find that the above competitive allocation is Pareto optimal (PO); that is, it is impossible to

rearrange the allocation of consumptions so that the utility of one agent is higher without diminishing the utility of the other agent.

One way to verify the optimality of the competitive allocation is to establish the precise conditions that must be satisfied for an allocation to be PO in the exchange economy context of our example. It is intuitively clear that the above Pareto superior real allocations will be impossible if the initial allocation maximizes the weighted sum of the two agents' utilities. That is, an allocation is optimal in our example if, for some weight λ it solves the following maximization problem.⁴

$$\begin{aligned} \max & U^1(c_0^1, c_1^1, c_2^1) + \lambda U^2(c_0^2, c_1^2, c_2^2) \\ & \{c_0^1, c_1^1, c_2^1\} \\ \text{s.t.} & \\ & c_0^1 + c_0^2 = 15; \quad c_1^1 + c_1^2 = 5; \quad c_2^1 + c_2^2 = 8, \\ & c_0^1, c_1^1, c_2^1, c_0^2, c_1^2, c_2^2 \geq 0 \end{aligned}$$

This problem can be interpreted as the problem of a benevolent central planner constrained by an economy's total endowment (15, 5, 8) and weighting the two agents' utilities according to a parameter λ , possibly equal to 1. The decision variables at his disposal are the consumption levels of the two agents in the two dates and the two states. With U_i^k denoting the derivative of agent k 's utility function with respect to c_i^k ($i = 1, 2, 3$), the FOCs for an interior solution to the above problem are found in Eq. (9.3).

$$\frac{U_0^1}{U_0^2} = \frac{U_1^1}{U_1^2} = \frac{U_2^1}{U_2^2} = \lambda \quad (9.3)$$

This condition states that, in a PO allocation, the ratio of the two agents' marginal utilities with respect to the three goods (i.e., the consumption good at date 0, the consumption good at date 1 if state 1, and the consumption good at date 1 if state 2) should be identical.⁵ In an exchange economy this condition, properly extended to take account of the possibility of a corner solution, together with the condition that the agents' consumption adds up to the endowment in each date-state, is necessary and sufficient.

⁴ In this discussion it is just as easy to work with the most general utility representation.

⁵ Check that Eq. (9.3) implies that the MRS between any two pair of goods is the same for the two agents and refer to the definition of the contract curve (the set of PO allocations) in the Appendix to Chapter 1.

It remains to check that Eq. (9.3) is satisfied at the equilibrium allocation. We can rewrite Eq. (9.3) for the parameters of our example:

$$\frac{\frac{1}{2}}{\frac{1}{2}} = \frac{(0.9)\frac{1}{3}\frac{1}{c_1^1}}{(0.9)\frac{1}{3}\frac{1}{c_1^2}} = \frac{(0.9)\frac{2}{3}\frac{1}{c_2^1}}{(0.9)\frac{2}{3}\frac{1}{c_2^2}}$$

It is clear that the condition in Eq. (9.3) is satisfied since $c_1^1 = c_1^2; c_2^1 = c_2^2$ at the competitive equilibrium, which thus corresponds to the Pareto optimum with equal weighting of the two agents' utilities: $\lambda = 1$, and all three ratios of marginal utilities are equal to 1. Note that other Pareto optima are feasible, for example, one where $\lambda = 2$. In that case, however, only the latter two equalities can be satisfied: the date 0 marginal utilities are constant, which implies that no matter how agent consumptions are redistributed by the market or by the central planner, the first ratio of marginal utilities in Eq. (9.3) cannot be made equal to 2. This is an example of a corner solution to the maximization problem leading to Eq. (9.3).

In our example, agents are able to purchase consumption in any date-state of nature. This is the case because there are enough Arrow–Debreu securities; specifically, there is an Arrow–Debreu security corresponding to each state of nature. If this were not the case, the attainable utility levels would decrease: at least one agent, possibly both of them, would be worse off. If we assume that only the state 1 Arrow–Debreu security is available, then there is no way to make the state 2 consumption of the agents differ from their endowments. It is easy to check that this constraint does not modify their demand for the state 1 contingent claim, nor its price. The post-trade allocation, in that situation, is found in Table 9.3.

The resulting post-trade utilities are

$$\text{Agent 1: } \frac{1}{2}(9.64) + 0.9\left(\frac{1}{3} \ln(2.5) + \frac{2}{3} \ln(2)\right) = 5.51$$

$$\text{Agent 2: } \frac{1}{2}(5.36) + 0.9\left(\frac{1}{3} \ln(2.5) + \frac{2}{3} \ln(6)\right) = 4.03$$

In the case with two state-contingent claim markets, the post-trade utilities are both higher (illustrating a reallocation of resources that is said to be *Pareto superior* to the no-trade allocation):

$$\text{Agent 1: } \frac{1}{2}(9.04) + 0.9\left(\frac{1}{3} \ln(2.5) + \frac{2}{3} \ln(4)\right) = 5.62$$

Table 9.3: The post-trade allocation

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	9.64	2.5	2
Agent 2	5.36	2.5	6
Total	15.00	5.0	8

$$\text{Agent 2: } \frac{1}{2}(5.96) + 0.9\left(\frac{1}{3}\ln(2.5) + \frac{2}{3}\ln(4)\right) = 4.09$$

When there is an Arrow–Debreu security available to trade corresponding to each state of nature, one says that the securities markets are complete.

9.4 Pareto Optimality and Risk Sharing

In this section and the next, we further explore the nexus between a competitive equilibrium in an Arrow–Debreu economy and Pareto optimality. We first discuss the risk-sharing properties of a PO allocation. We remain in the general framework of the example of the previous two sections but start with a different set of parameters. In particular, let the endowment matrix for the two agents be as shown in [Table 9.4](#).

Assume further that each state is now equally likely with probability $\frac{1}{2}$. As before, consumption in period 0 cannot be stored and carried over into period 1. In the absence of trade, agents clearly experience widely differing consumption and utility levels in period 1, depending on what state occurs (see [Table 9.5](#)).

How could agents' utilities be improved? By concavity (risk aversion), this must be accomplished by reducing the spread of the date 1 income possibilities, in other words, lowering the risk associated with date 1 income. Because of symmetry, all date 1 income fluctuations can, in fact, be eliminated if agent 2 agrees to transfer two units of the good in state 1 against the promise to receive two units from agent 1 if state 2 is realized (see [Table 9.6](#)).

Table 9.4: The new endowment matrix

		$t = 0$	$t = 1$	
			θ_1	θ_2
Agent 1	4	1	5	
	4	5	1	

Table 9.5: Agents' utility in the absence of trade

	State-Contingent Utility		Expected Utility in Period 1
	θ_1	θ_2	
Agent 1	$\ln(1) = 0$	$\ln(5) = 1.609$	$\frac{1}{2}\ln(1) + \frac{1}{2}\ln(5) = 0.8047$
Agent 2	$\ln(5) = 1.609$	$\ln(1) = 0$	$\frac{1}{2}\ln(1) + \frac{1}{2}\ln(5) = 0.8047$

Table 9.6: The desirable trades and post-trade consumptions

Date 1	Endowments Pre-trade		Consumption Post-trade	
	θ_1	θ_2	θ_1	θ_2
Agent 1	1	5 [$\downarrow 2$]	3	3
Agent 2	5 [$\uparrow 2$]	1	3	3

Now we can compare expected second-period utility levels before and after trade for both agents:

Before	After
0.8047	$\frac{1}{2} \ln(3) + \frac{1}{2} \ln(3) = 1.099 \cong 1.1$

In other words, expected utility has increased quite significantly, as anticipated.⁶

This feasible allocation is, in fact, *Pareto optimal*. In conformity with Eq. (9.3), the ratios of the two agents' marginal utilities are indeed equalized across states. More is accomplished in this perfectly symmetrical and equitable allocation: consumption levels and MU are equated across agents and states, but this is a coincidence resulting from the symmetry of the initial endowments.

Suppose the initial allocation is the one identified in Table 9.7.

Once again there is no aggregate risk: The total date 1 endowment is the same in the two states, but one agent is now richer than the other. Now consider the plausible trade outlined in Table 9.8.

Check that the new post-trade allocation is also PO. Although consumption levels and marginal utilities are not identical, the ratio of marginal utilities is the same across states (except at date 0 where, as before, we have a corner solution since the marginal utilities are given constants). Note that this PO allocation features perfect risk sharing as well. By that we mean that the two agents have constant date 1 consumption (two units for agent 1, four units for agent 2) independent of the realized state. This is a general characteristic of PO allocations in the absence of aggregate risk (and with risk-averse agents). If there is no aggregate risk, all PO allocations necessarily feature full mutual insurance.

This statement can be demonstrated, using the data of our problem. Equation (9.3) states that the ratio of the two agents' marginal utilities should be equated across states. This also

⁶ With the specified utility function, expected utility has increased by 37%. Such quantification is not, however, compatible with the observation that expected utility functions are defined only up to a linear transformation. Instead of using $\ln c$ for the period utility function, we could equally well have used $(b + \ln c)$ to represent the same preference ordering. The quantification of the increase in utility pre- and post-trade would be affected.

Table 9.7: Another set of initial allocations

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	4	1	3
Agent 2	4	5	3

Table 9.8: Plausible trades and post-trade consumptions

Date 1	Endowments Pre-trade		Consumption Post-trade	
	θ_1	θ_2	θ_1	θ_2
Agent 1	1	3 [$\downarrow 1$]	2	2
Agent 2	5 [$\uparrow 1$]	3	4	4

implies, however, that the MRS between state 1 and state 2 consumption must be the same for the two agents. In the case of log period utility:

$$\frac{1/c_1^1}{1/c_1^2} = \frac{1/c_2^1}{1/c_2^2} \Leftrightarrow \frac{1/c_1^1}{1/c_2^1} = \frac{1/c_1^2}{1/c_2^2}$$

The latter equality has the following implications:

1. If one of the two agents is fully insured—no variation in his date 1 consumption (i.e., MRS = 1)—the other must be as well.
2. More generally, if the MRS are to differ from 1, given that they must be equal between them, the low consumption–high MU state must be the same for both agents and similarly for the high consumption–low MU state. But this is impossible if there is no aggregate risk and total endowment is constant. Thus, as asserted, in the absence of aggregate risk, a PO allocation features perfectly insured individuals and MRS identically equal to 1.
3. If there is aggregate risk, however, the above reasoning also implies that, at a Pareto optimum, it is shared “proportionately.” This is literally true if agents’ preferences are homogeneous. Refer to the competitive equilibrium of [Section 9.3](#) for an illustration.
4. Finally, if agents are differentially risk averse, in a PO allocation the less risk averse will typically provide some insurance services to the more risk averse. This is most easily illustrated by assuming that one of the two agents, say agent 1, is risk neutral. By risk neutrality, agent one’s marginal utility is constant. But then the marginal utility of agent 2 should also be constant across states. For this to be the case, however, agent 2’s income uncertainty must be fully absorbed by agent 1, the risk-neutral agent.
5. More generally, optimal risk sharing dictates that the agent most tolerant of risk bears a disproportionate share of it.

9.5 Implementing PO Allocations: On the Possibility of Market Failure

Although to achieve the desired allocations, the agents of our previous section could just effect a handshake trade, real economic agents typically interact only through impersonal security markets or through deals involving financial intermediaries. One reason for this practice is that, in an organized security market, the contracts implied by the purchase or sale of a security are enforceable. This is important: without an enforceable contract, if state 1 occurs, agent 2 might retreat from his ex-ante commitment and refuse to give up the promised consumption to agent 1, and vice versa if state 2 occurs. Accordingly, we now address the following question: What securities could empower these agents to achieve the optimal allocation for themselves?

Consider the Arrow–Debreu security with payoff in state 1 and call it security Q to clarify the notation below. Denote its price by q_Q , and let us compute the demand by each agent for this security denoted $z_Q^i, i = 1, 2$. The price is expressed in terms of period 0 consumption. We otherwise maintain the setup of the preceding section. Thus,

$$\text{Agent 1 solves: } \max(4 - q_Q z_Q^1) + [\frac{1}{2} \ln(1 + z_Q^1) + \frac{1}{2} \ln(5)]$$

$$\text{s.t. } q_Q z_Q^1 \leq 4$$

$$\text{Agent 2 solves: } \max(4 - q_Q z_Q^2) + [\frac{1}{2} \ln(5 + z_Q^2) + \frac{1}{2} \ln(1)]$$

$$\text{s.t. } q_Q z_Q^2 \leq 4$$

Assuming an interior solution, the FOCs are, respectively,

$$-q_Q + \frac{1}{2} \left(\frac{1}{1 + z_Q^1} \right) = 0; \quad -q_Q + \frac{1}{2} \left(\frac{1}{5 + z_Q^2} \right) = 0 \Rightarrow \frac{1}{1 + z_Q^1} = \frac{1}{5 + z_Q^2};$$

also $z_Q^1 + z_Q^2 = 0$ in equilibrium, hence, $z_Q^1 = 2; z_Q^2 = -2$; these represent the holdings of each agent and $q_Q = (\frac{1}{2})(\frac{1}{3}) = \frac{1}{6}$. In effect, agent 1 gives up $q_Q z_Q^1 = (\frac{1}{6})(2) = \frac{1}{3}$ unit of consumption at date 0 to agent 2 in exchange for 2 units of consumption at date 1 if state 1 occurs. Both agents are better off as revealed by the computation of their expected utilities post-trade:

$$\text{Agent 1 expected utility: } 4 - \frac{1}{3} + \frac{1}{2} \ln 3 + \frac{1}{2} \ln 5 = 5.013$$

$$\text{Agent 2 expected utility: } 4 + \frac{1}{3} + \frac{1}{2} \ln 3 + \frac{1}{2} \ln 1 = 4.879,$$

Table 9.9: Market allocation when both securities are traded

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	4	3	3
Agent 2	4	3	3

though agent 2 only slightly so. Clearly agent 1 is made proportionately better off because security Q pays off in the state where his MU is highest. We may view agent 2 as the issuer of this security as it entails, for him, a future obligation.⁷

Let us denote R the other conceivable Arrow–Debreu security, one paying in state 2. By symmetry, it would also have a price of $\frac{1}{6}$, and the demand at this price would be $z_R^1 = -2$, $z_R^2 = +2$, respectively. Agent 2 would give up $\frac{1}{3}$ unit of period 1 consumption to agent 1 in exchange for 2 units of consumption in state 2.

Thus, if both security Q and R are traded, the market allocation will replicate the optimal allocation of risks, as seen in [Table 9.9](#).

In general, it is possible to achieve the optimal allocation of risks provided the number of linearly independent securities equals the number of states of nature. By linearly independent we mean, again, that there is no security whose payoff pattern across states and time periods can be duplicated by a portfolio of other securities. This important topic will be discussed at length in Chapter 11. Here let us simply take stock of the fact that our securities Q , R are the simplest pair of securities with this property.

Although a complete set of Arrow–Debreu securities is sufficient for optimal risk sharing, it is not necessary in the sense that it is possible, by coincidence, for the desirable trades to be effected with a simplified asset structure. For our simple example, one security would allow the agents to achieve that goal because of the essential symmetry of the problem. Consider security Z with payoffs:

Z	θ_1	θ_2
	2	-2

Clearly, if agent 1 purchases one unit of this security ($z_Z^1 = 1$) and agent 2 sells one unit of this security ($z_Z^2 = -1$), optimal risk sharing is achieved. (At what price would this security sell?)

⁷ In a noncompetitive situation, it is likely that agent 2 could extract a larger portion of the rent. Remember, however, that we maintain, throughout, the assumption of price-taking behavior for our two agents who are representatives of larger classes of similar individuals.

So far we have implicitly assumed that the creation of these securities is costless. In reality, the creation of a new security is an expensive proposition: disclosure documents, promotional materials, and so on, must be created, and the agents most likely to be interested in the security contacted. In this example, issuance will occur only if the cost of issuing Q and R does not exceed the (expected) utility gained from purchasing them. In this margin lies the investment banker's fee.

In the previous discussion we imagined each agent as issuing securities to the other simultaneously. More realistically, perhaps, we could think of the securities Q and R as being issued in sequence, one after the other (but both before period 1 uncertainty is resolved). Is there an advantage or disadvantage of going first, that is, of issuing the *first* security? Alternatively, we might be concerned with the fact that, although both agents benefit from the issuance of new securities, only the individual issuer pays the cost of establishing a new market. From this perspective, it is interesting to measure the net gains to trade for each agent. These quantities are summarized in [Table 9.10](#).

In our example, this computation tells us that the issuer of the security gains less than the other party in the future trade. If agent 2 goes first and issues security Q , his net expected utility gain is 0.0783, which also represents the most he would be willing to pay his investment bank in terms of period 0 consumption to manage the sale for him. By analogy, the marginal benefit to agent 1 of *then* issuing security R is 0.0780. The reverse assignments would have occurred if agent 1 had gone first, due to symmetry in the agent endowments. That these quantities represent the upper bounds on possible fees comes from the fact that period 0 utility of consumption is the level of consumption itself.

The impact of all this is that each investment bank will, out of desire to maximize its fee potential, advise its client to issue his security second. No one will want to go first. Alternatively, if the effective cost of setting up the market for security Q is anywhere between 0.0783 and 0.2936, there is a possibility of *market failure*, unless agent 2 finds a

**Table 9.10: The net gains from trade: expected utility levels and net trading gains
(Gain to issuer in bold)**

	No Trade	Trade Only Q		Trade Both Q and R	
	EU	EU	ΔEU^a	EU	ΔEU^b
Agent 1	4.8047	5.0206	0.2159	5.0986	0.0780
Agent 2	4.8047	4.883	0.0783	5.0986	0.2156
Total			0.2942		0.2936

^aDifference in EU when trading Q only, relative to no trade.

^bDifference in EU when trading both Q and R , relative to trading Q only.

way to have agent 1 share in the cost of establishing the market. We speak of market failure because the social benefit of setting up the market would be positive 0.2936 minus the cost itself—while the market might not go ahead if the private cost to agent 2 exceeds his own private benefit, measured at 0.0783 units of date 0 consumption. Of course, it might also be the case that the cost exceeds the total benefit. This is another reason for the market not to exist and, in general, for markets to be incomplete. But in this case, one would not talk of market failure. Whether the privately motivated decisions of individual agents lead to the socially optimal outcome—in this case the socially optimal set of securities—is a fundamental question in financial economics.

There is no guarantee that private incentives will suffice to create the social optimal set of markets. We have identified a problem of sequencing—the issuer of a security may not be the greatest beneficiary of the creation of the market—and as a result there may be a waiting game with suboptimal consequences. There is also a problem linked with sharing the cost of setting up a market. The benefits of a new market often are widely spread across a large number of potential participants, and it may be difficult to find an appropriate mechanism to have them share the initial setup cost, for example because of free rider or coordination problems. Note that in both cases, as well as in the situation where the cost of establishing a market exceeds the total benefit for individual agents, we anticipate that technical innovations leading to decreases in the cost of establishing markets will help alleviate the problem and foster a convergence toward a more complete set of markets.⁸

9.6 Risk-Neutral Valuations

The Arrow–Debreu pricing perspective allows us to introduce the “distorted probabilities” approach to asset valuation first mentioned in Chapter 2. In general, this asset pricing perspective goes under the title of “risk-neutral valuation.” While our introduction to the idea in the paragraphs below may appear at first to be nothing more than a clever algebraic manipulation—a “trick”—it is a concept that will turn out to be of paramount usefulness, facilitating such diverse activities as options pricing and the empirical assessment of our Arrow–Debreu cum VNM-expected utility asset pricing theory.

⁸ The larger issue here is whether making markets more complete uniformly improves agent welfare. By this we mean the following: consider our two period multi date-state setting and assume there is one state for which no state claim is traded; in that state all agents consume their endowments. Now open up claims trading to include one of the previously excluded states. Will the welfare of all the agents be improved? There is no guarantee this will happen. Transfers among agents could be affected, however, that would increase the welfare of everyone. But some agents may not agree to participate! See the Web Notes to this chapter for illustrative examples.

Table 9.11: Security payoffs and equilibrium date-state security prices

$t = 0$	$t = 1$		
	θ_1	θ_2	θ_3
$q_1 = 0.5$	1	0	0
$q_2 = 0.1$	0	1	0
$q_3 = 0.3$	0	0	1
$q^b = ?$	1	1	1
$q^e = ?$	2	1	4

While risk-neutral valuation does not require a VNM-expected utility preference specification (agents may be assumed to have ordinal representations $u^k(c_1^k, c_2^k, c_{\theta_1}^k, \dots, c_{\theta_N}^k)$, as noted in [Section 9.2](#)), it does demand that financial markets are complete: for every future date-state there must exist a traded security that pays exclusively in that date-state. In our discussion below we take these prices as given, side-stepping their equilibrium determination, in order to focus on the pricing of more complex, multistate cash flows. For clarity and simplicity we retain the two period setting of the present chapter; multiperiod generalizations are treated in later chapters.

Consider the simple date-state payoff uncertainty and pricing structure of [Table 9.11](#).

As earlier, q^b denotes the period $t = 0$ price of a risk-free bond paying one unit of consumption in each period 1 state, q^e is the price of some “equity security” with the indicated payoffs, and the q_θ , $\theta = 1, 2, 3$ represent the Arrow–Debreu state prices. Our objective is to study the equilibrium pricing of q^b and q^e . Since equilibrium tolerates no arbitrage opportunities, it must be that

$$q^b = q_1 + q_2 + q_3 = \sum_{i=1}^3 q_\theta = 0.9,$$

which in turn implies a risk-free rate as per

$$q^b = 0.9 = \frac{1}{(1 + r_f)}; \quad r_f = 0.11 \text{ or } 11\%.$$

In a like fashion, let us next price the “equity security”:

$$\begin{aligned} q^e &= 2q_1 + q_2 + 4q_3 (= 2(0.5) + 1(0.1) + 4(0.3) = 2.3) \\ &= q^b \left[2 \left(\frac{q_1}{q^b} \right) + 1 \left(\frac{q_2}{q^b} \right) + 4 \left(\frac{q_3}{q^b} \right) \right] \\ &= \frac{1}{(1 + r_f)} \left[2 \left(\frac{q_1}{q^b} \right) + 1 \left(\frac{q_2}{q^b} \right) + 4 \left(\frac{q_3}{q^b} \right) \right]. \end{aligned} \tag{9.4}$$

We pause here for a moment to make two observations. First, note that

$$\frac{q_\theta}{\sum_{\theta=1}^3 q_\theta} > 0, \forall \theta. \text{ Second note also that}$$

$$\sum_{\theta=1}^3 \left(\frac{q_\theta}{\sum_{\theta=1}^3 q_\theta} \right) = \frac{1}{q^b} \sum_{\theta=1}^3 q_\theta = \frac{1}{q^b} \cdot q^b = 1.$$

Taken together, these relationships imply that the quantities $\{(q_\theta)/(\sum_{\theta=1}^3 q_\theta)\}, i = 1, 2, 3$ have the structure of a probability density. Henceforth, we will refer to them as “risk-neutral” probabilities and employ the notation $\{\pi_\theta^{RN}\}_{\theta=1,2,3}$ where $\pi_\theta^{RN} = (q_\theta / \sum_{\theta=1}^3 q_\theta)$.

Equation (9.4) can then be rewritten as

$$\begin{aligned} q^e &= \frac{1}{(1 + r_f)} [2\pi_1^{RN} + 1\pi_2^{RN} + 4\pi_3^{RN}] \\ &= \frac{1}{(1 + r_f)} E_{\pi^{RN}} CF_\theta^e \end{aligned}$$

with $E_{\pi^{RN}}$ denoting the expectations operators taken with respect to the risk-neutral probability distribution. This accomplishes our goal: The asset is priced as its “distorted probability” expected “cash flow,” discounted at the risk-free rate. At this juncture we offer a number of clarifying comments.

1. The price of any asset whose payoffs lie in the span of the three Arrow–Debreu securities described in Table 9.11 (that is, any security whose payoffs can be replicated by some portfolio of these securities) can be represented as its expected cash flow, computed using the indicated risk-neutral probabilities, discounted at the risk-free rate.
2. No mention has been made of the objective probability distribution governing the future states because our development of the notion of risk-neutral valuation did not require that agents’ preferences take the VNM-expected utility form. As will become apparent in the next chapter these probabilities are embedded in the relative Arrow–Debreu state claims prices. Present and future state-dependent endowments and any other quantities relevant for agent security demands (e.g., agent risk aversions) are embedded in the Arrow–Debreu prices as well.
3. When expectations of future cash flows are taken with respect to the risk-neutral probability distribution, all assets in the span of the Arrow–Debreu securities are deemed to earn the risk-free rate of return, an assertion that seems strikingly

counterintuitive: risk appears to be “ignored.” In fact, this is not the case. When we impose greater economic structure (e.g., explicit objective probabilities and VNM-expected utility preferences) on the risk-neutral asset pricing theory, it will be shown that risk-neutral probabilities are “pessimistic” in the sense that they are larger than the objective probabilities for states where payoffs are small—the “bad” states—and smaller than the objective probabilities for states where payoffs are high, with the consequence that risk-neutral expected payoffs are numerically smaller than expected payoffs computed using the common objective probabilities. This adjustment compensates for discounting at the lower, not-risk-adjusted, risk-free rate in a risk-neutral pricing environment.

4. Our final comment concerns the origin of the title “risk-neutral” valuation given to the pricing perspective articulated above. Although there are a number of associations to which we could appeal to explore this designation, a principal one is as follows: as noted in the commentary following Eq. (9.1), if agents are risk neutral or if one agent is both risk neutral and has sufficient endowment to insure all other agents perfectly (no corner solution), then the Arrow–Debreu prices are directly proportional (by the common factor β) to their respective objective probabilities. In this case, the risk-neutral and objective probabilities coincide. Under either pricing perspective, expected asset cash flows are discounted at the risk-free rate. Note that these comments continue to presume the expected utility representation of agent preferences.

From what has been presented in Section 9.6, the usefulness of the risk-neutral valuation concept is not immediately apparent: it appears to be simply a reformulation of Arrow–Debreu pricing theory and we know Arrow–Debreu securities are generally not presently traded in organized financial markets. When more structure is imposed on the economy (e.g., VNM-expected utility preferences), however, the concept leads to the notion of a stochastic discount factor which is simply a more structured form of the risk-neutral probability distribution. The stochastic discount factor represents the ultimate discounting perspective in modern finance theory, and big financial firms spend millions attempting to estimate it. Its properties can also be related to certain macro data series thereby allowing empirical tests of the theory.

9.7 Conclusions

The Arrow–Debreu asset pricing theory presented in this chapter is in some sense the father of all asset pricing relationships. It is fully general and constitutes an extremely valuable reference formulation. Conceptually, its usefulness is unmatched, which justifies our investing more in its associated apparatus. At the same time, it is one of the most abstract theories, and its usefulness in practice is impaired by the difficulty in identifying individual states of nature and by the fact that, even when a set of states can be identified,

their actual realization cannot always be verified. As a result, it is difficult to write the appropriate conditional contracts. These problems go a long way in explaining why we do not see Arrow–Debreu securities being traded, a fact that does not strengthen the immediate applicability of the theory. In addition, as already mentioned, the static setting of the Arrow–Debreu theory is unrealistic for most applications. For all these reasons we cannot stop here, and we will explore a set of alternative, sometimes closely related, avenues for pricing assets in the following chapters.

References

- Arrow, K., 1951. An extension of the basic theorems of classical welfare economics. In: Neyman, J. (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 507–532.
- Debreu, G., 1959. Theory of Value. John Wiley & Sons, New York, NY.

The Consumption Capital Asset Pricing Model

Chapter Outline

10.1 Introduction	270
10.2 The Representative Agent Hypothesis and its Notion of Equilibrium	270
10.2.1 An Infinitely Lived Representative Agent	270
10.2.2 On the Concept of a “No-Trade” Equilibrium	271
10.3 An Exchange (Endowment) Economy	275
10.3.1 The Model	275
10.3.2 Interpreting the Exchange Equilibrium	278
10.3.3 The Formal CCAPM	281
10.4 Pricing Arrow–Debreu State-Contingent Claims with the CCAPM	281
10.4.1 The CCAPM and Risk-Neutral Valuation	285
10.5 Testing the CCAPM: The Equity Premium Puzzle	286
10.6 Testing the CCAPM: Hansen–Jagannathan Bounds	293
10.7 The SDF in Greater Generality	295
10.8 Some Extensions	297
10.8.1 Reviewing the Diagnosis	297
10.8.2 Adding a Disaster State	299
10.8.3 Habit Formation	302
10.8.4 The CCAPM with Epstein–Zin Utility	303
10.8.4.1 Bansal and Yaron (2004)	306
10.8.4.2 Collin-Dufresne et al. (2013)	308
10.8.5 Beyond a Representative Agent and Rational Expectations	313
10.8.5.1 Beyond a Representative Agent	313
10.8.5.2 Beyond Rational Expectations	316
10.9 Conclusions	317
References	317
Appendix 10.1: Solving the CCAPM with Growth	319
Appendix 10.2: Some Properties of the Lognormal Distribution	320

10.1 Introduction

Our asset pricing models thus far have been either one-period models such as the CAPM or multiperiod but essentially static, such as the Arrow–Debreu model. In the latter case, even if a large number of future periods are assumed, all decisions, including security trades, take place at date zero. It is in this sense that the Arrow–Debreu model is static. Reality is different, however. Assets are traded every period, as new information becomes available, and decisions are made sequentially, one period at a time, all the while keeping in mind the fact that today's decisions influence tomorrow's opportunities and constraints. Our objective in this chapter is to capture these dynamic features and to price assets in such an environment.

Besides adding an important dimension of realism, another advantage of a dynamic setup is to make it possible to begin to draw the link between the financial markets and the real side of the economy. Again, strictly speaking, this can be accomplished within an Arrow–Debreu economy: firms' present and future state-contingent production and investment decisions could also be modeled as occurring at date zero. However, as we will see, the main issues require a richer dynamic context where real production and consumption decisions are not made once and for all at the beginning of time, but progressively, as time evolves.

The present chapter begins this process by placing the production side of the economy temporarily in the background and studying the asset pricing implications of the resulting equilibrium consumption and dividend series. Under the consumption capital asset pricing model perspective, it is the properties of an economy's equilibrium per capita consumption series that ultimately determine asset pricing relationships. There is nothing unusual here: in our earlier chapters it was the desire of investors to maximize their expected utility of consumption that led to the formation of asset demands and equilibrium asset prices, those prices then being uniquely identified with investors' equilibrium consumption. As a by-product of these endeavors, we will also revisit the notion of risk-neutral valuation, specializing it in a way that allows the theory to be judged by actual economic data.

10.2 The Representative Agent Hypothesis and its Notion of Equilibrium

10.2.1 An Infinitely Lived Representative Agent

To accomplish these goals in a model of complete generality (in other words, with many different agents and firms) where asset prices can be tractably computed is beyond the present capability of economic science. As an alternative, we will make life simpler by postulating many *identical* infinitely lived consumers. This allows us to examine the decisions of a *representative*, stand-in consumer and explore their implications for asset

pricing. In particular, we will assume that the representative agent acts to maximize the expected present value of discounted utility of consumption over his entire, infinite lifetime:

$$\max E \left(\sum_{t=0}^{\infty} \delta^t U(\tilde{c}_t) \right)$$

where δ is his discount factor and $U(\cdot)$ his period utility function with $U_1(\cdot) > 0$, $U_{11}(\cdot) < 0$. This construct is the natural generalization to the case of infinite lifetimes of the preferences considered in our earlier two-period examples. Its use can be justified by the following considerations.

First, if we model the economy as ending at some terminal date T (as opposed to assuming an infinite horizon), then the agent's investment behavior will reflect this fact. In the last period of his life, in particular, the agent will stop saving, liquidate his portfolio, and consume its entire value. There is no real-world counterpart for this action as the real economy continues forever. Assuming an infinite horizon eliminates these terminal date complications. Second, it can be shown, under fairly general conditions, that an infinitely lived agent setup is formally equivalent to one in which agents live only a finite number of periods themselves, provided they derive utility from the well-being of their descendants (a bequest motive). This argument is detailed by [Barro \(1974\)](#).

Restrictive as it may seem, the identical agents' assumption can be justified by the fact that, in a competitive equilibrium with complete securities markets, there is an especially intuitive sense of a representative agent: one whose utility function is a weighted average of the utilities of the various agents in the economy. In [Box 10.1](#), we detail the precise way in which one can construct such a representative individual and discuss some of the issues at stake.

10.2.2 On the Concept of a “No-Trade” Equilibrium

In a representative agent economy, we must, of necessity, use a somewhat specialized notion of equilibrium—a *no-trade equilibrium*. If, indeed, for a particular model specification, some security is in positive net supply, the equilibrium price will be the price at which the representative agent is willing to hold that amount—the total supply—of the security. In other specifications, we will price securities that do not appear explicitly—securities that are said to be in zero net supply. The prototype of the latter is an IOU type of contract: in a one-agent economy, the total net supply of IOUs must, of course, be zero.¹ In this case, if at some price the representative agent wants to supply (sell) the security, since there is no one to demand it, supply exceeds demand. Conversely, if at some price, the representative agent wants to buy the security (and thus no one wants to supply it), demand exceeds supply. Financial markets are thus in equilibrium, if and only if, at the

¹ An IOU (I owe you) is simply a promise to pay a specific amount of money on a specific date.

BOX 10.1 Constructing a Representative Agent

In order to illustrate the procedure for constructing a representative agent, let us return to the two-period ($t = 0, 1$) Arrow–Debreu economy considered in Chapter 9 and assume markets are complete. Without loss of generality, assume K agents and N states of nature at $t = 1$. Each agent k , $k = 1, 2, \dots, K$, solves:

$$\begin{aligned} & \max U^k(c_0^k) + \delta^k \sum_{\theta=1}^N \pi_\theta U^k(c_\theta^k) \\ \text{s.t. } & c_0^k + \sum_{\theta=1}^N q_\theta c_\theta^k \leq e_0^k + \sum_{\theta=1}^N q_\theta e_\theta^k \end{aligned}$$

where the price of period 0 endowment is normalized to 1, and the endowments of a typical

agent k are described by the vector $\begin{pmatrix} e_0^k \\ e_1^k \\ \vdots \\ e_N^k \end{pmatrix}$

Under very standard assumptions (cf. Chapters 1 and 9), the equilibrium allocation is Pareto optimal, and, at the prevailing Arrow–Debreu prices $\{q_\theta\}_{\theta=1, \dots, N}$, supply equals demand in every market:

$$\begin{aligned} \sum_{k=1}^K c_0^k &= \sum_{k=1}^K e_0^k, \text{ and} \\ \sum_{k=1}^K c_\theta^k &= \sum_{k=1}^K e_\theta^k, \text{ for every state } \theta \end{aligned}$$

Since this competitive equilibrium allocation is Pareto optimal, no one can be better off without making someone else worse off. One important implication of this property for our problem is that there exists some set of weights $(\lambda_1, \dots, \lambda_K)$, which in general will depend on the distribution of initial endowments, such that the solution to the following problem gives an allocation that is identical to the equilibrium allocation:

$$\begin{aligned} & \max \sum_{k=1}^K \lambda_k \left\{ U^k(c_0^k) + \delta^k \sum_{\theta=1}^N \pi_\theta U^k(c_\theta^k) \right\}, \\ \text{s.t. } & \sum_{k=1}^K c_0^k = \sum_{k=1}^K e_0^k, \\ & \sum_{k=1}^K c_\theta^k = \sum_{k=1}^K e_\theta^k, \quad \forall \theta \\ & \sum_{k=1}^K \lambda_k = 1; \quad \lambda_k > 0, \quad \forall k \end{aligned} \tag{10.1}$$

(Continued)

BOX 10.1 Constructing a Representative Agent (Continued)

Maximization problem (10.1) is interpreted to represent the problem of a benevolent central planner attempting to allocate the aggregate resources of the economy so as to maximize the weighted sum of the utilities of the individual agents. We proposed a similar problem in Chapter 9 in order to identify the conditions characterizing a Pareto optimal allocation of resources. Here we see problem (10.1) as suggestive of the form the representative agent's preference ordering, defined over aggregate consumption, can take (the representative agent is denoted by the superscript A):

$$\begin{aligned} U^A(c_0^A, c_\theta^A) &= U_0^A(c_0^A) + \sum_{\theta=1}^N \pi_\theta U^A(c_\theta^A), \text{ where} \\ U_0^A(c_0^A) &= \sum_{k=1}^N \lambda_k U_0^k(c_\theta^A) \text{ with } \sum_{k=1}^K c_0^k = \sum_{k=1}^K e_0^k \equiv c_0^A \\ U^A(c_\theta^A) &= \sum_{k=1}^N \lambda_k \delta_k U^k(c_\theta^A) \text{ with } \sum_{k=1}^K c_\theta^k = \sum_{k=1}^K e_\theta^k \equiv c_\theta^A, \text{ for each state } \theta \end{aligned} \quad (10.2)$$

The above setup generalizes to as many periods as we like and, with certain minor modifications, to an infinite horizon. It accommodates future state-contingent endowments (e.g., representing uncertain labor income streams) as well as future production. Construct (10.2) presents an intuitive sense of a representative agent as one who constitutes a weighted average of all the economy's participants. Complete financial markets are a critical antecedent, as are expected utility preference representations. For more details, see Constantinides (1982), who first proposed this sense of a representative agent.

An important feature of construct (10.2) resides in the fact that, in general, the weights $\{\lambda_1, \lambda_2, \dots, \lambda_K\}$ will depend on the initial endowments: Loosely speaking, the agent with more wealth gets a bigger λ weight. In effect, this feature of the λ weights means that different initial wealth distributions may give rise to "different representative agents." Nothing in this statement is surprising: different initial wealth distributions will give rise to different competitive equilibria, which must then be matched, via the procedure above to different representative agents.

Rubinstein (1974) shows, however, that the utility function, constructed in Eq. (10.2), will, in addition, be independent of the initial endowment distribution if three further conditions are satisfied:

1. The subjective time preference parameter of every agent is the same (i.e., all agents δ 's are the same).
2. Agents' period preferences are identical and assume either of the two following forms:

$$U^k(c) = \frac{\gamma}{\gamma - 1} (\alpha^k + \gamma c)^{1-\frac{1}{\gamma}} \text{ or } U^k(c) = -e^{-\alpha^k c}$$

3. Agents receive no income or endowments in the period $t \geq 1$.

(Continued)

BOX 10.1 Constructing a Representative Agent (Continued)

If these conditions are met—i.e., by and large, if the agents' preferences can be represented by either a CRRA or a CARA utility function—then there exists a representative agent economy for which the equilibrium Arrow–Debreu security demands, and thus their equilibrium prices and rates of return, are the same as in the original heterogeneous agent economy.

In other words, the demand for the Arrow–Debreu security paying off in some arbitrary state $\hat{\theta}$ is of the form

$$z_{\hat{\theta}} \left(\{q_{\theta}\}_{\theta=1,2,\dots,N}, \sum_{k=1}^K e_0^k \right)$$

In this event, “demand aggregation” is said to result.

What does a representative agent “look like” under [Rubinstein \(1974\)](#) aggregation? All agents in his economy have the same utility function with the same parameter, γ , the same $\alpha = (1/K) \sum_{k=1}^K \alpha^k$, the same δ , the same initial wealth $(1/K) \sum_{k=1}^K e_0^k$, and, accordingly, the same equilibrium consumption. See the Web Notes for [Guvenen's \(2011\)](#) summary statement of [Rubinstein's \(1974\)](#) aggregation result.

[Constantinides's \(1982\)](#) notion of a “representative agent” is more general than [Rubinstein's \(1974\)](#), principally in two ways: (1) there is no assumption that agent preferences are identical, or of some specific form, and (2) exogenous endowments or endogenous production in future date-states is admitted. The cost of this generality is a loss of the aggregation property: different initial endowment distributions lead to different associated “representative agents,” and thus different equilibrium security prices and returns. In this chapter and in succeeding ones, our notion of a representative agent will be that of [Constantinides \(1982\)](#). Note that this choice demands both complete markets and VNM-expected utility preference representations.³⁵

³⁵ All these comments notwithstanding, the simplest sense of a representative agent occurs when all agents are assumed to have (i) identical preferences, (ii) identical initial wealth, and (iii) identical future income shocks (which must be, by definition, aggregate shocks). They then construct the same portfolios and at all times undertake the same savings and possess the same wealth. Idiosyncratic income shocks are not allowed under this interpretation.

prevailing price, supply equals demand *and* both are simultaneously zero. In all cases, the equilibrium price is that price at which the representative agent wishes to hold exactly the amount of the security present in the economy. Therefore, the essential question being asked is: What prices must securities assume so that the amount the representative agent *must* hold (for all markets to clear) exactly equals what he *wants* to hold? At these prices, further trade is not utility enhancing. In a more conventional multiagent economy, an identical state of affairs is verified post-trade. The representative agent class of models is not appropriate, of course, for the analysis of some issues in finance. For example, issues

linked with the volume of trade cannot be studied since, in a representative agent model, trading volume is, by construction, equal to zero.

10.3 An Exchange (Endowment) Economy

10.3.1 The Model

This economy will be directly analogous to the Arrow–Debreu exchange economies considered earlier: production decisions are in the background and abstracted away. It is, however, an economy that admits recursive trading, where investment decisions are made period by period (as opposed to being made once and for all at date 0).

There is one, perfectly divisible *share*, which we can think of as representing the market portfolio of the CAPM (later we shall relax this assumption). Ownership of this share entitles the owner to all the economy's output. (In this economy, all firms are publicly traded.) Output is viewed as arising exogenously and as being stochastically variable through time, although in a stationary fashion. This is the promised, though still remote, link with the real side of the economy. Indeed, we will use macroeconomic data to calibrate the model in the forthcoming sections. At this point, we can think of the output process as being governed by a large-number-of-states version of the three-state probability transition matrix found in [Table 10.1](#).

That is, we assume there are a given number of output states, levels of output that can be achieved at any given date, and the probabilities of the transition from one output state to another are constant and represented by entries in the matrix \mathbf{T} . The stationarity hypothesis embedded in this formulation may, at first sight, appear extraordinarily restrictive. The output levels defining the states may, however, be normalized variables, to allow for a constant rate of growth. Alternatively, the states could themselves be defined in terms of growth rates of output rather than output levels. See [Appendix 10.1](#) for a growth illustration.

Table 10.1: Three-state probability transition matrix

		Output in Period $t + 1$		
		Y^1	Y^2	Y^3
Output in Period t	Y^1	π_{11}	π_{12}	π_{13}
	Y^2	π_{21}	π_{22}	π_{23}
	Y^3	π_{31}	π_{32}	π_{33}
where $\pi_{ij} = \text{Prob}(Y_{t+1} = Y^j; Y_t = Y^i)$ for any t .				

In the continuous-state version of this perspective, the output process can be analogously described by a probability transition *function*

$$G(Y_{t+1}|Y_t) = \text{Prob}(Y_{t+1} \leq Y^j; Y_t = Y^i)$$

We imagine the security as representing ownership of a fruit tree where the (perishable) output (the quantity of fruit produced by the tree—the dividend) varies from year to year. This interpretation is often referred to as the *Lucas tree* economy in tribute to 1995 Nobel Prize winner, R.E. Lucas, Jr., who, in his 1978 article, first developed the consumption capital asset pricing model (CCAPM). The power of the approach, however, resides in the fact that any mechanism delivering a stochastic process on aggregate output, such as a full macroeconomic equilibrium model, can be grafted on the CCAPM, thus allowing an in-depth analysis of the rich relationships between the real and the financial sides of an economy.

Ours will be a *rational expectations economy*. By this expression, we mean that the representative agent's expectations will be on average correct and, in particular, will exhibit no systematic bias. In effect we assume, in line with a very large literature (and with most of what we have done implicitly so far), that the representative agent knows both the general structure of the economy and the exact output distribution as summarized by the matrix T . One possible justification is that this economy has been functioning for a long enough time to allow the agent to learn the probability process governing output and to understand the environment in which he operates. Accumulating such knowledge is clearly in his own interest if he wishes to maximize his expected utility.

The agent buys and sells securities (fractions of the single, perfectly divisible share) and consumes dividends. His security purchases solve:

$$\begin{aligned} & \max_{\{z_{t+1}\}} E \left(\sum_{t=0}^{\infty} \delta^t U(\tilde{c}_t) \right) \\ \text{s.t. } & c_t + q_t^e z_{t+1} + 1 \leq z_t Y_t + q_t^e z_t \\ & z_t \leq 1, \quad \forall t \end{aligned}$$

where q_t^e is the period t real price of the security in terms of consumption (the price of consumption is 1) and z_t is the agent's beginning-of-period t holdings of the security. Holding a fraction z_t of the security entitles the agent to the corresponding fraction of the distributed dividend Y_t , which, in an exchange economy without investment, equals total available output. The expectations operator applies across all possible values of Y feasible at each date t with transition probabilities provided by the matrix T .

Let us assume that the representative agent's period utility function is strictly concave with $\lim_{c_t \rightarrow 0} U_1(c_t) = \infty$. Making this latter assumption ensures that it is never optimal for the agent to select a zero consumption level. It thus normally ensures an interior solution to the

relevant maximization problem. The necessary and sufficient condition for the solution to this problem is then given by: For all t , z_{t+1} solves:

$$U_1(c_t)q_t^e = \delta E_t\{U_1(\tilde{c}_{t+1})(\tilde{q}_{t+1}^e + \tilde{Y}_{t+1})\} \quad (10.3)$$

where $c_t = (q_t^e z_t + z_t Y_t - q_t^e z_{t+1})$. Note that the expectations operator applies across possible output state levels; if we make explicit the functional dependence on the output state variables, Eq. (10.3) can be written (assuming Y^i is the current state):

$$U_1(c_t(Y^i))q_t^e(Y^i) = \delta \sum_j U_1(c_{t+1}(Y^j))(q_{t+1}^e(Y^j) + Y^j)\pi_{ij}$$

In Eq. (10.3), $U_1(c_t)q_t^e$ is the utility loss in period t associated with the purchase of an additional unit of the security, while $\delta U_1(c_{t+1})$ measures the units of marginal utility of an additional unit of consumption in period $t + 1$ and $(q_{t+1}^e + Y_{t+1})$ is the extra consumption (income) units obtained in period $t + 1$ from selling the additional unit of the security in addition to collecting its dividend entitlement. The RHS is thus the expected discounted gain in utility associated with buying the extra unit of the security. The agent is in equilibrium (utility maximizing) at the prevailing price q_t^e if the loss in utility today, which he would incur by buying one more unit of the security ($U_1(c_t)q_t^e$), is exactly offset by (equals) the expected gain in utility tomorrow ($\delta E_t U_1(\tilde{c}_{t+1})(\tilde{q}_{t+1}^e + \tilde{Y}_{t+1})$), which the ownership of that additional security will provide. If this equality is not satisfied, the agent will try either to increase or to decrease his holdings of securities.²

For the entire economy to be in equilibrium, it must, therefore, be true that:

- i. $z_t = z_{t+1} = z_{t+2} = \dots \equiv 1$, in other words, the representative agent owns the entire security;
- ii. $c_t = Y_t$, i.e., ownership of the entire security entitles the agent to all the economy's output;
- iii. $U_1(c_t)q_t^e = \delta E_t\{U_1(\tilde{c}_{t+1})(\tilde{q}_{t+1}^e + \tilde{Y}_{t+1})\}$, i.e., the agents' holdings of the security are optimal given the prevailing prices. Substituting (ii) into (iii) informs us that the equilibrium price must satisfy

$$U_1(Y_t)q_t^e = \delta E_t\{U_1(\tilde{Y}_{t+1})(\tilde{q}_{t+1}^e + \tilde{Y}_{t+1})\} \quad (10.4)$$

If there were many firms in this economy—say J firms, with firm j producing the (exogenous) output $\tilde{Y}_{j,t}$ —then the same equation would be satisfied for each firm's stock price, $q_{j,t}^e$, i.e.,

$$q_{j,t}^e U_1(c_t) = \delta E_t\{U_1(\tilde{c}_{t+1})(\tilde{q}_{j,t+1}^e + \tilde{Y}_{j,t+1})\} \quad (10.5)$$

where $c_t = \sum_{j=1}^J Y_{j,t}$ in equilibrium.

² In equilibrium, however, this is not possible since the supply of securities is fixed. Accordingly, the price will have to adjust until the equality in Eq. (10.3) is satisfied.

Equations (10.4) and (10.5) are the fundamental equations of the CCAPM.³ A recursive substitution of Eq. (10.2) into itself yields⁴

$$q_t^e = E_t \sum_{\tau=1}^{\infty} \delta^\tau \left[\frac{U_1(\tilde{Y}_{t+\tau})}{U_1(Y_t)} \tilde{Y}_{t+\tau} \right] \quad (10.6)$$

establishing the stock price as the sum of all expected discounted future dividends. Equation (10.6) resembles the standard discounting formula of elementary finance, but for the important observation that discounting takes place using the intertemporal marginal rates of substitution defined on the consumption sequence of the representative agent. If the utility function displays risk neutrality and the marginal utility is constant ($U_{11} = 0$), Eq. (10.6) reduces to

$$q_t^e = E_t \sum_{\tau=1}^{\infty} \delta^\tau [\tilde{Y}_{t+\tau}] = E_t \sum_{\tau=1}^{\infty} \left[\frac{\tilde{Y}_{t+\tau}}{(1+r_f)^\tau} \right] \quad (10.7)$$

which states that the stock price is the sum of expected future dividends discounted at the (constant) risk-free rate. The intuitive link between the discount factor and the risk-free rate leading to the second inequality in Eq. (10.7) will be formally established in Eq. (10.9). The difference between Eqs. (10.6) and (10.7) is the necessity, in a world of risk aversion, of discounting the flow of expected dividends at a rate higher than the risk-free rate, so as to include a risk premium. The question as to the appropriate risk premium constitutes a central issue in financial theory. Equation (10.6) proposes a definite, if not fully operational (due to the difficulty in measuring marginal rates of substitution), answer.

10.3.2 Interpreting the Exchange Equilibrium

To bring about a closer correspondence with traditional asset pricing formulas, we must first relate the asset prices derived previously to rates of return. In particular, we will want to understand, in this model context, what determines the amount by which the risky asset's expected return exceeds that of a risk-free asset. This basic question is also the one for which the standard CAPM provides such a simple, elegant answer ($E\tilde{r}_j = r_f + \beta_j(E\tilde{r}_M - r_f)$). Define the period t to $t+1$ return for security j as

$$1 + r_{j,t+1} = \frac{q_{j,t+1} + Y_{j,t+1}}{q_{j,t}}$$

³ The fact that the representative agent's consumption stream—via his MRS—is critical for asset pricing is true for all versions of this model, including ones with nontrivial production settings. More general versions of this model may not, however, display an identity between consumption and dividends. This will be the case, for example, if the agent receives wage income.

⁴ That is, update Eq. (10.4) with q_{t+1}^e on the LHS and q_{t+2}^e in the RHS and substitute the resulting RHS (which now contains a term in q_{t+2}^e) into the original Eq. (10.4); repeat for q_{t+2}^e, q_{t+3}^e , and so on, regroup terms, and extrapolate.

Equation (10.5) may then be rewritten as

$$1 = \delta E_t \left\{ \frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)} (1 + \tilde{r}_{j,t+1}) \right\} \quad (10.8)$$

Let q_t^b denote the price in period t of a one-period riskless discount bond in zero net supply. The bond pays one unit of consumption (income) in every state in the next period. By reasoning analogous to that presented previously,

$$q_t^b U_1(c_t) = \delta E_t \{U_1(\tilde{c}_{t+1})1\}$$

The price q_t^b is the equilibrium price at which the agent desires to hold zero units of the security, and thus supply equals demand. This is so because if he were to buy one unit of this security at a price q_t^b , the loss in utility today would exactly offset the gain in expected utility tomorrow. The representative agent is, therefore, content to hold zero units of the security.

Since the risk-free rate over the period from date t to $t+1$, denoted $r_{f,t+1}$, is defined by $q_t^b(1+r_{f,t+1}) = 1$, we have

$$\frac{1}{1 + r_{f,t+1}} = q_t^b = \delta E_t \left\{ \frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)} \right\}, \quad (10.9)$$

which formally establishes the link between the discount rate and the risk-free rate of return we have used in Eq. (10.7) under the risk neutrality hypothesis. Note that in the latter case ($U_{11} = 0$), Eq. (10.9) implies that the risk-free rate must be a constant.

Now we will combine Eqs. (10.8) and (10.9). Since, for any two random variables \tilde{x}, \tilde{y} , $E(\tilde{x} \cdot \tilde{y}) = E(\tilde{x}) \cdot E(\tilde{y}) + \text{cov}(\tilde{x} \cdot \tilde{y})$, Eq. (10.8) can be written in the form

$$1 = \delta E_t \left\{ \frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)} \right\} E_t \{1 + \tilde{r}_{j,t+1}\} + \delta \text{cov}_t \left\{ \frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)}, \tilde{r}_{j,t+1} \right\} \quad (10.10)$$

Let us make the identification $E_t \{1 + \tilde{r}_{j,t+1}\} = 1 + E(r_{j,t+1})$ while recognizing that the expectation remains conditional on period t information. Substituting Eq. (10.9) into Eq. (10.10) then gives

$$\begin{aligned} 1 &= \frac{1 + E\tilde{r}_{j,t+1}}{1 + r_{f,t+1}} + \delta \text{cov}_t \left(\frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)}, \tilde{r}_{j,t+1} \right), \text{ or, rearranging,} \\ \frac{1 + E\tilde{r}_{j,t+1}}{1 + r_{f,t+1}} &= 1 - \delta \text{cov}_t \left(\frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)}, \tilde{r}_{j,t+1} \right), \text{ or} \\ E\tilde{r}_{j,t+1} - r_{f,t+1} &= -\delta(1 + r_{f,t+1}) \text{cov}_t \left(\frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)}, \tilde{r}_{j,t+1} \right) \end{aligned} \quad (10.11)$$

[Equation \(10.11\)](#) is the central pricing relationship of the CCAPM, and we must consider its implications. The LHS of [Eq. \(10.11\)](#) is the risk premium on security j . [Equation \(10.11\)](#) tells us that the risk premium on a security will be large when $\text{cov}_t((U_1(\tilde{c}_{t+1}))/U_1(c_t)), \tilde{r}_{j,t+1})$ is large and negative, i.e., for those securities paying high returns when consumption is high (and thus when $U_1(c_{t+1})$ is low), and low returns when consumption is low (and $U_1(c_{t+1})$ is high). These securities are not very desirable for consumption risk reduction (consumption smoothing): they pay high returns when investors do not need them (consumption is already high) and low returns when they are most needed (consumption is low). Since they are not desirable, they have a low price and thus high expected returns relative to the risk-free security.

The CAPM tells us that a security is relatively undesirable and thus commands a high return when it covaries positively with the market portfolio, i.e., when its return is high precisely in those circumstances when the return on the market portfolio is also high, and conversely. The CCAPM is not in contradiction with this basic idea, but it adds some further degree of precision. From the viewpoint of smoothing *consumption* and risk diversification, an asset is desirable if it has a high return when consumption is low and vice versa.

When the portfolio and asset pricing problem is placed in its proper multiperiod context, the notion of the utility of end-of-period wealth (our paradigm of Chapters 5–8) is no longer relevant, and we have to go back to the more fundamental formulation in terms of the utility derived from consumption, $U(c_t)$. It then becomes clear that the possibility of expressing the objective as maximizing the utility of end-of-period wealth in the two-date setting has, in some sense, lured us down a false trail: fundamentally, the key to an asset's value is its return covariation with the marginal utility of consumption, not with the marginal utility of wealth.

[Equation \(10.11\)](#) has the unappealing feature that the risk premium is defined, in part, in terms of the marginal utility of consumption, which is not observable. To eliminate this feature, we shall make the following approximation.

Let $U(c_t) = ac_t - \frac{b}{2}c_t^2$ (i.e., a quadratic utility function or a truncated Taylor series expansion of a general $U()$), where $a > 0$, $b > 0$, and the usual restrictions apply on the range of consumption. It follows that $U_1(c_t) = a - bc_t$; substituting this into [Eq. \(10.11\)](#) gives

$$\begin{aligned} E\tilde{r}_{j,t+1} - r_{f,t+1} &= -\delta(1 + r_{f,t+1})\text{cov}_t\left(\tilde{r}_{j,t+1}, \frac{a - b\tilde{c}_{t+1}}{a - bc_t}\right) \\ &= -\delta(1 + r_{f,t+1})\frac{1}{a - bc_t}\text{cov}_t(\tilde{r}_{j,t+1}, \tilde{c}_{t+1})(-b), \text{ or} \\ E\tilde{r}_{j,t+1} - r_{f,t+1} &= \frac{\delta b(1 + r_{f,t+1})}{a - bc_t}\text{cov}_t(\tilde{r}_{j,t+1}, \tilde{c}_{t+1}). \\ E\tilde{r}_{j,t+1} - r_{f,t+1} &= \frac{\delta b(1 + r_{f,t+1})}{a - bc_t}\text{cov}_t(\tilde{r}_{j,t+1}, \tilde{c}_{t+1}) \end{aligned} \tag{10.12}$$

Equation (10.12) makes our earlier point as to the fundamental source of an asset's value easier to grasp: since the term in front of the covariance expression is necessarily positive, if next-period consumption covaries in a large positive way with $r_{j,t+1}$, then the risk premium on j will be high.

10.3.3 The Formal CCAPM

As a final step in our construction, let us denote the portfolio most highly correlated with consumption by the index $j = c$, and its expected rate of return for the period from t to $t + 1$ by $\tilde{r}_{c,t+1}$.

Equation (10.12) applies as well to this security, so we have

$$E\tilde{r}_{c,t+1} - r_{f,t+1} = \left[\frac{\delta b(1 + r_{f,t+1})}{a - bc_t} \right] \text{cov}_t(\tilde{r}_{c,t+1}, \tilde{c}_{t+1}) \quad (10.13)$$

Dividing Eq. (10.12) by Eq. (10.13) and thus eliminating the term $[(\delta(1 + r_{f,t+1})b)/(a - bc_t)]$, one obtains

$$\begin{aligned} \frac{E\tilde{r}_{j,t+1} - r_{f,t+1}}{E\tilde{r}_{c,t+1} - r_{f,t+1}} &= \frac{\text{cov}_t(\tilde{r}_{j,t+1}, \tilde{c}_{t+1})}{\text{cov}_t(\tilde{r}_{c,t+1}, \tilde{c}_{t+1})}, \text{ or} \\ \frac{E\tilde{r}_{j,t+1} - r_{f,t+1}}{E\tilde{r}_{c,t+1} - r_{f,t+1}} &= \frac{\frac{\text{cov}_t(\tilde{r}_{j,t+1}, \tilde{c}_{t+1})}{\text{var}(\tilde{c}_{t+1})}}{\frac{\text{cov}_t(\tilde{r}_{c,t+1}, \tilde{c}_{t+1})}{\text{var}(\tilde{c}_{t+1})}}, \text{ or} \\ E\tilde{r}_{j,t+1} - r_{f,t+1} &= \frac{\beta_{j,c_t}}{\beta_{c,c_t}} [E\tilde{r}_{c,t+1} - r_{f,t+1}] \end{aligned} \quad (10.14)$$

for $\beta_{j,c_t} = (\text{cov}_t(\tilde{r}_{j,t+1}, \tilde{c}_{t+1})) / (\text{var}(\tilde{c}_{t+1}))$, the consumption- β of asset j , and $(\text{cov}_t(\tilde{r}_{c,t+1}, \tilde{c}_{t+1})) / (\text{var}(\tilde{c}_{t+1}))$, the consumption- β of portfolio c . This equation defines the CCAPM.

If it is possible to construct a portfolio c such that $\beta_{c,c_t} = 1$, the direct analogue to the CAPM is obtained, with $\tilde{r}_{c,t+1}$ replacing the expected return on the market and β_{j,c_t} the relevant beta:

$$E\tilde{r}_{j,t+1} - r_{f,t+1} = \beta_{j,c_t} (E\tilde{r}_{c,t+1} - r_{f,t+1}) \quad (10.15)$$

10.4 Pricing Arrow–Debreu State-Contingent Claims with the CCAPM

Chapter 9 focused on the notion of an Arrow–Debreu state claim as the basic building block for all asset pricing, and it is important to understand what form these securities and their prices

assume in the CCAPM setting. Our treatment will be very general and will accommodate more complex settings where the state is characterized by more than one variable.

Whatever model we happen to use, let s_t denote the state in period t . In the prior sections, s_t coincided with the period t output, Y_t .

Given that we are in state s in period t , what is the price of an Arrow–Debreu security that pays one unit of consumption if and only if state s' occurs in period $t + 1$? We consider two cases:

1. Let the number of possible states be finite; denote the Arrow–Debreu price as

$$q(s_{t+1} = s'; s_t = s)$$

with the prime superscript referring to the value taken by the random state variable in the next period. Since this security is assumed to be in zero net supply,⁵ it must satisfy, in equilibrium,

$$\begin{aligned} U_1(c(s))q(s_{t+1} = s'; s_t = s) &= \delta U_1(c(s')) \text{ prob}(s_{t+1} = s'; s_t = s), \text{ or} \\ q(s_{t+1} = s'; s_t = s) &= \delta \frac{U_1(c(s'))}{U_1(c(s))} \text{ prob}(s_{t+1} = s'; s_t = s) \end{aligned}$$

As a consequence of our maintained stationarity hypothesis, the same price occurs when the economy is in state s and the claim pays one unit of consumption in the next period if and only if state s' occurs, whatever the current time period t . We may thus drop the time subscript and write

$$q(s'; s) = \delta \frac{U_1(c(s'))}{U_1(c(s))} \text{ prob}(s'; s) = \delta \frac{U_1(c(s'))}{U_1(c(s))} \pi_{ss'}$$

in the notation of our transition matrix representation. This is Eq. (9.1).

2. For a continuum of possible states, the analogous expression is

$$q(s'; s) = \delta \frac{U_1(c(s'))}{U_1(c(s))} f(s'; s)$$

where $f(s'; s)$ is the conditional density function on s_{t+1} given s , evaluated at s' .

⁵ Recall that the very existence of a representative agent required that the underlying multiagent economy possessed a complete financial market structure.

Note that *under risk neutrality*, we have a reconfirmation of our earlier identification of Arrow–Debreu prices as being proportional to the relevant state probabilities, with the proportionality factor corresponding to the time discount coefficient:

$$q(s'; s) = \delta f(s'; s) = \delta \pi_{ss}$$

If these prices are for one-period state-contingent claims, how is an N -period claim priced? They are priced exactly analogously:

$$q^N(s_{t+N} = s'; s_t = s) = \delta^N \frac{U_1(c(s'))}{U_1(c(s))} \text{prob}(s_{t+N} = s'; s_t = s)$$

The price of an N -period risk-free discount bound $q_t^{b,N}$ given state s is thus given by

$$q_t^{b,N}(s) = \delta^N \sum_{s'} \frac{U_1(c(s'))}{U_1(c(s))} \text{prob}(s_{t+N} = s'; s_t = s) \quad (10.16)$$

or, in the continuum of states notation,

$$q_t^{b,N}(s) = \delta^N \int_{s'} \frac{U_1(c(s'))}{U_1(c(s))} f_N(s'; s) ds' = E_s \left\{ \delta^N \frac{U_1(c_{t+N}(s'))}{U_1(c(s))} \right\}$$

where the expectation is taken over all possible states s' conditional on the current state being s .⁶

Now let us review Eq. (10.6) in the light of the expressions we have just derived.

$$\begin{aligned} q_t^e &= E_t \sum_{\tau=1}^{\infty} \delta^{\tau} \left[\frac{U_1(\tilde{c}_{t+\tau})}{U_1(c_t)} \tilde{Y}_{t+\tau} \right] \\ &= \sum_{\tau=1}^{\infty} \sum_{s'} \delta^{\tau} \left[\frac{U_1(c_{t+\tau}(s'))}{U_1(c_t(s))} Y_{t+\tau}(s') \right] \text{prob}(s_{t+\tau} = s'; s_t = s) \\ &= \sum_{\tau} \sum_{s'} q^{\tau}(s', s) Y_{t+\tau}(s') \end{aligned} \quad (10.17)$$

What this development tells us is that taking the appropriately discounted (at the intertemporal marginal rate of substitution (MRS)) sum of expected future dividends is simply valuing the stream of future dividends at the appropriate Arrow–Debreu prices! The fact that there are no restrictions in the present context in extracting the prices of Arrow–Debreu contingent claims is indicative of the fact that this economy is one of complete markets.

⁶ The corresponding state probabilities are given by the N th power of the matrix \mathbf{T} .

Applying the same substitution to Eq. (10.6) as employed to obtain Eq. (10.10) yields

$$\begin{aligned} q_t^e &= \sum_{\tau=1}^{\infty} \delta^{\tau} \left\{ E_t \left[\frac{U_1(\tilde{c}_{t+\tau})}{U_1(c_t)} \right] E_t[\tilde{Y}_{t+\tau}] + \text{cov} \left(\frac{U_1(\tilde{c}_{t+\tau})}{U_1(c_t)}, \tilde{Y}_{t+\tau} \right) \right\} \\ &= \sum_{\tau=1}^{\infty} \delta^{\tau} \left\{ E_t \left[\frac{U_1(\tilde{c}_{t+\tau})}{U_1(c_t)} \right] E_t[\tilde{Y}_{t+\tau}] \left[1 + \frac{\text{cov} \left(\frac{U_1(\tilde{c}_{t+\tau})}{U_1(c_t)}, \tilde{Y}_{t+\tau} \right)}{E_t \left[\frac{U_1(\tilde{c}_{t+\tau})}{U_1(c_t)} \right] E_t[\tilde{Y}_{t+\tau}]} \right] \right\} \end{aligned}$$

where the expectations operator applies across all possible values of the state output variable, with probabilities given on the line corresponding to the current state s_t in the matrix \mathbf{T} raised to the relevant power (the number of periods to the date of availability of the relevant cash flow).

Using the expression for the price of a risk-free discount bond of τ periods to maturity derived earlier and the fact that $(1+r_{f,t+\tau})^{\tau} q_t^{b,\tau} = 1$ we can rewrite this expression as

$$q_t^e = \sum_{\tau=1}^{\infty} \frac{\left\{ E_t[Y_{t+\tau}] \left\{ 1 + \frac{\text{cov}(U_1(\tilde{c}_{t+\tau}), \tilde{Y}_{t+\tau})}{E_t[U_1(\tilde{c}_{t+\tau})] E_t[\tilde{Y}_{t+\tau}]} \right\} \right\}}{(1+r_{f,t+\tau})^{\tau}} \quad (10.18)$$

The quantity being discounted (at the risk-free rate applicable to the relevant period) in the present value term is the equilibrium certainty equivalent of the real cash flow generated by the asset. This is the analogue for the CCAPM of the CAPM expression derived in Section 8.3.

If the cash flows exhibit no stochastic variation (i.e., they are risk free), then Eq. (10.18) reduces to

$$q_t^e = \sum_{\tau=1}^{\infty} \frac{Y_{t+\tau}}{(1+r_{f,t+\tau})^{\tau}}$$

This relationship will be derived again in Chapter 12 where we discount risk-free cash flows at the term structure of interest rates. If, on the other hand, the cash flows are risky, yet investors are risk neutral (constant marginal utility of consumption), Eq. (10.18) becomes

$$q_t^e = \sum_{\tau=1}^{\infty} \frac{E_t[\tilde{Y}_{t+\tau}]}{(1+r_{f,t+\tau})^{\tau}} \quad (10.19)$$

which is identical to Eq. (10.7) once we recall, from Eq. (10.9), that the risk-free rate must be constant under risk neutrality.

Equation (10.18) is fully in harmony with the intuition of Section 10.3: if the representative agent's consumption is highly positively correlated with the security's real cash flows, the

certainty equivalent values of these cash flows will be smaller than their expected values (namely, $\text{cov}(U_1(c_{t+\tau}), Y_{t+\tau}) < 0$). This is so because such a security is not very useful for hedging the agent's future consumption risk. As a result, it will have a low price and a high expected return. In fact, its price will be less than what it would be in an economy of risk-neutral agents (Eq. (10.19)). The opposite is true if the security's cash flows are negatively correlated with the agent's consumption.

10.4.1 The CCAPM and Risk-Neutral Valuation

While the center of our attention in the present chapter is and has been the CCAPM, it is useful to make the connection to Chapter 9's notion of risk-neutral valuation. To see the connection, let $\{\tilde{Y}_t\}$ be an arbitrary uncertain income stream to be priced. Using Eq. (10.17), we can express its price conditional on state s in period t , $q_t(s_t = s)$ as

$$\begin{aligned} q_t(s_t = s) &= E_t \sum_{\tau=1}^{\infty} \frac{\delta^\tau U_1(c_{t+\tau}(s'))}{U_1(c_t(s))} Y_{t+\tau}(s_{t+\tau} = s') \\ &= \sum_{\tau=1}^{\infty} \sum_{s'} \frac{\delta^\tau U_1(c_{t+\tau}(s'))}{U_1(c_t(s))} Y_{t+\tau}(s_{t+\tau} = s') \pi(s_{t+\tau} = s'; s_t = s) \end{aligned} \quad (10.20)$$

From Eq. (10.16), we also know that $(1 + r_{f,t+\tau}(s))^\tau q_t^{b,\tau}(s) = 1$, where $r_{f,t+\tau}(s)$ is the per period rate of return on a risk-free discount bond with price $q_t^{b,\tau}(s)$ paying one unit of the consumption good in every state that may materialize at time $t + \tau$. Accordingly,

Eq. (10.20) may be written as

$$\begin{aligned} q_t(s_t = s) &= \sum_{\tau=1}^{\infty} \sum_{s'} \frac{1}{(1 + r_{f,t+\tau}(s))^\tau} \left[\frac{\frac{\delta^\tau U_1(c_{t+\tau}(s'))}{U_1(c_t(s))}}{q_t^{b,\tau}(s)} \right] Y_{t+\tau}(s_{t+\tau} = s') \pi(s_{t+\tau} = s'; s_t = s) \\ &= \sum_{\tau=1}^{\infty} \sum_{s'} \frac{1}{(1 + r_{f,t+\tau}(s))^\tau} \left[\frac{\frac{\delta^\tau U_1(c_{t+\tau}(s')) \pi(s_{t+\tau} = s'; s_t = s)}{U_1(c_t(s))}}{\sum_{s'} \frac{\delta^\tau U_1(c_{t+\tau}(s')) \pi(s_{t+\tau} = s'; s_t = s)}{U_1(c_t(s))}} \right] Y_{t+\tau}(s_{t+\tau} = s') \\ &= \sum_{\tau=1}^{\infty} \sum_{s'} \frac{1}{(1 + r_{f,t+\tau}(s))^\tau} \pi^{RN}(s_{t+\tau} = s'; s_t = s) Y_{t+\tau}(s_{t+\tau} = s') \\ &= \sum_{\tau=1}^{\infty} \frac{1}{(1 + r_{f,t+\tau}(s))^\tau} E_{t+\tau}^{RN} Y_{t+\tau}(s_{t+\tau} = s') \end{aligned} \quad (10.21)$$

As in Chapter 9, the asset is once again priced equal to its expected future cash flows compiled using the risk-neutral probabilities, discounted at the risk-free rate. The applicable risk-free rate is endogenous to the underlying consumption process and may differ for different times to maturity. Note also that

$$\pi^{RN}(s_{t+\tau} = s'; s_t = s) = \left[\frac{\frac{U_1(c_{t+\tau}(s'))\pi(s_{t+\tau} = s'; s_t = s)}{U_1(c_t(s))}}{\sum_{s'} \frac{U_1(c_{t+\tau}(s'))\pi(s_{t+\tau} = s'; s_t = s)}{U_1(c_t(s))}} \right]$$

For those future period $t + \tau$ states for which consumption is very low relative to current consumption $c_t(s)$, the marginal utility of consumption is high relative to current consumption marginal utility. In this case, the risk-neutral probability of the low consumption (“bad”) state is accorded a weight $((U_1(c_{t+\tau}(s')))/(U_1(c_t(s))))$ greater than one relative to the true probability, while correspondingly high consumption states are premultiplied by numbers less than one. It is in this sense that risk-neutral probabilities are “pessimistic” (cf. Section 9.6) relative to the objective state probabilities.

Viewed in this light the CCAPM can be seen as translating the notion of risk-neutral valuation into a setting where there is a direct theoretical connection between the macroeconomy (aggregate consumption) and the financial markets where assets are priced. To evaluate the CCAPM, we will “feed” the observed stochastic process on US consumption into the model and explore whether the model implied asset pricing relationships and derived rates of return reasonably replicate what is observed in actual return data. If they do, we can be more confident that the CCAPM is a reasonable basis for understanding asset pricing relationships.

10.5 Testing the CCAPM: The Equity Premium Puzzle

In this section, we discuss the empirical validity of the CCAPM. Unfortunately, a set of simple and robust empirical observations has been put forward that falsifies this model in an unusually strong way. As a result, we are led to question the model’s underlying hypotheses and, a fortiori, those assumptions underlying some of the less sophisticated models seen before. In this instance, the recourse to sophisticated econometrics for drawing significant lessons about our approach to modeling financial markets is superfluous.

A few key empirical observations regarding financial returns in US markets are summarized in [Table 10.2](#), which shows that over a long period of observation the average *ex post* return on a diversified portfolio of US stocks (the market portfolio, as approximated in the United States by the S&P 500) has been close to 7% (in real terms, net of inflation) while the return on 1-year T-bills (taken to represent the return on the risk-free asset) has averaged

Table 10.2: Properties of US asset returns

	US Economy	
	(a)	(b)
R	6.98	16.54
r_f	0.80	5.67
$r - r_f$	6.18	16.67

(a) Annualized mean values in percent.

(b) Annualized standard deviation in percent.

Source: Data from [Mehra and Prescott \(1985\)](#).

less than 1%. These twin observations make up for an equity risk premium of 6.2%. This observation is robust in the sense that it has applied in the United States for a very long period and in several other important countries as well (see Chapter 2). Its meaning is not totally undisputed, however. [Goetzmann and Jorion \(1999\)](#), in particular, argue that the high return premium obtained for holding US equities is the exception rather than the rule.⁷

Here we will take the 6% equity premium at face value, as has much of the huge literature that followed the uncovering of the *equity premium puzzle* by [Mehra and Prescott \(1985\)](#). The puzzle is this: Mehra and Prescott argue that the CCAPM is completely unable, once reasonable parameter values are inserted in the model, to replicate such a high observed equity premium.

Let us illustrate their reasoning.⁸ According to the CCAPM, the only factors determining the characteristics of security returns are the representative agent's utility function, his subjective discount factor, and the process on consumption (which equals output or dividends in the exchange economy equilibrium). First, consider the utility function. It is natural in light of the development in Chapter 4 and the requirements for the existence of a representative agent ([Box 10.1](#)) to assume that the agent's period utility function displays constant relative risk aversion (CRRA); thus, let us set

$$U(c) = \frac{c^{1-\gamma}}{1-\gamma}$$

⁷ Using shorter, mostly postwar, data, premia close to or even higher than the US equity premium are obtained for France, Germany, the Netherlands, Sweden, Switzerland, and the United Kingdom (see, for example, [Campbell, 1998](#)). Goetzmann and Jorion, 1999, however, argue that such data samples do not correct for crashes and period of market interruptions, often associated with World War II and thus are not immune to survivorship bias. To correct for such a bias, they assemble long data series for all markets that existed during the twentieth century. They find that the United States has had "by far the highest uninterrupted real rate of appreciation of all countries, at about 5% annually. For other countries, the median appreciation rate is about 1.5%.

⁸ The discussion that follows in this section is based on work by Rajnish Mehra.

Empirical studies associated with this model have placed γ in the range of (1, 2). A convenient consequence of this utility specification is that the intertemporal MRS can be written as

$$\frac{U_1(c_{t+1})}{U_1(c_t)} = \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} \quad (10.22)$$

The second major ingredient is the consumption process. In our version of the model, consumption is a stationary process: It does not grow through time. In reality, however, consumption *is* growing through time. In a growing economy, the analogous notion to the variability of consumption is variability in the growth rate of consumption.

Let $g_{t+1} = (c_{t+1}/c_t)$ denote per capita consumption growth, and assume, for illustration, that g_t is independently and identically lognormally distributed through time. For the period 1889 through 1978, the US economy aggregate consumption has been growing at an average rate of 1.83% annually, with a standard deviation of 3.57% and a slightly negative measure of autocorrelation (-0.14) (cf. [Mehra and Prescott, 1985](#)).

The remaining item is the agent's subjective discount factor δ : What value should it assume? Time impatience requires, of course, that $\delta < 1$, but this is insufficiently precise. One logical route to its estimation is as follows: Roughly speaking, the equity in the CCAPM economy represents a claim to the aggregate income from the underlying economy's entire capital stock. We have just seen that, in the United States, equity claims to private capital flows average a 7% annual real return, while debt claims average 1%.⁹ Furthermore, the economywide debt-to-equity rates are not very different from 1. These facts together suggest an overall average real annual return to capital of about 4%.

If there were no uncertainty in the model, and if the constant growth rate of consumption were to equal its long-run historical average (1.0183), the asset pricing [Eq. \(10.8\)](#) would reduce to

$$1 = \delta E_t \left\{ \left(\frac{\tilde{c}_{t+1}}{c_t} \right)^{-\gamma} \tilde{R}_{t+1} \right\} = \delta(\bar{g})^{-\gamma} \bar{R} \quad (10.23)$$

where \tilde{R}_{t+1} is the gross rate of return on capital and the upper bars denote historical averages.¹⁰ For $\gamma = 1$, $\bar{g} = 1.0183$, and $\bar{R} = 1.04$, we can solve for the implied δ to obtain $\delta \cong 0.97$. Since we have used an annual estimate for \bar{g} , the resulting δ must be viewed as an annual or yearly subjective discount factor; on a quarterly basis it corresponds to $\delta \cong 0.99$. If, on the other hand, we want to assume $\gamma = 2$, [Eq. \(10.19\)](#) solves for $\delta \cong 0.99$ on an annual basis, yielding a quarterly δ even closer to 1. This reasoning demonstrates that assuming

⁹ Strictly speaking, these are the returns to publicly traded debt and equity claims. If private capital earns substantially different returns, however, capital is being inefficiently allocated; we assume this is not the case.

¹⁰ Time averages and expected values should coincide in a stationary model, provided the time series is of sufficient length.

higher rates of risk aversion would be incompatible with maintaining the hypothesis of a time discount factor less than 1. While in the case of positive consumption growth, we could technically entertain the possibility of a negative rate of time preference, and thus of a discount factor larger than 1, we rule it out on grounds of plausibility.

At the root of this difficulty is the low return on the risk-free asset (1%), which will haunt us in other ways. As we know, highly risk-averse individuals want to smooth consumption over time, meaning they want to transfer consumption from good times to bad times. When consumption is growing predictably, the good times lie in the future. Agents want to borrow now against their future income. In a representative agent model, it is difficult to reconcile growing consumption with a low rate on borrowing: everyone is on the same side of the market, a fact that inevitably forces a higher rate. This problem calls for an independent explanation for the abnormally low average risk-free rate (e.g., in terms of the liquidity advantage of short-term government debt as in [Bansal and Coleman, 1996](#)) or the acceptance of the possibility of a negative rate of time preference so that future consumption is given more weight than present consumption. We will not follow either of these routes here, but rather will, in the course of the present exercise, limit the coefficient of relative risk aversion to a maximum value of 2.

With these added assumptions we can manipulate the fundamental asset pricing [Eq. \(10.3\)](#) to yield two equations that can be used indirectly to test the model. The key step in the reasoning is to demonstrate that, in the context of these assumptions, the equity price formula takes the form

$$q_t^e = vY_t$$

where v is a constant coefficient. That is, the stock price at date t is proportional to the dividend paid at date t ([Box 10.2](#)).¹¹ To confirm this statement, we use a standard trick consisting of guessing that this is the form taken by the equilibrium pricing function, and then verifying that this guess is indeed borne out by the structure of the model. Under the $q_t^e = vY_t$ hypothesis, [Eq. \(10.3\)](#) becomes

$$vY_t = \delta E_t \left\{ (v\tilde{Y}_{t+1} + \tilde{Y}_{t+1}) \frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)} \right\}$$

Using [Eq. \(10.22\)](#) and dropping the conditional expectations operator, since \tilde{g}_t is independently and identically distributed (i.i.d.) through time (its mean is independent of time), we can rewrite this equation as

$$v = \delta E \left\{ (v + 1) \frac{\tilde{Y}_{t+1}}{Y_t} \tilde{g}_{t+1}^{-\gamma} \right\}$$

¹¹ Note that this property holds true as well for the example developed in [Box 10.2](#) as Eqs. (iv) and (v) attest.

BOX 10.2 Calculating the Equilibrium Price Function

[Equation \(10.4\)](#) implicitly defines the equilibrium price series. Can it be solved directly to produce the actual equilibrium prices $\{q(Y^j) : j = 1, 2, \dots, N\}$? The answer is positive. First, we must specify parameter values and functional forms. In particular, we need to select values for δ and for the various output levels Y^j , to specify the probability transition matrix \mathbf{T} and the form of the representative agent's period utility function (a CRRA function of the form $U(c) = (c^{1-\gamma})/(1 - \gamma)$ is a natural choice). We may then proceed as follows.

Solve for the $\{q(Y^j) : j = 1, 2, \dots, N\}$ as the solution to a system of linear equations. Note that [Eq. \(10.4\)](#) can be written as the following system of linear equations (one for each of the N possible current states Y^j):

$$\begin{aligned} U_1(Y^1)q(Y^1) &= \delta \sum_{j=1}^N \pi_{1j} U_1(Y^j) Y^j + \delta \sum_{j=1}^N \pi_{1j} U_1(Y^j) q(Y^j) \\ &\quad \vdots \quad \vdots \\ U_1(Y^N)q(Y^N) &= \delta \sum_{j=1}^N \pi_{Nj} U_1(Y^j) Y^j + \delta \sum_{j=1}^N \pi_{Nj} U_1(Y^j) q(Y^j) \end{aligned}$$

with unknowns $q(Y^1), q(Y^2), \dots, q(Y^N)$. Note that for each of these equations, the first term on the RHS is simply a number, while the second term is a linear combination of the $q(Y^j)$ s. Barring a very unusual output process, this system will have a solution: one price for each Y^j , i.e., the equilibrium price function.

Let us illustrate: Suppose $U(c) = \ln(c)$, $\delta = 0.96$ and $(Y^1, Y^2, Y^3) = (1.5, 1, 0.5)$ —an exaggeration of *boom*, *normal*, and *depression* times. The transition matrix is taken to be as found in [Table 10.3](#).

The equilibrium conditions implicit in [Eq. \(10.4\)](#) then reduce to

$$\begin{aligned} Y^1 : \frac{2}{3}q(1.5) &= 0.96 + 0.96\left\{\frac{1}{3}q(1.5) + \frac{1}{4}q(1) + \frac{1}{2}q(0.5)\right\} \\ Y^2 : q(1) &= 0.96 + 0.96\left\{\frac{1}{6}q(1.5) + \frac{1}{2}q(1) + \frac{1}{2}q(0.5)\right\} \\ Y^3 : 2q(0.5) &= 0.96 + 0.96\left\{\frac{1}{6}q(1.5) + \frac{1}{4}q(1) + 1q(0.5)\right\} \end{aligned} \tag{i}$$

or,

Table 10.3: Transition matrix

1.5	1	0.5
1.5	0.5	0.25
1	0.25	0.5
0.5	0.25	0.25

(Continued)

BOX 10.2 Calculating the Equilibrium Price Function (Continued)

$$Y^1 : 0 = 0.96 - 0.347q(1.5) + 0.24q(1) + 0.48q(0.5)$$

$$Y^2 : 0 = 0.96 + 0.16q(1.5) - 0.52q(1) + 0.48q(0.5) \quad (\text{ii})$$

$$Y^3 : 0 = 0.96 + 0.16q(1.5) + 0.24q(1) - 1.04q(0.5) \quad (\text{iii})$$

$$(\text{i}) - (\text{ii}) \text{ yields : } q(1.5) = \frac{0.76}{0.507}q(1) = 1.5q(1) \quad (\text{iv})$$

$$(\text{ii}) - (\text{iii}) \text{ gives : } q(0.5) = \frac{0.76}{1.52}q(1) = 1/2q(1) \quad (\text{v})$$

substituting Eqs. (iv) and (v) into Eq. (i) to solve for $q(1)$ yields $q(1) = 24$; $q(1.5) = 36$ and $q(0.5) = 12$ follow.

The market-clearing condition implies that $(Y_{t+1}/Y_t) = g_{t+1}$, thus

$$\begin{aligned} v &= \delta E\{(v+1)\tilde{g}^{1-\gamma}\} \\ &= \frac{\delta E\{\tilde{g}^{1-\gamma}\}}{1 - \delta E\{\tilde{g}^{1-\gamma}\}} \end{aligned}$$

This is indeed a constant, and our initial guess is thus confirmed!

Taking advantage of the validated pricing hypothesis, we can rewrite the equity return as

$$\tilde{R}_{t+1} = 1 + \tilde{r}_{t+1} = \frac{\tilde{q}_{t+1}^e + \tilde{Y}_{t+1}}{q_t^e} = \frac{v+1}{v} \frac{\tilde{Y}_{t+1}}{Y_t} = \frac{v+1}{v} \tilde{g}_{t+1}$$

Taking expectations, we obtain

$$E_t(\tilde{R}_{t+1}) = E(\tilde{R}_{t+1}) = \frac{v+1}{v} E(\tilde{g}_{t+1}) = \frac{E(\tilde{g})}{\delta E\{\tilde{g}^{1-\gamma}\}}$$

The risk-free rate is (Eq. (10.9))

$$R_{f,t+1} \equiv \frac{1}{q_t^b} = \left[\delta E_t \left\{ \frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)} \right\} \right]^{-1} = \frac{1}{\delta E\{\tilde{g}^{1-\gamma}\}} \quad (10.24)$$

which is seen to be constant under our current hypotheses.

Taking advantage of the lognormality assumption, we can express the ratio of the two preceding equations as (see Appendix 10.2 for details)

$$\frac{E(\tilde{R}_{t+1})}{R_f} = \frac{E(\tilde{g})E(\tilde{g}^{-\gamma})}{E(\tilde{g}^{1-\gamma})} = \exp[\gamma\sigma_g^2] \quad (10.25)$$

where σ_x^2 is the variance of $\ln x$. Taking logs, we finally obtain

$$\ln(ER) - \ln(R_f) = \gamma\sigma_g^2 \quad (10.26)$$

Now, we are in a position to confront the model with the data. Let us start with Eq. (10.26). Feeding in the return characteristics of the US economy and solving for γ , we obtain (see Appendix 10.2 for the computation of σ_g^2),

$$\frac{\ln(ER) - \ln(E_{rf})}{\sigma_x^2} = \frac{1.0698 - 1.008}{(0.0357)^2} = 50.24 = \gamma$$

Alternatively, if we assume $\gamma = 2$ and multiply by σ_g^2 as per Eq. (10.26), one obtains an equity premium of

$$2(0.00123) = 0.002 = (\ln(ER) - \ln(E_{rf})) \cong ER - E_{rf} \quad (10.27)$$

In either case, this reasoning identifies a major discrepancy between model prediction and reality. The observed equity premium can only be explained by assuming an extremely high coefficient of relative risk aversion ($\cong 50$), one that is completely at variance with independent estimates. An agent with risk aversion of this level would be too fearful to take a bath (many accidents involve falling in a bathtub!) or to cross the street. On the other hand, insisting on a more reasonable coefficient of risk aversion of 2 leads to predicting a minuscule premium of 0.2%, much below the 6.2% that has been historically observed over long periods.¹²

Similarly, it is shown in Appendix 10.2 that $E\{g_t^{-\gamma}\} = 0.97$ for $\gamma = 2$; Eq. (10.24) and the observed value for R_f (1.008) then implies that δ should be larger than 1 (1.02). This problem was to be anticipated from our discussion of the calibration of δ , which was based on reasoning similar to that underlying Eq. (10.24). Here the problem is compounded by the fact that we are using an even lower risk-free rate (0.8%) rather than the steady-state rate of return on capital of 4% used in the prior reasoning. In the present context, this difficulty in calibrating δ or, equivalently, in explaining the low rate of return on the risk-free asset has been dubbed the *risk-free rate puzzle* by Weil (1989). As noted previously, we read this

¹² Azeredo (2012) argues that in the pre-1929 period of the Mehra and Prescott (1985) data set, per capita consumption was mismeasured in the standard reported statistics. If it is measured in a way that is more in keeping with practice in the post-1929 period, the serial correlation in consumption growth over the entire sample period (1889–1978) is 0.42, not –0.14 as Mehra and Prescott (1985) report. When their model is recalibrated to incorporate this high, positive serial correlation in consumption, the model-generated equity premium turns progressively more negative as the CRRAs increase in excess of 2.2. High persistent consumption growth causes even a severely risk averse investor to view the equity security as less risky—and more valuable—than the risk-free one. Once again it is not “cash flow” variation *per se* that measures risk but the pattern of that variation relative to an investor’s consumption.

result as calling for a specific explanation for the observed low return on the risk-free asset, one that the CCAPM is not designed to provide.

10.6 Testing the CCAPM: Hansen–Jagannathan Bounds

Another, parallel perspective on the puzzle is provided by the Hansen–Jagannathan (1991) bound. The idea is very similar to our prior test, and the end result is the same. The underlying reasoning, however, postpones as long as possible making specific modeling assumptions. It is thus more general than a test of a specific version of the CCAPM. The bound proposed by Hansen and Jagannathan potentially applies to other asset pricing formulations and it similarly leads to a falsification of the standard CCAPM.

The reasoning goes as follows: Let q_t denote the price of an arbitrary asset. For all homogeneous agent economies, the fundamental equilibrium asset pricing Eq. (10.3) can be expressed as

$$q(s_t) = E_t[m_{t+1}(\tilde{s}_{t+1})X_{t+1}(\tilde{s}_{t+1}); s_t] \quad (10.28)$$

where s_t is the state today (it may be today's output in the context of a simple exchange economy, or it may be something more elaborate as in the case of a production economy), $X_{t+1}(\tilde{s}_{t+1})$ is the total (consumption) return in the next period to the asset owner (e.g., in the case of an exchange economy this equals $(\tilde{q}_{t+1} + \tilde{Y}_{t+1})$) and $m_{t+1}(\tilde{s}_{t+1})$ is the *equilibrium pricing kernel*, also known as the *stochastic discount factor* (SDF):

$$m_{t+1}(\tilde{s}_{t+1}) = \frac{\delta U_1(c_{t+1}(\tilde{s}_{t+1}))}{U_1(c_t)}$$

As before $U_1()$ is the marginal utility of the representative agent and c_t is his equilibrium consumption. Equation (10.28) is thus the general statement that the price of an asset today must equal the expectation of its total payout tomorrow multiplied by the appropriate pricing kernel. For notational simplicity, let us suppress the state dependence, leaving it as understood, and write Eq. (10.28) as

$$q_t = E_t[\tilde{m}_{t+1}\tilde{X}_{t+1}] \quad (10.29i)$$

This is equivalent to

$$1 = E_t[\tilde{m}_{t+1}\tilde{R}_{t+1}] \quad (10.29ii)$$

where \tilde{R}_{t+1} is the gross rate of return to ownership of the asset. Since Eq. (10.29ii) holds for each state s_t , it also holds unconditionally; we thus can also write

$$1 = E[\tilde{m}\tilde{R}]$$

where E denotes the unconditional expectation. For any two assets i and j (to be viewed shortly as the return on the market portfolio and the risk-free return, respectively) it must, therefore, be the case that

$$E[\tilde{m}(\tilde{R}_i - \tilde{R}_j)] = 0, \text{ or}$$

$$E[\tilde{m}\tilde{R}_{i-j}] = 0$$

where, again for notational convenience, we substitute \tilde{R}_{i-j} for $\tilde{R}_i - \tilde{R}_j$. This latter expression furthermore implies the following series of relationships:

$$\begin{aligned} E\tilde{m}E\tilde{R}_{i-j} + \text{cov}(\tilde{m}, \tilde{R}_{i-j}) &= 0, \text{ or} \\ E\tilde{m}E\tilde{R}_{i-j} + \rho(\tilde{m}, \tilde{R}_{i-j})\sigma_m\sigma_{R_{i-j}} &= 0, \text{ or} \\ \frac{E\tilde{R}_{i-j}}{\sigma_{R_{i-j}}} + \rho(\tilde{m}, \tilde{R}_{i-j})\frac{\sigma_m}{E\tilde{m}} &= 0, \text{ or} \\ \frac{E\tilde{R}_{i-j}}{\sigma_{R_{i-j}}} &= -\rho(\tilde{m}, \tilde{R}_{i-j})\frac{\sigma_m}{E\tilde{m}} \end{aligned} \quad (10.30)$$

It follows from Eq. (10.30) and the fact that a correlation is never larger than 1 that

$$\frac{\sigma_m}{E\tilde{m}} > \frac{|E\tilde{R}_{i-j}|}{\sigma_{R_{i-j}}} \quad (10.31)$$

The inequality in expression (10.31) is referred to as the Hansen–Jagannathan lower bound on the pricing kernel. If, as noted earlier, we designate asset i as the market portfolio and asset j as the risk-free return, then the data from Table 10.2 and Eq. (10.31) together imply (for the US economy):

$$\frac{\sigma_m}{E\tilde{m}} > \frac{|E(\tilde{r}_M - r_f)|}{\sigma_{r_M - r_f}} = \frac{0.062}{0.167} = 0.37$$

Let us check whether this bound is satisfied for our model. From Eq. (10.22), $\tilde{m}(\tilde{c}_{t+1}, c_t) = \delta(g_{t+1})^{-\gamma}$, the expectation of which can be computed (see Appendix 10.2) to be

$$E\tilde{m} = \delta \exp\left(-\gamma\mu_g + \frac{1}{2}\gamma^2\sigma_g^2\right) = 0.99(0.967945) = 0.96 \text{ for } \gamma = 2$$

In fact, Eq. (10.28) reminds us that $E\tilde{m}$ is simply the expected value of the price of a one-period risk-free discount bound, which cannot be very far away from 1. This implies that for the Hansen–Jagannathan bound to be satisfied, the standard deviation of the pricing kernel cannot be much lower than 0.3; given the information we have on \tilde{g}_t , it is a short

step to estimate this parameter numerically under the assumption of lognormality. When we perform the calculation (again, see [Appendix 10.2](#)), we obtain an estimate for $\sigma_m = 0.002$, which is an order of magnitude lower than what is required for [Eq. \(10.31\)](#) to be satisfied. The message is that it is very difficult to get the equilibrium pricing kernel volatility to be anywhere near the required level. In a homogeneous agent, complete market model with standard preferences, where the variation in equilibrium consumption matches the data, consumption is just too smooth. As a result, the marginal utility of consumption does not vary sufficiently to satisfy the bound implied by the data unless the curvature of the utility function—the degree of risk aversion—is assumed to be astronomically high, an assumption which, as we have seen, raises problems of its own.

10.7 The SDF in Greater Generality

The notion of a stochastic discount factor/pricing kernel \tilde{m}_t turns out to be a very general idea with the form $\delta((U_1(c_{t+1}(\tilde{s}_{t+1}))) / (U_1(c_t(s_t))))$ being only a very high-profile, special case with a particularly well-defined economic story. The more general perspective may be expressed as follows: For a sufficiently rich set of security payoffs \tilde{X}_t and their associated prices $q(\tilde{X}_t)$, there always exists a unique pricing kernel \tilde{m}_t , such that

$q(\tilde{X}_t) = E\tilde{m}_{t+1}\tilde{X}_{t+1}$. In the present section, we propose to introduce the idea and illustrate its use. For simplicity of presentation our focus will be on pricing assets with one-period payoffs and we drop the time subscript.

Payoffs \tilde{X} are random variables defined on some probability space of events \mathcal{P} . Let the set of eligible security payoffs χ be defined as:

$$\chi = \{\tilde{X}: E\tilde{X}^2 < \infty, \tilde{X} \text{ a random variable defined on } \mathcal{P}\}$$

with the expectations operator E defined with respect to the probability space \mathcal{P} .

The set χ has the property that if \tilde{X} and \tilde{Z} are two distinct asset payoffs in χ , and if a and b are any two real numbers, then $a\tilde{X} + b\tilde{Z} \in \chi$. With this property, χ is said to constitute a linear space.

Now consider an arbitrary pricing function $q: \chi \mapsto R$ such that for any payoff $\tilde{X} \in \chi$, $q(\)$ assigns a price to this payoff, $q(\tilde{X})$.¹³ It is natural to assume that $q(\)$ respects the law of one price (LOP), which is simply to say that $q(\)$ assigns only one price to each $\tilde{X} \in \chi$. An absence of arbitrage opportunities is sufficient for this to be true. We are then led to the following straightforward preliminary result.

¹³ The expression “pricing function” is just “formalism”: $q(\tilde{X})$ is simply the observed price in the securities markets of the asset with claim to (\tilde{X}) .

Theorem 10.1 The LOP is satisfied for a pricing function $q(\cdot):\chi\mapsto R$ if and only if $q(\cdot)$ is a linear function defined on X .

Proof \Rightarrow Consider two distinct payoffs \tilde{X} and \tilde{Z} in χ with corresponding prices $q(\tilde{X})$ and $q(\tilde{Z})$. Since χ is a linear space $\tilde{W} \equiv a\tilde{X} + b\tilde{Z} \in \chi$ for any real numbers a and b . Let $q(\tilde{W})$ be the price of \tilde{W} .

The payoff \tilde{W} can be created by purchasing a units of the payoff \tilde{X} and b units of the payoff \tilde{Z} , with a price of $aq(\tilde{X}) + bq(\tilde{Z})$. Since $q(\cdot)$ respects the LOP, it must be that $q(\tilde{W}) = aq(\tilde{W}) + bq(\tilde{Z})$. We conclude that the pricing function $q(\cdot)$ is linear on χ .

\Leftarrow Suppose that $q(\cdot)$ is a linear pricing function on χ , but that there exists an $\tilde{X} \in \chi$ for which $q(\tilde{X}) = q_1$ and $q(\tilde{X}) = q_2$. By the linearity property of $q(\cdot)$, $q_1 = q(\tilde{X}) = q\left(\frac{1}{2}\tilde{X} + \frac{1}{2}\tilde{X}\right) = \frac{1}{2}q(\tilde{X}) + \frac{1}{2}q(\tilde{X}) = \frac{1}{2}q_1 + \frac{1}{2}q_2$. It follows from the string of equalities that $q_1 = q_2$ and the LOP holds under the linear pricing function $q(\cdot)$.

The small observation made in [Theorem 10.1](#) has very large implications when seen as a part of [Theorem 10.2](#).

Theorem 10.2 For any linear pricing function $q(\cdot):\chi\mapsto R$, there is a unique $m^* \in \chi$ such that $q(\tilde{X}) = Em^*\tilde{X}$ for all $\tilde{X} \in \chi$. Furthermore, if there are no arbitrage opportunities under $q(\cdot)$, $m^* > 0$. If the underlying financial market is complete m^* is unique.

Proof Application of the Reitz representation theorem; the proof and examples are found in the Web Notes to this chapter.

[Theorem 10.2](#) merits attention for a number of subtle reasons. First, it is very general: there is no mention of a representative agent whose “preference identity” is not easy to come by (the data certainly suggests that it is not captured by the utility function $U(c) = (c^{1-\gamma})/(1-\gamma)$). All that we know is that the pricing kernel m^* is of the same form as the asset payoffs: if X is the set of normally distributed payoffs, m^* will assume the form of a normal distribution. More generally, there is no specific equilibrium model underlying the conclusion to [Theorem 10.2](#), a feature that is both a strength and a weakness.

It is a strength because it suggests that the overall general equilibrium pricing perspective adopted in this chapter is a reasonable one since it leads to a pricing relationship that must exist theoretically. The obvious downside to [Theorem 10.2](#) is that it is totally uninformative as to how the economic structures (preferences, technologies, and markets) give rise to the asset prices we see arising out of the financial markets. In this sense, it does not contribute to the goal of this text!

Second, as we will see in later chapters, there are useful contexts where the relevant m^* can be precisely calculated from no arbitrage relationships, without reference to an economic model. Finance professionals welcome this feature. Note that the identification of the

pricing kernel m^* of [Theorem 10.2](#) is essentially the same as identifying the appropriate risk-neutral probabilities or Arrow–Debreu state prices.

There is one refinement of the conclusion to [Theorem 10.2](#) which concerns the form of the postulated m^* and recalls the by-now-customary notion of an efficient portfolio. We present this refinement in [Corollary 10.1](#).

Corollary 10.1 Consider the pricing kernel m^* of [Theorem 10.2](#). Then there exist associated (with m^*) constants a and b and an associated MV-efficient portfolio $p(m^*)$ such that

$$m^* = a + br_{p(m^*)}$$

where $r_{p(m^*)}$ is the return on that portfolio.

The proof of [Corollary 10.1](#) is constructive and not immediately intuitive so we stop the discussion at this point and invite the reader to refer to the Web Notes where the proof is taken up in detail.¹⁴

10.8 Some Extensions

10.8.1 Reviewing the Diagnosis

Our first dynamic general equilibrium model thus fails when confronted with actual data. Let us review the source of this failure. Recall our original pricing [Eq. \(10.9\)](#), specialized for a single asset, the market portfolio:

$$\begin{aligned} \bar{r}_{M,t+1} - r_{f,t+1} &= \delta(1 + r_{f,t+1})\text{cov}_t\left(\frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)}, \tilde{r}_{M,t+1}\right) \\ &= -\delta(1 + r_{f,t+1})\rho\left(\frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)}, \tilde{r}_{M,t+1}\right)\sigma\left(\frac{U_1(\tilde{c}_{t+1})}{U_1(c_t)}\right)\sigma(\tilde{r}_{M,t+1}) \\ &= -(1 + r_{f,t+1})\rho(\tilde{m}_t, \tilde{r}_{M,t+1})\sigma(\tilde{m}_t)\sigma(\tilde{r}_{M,t+1}) \end{aligned}$$

Written in this way, it is clear that the equity premium depends upon the standard deviation of the MRS (or, equivalently, the stochastic discount factor), the standard deviation of the return on the market portfolio, and the correlation between these quantities. For the United States, and most other industrial countries, the problem with a model in which pricing and return relationships depend so much on consumption (and thus MRS) variation is that average per

¹⁴ The statement of the Corollary becomes a bit more believable if we write $q(\tilde{X}) = E\tilde{m}^*\tilde{X}, \tilde{X} \in \chi$ as $1 = E\tilde{m}^*(\tilde{X}/q(\tilde{X})) = E\tilde{m}^*\tilde{R}_{\tilde{X}}$ where $\tilde{R}_{\tilde{X}}$ is the gross return on the asset with payoff $\tilde{X} \in \chi$. Written this way, with $\tilde{R}_{\tilde{X}}$ a return function, so also must \tilde{m}^* be a return function by [Theorem 10.2](#).

capita consumption does not vary much at all. If this model is to have any hope of matching the data, we must modify it in a way that will increase the standard deviation of the relevant MRS, or the variability of the dividend being priced (and thus the $\sigma(r_{M,t+1})$). We do not have complete freedom over this latter quantity, however, as it must be matched to the data as well.

These thoughts suggest a number of “strategies” for resolving the joint equity premium and risk-free rate puzzles:

1. Modify the representative agent’s utility function in such a way that he is much more sensitive to consumption variation than is evident in the CRRA specification. Most contemporary theories go down this route in some way. We are reminded, however, the [Constantinides’ \(1982\)](#) construction of a representative agent presumes that investor preferences are VNM-expected utility, a requirement that constrains “creativity” along the preference dimension.
2. Uncover some feature of the consumption growth process which, although consistent with the basic consumption growth data in [Section 10.5](#), is nevertheless highly objectionable to the representative agent, thus requiring an offsetting additional risk premium on any asset whose payout is closely aligned with this feature. It is “additional” in the sense that it goes beyond the premium for quarterly aggregate dividend risk.
3. Argue that it is not the “representative agent” who bears stock market risk, but the fraction of the population that owns stocks and actually trades them. Government data reveals that equity ownership is highly concentrated: for the US 20% of the population owned 92% of all stock in the year 2010. It is thus the consumption and income processes of this group alone that should be relevant to CCAPM modeling.

To date, most efforts to resolve the equity premium cum risk-free rate puzzles involve some combination of the first and second strategies. At the same time, the set of time series statistics that a model is challenged to explain has expanded. Presently, any CCAPM-style model, if it is to be accorded much credibility, must explain not only the mean market equity return and risk-free rate (and thus also the equity premium), but also the volatilities of these return series as well. Furthermore, all these statistics must be replicated in an environment where the mean and SD of the growth rates of consumption, \tilde{g} , and dividends, \tilde{g}_{dir} , also well approximate their empirical counterparts. Using data from [Bansal and Yaron \(2004\)](#), whose work we will review shortly, [Table 10.4](#) refreshes our recollection as to the magnitudes of the relevant quantities involved.

The list now includes such other concerns as demonstrating predictability patterns in stock return data.¹⁵

¹⁵ The basic predictability phenomenon is most simply laid out in Cochrane (2011). It is the observation that regressions of the form $R_{t+j} = \alpha_j + \beta_j(D_t/P_t) + \varepsilon_j$ have substantial R^2 and highly significant β_j coefficients for $j = 5$. The expression (D_t/P_t) is the market aggregate dividend/price ratio in period t and R_{t+j} is the cumulative market return over the subsequent j periods. It is in the context of regressions of this type that dividend price ratios are said to predict returns.

Table 10.4: Financial statistics^a

	Mean	SD
r^e	7.19	19.42
r^f	0.86	0.97
r^P	6.33	19.2
G	1.8	2.93
g_{div}	1.8	11.49
$(P/D)_t$	15.5 ^b	25.56

^aSource: Bansal and Yaron (2004); based on data for the period 1928–1998. All numbers measured in percent, annualized.

^bFor the period 1871 to the present time (March 5, 2014) the average (P/D) ratio for the S&P₅₀₀ (or its predecessors) is the figure indicated. As of this writing (March 5, 2014) the actual (P/D) ratio is 19.46.

In the remaining sections of this chapter, we review a number of modeling approaches to explaining at least the basic return statistics of Table 10.4¹⁶ Most are straightforward adaptations of the basic Lucas (1978) tree model of Section 10.3 but generalized to reflect the observed growth rates of consumption and dividends (see Appendix 10.1). Explaining the sources of these most basic return regularities is important not only for progress in financial economics, but also in business cycle and growth theory which are based on the same CCAPM paradigm enriched with a production sector so that the equilibrium consumption/dividend series become endogenous to the model.

10.8.2 Adding a Disaster State

Here the story goes back to Reitz (1988), but remains, in its various manifestations, under active consideration. The notion is as follows: suppose there is a very low probability state of the world (e.g., the Great Depression of the 1930s—a once per 100 years event) in which consumption growth (simultaneously dividend and output growth) turns negative. Such an event will have enormous consequences for investor welfare, *ceteris paribus*, since *it will lead to lower future consumption levels in all future periods*. Accordingly, equity securities, whose returns are highly correlated with consumption growth, will be very unattractive to investors and must pay high average returns if investors are to be persuaded to hold them. So the logic goes. Adding a low probability disaster state is one example of strategy (2).

To explore the power of this idea Reitz (1988) proposed a very modest generalization of the original Mehra and Prescott (1985) model to accommodate a “disaster state.” In particular,

¹⁶ See Lengwiler (2004) for a complementary, and very thorough, discussion of the earlier literature. See also Mehra (2012) for a review of more recent developments.

he describes the evolution of consumption growth by a *three-state* Markov transition matrix of the form:

$$\begin{array}{ccc} & g_1 & g_2 & g_3 \\ \begin{matrix} g_1 = 1 + \bar{g} + \sigma_g \\ g_2 = 1 + \bar{g} - \sigma_g \\ g_3 = \frac{1 + \bar{g}}{2} \end{matrix} & \left[\begin{matrix} \phi & 1 - \phi - \eta & \eta \\ 1 - \phi - \eta & \phi & \eta \\ 1/2 & 1/2 & 0 \end{matrix} \right] = T_1 \end{array}$$

where η is the probability of entering the disaster growth state g_3 from either of the “normal” growth states (g_1, g_2). There is equal probability (1/2) of exiting the disaster state to either of the normal states, and zero probability of persisting in it. With $\eta = 0.003$, $\phi = 0.47$, $\gamma = 5.3$, $\delta = 0.98$, $\bar{g} = 0.018$ and $\sigma_g = 0.036$, e.g., [Reitz \(1988\)](#) obtains $Er^e = 6.15\%$ and $Er_f = 0.89\%$ (he does not report σ_{r^e} or σ_{r_f}). With this set of parameters, the long-run likelihood of being in a disaster state is 0.0029, which is small. With consumption falling to roughly one-half its prior level when the disaster ensues, however, 38 years of average growth are required for consumption to recover only to its precrash level. This is a severe disaster indeed.^{17,18}

This result is suggestive of the power of the “disaster scenario.” It is also an attractive generalization of [Mehra and Prescott \(1985\)](#), since it is entirely consistent with the complete markets-expected utility-representative agent modeling perspective underlying the entirety of this chapter. The subsequent literature is thus principally focused on its

¹⁷ As [Reitz \(1988\)](#) notes, per capita consumption in the United States “only” fell to 78% of its prior level from the period 1929–1933. With $g_3 = 0.75$, he needs a higher disaster probability ($\eta = .008$), higher CRRA ($\gamma = 10$, $\delta = 0.992$, all other parameters unchanged from the case above) to achieve the following results: $Er^e = 6.37\%$, and $Er_f = 2.97\%$. While the average return on equity is about unchanged from the prior example, the risk-free security is in much less relative demand. As a result, its equilibrium price is lower. This outcome is to be expected: the disaster-induced consumption decline is so much less and, despite the increased frequency of entering the disaster state, its long-run likelihood remains less than 1%.

¹⁸ The notion of a “peso problem” is closely related to the concept of a disaster state. A peso problem arises when security returns reflect the possibility of a disaster event even though one has not yet occurred. Suppose a long time series of dividend (equivalently, consumption) growth rates was generated using the [Reitz \(1988\)](#) transition matrix specification T_1 , yet it just so happened that the series displayed no occasions of the disaster state (a low-probability event, certainly, but not an impossible one). A representative agent knowing the form T_1 would price the dividend stream (in the manner of Eq. (10.6)) accordingly to reflect the disaster possibility. To an outside observer with the same utility function as the agent but with only information from the observed time series to guide his calculations, it would appear that the equilibrium price was too low. This (seeming) mispricing phenomenon is referred to as a “peso problem.” The name comes from a set of events in the period July 1974–July 1976 where the futures price of the Mexican peso in terms of US dollars indicated the possibility of a peso devaluation even though none had been observed. The Mexican peso was subsequently devalued relative to the US dollar in August 1976. For a simple discussion of “peso problem” implications for asset pricing, see [Danthine and Donaldson \(1999\)](#).

calibration: Are there actual events representable by a [Reitz \(1988\)](#) disaster; how does the severity of actual disasters compare to the [Reitz \(1988\)](#) parameterizations, what is the comparative duration, and what is the nature of the relative recoveries (note that a [Reitz, 1988](#) disaster is short-lived)? Mehra and Prescott (1988), in particular, find the relevant [Reitz \(1988\)](#) calibrations basically implausible.

Barro (2006) uses the “disaster experiences” of a sample of 35 countries in the 20th century to calibrate a generalized version of the [Reitz \(1988\)](#) model. Within this sample of countries, he finds 60 occasions where GDP per capita fell by at least 15% which becomes his criterion for a “disaster” (the maximum decline occurred in Germany, 64%, during the 1944–1946 period), from which he estimates a constant disaster probability of 1.7% per year. In the [Reitz \(1988\)](#) formulation, this would amount to a $\eta = 0.017$. When a disaster occurs, Barro (2006) assumes, however, that the extent of the disaster (the percentage output decline) is governed by a probability distribution calibrated to match the historical frequency of disaster severities, with a mean value of 0.29. Barro (2006) also allows for a partial bond default, which is modeled as an event that only occurs conditional on a disaster experience with conditional probability of 0.4. The conditional severity of the bondholders’ haircut is assumed to have the same probability density as the disaster severity itself. Basically, Barro (2006) has less severe “consumption disasters” than [Reitz \(1988\)](#), but they are more frequent and affect both bond and stock returns. Under his calibration, Barro (2006) is able to match the equity premium, though at the cost of somewhat excessive risk-free rate estimates. He does not report the volatilities of returns or consumption growth, however.

Barro’s (2006) paper rehabilitated the “disaster scenario” of [Reitz \(1988\)](#) by presenting a model calibration that is more empirically plausible, and most of the subsequent “disaster” literature has been focused on generalizing his model in various ways. Gabaix (2012) adds the feature that both the severity of the disaster and the expected recovery rates are themselves time varying, thereby introducing another source of risk. Note that in [Reitz \(1988\)](#) both the disaster severity (g_3) and recovery rate (immediate) are fixed; in Barrow (2006), the disaster severity is a fixed probability distribution and recovery is also immediate. Gabaix (2012) is able to match a very wide class of both equity regularities (e.g., the premium, the SD of the price dividend ratio) and regularities in the bond market.

Nakamura et al. (2013) also takes off from Barro (2006) along similar dimensions: (1) they assume the representative agent’s preferences are [Epstein–Zin \(1989\)](#)—see Section 5; and they allow (2) disasters and (3) recoveries to evolve over multiple periods (recall that in [Reitz, 1988](#), and Barro, 2006, disasters and recoveries are both of one-period duration). Essentially their focus is on a richer and more realistic evolution of per capita consumption over disasters and recoveries. In their baseline estimation, they are able to achieve a 4.8% premium and a risk-free rate of 0.10% with an estimated CRRA of 6.4. Volatilities are not reported, however.

10.8.3 Habit Formation

Another device for resolving the equity premium puzzle has been to admit utility functions that exhibit higher rates of risk aversion at the margin and thus can translate small variations in consumption into a large variability of the pricing kernel.¹⁹ One way to achieve this objective without being confronted with the risk-free rate puzzle—which is exacerbated if we simply decide to postulate a higher γ —is to admit some form of *habit formation*. This is the notion that the agent’s utility today is determined not only by her absolute consumption level, but also by the relative position of her current consumption vis-à-vis what can be viewed as a *stock of habit*, the latter summarizing either her past consumption history (with more or less weight placed on more distant consumption levels) or the history of aggregate consumption (summarizing in a sense the consumption habits of her neighbors, a “keeping up with the Joneses” effect: see [Abel, 1990](#)). This modeling perspective takes the view that an investor’s utility of consumption is principally affected by departures from his prior consumption history, either his own or that of a social reference group, i.e., departures from what he may have been accustomed to consume or what he may hold as a socially accepted level of consumption. This concept is open to a variety of different specifications, with diverse implications for behavior and asset pricing. The interested reader is invited to consult [Campbell and Cochrane \(1999\)](#) for a review. Here we will be content to illustrate the underlying working principle. To that end, we specify the representative agent’s period preference ordering to be of the form

$$U(c_t, c_{t-1}) \equiv \frac{(c_t - \chi c_{t-1})^{1-\gamma}}{1 - \gamma}$$

where $\chi \leq 1$ is a parameter. In an extreme case, $\chi = 1$, the period utility depends only upon the deviation of current period t consumption from the prior period’s consumption. As we noted earlier, actual data indicate that per capita consumption for the United States and most other developed countries is very smooth. This implies that (c_t/c_{t-1}) is likely to be close to 1 much of the time. For this specification, the agent’s effective (marginal) relative risk aversion reduces to $R_R(c_t) = (\gamma/(1 - (c_{t-1}/c_t)))$; with $c_t \approx c_{t-1}$, the effective $R_R(c)$ will thus be very high, even with a low γ , and the representative agent will behave as though he is very risk averse to consumption variation. With a careful choice of the habit specification, the risk-free rate puzzle can be avoided (see [Constantinides, 1990](#); [Campbell and Cochrane, 1999](#)), and the equity premium increased.

¹⁹ There is a large literature which seeks to find preference representations more amenable to the resolution of the “puzzles” than the basic CRRA family, and we cannot detail that enormous literature here. The interested reader is referred to [Donaldson and Mehra \(2008\)](#) for a reasonably comprehensive review. Many of the proposed orderings are vulnerable to the same criticisms as will be leveled at “habit-formation preferences.”

“Habit-formation preferences” are by now a standard feature of many finance and macrofinance models. Its drawbacks as a modeling device are also well known. In some models, the marginal CRRA exceeds 100 in order for a good match to the basic stylized facts to be achieved, a figure that seems excessive especially for wealthy investors who own a preponderance of outstanding stock.²⁰ In addition, if the growth rate of consumption is modeled as an i.i.d. process, a reasonable first approximation to the data, habit formation can lead to an excessively volatile equilibrium risk-free rate. [Campbell and Cochrane \(1999\)](#) are able to deal with this later drawback but in a habit model variant in which it, curiously, can pay to destroy consumption: the loss in utility during the period in which the consumption is destroyed is more than offset by the enhanced future utility due to a reduced habit! (see [Ljungqvist and Uhlig, 2009](#)). In more general models where consumption is endogenous, habit-formation preferences often end up generating equilibrium consumption sequences which are too smooth relative to data. Perhaps the most basic criticism of the habit construct is that to date we lack any choice-theoretic axiomatic foundations for habit formation within the domain of VNM-expected utility preferences, a fact that may mean they are inconsistent with Constantides’s (1982) representative agent construct. In this sense, habit-formation preferences are essentially “behavioral” in nature, and the need for an underlying theory is recognized. See Rozen (2010) and [Chetty and Szeidl \(2004\)](#) for progress in this direction. Habit-formation preferences are an illustration of strategy (1) for resolving the puzzle.

10.8.4 The CCAPM with Epstein–Zin Utility

At this stage it is interesting to inquire whether, in addition to its intellectual appeal on grounds of generality, [Epstein and Zin’s \(1989\)](#) separation of time and risk preferences might contribute a solution to the equity premium puzzle and more generally, alter our vision of the CCAPM and its message. The work by Nakamura et al. (2013) mentioned earlier suggests this possibility.

Let us start by looking specifically at the equity premium puzzle. It will facilitate our discussion to repeat Eqs. (5.14) and (5.15) defining the Epstein–Zin preference representation (refer to Chapter 5 for a discussion and for the log case):

$$U(c_t, CE_{t+1}) = [(1 - \delta)c_t^{1-\rho} + \delta CE_{t+1}^{1-\rho}]^{\frac{1}{1-\rho}}, \text{ where}$$

$$[CE(\tilde{U}_{t+1})]^{1-\gamma} = E_t(\tilde{U}_{t+1})^{1-\gamma}$$

²⁰ In the United States during the 1990s households in the top 20% of the wealth distribution own 98% of all outstanding stocks.

Weil (1989) uses these preferences in a setting otherwise identical to that of Mehra and Prescott (1985). Asset prices and returns are computed similarly. What he finds, however, is that this greater generality, *per se*, does not resolve the *equity premium puzzle*, but rather tends to underscore what we have already introduced as the *risk-free rate puzzle*.

The Epstein–Zin (1989, 1991) preference representation does not really innovate along the risk dimension, with the parameter γ alone capturing risk aversion in a manner very similar to the standard case. It is, therefore, not surprising that Weil (1989) finds that only if this parameter is fixed at implausibly high levels ($\gamma \approx 45$) can a properly calibrated model replicate the premium—the Mehra and Prescott (1985) result in a different setting. With respect to time preferences, if ρ is calibrated to respect empirical studies, ($\frac{1}{\rho}$, the intertemporal elasticity of substitution is estimated to be about one), then the model also predicts a risk-free rate that is much too high. The reason for this is the same as the one outlined at the end of Section 10.5: separately calibrating the intertemporal substitution parameter ρ tends to strengthen the assumption that the representative agent is highly desirous of a smooth intertemporal consumption stream. With consumption growing on average at 1.8% per year, the agent must be offered a very high risk-free rate in order to be induced to save more, thus making his consumption tomorrow even more in excess of what it is today (less smoothing).

Although Epstein and Zin preferences do not help solve the equity premium puzzle, it is interesting to study a version of the CCAPM with these generalized preferences. The setting is once again a Lucas (1978) style economy with N assets, with the return on the equilibrium portfolio of all assets representing the return on the market portfolio. Using an elaborate dynamic programming argument, Epstein and Zin (1989, 1991) derive an asset pricing equation of the form

$$E_t \left\{ \left[\delta \left(\frac{\tilde{c}_{t+1}}{c_t} \right)^{-\rho} \right]^\theta \left[\frac{1}{1 + \tilde{r}_{M,t+1}} \right]^{1-\theta} (1 + \tilde{r}_{j,t+1}) \right\} \equiv 1 \quad (10.32)$$

where $\tilde{r}_{M,t}$ denotes the period t return on the market portfolio, $r_{j,t}$ the period t return on some asset in it, and $\theta = ((1 - \gamma)/(1 - \rho))$, $0 < \delta < 1$, $1 \neq \gamma > 0$, $\rho > 0$. Note that when time and risk preferences coincide ($\gamma = \rho$, $\theta = 1$), Eq. (10.32) reduces to the pricing equation of the standard time-separable CCAPM case.

The pricing kernel itself is of the form

$$\left[\delta \left(\frac{\tilde{c}_{t+1}}{c_t} \right)^{-\rho} \right]^\theta \left[\frac{1}{1 + \tilde{r}_{M,t+1}} \right]^{1-\theta} \quad (10.33)$$

which is a geometric average (with weights θ and $1 - \theta$, respectively) of the pricing kernel of the standard CCAPM, $[\delta(\tilde{c}_{t+1}/c_t)^{-\rho}]$, and the pricing kernel for the $\log(\rho = 0)$ case, $[1/(1 + \tilde{r}_{M,t+1})]$.

Epstein and Zin (1991) next consider a linear approximation to the geometric average in Eq. (10.33),

$$\theta \left[\delta \left(\frac{\tilde{c}_{t+1}}{c_t} \right)^{-\rho} \right] + (1 - \theta) \left[\frac{1}{1 + \tilde{r}_{M,t+1}} \right] \quad (10.34)$$

Substituting Eq. (10.34) into Eq. (10.32) gives

$$E_t \left\{ \left\{ \theta \left[\delta \left(\frac{\tilde{c}_{t+1}}{c_t} \right)^{-\rho} \right] + (1 - \theta) \left[\frac{1}{1 + \tilde{r}_{M,t+1}} \right] \right\} (1 + \tilde{r}_{j,t+1}) \right\} \approx 1, \text{ or} \\ E_t \left\{ \theta \left[\delta \left(\frac{\tilde{c}_{t+1}}{c_t} \right)^{-\rho} \right] (1 + \tilde{r}_{j,t+1}) + (1 - \theta) \left[\frac{1}{1 + \tilde{r}_{M,t+1}} \right] (1 + \tilde{r}_{j,t+1}) \right\} \approx 1 \quad (10.35)$$

Equation (10.35) is revealing. As we noted earlier, the standard CAPM relates the (essential, nondiversifiable) risk of an asset to the covariance of its returns with \tilde{r}_M , while the CCAPM relates its riskiness to the covariance of its returns with the growth rate of consumption. With separate time and risk preferences, Eq. (10.35) suggests that both covariances matter for an asset's return pattern.²¹ But why do these covariance effects enter separately? The second term leads to the covariance of an asset's return with M which captures its atemporal, nondiversifiable risk (as in the static CAPM model). The first term leads to a covariance of the asset's return with the growth rate of consumption which captures its risk across successive time periods. When risk and time preferences are separated, it is not surprising that both sources of risk should be individually present and that there should exist risk premia associated with each. This relationship is more striking if we assume joint lognormality and heteroskedasticity in consumption and asset returns. Campbell et al. (1997) are able to express Eq. (10.35) in a form whereby the risk premium on asset j satisfies:

$$E_t(\tilde{r}_{j,t+1}) - r_{f,t+1} = \delta \frac{\sigma_{j,c}}{\psi} + (1 - \delta) \sigma_{j,M} - \frac{\sigma_j^2}{2} \quad (10.36)$$

where $\sigma_{j,c} = \text{cov}(\tilde{r}_{j,t}, \tilde{c}_t/(c_{t-1}))$, and $\sigma_{j,M} = \text{cov}(\tilde{r}_{j,t}, \tilde{r}_{M,t})$. Both sources of risk are clearly present.

There are two important applications of Epstein–Zin (1989) to which we now turn.

²¹ To see this, recall that for two random variables \tilde{x} and \tilde{y} , $E(\tilde{x}\tilde{y}) = E(\tilde{x})E(\tilde{y}) + \text{cov}(\tilde{x}, \tilde{y})$, and employ this substitution in both terms on the LHS of Eq. (10.35).

10.8.4.1 Bansal and Yaron (2004)

To prepare for a description of the [Bansal and Yaron \(2004\)](#) model, let us return for a moment to the i.i.d. consumption growth paradigm of [Section 10.5](#). It is permissible to express this process as

$$\tilde{g}_t = \bar{g} + \sigma_g \tilde{\varepsilon}_t \quad (10.37)$$

where $\bar{g} = 0.018$, $\sigma_g = 0.036$, and $\{\tilde{\varepsilon}_t\}$ is a sequence of i.i.d. normal random variables with $\{\tilde{\varepsilon}_t\} \sim N(0, 1)$ for all t . Under the [Lucas \(1978\)–Mehra and Prescott \(1985\)](#) paradigm the growth process on dividends, $\tilde{g}_{\text{div},t}$, is the same: $\tilde{g}_t = \tilde{g}_{\text{div},t}$. With representation (10.37) in mind as our baseline formulation, [Bansal and Yaron \(2004\)](#) propose a generalization with four dimensions.

- (i) The stochastic processes on the growth rates of consumption and dividends are specified independently of one another. Among other advantages, this separation allows a greater volatility to be assigned to the dividend series than to the per capita consumption series, as is evident in the data.²² See [Table 10.4](#).
- (ii) Both the dividend growth and consumption growth series share a small, highly persistent long-run component. It is this feature that has led to the sobriquet “long-run risks” model. In other words, the [Bansal and Yaron \(2004\)](#) economy can persist in a regime of low growth realizations for many period (both consumption and dividends), before moving to a high growth regime and vice versa. As such, the growth rate uncertainty (captured, say, by the expected time to transition from one regime to another) is only resolved very slowly, after the passage, on average, of many periods. Since this slowly resolving risk component is shared by both the dividend and consumption series, the covariance of the investor’s consumption growth with equity returns can be large.
- (iii) The utility specification for the representative agent is Epstein–Zin with parameters $\gamma > 1/\rho$ (risk aversion parameter exceeds the EIS - the elasticity of intertemporal substitution). We recall from our earlier discussion (see [Section 5.7.3](#)) that investors with these preference parameters prefer the early resolution of uncertainty. This desired property is exactly what the dividend series of the equity security (the “market portfolio”) does *not* provide in the [Bansal and Yaron \(2004\)](#) model because of its long-run risk component.
- (iv) Short-run dividend and consumption volatility (these will be risks analogous to the $\sigma_g \tilde{\varepsilon}_t$ term in the baseline representation (10.37)) are time varying. In this way, yet another source of uncertainty is introduced. It is referred to as “stochastic volatility.”

²² In the [Bansal and Yaron \(2004\)](#) model, the difference in these series is labor income.

Taken together, there are three sources of uncertainty faced by an investor participating in the [Bansal and Yaron \(2004\)](#) economy: short-run volatility, long-run volatility, and changing short-run volatility. In equilibrium, each will have an associated risk premium.

To gain a bit more intuition as to the workings of the [Bansal and Yaron \(2004\)](#) model, let us make explicit the growth rate processes discussed above:

$$\tilde{g}_{t+1} = \bar{g} + x_t + \sigma_t \tilde{\eta}_{t+1} \quad (10.38i)$$

$$\tilde{g}_{\text{div},t+1} = \bar{g}_d + \varphi x_t + \varphi_d \sigma_t \tilde{u}_{t+1} \quad (10.38ii)$$

where \tilde{x}_t denotes the long-run risks component and is itself governed by

$$\tilde{x}_{t+1} = \bar{\rho} x_t + \varphi_e \sigma_t \tilde{e}_{t+1} \quad (10.38iii)$$

Note that the choice of autocorrelation coefficient $\bar{\rho}$ determines the persistence of the growth component while the choices of φ and φ_d calibrate (1) the volatility of dividends relative to consumption and (2) their respective correlations with per capita consumption.²³ The shocks $\tilde{\eta}_{t+1}$, \tilde{u}_{t+1} , and \tilde{e}_{t+1} are each distributed $N(0, 1)$, and all are assumed to be statistically independent of one another.

Lastly, the volatility of the short-run risks component, σ_t^2 , is itself given by a mean reverting stochastic process

$$\tilde{\sigma}_{t+1}^2 = \sigma^2 + v(\sigma_t^2 - \sigma^2) + \sigma_w \tilde{w}_{t+1} \quad (10.38iv)$$

Again, \tilde{w}_{t+1} is i.i.d. $N(0,1)$, while $v > 0$ governs the extent of mean reversion to the long-run average, σ^2 .

Although we will not detail the solution technique here, or the approximations that make it feasible, the net effect of the three risk sources is to lead to an equilibrium return expression parallel to [\(10.36\)](#)

$$\begin{aligned} E_t(\tilde{r}_{M,t+1}) - r_{f,t+1} &= \beta_{M,\eta} \gamma \sigma_t^2 + \beta_{M,e} \left(\gamma - \frac{1}{\bar{\rho}} \right) \bar{K}_1 \sigma_t^2 \\ &\quad + \beta_{M,w} \left(\gamma - \frac{1}{\bar{\rho}} \right) (1 - \gamma) \bar{K}_2 \sigma_w^2 - \frac{1}{2} \text{var}_t(\tilde{r}_{M,t+1}) \end{aligned} \quad (10.39)$$

where \bar{K}_1 and \bar{K}_2 are constants determined by $\bar{\rho}$, φ_e , etc. Note that each of the three distinct risk manifestations has its own “designated premium.” The model attains an excellent fit to the data, not only insofar as replicating the equity premium and risk-free rate, but also as regards their respective return volatilities and the volatility of the dividend price ratio. The importance

²³ Bansal and Yaron set $\bar{\rho} = 0.979$.

of the long-run risks component \tilde{x}_t is forcefully present by the fact that if $\rho = 0$ (so that consumption and dividend risk become i.i.d.), the premium declines essentially to zero.²⁴

The [Bansal and Yaron \(2004\)](#) model is a complex one. Nevertheless, it remains a direct descendant of the basic CCAPM of this chapter. While there are more sources of risk, each one commands a recognizable premium. Agent preferences are hypothesized to create high aversion to the new sources of risk. As such, [Bansal and Yaron \(2004\)](#) follow strategies (1) and (2) presented at the close of [Section 10.8.1](#).

Unfortunately, the long-run risks model of consumption growth is disputed along a number of notable dimensions vis-à-vis the data. [Beeler and Campbell \(2012\)](#), in particular, argue that the long-run risks model, as calibrated in [Bansal and Yaron \(2004\)](#), leads to excessive persistence in consumption and dividend growth and excessive predictability (by equity prices) of these quantities relative to what is found in the data. A subsequent calibration ([Bansal et al., 2011](#)) undertaken to resolve the aforementioned shortcomings, while successful in its immediate goal, has the unfortunate implication that the model's equilibrium real term structure is downward sloping, which is also not generally observed. It appears that [Bansal and Yaron \(2004\)](#) may not be the last word on the equity premium puzzle.

We are thus left with two models of the economy, each of which can be argued reflects the economy reasonably well: the traditional i.i.d. consumption growth model and the [Bansal and Yaron \(2004\)](#) model of consumption growth with a long-run risks component. Which of these “world views” should the representative agent–investor adopt? On the one hand, the long-run risks model does a much better job of explaining financial data. On the other hand, the representative agent’s welfare—the present value of his discount future expected utility—is much higher under the i.i.d. consumption growth perspective. Under a “robust control” perspective, the representative agent–investor would choose between the two consumption growth models based on the “worst-case” scenario, and thus act as though he believed the long-run risks model was fact.

10.8.4.2 Collin-Dufresne et al. (2013)

Thus far in this chapter, all the models discussed have the feature that the representative agent either knows the true parameter values of the stochastic environment in which he operates or has subjective beliefs as to those values but does not change his beliefs as events unfold. In the standard [Lucas \(1978\)](#)–[Mehra and Prescott \(1985\)](#) paradigm, for example, the representative agent is assumed to know the actual parameters of the consumption growth process, \bar{g} and σ_g . Even in the case where the modeling exercise includes the estimation of model parameters (such as in Nakamura et al., 2013), these very

²⁴ This result confirms the earlier work of [Weil \(1989\)](#) who argued that Epstein–Zin preferences, *per se*, will not guarantee a resolution of the equity premium and risk-free rate puzzles.

parameters are then held to be fixed and accurately measured by the representative agent as the economy evolves and the model time series of asset prices and returns are generated.

[Collin-Dufresne et al. \(2013\)](#) abandon this perspective and explore the asset pricing implications of allowing the representative agent to learn the pertinent stochastic process parameters on, for example, consumption growth. To illustrate the consequences of this innovative perspective, consider a standard [Lucas \(1978\)–Mehra and Prescott \(1985\)](#) setting (consumption equals dividends) with consumption growth evolving in the customary way ($\tilde{g}_{t+1} = \bar{g} + \sigma_g \tilde{\varepsilon}_{t+1}$, where $\{\tilde{\varepsilon}_t\}$ is i.i.d., $N(0,1)$) modified by assuming the representative agent does not know \bar{g} . Rather, he “begins the day” with a “prior” estimate $\bar{g} \sim N(\bar{g}_0, A_0 \sigma_g^2)$.^{25,26}

[Collin-Dufresne et al. \(2013\)](#) assume the agent employs Bayes’ rule to update his prior distribution recursively as time passes and he gains additional information from the actual, observed g_t realizations. Accordingly, after t realizations have been observed, his estimate for the distribution governing \bar{g} evolves as per

$$\bar{g} \sim N(\bar{g}_t, A_t \sigma_g^2) \text{ where} \quad (10.39i)$$

$$A_{t+1} = \frac{A_t}{1 + A_t} < A_t, A_0 = 1, \text{ and} \quad (10.39ii)$$

$$\bar{g}_{t+1} = \left(\frac{A_t}{1 + A_t} \right) g_t + \left(\frac{1}{1 + A_t} \right) \bar{g}_t \quad (10.39iii)$$

Note that the updated mean of the subjective distribution is computed as a weighted average of the prior period’s estimate and the current period’s growth realization.²⁷

From the representative agent’s (subjective) perspective, the consumption dynamics he faces become

$$\tilde{g}_{t+1} = \bar{g}_t + (\sqrt{1 + A_t}) \sigma_g \tilde{\varepsilon}_{t+1}, \tilde{\varepsilon}_{t+1} \sim N(0, 1) \text{ and} \quad (10.40)$$

$$\bar{g}_{t+1} = \bar{g}_t + \frac{A_t}{\sqrt{1 + A_t}} \sigma_g \tilde{\varepsilon}_{t+1} \quad (10.41)$$

The import of relationships (10.40) and (10.41) is that changes (updates) in \bar{g}_t will have permanent effects as regards the agent’s continuation utility and thus exercise substantial

²⁵ Our discussion going forward follows directly from [Collin-Dufresne et al. \(2013\)](#).

²⁶ For simplicity the representative agent is assumed to know σ . We know, at least, that σ can be estimated more precisely relative to the mean.

²⁷ See the Web Notes on Bayesian updating.

influence on his asset demands and the economy's equilibrium asset prices and returns.²⁸ Note that changes in the estimated \bar{g}_t become a source of “long-run risk” for the agent. The origins of this risk component, however, are very different from the long-run risk component in [Bansal and Yaron \(2004\)](#). In that latter study, the source of long-run risk lay in the presumed, highly persistent consumption growth component. In [Collin-Dufresne et al. \(2013\)](#), the source of long-run risk lies in the changing estimate of the true mean consumption growth rate.

From our discussion of [Bansal and Yaron \(2004\)](#), we have learned that long-run risks are relevant for asset pricing only when the representative agent dislikes them. Accordingly, the second pillar of the [Collin-Dufresne et al. \(2013\)](#) model is the assumption of Epstein–Zin preferences for the representative agent with $\gamma > \frac{1}{\rho}$: a preference for the early resolution of uncertainty. In the event that $\rho = 1$ (see Eq. (5.14b)), these authors are able to solve directly for the market risk premium:

$$E\tilde{r}_{M,t} - r_{f,t} = \gamma\sigma_g^2 + \left(\frac{\gamma-1}{\tilde{\delta}}\right)A_t\sigma_g^2 \quad (10.42)$$

where $\tilde{\delta} = -\ln \delta$, $\tilde{\delta}$ is the representative agent's time subjective time preference parameter and γ his Epstein–Zin risk parameter. Note that in Eq. (10.42), the time index t denotes the t th period in the learning process.

If the agent is indifferent to the timing of uncertainty resolution so that $\gamma = \frac{1}{\rho} = 1$, then the long-run risk component of the premium, $((\gamma-1)/\tilde{\delta})A_t\sigma_g^2$, disappears, and the premium reverts to its value in the standard CRRA utility case (see Eq. (10.26)).

As $t \rightarrow \infty$, $A_t \rightarrow 0$ and the long-run risks premium gradually disappears. How quickly does it happen? Using parameter values taken from [Bansal and Yaron \(2004\)](#) ($\gamma = 10$, $\delta = 0.994$) together with $A_0 = 1$, [Collin-Dufresne et al. \(2013\)](#) find that the equity risk premium with learning, while it declines rapidly with the passage of time, is still significantly larger than the premium under the known-parameter scenario:

$$\frac{\text{equity risk premium with unknown mean growth parameter}}{\text{equity risk premium with known mean growth parameter}} = \frac{\gamma\sigma_g^2 + \left(\frac{\gamma-1}{-\ln\delta}\right)A_t\sigma_g^2}{\gamma\sigma_g^2} = \begin{cases} 151 & t = 0 \\ 1.37 & t = 100 \text{ years} \\ 1.19 & t = 200 \text{ years} \end{cases}$$

²⁸ Note that for this system and $A_0 = 1$ (since the agent knows the true SD), $A_t \rightarrow 0$ and $\bar{g}_t \rightarrow \bar{g}$.

These results suggest that even after 100 years of learning, when the variance of the agent's subjective distribution of the mean growth rate has been substantially diminished, the effect of the remaining uncertainty on the representative agent's continuation utility is still large, with an elevated equity premium as the result.²⁹

An especially interesting analysis in [Collin-Dufresne et al. \(2013\)](#) concerns the representative agent learning the probability of the economy's transition to a "disaster state." Here we have a situation where the "disaster effect" and the "parameter-learning effect" can, potentially, work together to resolve the various puzzles.

The basic innovation here is to assume that the true model of consumption growth conforms to a process

$$\tilde{g}_t = \bar{g}(\tilde{s}_t) + \sigma(\tilde{s}_t)\tilde{\varepsilon}_t$$

where the "state," \tilde{s}_t , follows a two-state Markov chain parameterized as follows:

$$\begin{aligned} & (\bar{g}(s_1), \sigma(s_1)) \quad (\bar{g}(s_2), \sigma(s_2)) \\ & \begin{pmatrix} \bar{g}(s_1), \sigma(s_1) \\ \bar{g}(s_2), \sigma(s_2) \end{pmatrix} \begin{bmatrix} \pi_{11} & 1 - \pi_{11} \\ \pi_{21} & \pi_{22} \end{bmatrix} \\ & (\bar{g}(s_1), \sigma(s_1)) = (0.54\%, 0.98\%) \quad (\text{quarterly}) \\ & (\bar{g}(s_2), \sigma(s_2)) = (-1.15\%, 1.47\%) \\ & \pi_{11} = 0.9975 \quad (\text{quarterly}) \\ & \pi_{22} = 0.9325 \end{aligned}$$

The disaster event $(\bar{g}(s_2), \sigma(s_2))$ is intended to replicate the fact that in the Great Depression, consumption growth declined at an average annual rate of -4.6% from 1929 to 1933 while the transition probabilities are chosen so as to result, on average, in one 4-year depression (a succession of 16 disaster states) per 100-year time horizon. [Collin-Dufresne et al. \(2013\)](#) also break the dividend–output–consumption equivalence of [Mehra and Prescott \(1985\)](#) by postulating an independent process on dividends of the form

²⁹ What about uncertainty in the σ_g^2 parameter? [Collin-Dufresne et al. \(2013\)](#), page 12 give a succinct answer: "... as it is easier to learn a constant volatility parameter than a mean parameter and since volatility is a second-order effect in terms of utility the asset pricing effects of learning about the variance of shocks (even in the case of Epstein–Zin preferences) are very quickly negligible."

Table 10.5: Summary financial statistics learning model^a

	Data		Model	
	Mean	SD	Mean	SD ^b
\tilde{r}^e	5.96		6.58	
\tilde{r}_f	0.86	0.97	0.91	0.63
\tilde{r}_p	5.10	20.21	5.67	16.23

^aSource: Table entries are from [Collin-Dufresne et al. \(2013\)](#), Table 1, Panel A.

^bSD (r^e) not reported; data value from [Bansal and Yaron \(2004\)](#).

$$g_{\text{div},t} = \bar{g} + \lambda(g - \bar{g}) + \sigma_{\text{div}}\tilde{\eta}_{t+1}$$

where $\{\eta_t\}$ is i.i.d., $N(0, 1)$, and $\text{cov}(\tilde{\eta}_t, \tilde{\varepsilon}_t) = 0$. The parameter λ is included to reflect the fact that dividend growth reacts more dramatically to business cycle variations than consumption growth, while σ_{div} is chosen to match real dividend growth volatility over their data period (11.5%).³⁰

The model is then simulated and return statistics based on successively generating 20,000 sample paths of artificial return data, each 400 model periods (100 years for a quarterly calibration) in length. The return statistics reported in [Table 10.5](#) represent averages of the return statistics computed across all of the 20,000 artificially generated sample paths. As part of the data generating strategy, agents learn about the parameters for 400 model periods prior to the initiation of data collection. To be clear, the data from which each sample path's statistics are computed is the equilibrium consequence of agents' continued learning after the completion of a prior 100-year learning period.³¹ The results of this extensive exercise represent an excellent replication of the basic financial stylized facts (See [Table 10.5](#)).

Many other experiments are reported in this article, with the results largely supportive of their modeling hypothesis.

Let us review where we stand. As with [Reitz \(1988\)](#), the background context is a much-feared disaster state in which the agent's marginal utility of consumption is high: he desperately wishes to consume more. In [Reitz \(1988\)](#), the probability of entering the disaster state is precisely known to the agent, and he exits the state immediately. In [Collin-Dufresne et al. \(2013\)](#), the agent slowly learns the probabilities of entering the disaster state and exiting from it (after, potentially, an extended period of time). Each time period, new

³⁰ The parameter λ is estimated by regressing annual real dividend growth on annual consumption growth using data from the period 1929–2011. In particular, the authors find $\lambda = 2.5$ and subsequently use this figure. See the next section of this chapter for a partial explanation for the behavior of dividends relative to aggregate consumption.

³¹ It is as if the agents learned about the disaster likelihood using nineteenth century economic data before entering the twentieth century.

information is learned (a transition event is observed or not) and the relevant probabilities updated. A changing estimate of the probability of entering the disaster state can have large consequences for the agent's continuation utility because of the permanent, consumption level consequences of altered future growth rate perceptions. As a result, the agent perceives the claim to the dividend stream to be a highly risky asset.

As the agent continues to learn, he gradually becomes aware of the model's exact parameter values, in which case the premium declines to roughly 1%, and the risk-free rate climbs to roughly 3%: in a world with known risks with mean consumption growing, agents need a substantial return to postpone consumption thereby making it less smooth. All in all, this story strikes us as a plausible one.

Looking across all the disaster-related models we have discussed so far, [Reitz \(1988\)](#), Barro (2006), Nakamura et al. (2013), and [Collin-Dufresne et al. \(2013\)](#), the true parameters of the evolution of uncertainty in the model economy are assumed to be known: either they are subjectively held or they are assumed to have been properly estimated from historical data. In the case of [Collin-Dufresne et al. \(2013\)](#), the agent learns the parameter values that are assumed to be known if only to us, the readers: we know exactly what the agent is learning about but he himself does not. What if the parameters (such as σ_g) are not known, but are governed themselves by a hypothetical stochastic process. This issue is dealt with in [Weitzman \(2007\)](#) to follow.

10.8.5 Beyond a Representative Agent and Rational Expectations

10.8.5.1 Beyond a Representative Agent

Another approach to addressing the outstanding financial puzzles focuses on recognizing that only a small fraction of the population holds substantial financial assets, stocks in particular.¹⁵ This fact implies that only the variability of the consumption stream of the stockholding class should matter for pricing risky assets. There are reasons to believe that the consumption patterns of this class of the population are both more variable and more highly correlated with stock returns than average per capita consumption.³² Observing, furthermore, that wages are very stable and that the aggregate wage share is countercyclical (i.e., proportionately larger in bad times when aggregate income is relatively low), it is reasonable to suggest that firms, and thus their owners, the shareholders, implicitly insure workers against income fluctuations associated with the business cycle. If this is a significant feature of the real world, it should have implications for asset pricing, as we presently demonstrate.

³² [Mankiw and Zeldes \(1991\)](#) confirm this conjecture. They indeed find that shareholder consumption is 2.5 times as variable as nonshareholder consumption. Data problems, however, preclude taking their results as more than indicative.

Before trying to incorporate such a feature into a CCAPM-type model, it is useful first to recall the notion of risk sharing. Consider the problem of allocating an uncertain income (consumption) stream between two agents so as to maximize overall utility. Assume, furthermore, that these income shares are not fixed across all states but can be allocated on a state-by-state basis. This task can be summarized by the allocation problem

$$\begin{aligned} & \max_{c_1(\tilde{\theta}), c_2(\tilde{\theta})} U(c_1(\tilde{\theta})) + \lambda V(c_2(\tilde{\theta})), \text{ s.t.} \\ & c_1(\tilde{\theta}) + c_2(\tilde{\theta}) \leq Y(\tilde{\theta}) \end{aligned}$$

where $U(\cdot)$, $V(\cdot)$ are, respectively, the two agents' utility functions, $c_1(\tilde{\theta})$ and $c_2(\tilde{\theta})$ their respective income assignments, $Y(\tilde{\theta})$ the economy-wide state-dependent aggregate income stream, and λ their relative weight.

The necessary and sufficient first-order condition for this problem is

$$U_1(c_1(\tilde{\theta})) = \mu V_1(c_2(\tilde{\theta})) \quad (10.43)$$

[Equation \(10.43\)](#) states that the ratio of the marginal utilities of the two agents should be constant. We have seen it before as Eq. (8.3). As we saw there, it can be interpreted as an optimal risk-sharing condition in the sense that it implicitly assigns more of the income risk to the less risk-averse (more risk-tolerant) agent. To see this, take the extreme case where one of the agents, say the one with utility function $V(\cdot)$, is risk neutral—indifferent to risk. According to [Eq. \(10.43\)](#), it will then be optimal for the other agent's income stream to be constant across all states: he will be perfectly insured. Agent $V(\cdot)$ will thus absorb all the risk (in exchange for a higher average income share).

To understand the potential place of these ideas in the CCAPM setting, let $V(\cdot)$ now denote the period utility function of the representative shareholder, and $U(\cdot)$ the period utility function of the representative worker who is assumed to hold no financial assets and who consequently consumes his wage w_t . As before, let Y_t be the uncertain (exogenously given) output. The investment problem of the shareholders—the maximization problem with which we began this chapter—now becomes

$$\begin{aligned} & \max_{\{z_t\}} E \left(\sum_{t=0}^{\infty} \delta^t V(\tilde{c}_t) \right) \\ & \text{s.t.} \\ & c_t + p_t z_{t+1} \leq z_t d_t + p_t z_t \\ & d_t = Y_t - w_t \\ & U_1(w_t) = \lambda V_1(d_t), \\ & z_t \leq 1, \forall t \end{aligned}$$

Here we simply introduce a distinction between the output of the tree, Y_t , and the dividends paid to its owners, d_t , on the plausible grounds that workers need to be paid to

take care of the trees and collect the fruits. This payment is w_t . Moreover, we introduce the idea that the wage bill may incorporate a risk insurance component, which we formalize by assuming that the variability of wage payments is determined by an optimal risk sharing rule equivalent to Eq. (10.43). One key parameter is the income share, μ , which may be interpreted as reflecting the relative bargaining strengths of the two groups. Indeed, a larger μ gives more income to the worker.

Assets in this economy are priced as before, with Eq. (10.1) becoming

$$V_1(c_t)p_t = \delta E_t\{V_1(\tilde{c}_{t+1})(\tilde{p}_{t+1} + \tilde{d}_{t+1})\} \quad (10.44)$$

While the differences between Eqs. (10.1) and (10.38) may appear purely notational, their importance cannot be overstated. First, the pricing kernel derived from Eq. (10.44) will build on the firm owners' MRS, defined over shareholder consumption (dividend) growth rather than the growth in average per capita consumption. Moreover, the definition of dividends as output minus a stabilized stream of wage payments opens up the possibility that the flow of payments to which firm owners are entitled is effectively not only much more variable than consumption but more variable than output as well. Therein lies a concept of leverage, one that has been dubbed *operating leverage*, similar to the familiar notion of financial leverage. In the same way that bondholders come first and are entitled to a fixed, noncontingent interest payment, workers also have priority claims to the income stream of the firm, and macroeconomic data on the cyclical behavior of the wage share confirm that wage payments are more stable than aggregate income.

These ideas are most simply laid out in Danthine and Donaldson (2002) to which the reader is referred for details, and we find that this class of models can generate significantly increased equity premia. When an extra notion of distributional risk, associated with the possibility that μ varies stochastically, is added, in a way that permits better accounting of the observed behavior of the wage share over the medium run, Danthine and Donaldson (2002) find the premium approaches 6%, a fact that is not entirely surprising since a fundamentally new source of risk with its own premium is being introduced. Favilukis and Lin (2013) argue, in a production economy context, that it is impossible to replicate the financial stylized facts without an operating leverage component, while Santos and Veronesi (2006) demonstrate that the ratio of labor income to consumption is helpful in explaining long-horizon stock returns.

Aside from the worker-firm owner dichotomy, heterogeneous agent models have generally not been prominent in the equity premium resolution literature.³³

³³ An exception to this statement is Dumas (1989) who considers an economy with two agents of differing CRRAs. In equilibrium, the less risk averse agent owns a larger fraction of the wealth than in bad times. The corresponding (wealth distribution dependent) representative agent exhibits time varying risk aversion. This feature also improves model performance.

10.8.5.2 Beyond Rational Expectations

Let us review the modeling philosophy that has governed this chapter's discussion thus far. In the case of [Mehra and Prescott \(1985\)](#), Barro (2006), [Bansal and Yaron \(2004\)](#), and most others, the representative agent is assumed to know the identity of the stochastic process governing the evolution of the uncertain quantities of relevance to him and the precise parameters governing that process. While the parameters may be the result of an estimation procedure (e.g., Nakamura et al., 2013), the representative agent in the model is presumed to take the values on faith. The case of [Collin-Dufresne et al. \(2013\)](#) is a bit different but in the same spirit: the model builder is presumed to know the relevant model parameters about which the representative agent learns as the model's time path evolves. This overall perspective is a reflection of the idea that the economy is stationary and has been operating for a very long time, sufficiently long for economic participants to have learned all the relevant stochastic parameters. As a modeling philosophy, it is loosely labeled as the "rational expectations" view of the world.

[Weitzman \(2007\)](#) forcefully disputes this view of the world, arguing that there will always exist "fundamental structural uncertainty" associated, in particular, with the consumption growth rate parameter σ_g . Suppose that $\tilde{\sigma}_{g,t}$ is time dependent and evolves stochastically in a nonergodic way (it does not converge to a fixed σ_g^*), and that the representative agent employs Bayesian methods to learn about $\tilde{\sigma}_{g,t}$. [Weitzman \(2007\)](#) points out that the Bayesian learning will lead to fat tailed predictive student t probability distributions governing consumption growth. Under this probability distribution, [Geweke \(2001\)](#) has shown earlier that expected utility may not exist for standard CRRA utility, a tendency that must be held in check by postulating a reasonable prior distribution at the start of the learning process. [Weitzman \(2007\)](#) refers to his perspective as the "Bayesian evolutionary-learning thickened-tail explanation" of the equity premium. [Weitzman \(2007\)](#) argues that standard methodologies for learning in an evolutionary context lead to a reversal of the standard baseline financial anomalies. The model derived equity premium will be too large, the equity return volatility too large, and the risk-free rate too low (potentially negative) relative to data. In [Weitzman \(2007\)](#), the puzzle is not that the observed equity premium is as large as it is, but that it is so small relative to what it might be under his Bayesian evolutionary learning perspective. We close with a series of quotes taken from Weitzman's paper:

Weitzman (2007)... "formalizes the idea that non ergodic parameter uncertainty leads to permanently tail thickened distribution of growth rates that can cause expected marginal utility to blow up—and shows a rigorous series in which 'containing the student-t explosion' necessitates an unavoidable dependence of asset prices upon some form or another of exogenously imposed subjective beliefs."

"This potentially explosive outcome remains the mathematical driving force behind the scene, which imparts the statistical illusion of an enormous equity premium incompatible

with the standard neoclassical paradigm. When people are peering forward into the future they are also looking backward at their own prior and what they are seeing there is a spooky reflection of their own present insecurity in not being able to judge accurately the possibility of bad evolutionary mutations of future history t might conceivably ruin equity investors by wiping out their stock market holdings at a time t when their world has already taken a very bad turn.”

“... for asset pricing applications it is not at all unscientific to adhere to the non REE idea that no amount of past data can be... large enough to identify the relevant structural uncertainty concerning future economic growth. Moreover, as a corollary, REE calibrations ignoring this basic principle of learning about hidden evolutionary parameters may very well end up badly underestimating the comparative utility risk of a real-world gamble in the unknown structural potential economic growth, relative to a nearly safe investment in a nearly sure thing.”

10.9 Conclusions

The various modifications considered in the previous sections represent many of the most recent contributions to the equity premium literature. As is apparent, there is no consensus and active research continues. Nevertheless, the overall equilibrium stochastic discount factor seems secure because of its natural connection to the data. Even Weitzman (2007) does not dispute the overall research program.

At this juncture, one may nevertheless be led to the view that structural asset pricing theory, based on rigorous dynamic general equilibrium models, provides limited operational support in our quest for the understanding of time series financial market phenomena. This state of affairs perhaps explains the popularity of less encompassing approaches based on the concept of arbitrage to be reviewed in succeeding chapters.

References

- Abel, A., 1990. Asset prices under habit formation and catching up with the joneses. *Am. Econ. Rev.* 80, 38–42.
- Azeredo, F., 2012. The equity premium: a deeper puzzle. Discussion Paper, Navigant Economics.
- Bansal, R., Coleman, W.J., 1996. A monetary explanation of the equity premium, term premium and risk-free rates puzzles. *J. Polit. Econ.* 104, 1135–1171.
- Bansal, R., Yaron, A., 2004. Risks for the long run: a potential resolution of asset pricing puzzles. *J. Finan.* 59, 1481–1509.
- Barro, R., 2006. Rare disasters and asset markets in the twentieth century. *Q. J. Econ.* 121, 823–866.
- Barro, R.J., 1974. Are government bonds net wealth? *J. Polit. Econ.* 82, 1095–1117.
- Bansal, J., Kiku, D., Yaron, A., 2011. An empirical evaluation of the long-run risks model of asset prices. Working Paper, Duke University and the University of Pennsylvania.
- Beeler, J., Campbell, J., 2012. The long-run risks model and aggregate asset prices: an empirical assessment. *Crit. Finan. Rev.* 1, 141–182.

- Campbell, J., Lo, A., MacKinlay, A.C., 1997. *The econometrics of financial markets*. Princeton University Press, Princeton, NJ.
- Campbell, J.Y., 1998. Asset prices, consumption, and the business cycle. NBER Working Paper 6485, March 1998, forthcoming in the *Handbook of Macroeconomics*, Amsterdam, North Holland.
- Campbell, J.Y., Cochrane, J.H., 1999. By force of habit: a consumption-based explanation of aggregate stock market behavior. *J. Polit. Econ.* 107, 205–251.
- Chetty, R., Szeidl, A., 2004. Consumption commitments: neoclassical foundations for habit formation. NBER Working Paper #10970.
- Collin-Dufresne, P., Johannes, M., Lochstoer, L., 2013. Parameter learning in general equilibrium. Working Paper, Columbia University, Graduate School of Business.
- Constantinides, G.M., 1982. Intertemporal asset pricing with heterogeneous consumers and without demand aggregation. *J. Bus.* 55, 253–267.
- Constantinides, G.M., 1990. Habit formation: a resolution of the equity premium puzzle. *J. Polit. Econ.* 98, 519–543.
- Danthine, J.P., Donaldson, J., 2002. Labor relations and asset returns. *Rev. Econ. Stud.* 69, 41–64.
- Danthine, J.P., Donaldson, J., 1999. Non-falsified expectations and general equilibrium asset pricing. *Econ. J.* 109, 607–635.
- Donaldson, J., Mehra, R., 2008. Risk-based explanations of the equity premium. *Handbook of the Equity Premium*. Elsevier, New York and Amsterdam.
- Dumas, B., 1989. Two person dynamic equilibrium in the capital market. *Rev. Finan. Stud.* 2, 157–188.
- Epstein, L., Zin, S., 1989. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework. *Econometrica*. 57, 937–969.
- Epstein, L., Zin, S., 1991. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: an empirical analysis. *J. Polit. Econ.* 99, 263–286.
- Favilukis, J., Lin, X., 2013. Wage rigidity: a quantitative solution to several asset pricing puzzles. Working Paper, London School of Economics and Ohio State University.
- Gabaix, X., 2012. Variable rare disasters: an exactly solved framework for ten puzzles in macro-finance. *Q. J. Econ.* 127, 645–700.
- Geweke, J., 2001. A note on some limitations of CRRA utility. *Econ. Lett.* 71, 341–345.
- Goetzmann, W., Jorion, P., 1999. A century of global stock markets. *J. Finan.* 55, 953–980.
- Guvenen, F., 2011. Macroeconomics with heterogeneity: a practical guide. Working Paper, University of Minnesota.
- Hansen, L., Jagannathan, R., 1991. Implications of security market data for models of dynamic economies. *J. Polit. Econ.* 99, 225–262.
- Lengwiler, Y., 2004. *Microfoundations of Financial Economics*. Princeton University Press, Princeton, NJ.
- Lucas, R.E., 1978. Asset pricing in an exchange economy. *Econometrica*. 46, 1429–1445.
- Ljungqvist, L., Uhlig, H., 2009. Optimal endowment destruction under Campbell–Cochrane. NBER Working Paper #14772.
- Mankiw, G., Zeldes, S., 1991. The consumption of stockholders and non-stockholders. *J. Finan. Econ.* 29, 97–112.
- Mehra, R., 2012. Consumption-based asset pricing models. *Annu. Rev. Financ. Econ.* 4, 385–409.
- Mehra, R., Prescott, E.C., 1985. The equity premium: a puzzle. *J. Monet. Econ.* 15, 145–161.
- Mehra, R., Prescott, E.C., 1988. The equity risk premium: a solution? *J. Monet. Econ.* 22, 133–136.
- Nakamura, E., Steinsson, J., Barro, R., Ursúa, J., 2013. Crises and recoveries in an empirical model of consumption disasters. *Am. Econ. J.: Macroecon.* 5, 35–74.
- Reitz, T., 1988. The equity premium: a solution. *J. Monet. Econ.* 22, 117–131.
- Rozen, K., 2010. Foundations of intrinsic habit formation. *Econometrica*. 78, 1341–1373.
- Rubinstein, M., 1974. An aggregation theorem for securities markets. *J. Finan. Econ.* 1, 225–244.
- Santos, T., Veronesi, P., 2006. Labor income and predictable stock returns. *Rev. Econ. Stud.* 19, 1–44.
- Weil, P.h., 1989. The equity premium puzzle and the risk-free rate puzzle. *J. Monet. Econ.* 24, 401–421.
- Weitzman, M., 2007. Subjective expectations and asset return puzzles. *Am. Econ. Rev.* 97, 1102–1130.

Appendix 10.1: Solving the CCAPM with Growth³⁴

Assume that there is a finite set of possible growth rates $\{g_1, \dots, g_N\}$ whose realizations are governed by a Markov process with transition matrix \mathbf{T} and entries π_{ij} . Then, for whatever g_i is realized in period $t + 1$,

$$d_{t+1} = g_{t+1}Y_t = g_{t+1}c_t = g_i c_t$$

Under the usual utility specification, $U(c) = (c^{1-\gamma})/(1-\gamma)$, the basic equilibrium ($Y_t = c_t$) asset pricing equation reduces to

$$\begin{aligned} Y_t^{-\gamma} q(Y_t, g_i) &= \delta \sum_{j=1}^N \pi_{ij} (g_j Y_t)^{-\gamma} [Y_t g_j + q(g_j Y_t, g_j)] \text{ or} \\ q(Y_t, g_i) &= \delta \sum_{j=1}^N \pi_{ij} \left(\frac{g_j Y_t}{Y_t} \right)^{-\gamma} [c_t g_j + q(g_j Y_t, g_j)] \end{aligned}$$

Notice that the MRS is determined exclusively by the consumption growth rate.

The essential insight of Mehra and Prescott (1985) was to observe that a solution to this linear system has the form

$$q(Y_t, g_i) = v_i Y_t$$

for a set of constants $\{v_1, \dots, v_N\}$, each identified with the corresponding growth rate.

With this functional form, the asset pricing equation reduces to

$$\begin{aligned} v_i Y_t &= \delta \sum_{j=1}^N \pi_{ij} (g_j Y_t)^{-\gamma} [g_j Y_t + v_j g_j Y_t] \text{ or} \\ v_i &= \delta \sum_{j=1}^N \pi_{ij} (g_j)^{1-\gamma} [1 + v_j] \end{aligned} \tag{10.45}$$

This is again a system of linear equations in the N unknowns $\{v_1, \dots, v_N\}$. Provided the growth rates are not too large (so that the agent's utility is not unbounded), a solution exists—a set of $\{v_1^*, \dots, v_N^*\}$ that solves the system of Eqs. (10.45).

Thus, for any state $(Y, g_j) = (c, g_j)$, the equilibrium equity asset price is

$$q(Y, g_j) = v_j^* Y$$

If we suppose the current state is (Y, g_i) while next period it is $(g_j Y, g_j)$, then the one-period return earned by the equity security over this period is

³⁴ This appendix is based on Mehra and Prescott (1985).

$$\begin{aligned}
r_{ij} &= \frac{q(g_j Y, g_j) + g_j Y - q(Y, g_i)}{q(Y, g_i)} \\
&= \frac{v_j^* g_j Y + g_j Y - v_i^* Y}{v_i^* Y} \\
&= \frac{g_i(v_j^* + 1)}{v_i^*} - 1
\end{aligned}$$

and the mean or expected return, conditional on state i , is

$$r_i = \sum_{j=1}^N \pi_{ij} r_j$$

The unconditional equity return is thus given by

$$Er = \sum_{j=1}^N \hat{\pi}_j r_j$$

where $\hat{\pi}_j$ are the long-run stationary probabilities of each state.

The risk-free security is analogously priced as

$$q^{\text{rf}}(c, g_i) = \delta \sum_{j=1}^N \pi_{ij}(g_j)^{-\gamma}, \text{etc.}$$

Appendix 10.2: Some Properties of the Lognormal Distribution

Definition A10.1 A variable x is said to follow a lognormal distribution if $\ln x$ is normally distributed. Let $\ln x \sim N(\mu_x, \sigma_x^2)$. If this is the case,

$$E(x) = \exp \left\{ \mu_x + \frac{1}{2} \sigma_x^2 \right\}$$

$$E(x^a) = \exp \left\{ a\mu_x + \frac{1}{2} a^2 \sigma_x^2 \right\}$$

$$\text{var}(x) = \exp\{2\mu_x + \sigma_x^2\}(\exp \sigma_x^2 - 1)$$

Suppose furthermore that x and y are two variables that are independently and identically lognormally distributed, then we also have

$$E(x^a y^b) = \exp \left\{ a\mu_x + b\mu_y + \frac{1}{2}(a^2\sigma_x^2 + b^2\sigma_y^2) + 2\rho ab\sigma_x\sigma_y \right\}$$

where ρ is the correlation coefficient between $\ln x$ and $\ln y$.

Let us apply these relationships to consumption growth: g_t is lognormally distributed, i.e., $\ln g_t \sim N(\mu_g, \sigma_g^2)$.

We know that $E(g_t) = 1.0183$ and $\text{var}(g_t) = (0.0357)^2$. To identify (μ_g, σ_g^2) , we need to find the solutions of

$$\begin{aligned} 1.0183 &= \exp \left\{ \mu_g + \frac{1}{2}\sigma_g^2 \right\} \\ (0.0357)^2 &= \exp\{2\mu_g + \sigma_g^2\}(\exp\sigma_g^2 - 1) \end{aligned}$$

Substituting the first equation squared into the second by virtue of the fact that $[\exp(y)]^2 = \exp(2y)$ and solving for σ_g^2 , one obtains

$$\sigma_g^2 = 0.00123$$

Substituting this value in the equation for μ_g , one solves for $\mu_g = 0.01752$.

We can directly use these values to solve Eq. (10.20):

$$E\{g_t^{-\gamma}\} = \exp \left\{ -\gamma\mu_g + \frac{1}{2}\gamma^2\sigma_g^2 \right\} = \exp\{-0.03258\} = 0.967945$$

thus $\delta = 1.024$.

Focusing now on the numerator of Eq. (10.21), one has

$$\exp \left\{ \mu_g + \frac{1}{2}\sigma_g^2 \right\} \exp \left\{ -\gamma\mu_g + \frac{1}{2}\gamma^2\sigma_g^2 \right\}$$

while the denominator is

$$\exp \left\{ (1-\gamma)\mu_g + \frac{1}{2}(1-\gamma)^2\sigma_g^2 \right\}$$

It remains to recall that $(\exp(a)\exp(b))/(\exp(c)) = \exp(a+b-c)$ to obtain Eq. (10.22).

Another application is as follows: The standard deviation of the pricing kernel $m_t = g_t^{-\gamma}$ where consumption growth g_t is lognormally distributed. Given that $E m_t$ is as derived in Section 10.6, one estimates

$$\sigma^2(m_t) \cong \frac{1}{k} \left\{ \sum_{i=1}^k [\delta(g_i)^{-\gamma} - Em_t]^2 \right\}$$

for $\ln g_i$ drawn from $N(0.01752; 0.00123)$ and k sufficiently large (say $k = 10,000$). For $\gamma = 2$, one obtains $\sigma^2(m_t) = (0.00234)^2$, which yields

$$\frac{\sigma_m}{E\tilde{m}} \cong \frac{0.00234}{0.9559} = 0.00245$$

Arrow–Debreu Pricing, Part II

Chapter Outline

11.1 Introduction	325
11.2 Market Completeness and Complex Securities	326
11.3 Constructing State-Contingent Claims Prices in a Risk-Free World:	
Deriving the Term Structure	330
11.4 The Value Additivity Theorem	335
11.5 Using Options to Complete the Market: An Abstract Setting	337
11.6 Synthesizing State-Contingent Claims: A First Approximation	343
11.7 Recovering Arrow–Debreu Prices from Options Prices: A Generalization	345
11.8 Arrow–Debreu Pricing in a Multiperiod Setting	352
11.9 Conclusions	357
References	358
Appendix 11.1: Forward Prices and Forward Rates	358

11.1 *Introduction*

Chapter 9 presented the Arrow–Debreu asset pricing theory from the equilibrium perspective. With the help of a number of modeling hypotheses and building on the concept of market equilibrium, we showed that the price of a future contingent dollar can appropriately be viewed as the product of three main components: a pure time discount factor, the probability of the relevant state of nature, and an intertemporal marginal rate of substitution reflecting the collective (market) assessment of the scarcity of consumption in the future relative to today. This important message is one that we confirmed with the CCAPM of Chapter 10. Here, however, we adopt the alternative arbitrage perspective and revisit the same Arrow–Debreu pricing theory. Doing so is productive precisely because, as we have stressed before, the design of an Arrow–Debreu security is such that once its price is available, whatever its origin and makeup, it provides the answer to the key valuation question: what is a unit of the future state-contingent numeraire worth today? As a result, it constitutes the essential piece of information necessary to price arbitrary cash flows. Even if the equilibrium theory of Chapter 9 were all wrong, in the sense that the hypotheses made there turn out to be a very poor description of reality and that, as a consequence, the

prices of Arrow–Debreu securities are not well described by Eq. (9.1), it remains true that if such securities are traded, their prices constitute the essential building blocks (in the sense of our Chapter 2 bicycle pricing analogy) for valuing any arbitrary risky cash flow.

Section 11.2 develops this message and goes further, arguing that the detour via Arrow–Debreu securities is useful even if no such security is actually traded. In making this argument, we extend the definition of the complete market concept. Section 11.3 illustrates the approach in the abstract context of a risk-free world where we argue that any *risk-free* cash flow can be easily and straightforwardly priced as an equivalent portfolio of *date-contingent* claims. These latter instruments are, in effect, discount bonds of various maturities.

Our main interest, of course, is to extend this approach to the evaluation of *risky* cash flows. To do so requires, by analogy, that for each future date-state the corresponding contingent cash flow be priced. This, in turn, requires that we know, for each future *date-state*, the price today of a security that pays off in that date-state and only in that date-state. This latter statement is equivalent to the assumption of market completeness.

In this chapter, we take on the issue of completeness in the context of securities known as options. Our goal is twofold. First, we want to give the reader an opportunity to review an important element of financial theory—the theory of options. A special appendix to this chapter, available on this text’s website, describes the essentials for the reader in need of a refresher. Second, we want to provide a concrete illustration of the view that the recent expansion of derivative markets constitutes a major step in the quest for the “Holy Grail” of achieving a complete securities market structure. We will see, indeed, that options can, in principle, be used relatively straightforwardly to complete the markets. Furthermore, even in situations where this is not practicable, we can use option pricing theory to value risky cash flows in a manner as though the financial markets were complete.

Our discussion will follow the outline suggested by the following two questions:

1. How can options be used to complete the financial markets? We will first answer this question in a simple, highly abstract setting. Our discussion closely follows Ross (1976).
2. What is the link between the prices of market-quoted options and the prices of Arrow–Debreu securities? We will see that it is indeed possible to infer Arrow–Debreu prices from option prices in a practical setting conducive to the valuation of an actual cash-flow stream. Here our discussion follows Banz and Miller (1978) and Breeden and Litzenberger (1978).

11.2 Market Completeness and Complex Securities

In this section we pursue, more systematically, the important issue of market completeness first addressed in Chapter 1 when we discussed the optimality property of a general competitive equilibrium. Let us start with two definitions.

1. **Completeness.** Financial markets are said to be *complete* if, for each state of nature θ , there exists a market for contingent claim or Arrow–Debreu security θ —in other words, for a claim promising delivery of one unit of the consumption good (or, more generally, the numeraire) if state θ is realized, and nothing otherwise. Note that this definition takes a form specifically appropriate to models where there is only one consumption good and several date-states. This is the usual context in which financial issues are addressed.
2. **Complex security.** A complex security is one that pays off in more than one state of nature.

Suppose the number of states of nature $N = 4$; an example of a complex security is $S = (5, 2, 0, 6)$ with payoffs 5, 2, 0, and 6, respectively, in states of nature 1, 2, 3, and 4. If markets are complete, we can immediately price such a security since

$$(5, 2, 0, 6) = 5(1, 0, 0, 0) + 2(0, 1, 0, 0) + 0(0, 0, 1, 0) + 6(0, 0, 0, 1),$$

in other words, since the complex security can be replicated by a portfolio of Arrow–Debreu securities, the price of security S , q_S , must be

$$q_S = 5q_1 + 2q_2 + 6q_4.$$

We are appealing here to the law of one price or, equivalently, to a condition of no arbitrage.¹ This is the first instance of our using the second main approach to asset pricing, the arbitrage approach, which is our exclusive focus in Chapters 11–13. We are pricing the complex security on the basis of our knowledge of the prices of its components. The relevance of the Arrow–Debreu pricing theory resides in the fact that it provides the prices for what can be argued are the essential components of any asset or cash flow.

Effectively, the argument can be stated in the following proposition.

Proposition 11.1 If markets are complete, any complex security or any cash-flow stream can be replicated as a portfolio of Arrow–Debreu securities.

If markets are complete in the sense that prices exist for all the relevant Arrow–Debreu securities, then the “no arbitrage” condition implies that any complex security or cash flow can also be priced using Arrow–Debreu prices as fundamental elements. The portfolio, which is easily priced using the (Arrow–Debreu) prices of its individual components, is essentially the same good as the cash flow or the security it replicates: it pays the same amount of the consumption good in each and every state. Therefore, it should bear the same

¹ This is stating that the equilibrium prices of two separate units of what is essentially the same good should be identical. If this were not the case, a riskless and costless arbitrage opportunity would open up: buy extremely large amounts at the low price and sell them at the high price, forcing the two prices to converge. When applied across two different geographical locations (which is not the case here: our world is a point in space), the law of one price may not hold because of transport costs rendering the arbitrage costly.

price. This is a key result underlying much of what we do in the remainder of this chapter and our interest in Arrow–Debreu pricing.

If this equivalence is not observed, an arbitrage opportunity—the ability to make unlimited profits with no initial investment—will exist. By taking positions to benefit from the arbitrage opportunity, however, investors will expeditiously eliminate it, thereby forcing the price relationships implicitly asserted in [Proposition 11.1](#). To illustrate how this would work, let us consider the prior example and postulate the following set of prices:

$q_1 = \$0.86$, $q_2 = \$0.94$, $q_3 = \$0.93$, $q_4 = \$0.90$, and $q_{(5,2,0,6)} = \$9.80$. At these prices, the law of one price fails, since the price of the portfolio of state claims that exactly replicates the payoff to the complex security does not coincide with the complex’s security’s price:

$$q_{(5,2,0,6)} = \$9.80 < \$11.58 = 5q_1 + 2q_2 + 6q_4.$$

We see that the complex security is relatively undervalued vis-à-vis the state claim prices. This suggests acquiring a positive amount of the complex security while selling (short) the replicating portfolio of state claims. [Table 11.1](#) illustrates a possible combination.

So the arbitrageur walks away with \$1.78 while (1) having made no investment of her own wealth and (2) without incurring any future obligation (perfectly hedged). She will thus replicate this portfolio as much as she can. But the added demand for the complex security will, *ceteris paribus*, tend to increase its price while the short sales of the state claims will depress their prices. This will continue (the arbitrage opportunity will persist) as long as the pricing relationships are not in perfect alignment.

Suppose now that only complex securities are traded and that there are M of them and N states. The following is true.

Proposition 11.2 If $M = N$, and all the M complex securities are linearly independent, then (i) it is possible to infer the prices of the Arrow–Debreu state-contingent claims from the complex securities’ prices and (ii) markets are effectively complete.²

Table 11.1: An arbitrage portfolio

$t = 0$	$t = 1$ Payoffs				
Security	Cost	θ_1	θ_2	θ_3	θ_4
Buy 1 complex security	-\$9.80	5	2	0	6
Sell short 5 (1,0,0,0) securities	\$4.30	-5	0	0	0
Sell short 2 (0,1,0,0) securities	\$1.88	0	-2	0	0
Sell short 6 (0,0,0,1) securities	\$5.40	0	0	0	-6
Net	\$1.78	0	0	0	0

² When we use the language “linearly dependent,” we are implicitly regarding securities as N -vectors of payoffs.

The hypothesis of linear independence can be interpreted as a requirement that there exist N truly different securities for completeness to be achieved. Thus, it is easy to understand that if among the N complex securities available, one security, A , pays $(1, 2, 3)$ in the three relevant states of nature, and the other, B , pays $(2, 4, 6)$, only $N - 1$ truly distinct securities are available: B does not permit any different redistribution of purchasing power across states than A permits. More generally, the linear independence hypothesis requires that no one complex security can be replicated as a portfolio of some of the other complex securities. The reader will remember that we made the same hypothesis at the beginning of Section 8.4.

Suppose the following securities are traded:

$$(3, 2, 0) \quad (1, 1, 1) \quad (2, 0, 2)$$

at equilibrium prices \$1.00, \$0.60, and \$0.80, respectively. It is easy to verify that these three securities are linearly independent. We can then construct the Arrow–Debreu prices as follows. Consider, for example, the security $(1, 0, 0)$:

$$(1, 0, 0) = w_1(3, 2, 0) + w_2(1, 1, 1) + w_3(2, 0, 2)$$

$$\text{Thus, } 1 = 3w_1 + w_2 + 2w_3$$

$$0 = 2w_1 + w_2$$

$$0 = w_2 + 2w_3$$

Solve: $w_1 = \frac{1}{3}$, $w_2 = -\frac{2}{3}$, $w_3 = \frac{1}{3}$, and $q_{(1,0,0)} = \frac{1}{3}(1.00) + (-\frac{2}{3})(0.60) + \frac{1}{3}(0.80) = 0.1966$

Similarly, we could replicate $(0, 1, 0)$ and $(0, 0, 1)$ with portfolios $(w_1 = 0, w_2 = 1, w_3 = -\frac{1}{2})$ and $(w_1 = -\frac{1}{3}, w_2 = \frac{2}{3}, w_3 = \frac{1}{6})$, respectively, and price them accordingly.

Expressed in a more general way, the reasoning just completed amounts to searching for a solution of the following system of equations:

$$\begin{pmatrix} 3 & 1 & 2 \\ 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} w_1^1 & w_1^2 & w_1^3 \\ w_2^1 & w_2^2 & w_2^3 \\ w_3^1 & w_3^2 & w_3^3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Of course, this system has solution $\begin{pmatrix} 3 & 1 & 2 \\ 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ only if the matrix of

security payoffs can be inverted, which requires that it be of full rank, or that its determinant be nonzero, or that all its lines or columns be linearly independent.

Now suppose the number of linearly independent securities is strictly less than the number of states. In this case the securities markets are fundamentally incomplete: there may be some assets that cannot be unambiguously priced (see Chapter 12 upcoming). Furthermore, risk-sharing opportunities are less than if the securities markets were complete and, in

general, social welfare is lower than what it would be under complete markets: some gains from exchange cannot be exploited due to the lack of instruments permitting these exchanges to take place.

We conclude this section by revisiting the project valuation problem. In the light of the Arrow–Debreu pricing approach, how should we value an uncertain cash-flow stream such as:

$$\begin{array}{ccccccc} t = & 0 & 1 & 2 & 3 & \dots & T ? \\ & -I_0 & \tilde{C}F_1 & \tilde{C}F_2 & \tilde{C}F_3 & \dots & \tilde{C}F_T \end{array}$$

This cash-flow stream is akin to a complex security since it pays in multiple states of the world. Let us specifically assume that there are N states at each date t , $t = 1, \dots, T$ and let us denote $q_{t,\theta}$ the price of the Arrow–Debreu security promising delivery of one unit of the numeraire if state θ is realized at date t . Similarly, let us identify as $CF_{t,\theta}$ the cash flow associated with the project in the same occurrence. Then pricing the complex security in the manner of Arrow–Debreu pricing means valuing the project as in [Eq. \(11.1\)](#).

$$NPV = -I_0 + \sum_{t=1}^T \sum_{\theta=1}^N q_{t,\theta} CF_{t,\theta}. \quad (11.1)$$

Although this is a demanding procedure, it is a pricing approach that is fully general and involves no approximation. For this reason it constitutes an extremely useful benchmark.

In a risk-free setting, the concept of the state-contingent claim has a very familiar real-world counterpart. In fact, the notion of the term structure is simply a reflection of “date-contingent” claims prices. We pursue this idea in the next section.

11.3 Constructing State-Contingent Claims Prices in a Risk-Free World: Deriving the Term Structure

Suppose we are considering risk-free investments and risk-free securities exclusively. In this setting—where we ignore risk—the “states of nature” about which we have been speaking correspond to *future time periods*. This section shows that the process of computing the term structure from the prices of coupon bonds is akin to recovering Arrow–Debreu prices from the prices of complex securities.

Under this interpretation, the Arrow–Debreu state-contingent claims correspond to risk-free discount bonds of various maturities, as seen in [Table 11.2](#).

These are Arrow–Debreu securities because they pay off in one state (the period of maturity) and zero in all other time periods (states).

Table 11.2: Risk-free discount bonds as Arrow–Debreu securities

Current Bond Price		Future Cash Flows					
$t = 0$	1	2	3	4	...	T	
$-q_1$	\$1000						
$-q_2$		\$1000					
...							
$-q_T$							\$1000

The cash flow of a “ j -period discount bond” is given by:

$t = 0$	1	...	j	$j + 1$...	T
$-q_j$	0	0	\$1000	0	0	0

Table 11.3: Present and future cash flows for two coupon bonds

Bond Type	Cash Flow at Time t					
	$t = 0$	1	2	3	4	5
$\frac{7}{8}\%$ bond:	−1097.8125	78.75	78.75	78.75	78.75	1078.75
$\frac{5}{8}\%$ bond:	−1002.8125	56.25	56.25	56.25	56.25	1056.25

In the United States at least, securities of this type are not issued for maturities longer than 1 year. Rather, only interest-bearing or coupon bonds are issued for longer maturities. These are complex securities by our definition: they pay off in many states of nature. But we know that if we have enough distinct complex securities we can compute the prices of the Arrow–Debreu securities even if they are not explicitly traded. So we can also compute the prices of these zero coupon or discount bonds from the prices of the coupon or interest-bearing bonds, assuming no arbitrage opportunities in the bond market.

For example, suppose we wanted to price a 5-year discount bond coming due in November 2019 (we view $t = 0$ as November 2014), and we observe two coupon bonds being traded that mature at the same time:

- i. $\frac{7}{8}\%$ bond priced at $109\frac{25}{32}$ or \$1097.8125/\$1000 of face value
- ii. $\frac{5}{8}\%$ bond priced at $100\frac{9}{32}$ or \$1002.8125/\$1000 of face value

The coupons of these bonds are, respectively,

$$\begin{aligned} 0.07875 \times \$1000 &= \$78.75/\text{year} \\ 0.05625 \times \$1000 &= \$56.25/\text{year}^3 \end{aligned}$$

and their cash flows are shown in [Table 11.3](#).

³ In fact, interest is paid every 6 months on this sort of bond, a refinement that would double the number of periods without altering the argument in any way. Actual default free rates are of course much lower in 2014 than these numbers suggest.

Table 11.4: Eliminating intermediate payments

Bond	Cash Flow at Time t					
	$t = 0$	1	2	3	4	5
$-1 \times 7\frac{7}{8}\% :$	+1097.8125	-78.75	-78.75	-78.75	-78.75	-1078.75
$+1.4 \times 5\frac{5}{8}\% :$	-1403.9375	78.75	78.75	78.75	78.75	1478.75
Difference:	-306.125	0	0	0	0	400.00

Note that we want somehow to eliminate the interest payments (to create a discount bond). Notice that $78.75/56.25 = 1.4$, and consider the following strategy: sell one $7\frac{7}{8}\%$ bond while simultaneously buying 1.4 unit of $5\frac{5}{8}\%$ bonds. The corresponding cash flows are found in Table 11.4.

The net cash flow associated with this strategy thus indicates that the $t = 0$ price of a \$400 payment in 5 years is \$306.25. This price is implicit in the pricing of our two original coupon bonds. Consequently, the price of \$1000 in 5 years must be

$$\$306.125 \times \frac{1000}{400} = \$765.3125$$

Alternatively, the price today of \$1.00 in 5 years is \$0.7653125. In the notation of our earlier discussion we have the following securities:

$$\begin{aligned} \theta_1 & \left[\begin{array}{c} 78.75 \\ 78.75 \\ 78.75 \\ 78.75 \\ 1078.75 \end{array} \right] & \left[\begin{array}{c} 56.25 \\ 56.25 \\ 56.25 \\ 56.25 \\ 1056.25 \end{array} \right] \\ \theta_2 & \text{and} & \text{and we consider} \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{aligned}$$

$$-\frac{1}{400} \left[\begin{array}{c} 78.75 \\ 78.75 \\ 78.75 \\ 78.75 \\ 1078.75 \end{array} \right] + \frac{1.4}{400} \left[\begin{array}{c} 56.25 \\ 56.25 \\ 56.25 \\ 56.25 \\ 1056.25 \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{array} \right]$$

This is an Arrow–Debreu security in the riskless context we are considering in this section.

If there are enough coupon bonds with different maturities with pairs coming due at the same time and with different coupons, we can thus construct a complete set of Arrow–Debreu securities and their implicit prices. Notice that the payoff patterns of

the two bonds are fundamentally different: They are linearly independent of one another. This is a requirement, as per our earlier discussion, for being able to use them to construct a fundamentally new payoff pattern, in this case, the discount bond.

Implicit in every discount bond price is a well-defined rate of return notion. In the case of the prior illustration, for example, the implied 5-year compound risk-free rate is given by

$$\begin{aligned} \$765.3125(1+r_5)^5 &= \$1000, \text{ or} \\ r_5 &= 0.0549 \end{aligned}$$

This observation suggests an intimate relationship between discounting and Arrow–Debreu date pricing. Just as a full set of date claims prices should allow us to price any risk-free cash flow, the rates of return implicit in the Arrow–Debreu prices must allow us to obtain the same price by discounting at the equivalent family of rates. This family of rates is referred to as *term structure of interest rates*.

Definition 11.1 The term structure of interest rates r_1, r_2, \dots is the family of interest rates corresponding to risk-free discount bonds of successively greater maturity; that is, r_i is the rate of return on a risk-free discount bond maturing i periods from the present.

We can systematically recover the term structure from coupon bond prices provided we know the prices of coupon bonds of all different maturities. To illustrate, suppose we observe risk-free government bonds of 1,2,3,4-year maturities all selling at par with coupons, respectively, of 6%, 6.5%, 7.2%, and 9.5%. We can construct the term structure as follows:⁴

r_1 : Since the 1-year bond sells at par, we have $r_1 = 6\%$;

r_2 : By definition, we know that the 2-year bond is priced such that

$1000 = 65/(1+r_1) + 1065/(1+r_2)^2$ which, given that $r_1 = 6\%$, solves for
 $r_2 = 6.5113\%$.

r_3 : This is derived accordingly as the solution to

$$1000 = \frac{75}{(1+r_1)} + \frac{72}{(1+r_2)^2} + \frac{1072}{(1+r_3)^3}$$

With $r_1 = 6\%$ and $r_2 = 6.5113\%$, the solution is $r_3 = 7.2644\%$. Finally, given these values for r_1 to r_3 , r_4 solves:

$$1000 = \frac{95}{(1+r_1)} + \frac{95}{(1+r_2)^2} + \frac{95}{(1+r_3)^3} + \frac{1072}{(1+r_4)^4}, \text{ that is, } r_4 = 9.935\%.$$

⁴ A bond selling at par is selling at its face value, typically \$1000.

Note that these rates are the counterpart to the date-contingent claim prices (see Table 11.5).

Once we have the discount bond prices (the prices of the Arrow–Debreu claims) we can then price all other risk-free securities; for example, suppose we wished to price a 4-year 8% bond:

$t = 0$	1	2	3	4
$-q_0^{8\% \text{ bond}(?)}$	80	80	80	1080

and suppose also that we have available the discount bonds described in Tables 11.5 and 11.6.

Then the portfolio of discount bonds (Arrow–Debreu claims) which replicates the 8% bond cash flow is (Table 11.7):

$$\{0.08 \times 1\text{-yr bond}, 0.08 \times 2\text{-yr bond}, 0.08 \times 3\text{-yr bond}, 1.08 \times 4\text{-yr bond}\}.$$

Table 11.5: Date claim prices versus discount bond prices

	Price of an N Year Claim	Corresponding Discount Bond Price (\$1000 Denomination)
$N = 1$	$q_1 = \$1/1.06 = \0.94339	\$943.39
$N = 2$	$q_2 = \$1/(1.065113)^2 = \0.88147	\$881.47
$N = 3$	$q_3 = \$1/(1.072644)^3 = \0.81027	\$810.27
$N = 4$	$q_4 = \$1/(1.09935)^4 = \0.68463	\$684.63

Table 11.6: Discount bonds as Arrow–Debreu claims

Bond	Price ($t = 0$)	CF Pattern			
		$t = 1$	2	3	4
1-yr discount	-\$943.39	\$1000			
2-yr discount	-\$881.47		\$1000		
3-yr discount	-\$810.27			\$1000	
4-yr discount	-\$684.63				\$1000

Table 11.7: Replicating the discount bond cash flow

Bond	Price ($t = 0$)	CF Pattern			
		$t = 1$	2	3	4
0.08 1-yr discount	$(0.08)(-943.39) = -\$75.47$				
0.08 2-yr discount	$(0.08)(-881.47) = -\$70.52$				
0.08 3-yr discount	$(0.08)(-810.27) = -\$64.82$				
1.08 4-yr discount	$(1.08)(-684.63) = -\$739.40$				
		\$80 (80 date 1 A–D claims)	\$80 (80 date 2 A–D claims)	\$80	\$1080

Thus:

$$q_0^{8\% \text{ bond}} = 0.08(\$943.39) + 0.08(\$881.47) + 0.08(\$810.27) + 1.08(\$684.63) = \$950.21.$$

Notice that we are emphasizing, in effect, the equivalence of the term structure of interest rates with the prices of date-contingent claims. Each defines the other. This is especially apparent in [Table 11.5](#).

Let us now extend the above discussion to consider the evaluation of arbitrary risk-free cash flows: any such cash flow can be evaluated as a portfolio of Arrow–Debreu securities.

For example:

$$\begin{array}{ccccc} t=0 & 1 & 2 & 3 & 4 \\ 60 & 25 & 150 & 300 \end{array}$$

We want to price this cash flow today ($t = 0$) using the Arrow–Debreu prices we have calculated in [Table 11.5](#).

$$\begin{aligned} q_0 &= (\$60 \text{ at } t=1) \left(\frac{\$0.94339 \text{ at } t=0}{\$1 \text{ at } t=1} \right) + (\$25 \text{ at } t=2) \left(\frac{\$0.88147 \text{ at } t=0}{\$1 \text{ at } t=2} \right) + \dots \\ &= (\$60) \frac{1.00}{1+r_1} + (\$25) \frac{1.00}{(1+r_2)} + \dots \\ &= (\$60) \frac{1.00}{1.06} + (\$25) \frac{1.00}{(1.065113)^2} + \dots \end{aligned}$$

The second equality underlines the fact that *evaluating risk-free projects as portfolios of Arrow–Debreu state-contingent securities is equivalent to discounting at the term structure:*

$$= \frac{60}{(1+r_1)} + \frac{25}{(1+r_2)} + \frac{150}{(1+r_3)} + \dots, \text{ etc.}$$

In effect, we treat a risk-free project as a risk-free coupon bond with (potentially) differing coupons. There is an analogous notion of forward prices and its more familiar counterpart, the forward rate. We discuss this extension in [Appendix 11.1](#).

11.4 The Value Additivity Theorem

In this section, we present an important result illustrating the power of the Arrow–Debreu pricing apparatus to generate one of the main lessons of the CAPM. Let there be two assets (complex securities) a and b with corresponding date 1 payoffs \tilde{z}_a and \tilde{z}_b ,

respectively, and equilibrium ($t = 0$) prices q_a and q_b . Suppose a third asset, c , turns out to be a linear combination of a and b . By that we mean that the payoff to c can be replicated by a portfolio of a and b . One can thus write

$$\tilde{z}_c = A\tilde{z}_a + B\tilde{z}_b, \text{ for some constant coefficients } A \text{ and } B \quad (11.2)$$

Then the proposition known as the **Value Additivity Theorem** asserts that the same linear relationship must hold for the date 0 prices of the three assets:

$$q_c = Aq_a + Bq_b$$

Let us first prove this result and then discuss its implications. The proof easily follows from our discussion in [Section 11.2](#) on the pricing of complex securities in a complete market Arrow–Debreu world. Indeed, for our two securities a, b , one must have:

$$q_i = \sum_s q_s z_{si}, \quad i = a, b \quad (11.3)$$

where q_s is the price of an Arrow–Debreu security that pays one unit of consumption in state s (and zero otherwise) and z_{si} is the payoff of asset i in state s

But then, the pricing of c must respect the following relationships:

$$q_c = \sum_s q_s z_{sc} = \sum_s q_s (A z_{sa} + B z_{sb}) = \sum_s (A q_s z_{sa} + B q_s z_{sb}) = Aq_a + Bq_b$$

The first equality follows from the fact that c is itself a complex security and can thus be priced using Arrow–Debreu prices (i.e., an equation such as [Eq. \(11.3\)](#) applies); the second directly follows from [Eq. \(11.2\)](#); the third is a pure algebraic expansion that is feasible because our pricing relationships are fundamentally linear; the fourth again follows from [Eq. \(11.3\)](#).

Now this is easy enough, but why is it interesting? Think of a and b as being two stocks with negatively correlated returns; we know that c , a portfolio of these two stocks, is much less risky than either one of them. But q_c is a linear combination of q_a and q_b . Thus, the fact that they can be combined in a less risky portfolio has implications for the pricing of the two independently riskier securities and their equilibrium returns. Specifically, it cannot be the case that q_c would be *high* because it corresponds to a desirable, riskless, claim while the q_a and q_b would be *low* because they are risky.

To see this more clearly, let us take an extreme example. Suppose that a and b are *perfectly* negatively correlated. For an appropriate choice of A and B , say A^* and B^* , the resulting portfolio, call it d , will have zero risk; that is, it will pay a constant amount in each and every state of nature. What should the price of this riskless portfolio be? Intuitively, its price must be such that purchasing d units at q_d will earn the riskless rate of return.

But how could the risk of a and b be remunerated while, simultaneously, d would earn the riskless rate and the Value Additivity Theorem would hold? The answer is that this is not possible. Therefore, there cannot be any remuneration for risk in the pricing of a and b . The prices q_a and q_b must be such that the expected return on a and b is the riskless rate. This is true despite the fact that a and b are two risky assets (they do not pay the same amount in each state of nature).

In formal terms, we have just asserted that the two terms of the Value Additivity Theorem $\tilde{z}_d = A^* \tilde{z}_a + B^* \tilde{z}_b$ and $q_d = A^* q_a + B^* q_b$, together with the fact that d is risk free,

$$\frac{E\tilde{z}_d}{q_d} = 1 + r_f, \text{ force}$$

$$\frac{E\tilde{z}_a}{q_a} = \frac{E\tilde{z}_b}{q_b} = 1 + r_f.$$

What we have obtained in this very general context is a confirmation of one of the main results of the CAPM: diversifiable risk is not priced. If risky assets a and b can be combined in a riskless portfolio, that is, if their risk can be diversified away, their return cannot exceed the risk-free return. Note that we have made no assumption here on utility functions or on the return expectations held by agents. On the other hand, we have explicitly assumed that markets are complete and that, consequently, each and every complex security can be priced (by arbitrage) as a portfolio of Arrow–Debreu securities.

It thus behooves us to describe how Arrow–Debreu state claim prices might actually be obtained in practice. This is the subject of the remaining sections of Chapter 11.

11.5 Using Options to Complete the Market: An Abstract Setting

Let us assume a finite number of possible future date-states indexed $i = 1, 2, \dots, N$. Suppose, for a start, that three states of the world are possible in date $T = 1$, yet only one security (a stock) is traded. The single security's payoffs are as follows:

State	Payoff
θ_1	[1]
θ_2	[2]
θ_3	[3]

Clearly, this unique asset is not equivalent to a complete set of state-contingent claims. Note that we can identify the payoffs with the ex-post price of the security in each of the three states: the security pays two units of the numeraire commodity in state 2, and we decide that its price then is \$2. This amounts to normalizing the ex-post, date 1, price of the commodity to \$1, much as we have done at date 0. On that basis, we can consider call

options written on this asset with exercise prices \$1 and \$2, respectively. These securities are contracts giving the right (but not the obligation) to purchase the underlying security tomorrow at prices \$1 and \$2, respectively. They are contingent securities in the sense that the right they entail is valuable only if the price of the underlying security exceeds the exercise price at expiration, and they are valueless otherwise. We think of the option expiring at $T = 1$, when the state of nature is revealed.⁵ The *states of nature* structure enables us to be specific regarding what these contracts effectively promise to pay. Take the call option with exercise price \$1. If state 1 is realized, that option is a right to buy at \$1 the underlying security whose value is exactly \$1. The option is said to be *at the money*, and, in this case, the right in question is valueless. If state 2 is realized, however, the stock is worth \$2. The right to buy, at a price of \$1, something one can immediately resell for \$2 naturally has a market value of \$1. In this case, the option is said to be *in the money*. In other words, at $T = 1$, when the state of nature is revealed, an option is worth the difference between the value of the underlying asset and its exercise price, if this difference is positive, and zero otherwise. The complete payoff vectors of these options at expiration are as follows:

$$C_T([1, 2, 3]; 1) = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} \quad \begin{array}{l} \theta_1 \\ \theta_2 \\ \theta_3 \end{array} \quad \left\{ \begin{array}{l} \text{at the money} \\ \text{in the money} \\ \text{in the money} \end{array} \right\}.$$

Similarly,

$$C_T([1, 2, 3]; 2) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \begin{array}{l} \theta_1 \\ \theta_2 \\ \theta_3 \end{array}$$

In our notation, $C_T(S; K)$ is the payoff to a call option written on security S with exercise price K at expiration date T . We use $C_t(S; K)$ to denote the option's market price at time $t \leq T$. We frequently drop the time subscript to simplify notation when there is no ambiguity.

It remains now to convince ourselves that the three traded assets (the underlying stock and the two call options, each denoted by its payoff vector at $T = 1$)

$$\begin{array}{ll} \theta_1 & \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \\ \theta_2 & \\ \theta_3 & \end{array}$$

⁵ In our simple two-date world there is no difference between an American option, which can be exercised at any date before the expiration date, and a European option, which can be exercised only at expiration.

constitute a complete set of securities markets for states $(\theta_1, \theta_2, \theta_3)$. This is so because

we can use them to create all the state claims. Clearly $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ is present.

To create $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, observe that

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = w_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + w_2 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} + w_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

where $w_1 = 0$, $w_2 = 1$, and $w_3 = -2$.

The vector $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ can be similarly created.

We have thus illustrated one of the main ideas of this chapter, and we need to discuss how general and applicable it is in more realistic settings. A preliminary issue is why trading call option securities $C([1,2,3]; 1)$ and $C([1,2,3]; 2)$ might be the preferred approach to completing the market, relative to the alternative possibility of directly issuing the Arrow–Debreu securities [1,0,0] and [0,1,0]? In the simplified world of our example, in the absence of transactions costs, there is, of course, no advantage to creating the options markets. In the real world, however, if a new security is to be issued, its issuance must be accompanied by costly disclosure as to its characteristics; in our parlance, the issuer must disclose as much as possible about the security's payoff in the various states. As there may be no agreement as to what the relevant future states are—let alone what the payoffs will be—this disclosure is difficult. And if there is no consensus as to its payoff pattern (i.e., its basic structure of payoffs), investors will not want to hold it, and it will not trade. But the payoff pattern of an option on an already traded asset is obvious and verifiable to everyone. For this reason, it is, in principle, a much less expensive new security to issue. Another way to describe the advantage of options is to observe that it is useful conceptually, but difficult in practice, to define and identify a single state of nature. It is more practical to define contracts contingent on a well-defined *range* of states. The fact that these states are themselves defined in terms of, or revealed through, market prices is another advantage of this type of contract.

Note that options are by definition in zero net supply; that is, in this context

$$\sum_j C_t^j([1, 2, 3]; K) = 0$$

where $C_t^j([1, 2, 3]; K)$ is the value of call options with exercise price K , held by agent j at time $t \leq T$. This means that there must exist a group of agents with negative positions

serving as the counterparty to the subset of agents with positive holdings. We naturally interpret those agents as agents who have *written* the call options.

We have illustrated the property that markets can be completed using call options. Now let us explore the generality of this result. Can call options always be used to complete the market in this way? The answer is not necessarily. It depends on the payoff to the underlying fundamental assets. Consider the asset:

$$\begin{matrix} \theta_1 & \begin{bmatrix} 2 \\ 2 \end{bmatrix} \\ \theta_2 & \begin{bmatrix} 2 \\ 2 \end{bmatrix} \\ \theta_3 & \begin{bmatrix} 3 \\ \end{bmatrix} \end{matrix}$$

For any exercise price K , all options written on this security must have payoffs of the form:

$$C([2, 2, 3]; K) = \begin{cases} \begin{bmatrix} 2 - K \\ 2 - K \\ 3 - K \end{bmatrix} & \text{if } K \leq 2 \\ \begin{bmatrix} 0 \\ 0 \\ 3 - K \end{bmatrix} & \text{if } 2 < K \leq 3 \end{cases}$$

Clearly, for any K ,

$$\begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} \text{ and } \begin{bmatrix} 2 - K \\ 2 - K \\ 3 - K \end{bmatrix}$$

have identical payoffs in state θ_1 and θ_2 , and, therefore, they cannot be used to generate Arrow–Debreu securities

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

There is no way to complete the markets with options in the case of this underlying asset. This illustrates the following truth: it is not possible to write options that distinguish between two states if the underlying assets pay identical returns in those states.

The problem just illustrated can sometimes be solved *if we permit options to be written on portfolios of the basic underlying assets*. Consider the case of four possible states at $T = 1$, and suppose that the only assets currently traded are

$$\theta_1 \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix} \text{ and } \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

It can be shown that it is not possible, using call options, to generate a complete set of securities markets using only these underlying securities. Consider, however, the portfolio composed of two units of the first asset and one unit of the second:

$$2 \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix} + 1 \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

The portfolio pays a different return in each state of nature. Options written on the portfolio alone can thus be used to construct a complete set of traded Arrow–Debreu securities. The example illustrates a second general truth, which we will enumerate as [Proposition 11.3](#).

Proposition 11.3 A necessary as well as sufficient condition for the creation of a complete set of Arrow–Debreu securities is that there exists a single portfolio with the property that options can be written on it and such that its payoff pattern distinguishes among all states of nature.

Returning to the example immediately above, we easily see that the created portfolio and the three calls to be written on it,

$$\begin{bmatrix} 3 \\ 4 \\ 5 \\ 6 \end{bmatrix} \text{ plus } \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 0 \\ 1 \\ 2 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$(K = 3)$ $(K = 4)$ $(K = 5)$

are sufficient (i.e., constitute a complete set of markets in our four-state world). Combinations of the $(K = 5)$ and $(K = 4)$ vectors can create:

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Combinations of this vector, and the ($K = 5$) and ($K = 3$) vectors can then create:

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \text{ etc.}$$

Probing further we may inquire whether *the writing of calls on the underlying assets is always sufficient*, or whether there are circumstances under which other types of options may be necessary. Again, suppose there are four states of nature, and consider the following set of *primitive securities*:

$$\begin{array}{c} \theta_1 \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\ \theta_2 \quad \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\ \theta_3 \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \\ \theta_4 \quad \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{array}$$

Because these assets pay either one or zero in each state, calls written on them will either replicate the asset itself, or give the zero payoff vector. The writing of call options will not help because they cannot further discriminate among states. But suppose we write a put option on the first asset with exercise price 1. A put is a contract giving the right, but not the obligation, to *sell* an underlying security at a prespecified exercise price on a given expiration date. The put option with exercise price 1 has positive value at $T = 1$ in those states where the underlying security has value less than 1. The put on the first asset with exercise price = \$1 thus has the following payoff:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = P_T([0, 0, 0, 1]; 1).$$

You can confirm that the securities plus the put are sufficient to allow us to construct (as portfolios of them) a complete set of Arrow–Debreu securities for the indicated four states. In general, one can prove [Proposition 11.4](#).

Proposition 11.4 If it is possible to create, using options, a complete set of traded securities, simple put and call options written on the underlying assets are sufficient to accomplish this goal.

That is, portfolios of options are not required.

11.6 Synthesizing State-Contingent Claims: A First Approximation

The abstract setting of the preceding discussion aimed at conveying the message that options are natural instruments for completing the markets. In this section, we show how we can directly create a set of state-contingent claims, *as well as their equilibrium prices*, using option prices or option pricing formulas in a more realistic setting. The interest in doing so is, of course, to exploit the possibility, inherent in Arrow–Debreu prices, of pricing any complex security. In this section, we first approach the problem under the hypothesis that the price of the underlying security or portfolio can take only discrete values.

Assume that a risky asset is traded with current price S and future price S_T . It is assumed that S_T discriminates across all states of nature so that [Proposition 11.3](#) applies. Without loss of generality, we may assume that S_T takes the following set of values:

$$S_1 < S_2 < \dots < S_\theta < \dots < S_N,$$

where S_θ denotes the price of this complex security if state θ is realized at date T . Assume also that call options are written on this asset with all possible exercise prices, and that these options are traded. Let us also assume that $S_\theta = S_{\theta-1} + \delta$ for every state θ . (This is not so unreasonable as stocks, say, are traded at prices that can differ only in multiples of a minimum price change).⁶ Throughout the discussion we will fix the time to expiration and will not denote it notationally.

Consider, for any state $\hat{\theta}$, the following portfolio P :

- Buy one call with $K = S_{\hat{\theta}-1}$
- Sell two calls with $K = S_{\hat{\theta}}$
- Buy one call with $K = S_{\hat{\theta}+1}$

At any point in time, the value of this portfolio, V_P , is

$$V_P = C(S, K = S_{\hat{\theta}-1}) - 2C(S, K = S_{\hat{\theta}}) + C(S, K = S_{\hat{\theta}+1}).$$

To see what this portfolio represents, let us examine its payoff *at expiration* (refer to [Figure 11.1](#)).

⁶ Until recently, the minimum price change was equal to $\$ \frac{1}{16}$ on the NYSE. At the end of 2000, *decimal pricing* was introduced whereby the prices are quoted to the nearest $\$ \frac{1}{100}$ (1 cent).

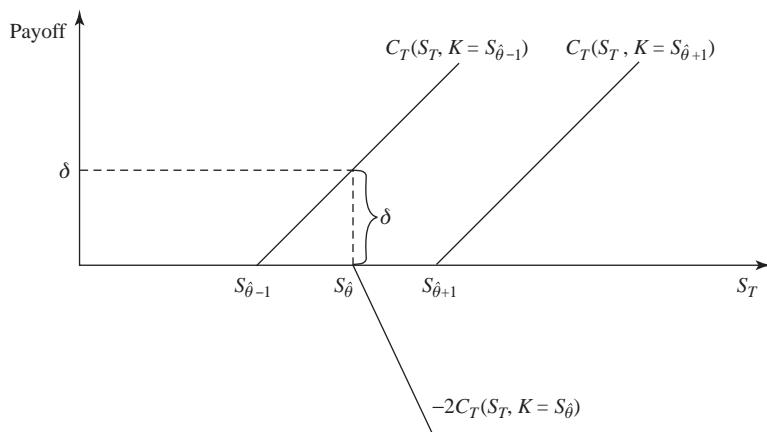


Figure 11.1
Payoff diagram for all options in the portfolio P .

For $S_T \leq S_{\hat{\theta}-1}$, the value of our options portfolio, P , is zero. A similar situation exists for $S_T \geq S_{\hat{\theta}+1}$ since the loss on the two written calls with $K = S_{\hat{\theta}}$ exactly offsets the gains on the other two calls. In state $\hat{\theta}$, the value of the portfolio is δ corresponding to the value of $C_T(S_{\hat{\theta}}, K = S_{\hat{\theta}-1})$, the other two options being worthless when the underlying security takes value $S_{\hat{\theta}}$. The payoff from such a portfolio thus equals:

$$\text{Payoff to } P = \begin{cases} 0 & \text{if } S_T < S_{\hat{\theta}} \\ \delta & \text{if } S_T = S_{\hat{\theta}} \\ 0 & \text{if } S_T > S_{\hat{\theta}}; \end{cases}$$

in other words, it pays a positive amount δ in state $\hat{\theta}$, and nothing otherwise. That is, it replicates the payoff of the Arrow–Debreu security associated with state $\hat{\theta}$ up to a factor (in the sense that it pays δ instead of 1). Consequently, the current price of the state $\hat{\theta}$ contingent claim (i.e., one that pays \$1 if state $\hat{\theta}$ is realized and nothing otherwise) must be

$$q_{\hat{\theta}} = \frac{1}{\delta} [C(S, K = S_{\hat{\theta}-1}) + C(S, K = S_{\hat{\theta}+1}) - 2C(S, K = S_{\hat{\theta}})].$$

Even if these calls are not traded, if we identify our relevant states with the prices of some security—say the market portfolio—then we can use readily available option pricing formulas (such as the famous Black and Scholes formula) to obtain the necessary call prices and, from them, compute the price of the state-contingent claim. We explore this idea further in the next section.

11.7 Recovering Arrow–Debreu Prices from Options Prices: A Generalization

By the CAPM, the only relevant risk is systematic risk. We may interpret this to mean that the only states of nature that are economically or financially relevant are those that can be identified with different values of the market portfolio.⁷ The market portfolio thus may be selected to be the complex security on which we write options, portfolios of which will be used to replicate state-contingent payoffs. The conditions of [Proposition 11.1](#) are satisfied, guaranteeing the possibility of completing the market structure.

In [Section 11.6](#), we considered the case where the underlying asset assumed a discrete set of values. If the underlying asset is the market index quoted in \$.01 units, the number of potential discrete values may be quite large, and the discrete calculations accordingly involved. Can economies of procedure be achieved by assuming the underlying asset assumes a continuum of values? How to accommodate this generalization is discussed below.

1. Suppose that S_T , the price of the underlying portfolio (we may think of it as a proxy for M), assumes a *continuum* of possible values. We want to price an Arrow–Debreu security that pays \$1 if $\tilde{S}_T \in [-\delta/2 + \hat{S}_T, \hat{S}_T + \delta/2]$, in other words, if S_T assumes any value in a range of width δ , centered on \hat{S}_T . We are thus identifying our states of nature with **ranges of possible values** for the market portfolio. Here the subscript T refers to the future date at which the Arrow–Debreu security is to pay \$1 if the relevant state is realized.
2. Let us construct the following portfolio for some small positive number $\varepsilon > 0$,

$$\text{Buy one call with } K = \hat{S}_T - \frac{\delta}{2} - \varepsilon$$

$$\text{Sell one call with } K = \hat{S}_T - \frac{\delta}{2}$$

$$\text{Sell one call with } K = \hat{S}_T + \frac{\delta}{2}$$

$$\text{Buy one call with } K = \hat{S}_T + \frac{\delta}{2} + \varepsilon.$$

[Figure 11.2](#) depicts what this portfolio pays *at expiration*.⁸

⁷ That is, diversifiable risks have zero market value (see Chapter 8 and [Section 11.4](#)). At an individual level, personal risks are, of course, also relevant. They can, however, be insured or diversified away. Insurance contracts are often the most appropriate to cover these risks. Recall our discussion of this issue in Chapter 1.

⁸ The option position corresponding to this portfolio is known as a *butterfly spread* in the jargon of options traders.

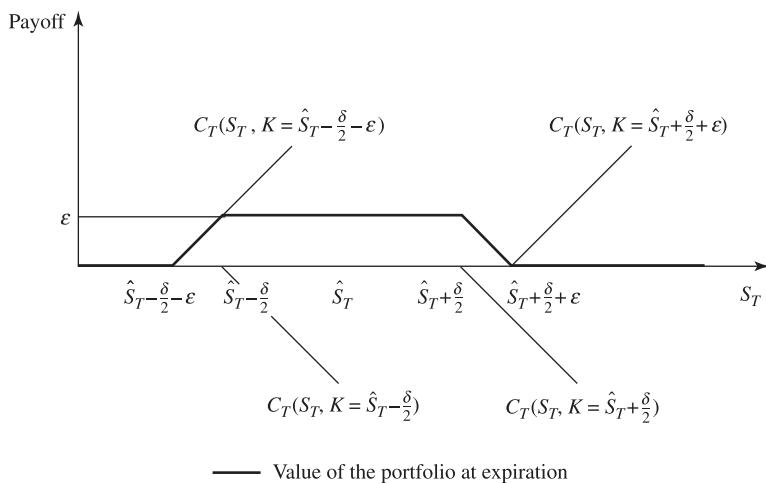


Figure 11.2
Payoff diagram: portfolio of options.

Observe that our portfolio pays ε on a range of states and 0 almost everywhere else. By purchasing $1/\varepsilon$ units of the portfolio, we will mimic the payoff of an Arrow–Debreu security, except for the two small diagonal sections of the payoff line where the portfolio pays something between 0 and ε . This undesirable feature (since our objective is to replicate an Arrow–Debreu security) will be dealt with by using a standard mathematical trick involving taking limits.

3. Let us thus consider buying $1/\varepsilon$ units of the portfolio. The total payment, when $\hat{S}_T - \delta/2 \leq S_T \leq \hat{S}_T + \delta/2$, is $\varepsilon \cdot 1/\varepsilon \equiv 1$, for any choice of ε . We want to let $\varepsilon \mapsto 0$, so as to eliminate payments in the ranges $S_T \in [\hat{S}_T - \delta/2 - \varepsilon, \hat{S}_T - \delta/2]$ and $S_T \in (\hat{S}_T + \delta/2, \hat{S}_T + \delta/2 + \varepsilon]$. The value of $1/\varepsilon$ units of this portfolio is:

$$\begin{aligned} & \frac{1}{\varepsilon} \left\{ C\left(S, K = \hat{S}_T - \frac{\delta}{2} - \varepsilon\right) - C\left(S, K = \hat{S}_T - \frac{\delta}{2}\right) \right. \\ & \left. - \left[C\left(S, K = \hat{S}_T + \frac{\delta}{2}\right) - C\left(S, K = \hat{S}_T + \frac{\delta}{2} + \varepsilon\right) \right] \right\}, \end{aligned}$$

where a minus sign indicates that the call was sold (thereby reducing the cost of the portfolio by its sale price). On balance, the portfolio will have a positive price as it represents a claim on a positive cash flow in certain states of nature. Let us assume that the pricing function for a call with respect to changes in the exercise price can

be differentiated. (This property is true, in particular, in the case of the Black and Scholes option pricing formula.) We then have:

$$\begin{aligned}
 & \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \left\{ C\left(S, K = \hat{S}_T - \frac{\delta}{2} - \varepsilon\right) - C\left(S, K = \hat{S}_T - \frac{\delta}{2}\right) \right. \\
 & \quad \left. - \left[C\left(S, K = \hat{S}_T + \frac{\delta}{2}\right) - C\left(S, K = \hat{S}_T + \frac{\delta}{2} + \varepsilon\right) \right] \right\} \\
 &= -\lim_{\varepsilon \rightarrow 0} \left\{ \underbrace{\frac{C\left(S, K = \hat{S}_T - \frac{\delta}{2} - \varepsilon\right) - C\left(S, K = \hat{S}_T - \frac{\delta}{2}\right)}{-\varepsilon}}_{\leq 0} \right\} \\
 &+ \lim_{\varepsilon \rightarrow 0} \left\{ \underbrace{\frac{C\left(S, K = \hat{S}_T + \frac{\delta}{2} + \varepsilon\right) - C\left(S, K = \hat{S}_T + \frac{\delta}{2}\right)}{\varepsilon}}_{\leq 0} \right\} \\
 &= C_2\left(S, K = \hat{S}_T + \frac{\delta}{2}\right) - C_2\left(S, K = \hat{S}_T - \frac{\delta}{2}\right).
 \end{aligned}$$

Here the subscript 2 indicates the partial derivative with respect to the second argument (K), evaluated at the indicated exercise prices. In summary, the limiting portfolio has a payoff at expiration as represented in [Figure 11.3](#) and a (current) price $C_2(S, K = \hat{S}_T + \delta/2) - C_2(S, K = \hat{S}_T - \delta/2)$ that is positive since the payoff is positive. We have thus priced an Arrow–Debreu state-contingent claim one period ahead, given

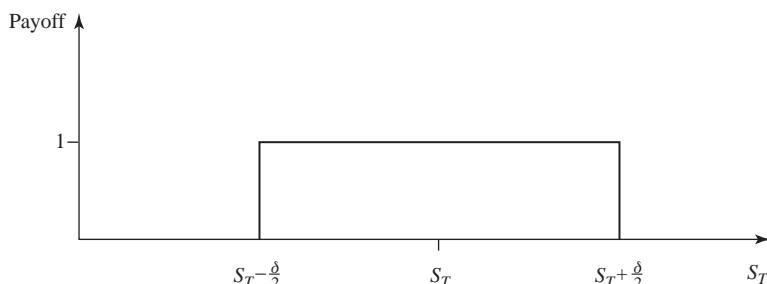


Figure 11.3
Payoff diagram for the limiting portfolio.

that we define states of the world as coincident with ranges of a proxy for the market portfolio.

4. Suppose, for example, we have an uncertain payment with the following payoff at time T :

$$CF_T = \begin{cases} 0 & \text{if } S_T \notin \left[\hat{S}_T - \frac{\delta}{2}, \hat{S}_T + \frac{\delta}{2} \right] \\ 50,000 & \text{if } S_T \in \left[\hat{S}_T - \frac{\delta}{2}, \hat{S}_T + \frac{\delta}{2} \right] \end{cases}$$

The value today of this cash flow is:

$$50,000 \cdot \left[C_2\left(S, K = \hat{S}_T + \frac{\delta}{2}\right) - C_2\left(S, K = \hat{S}_T - \frac{\delta}{2}\right) \right].$$

The formula we have developed is very general. In particular, for any arbitrary values S_T^1 and S_T^2 , the price of an Arrow–Debreu contingent claim that pays \$1 if the underlying market portfolio assumes a value $S_T \in [S_T^1, S_T^2]$, is given by

$$q(S_T^1, S_T^2) = C_2(S, K = S_T^2) - C_2(S, K = S_T^1). \quad (11.4)$$

We value this quantity in [Box 11.1](#) for a particular set of parameters making explicit use of the Black–Scholes option pricing formula.

BOX 11.1 Pricing A–D Securities with Black–Scholes

For calls priced according to the Black–Scholes option pricing formula, [Breeden and Litzenberger \(1978\)](#) prove that

$$\begin{aligned} q(S_T^1, S_T^2) &= C_2(S, K = S_T^2) - C_2(S, K = S_T^1) \\ &= e^{-rT} \{N(d_2(S_T^1)) - N(d_2(S_T^2))\} \end{aligned}$$

where

$$d_2(S_T^i) = \frac{\left[\ln\left(\frac{S_0}{S_T^i}\right) + \left(r_f - \gamma - \frac{\sigma^2}{2}\right)T \right]}{\sigma\sqrt{T}}$$

In this expression, T is the time to expiration, r_f the annualized continuously compounded riskless rate over that period, γ the continuous annualized portfolio dividend yield, σ the standard deviation of the continuously compounded rate of return on the underlying index portfolio, $N()$ the standard normal distribution, and S_0 the current value of the index.

(Continued)

BOX 11.1 Pricing A–D Securities with Black–Scholes (Continued)

Suppose the not-continuously compounded risk-free rate is 0.06, the not-continuously compounded dividend yield is $\delta = 0.02$, $T = 0.5$ year, $S_0 = 1500$, $S_T^2 = 1700$, $S_T^1 = 1600$, $\sigma = 0.20$; then

$$\begin{aligned} d_2(S_T^1) &= \frac{\left\{ \ln\left(\frac{1500}{1600}\right) + \left[\ln(1.06) - \ln(1.02) - \frac{(0.20)^2}{2} \right] (0.5) \right\}}{0.20\sqrt{0.5}} \\ &= \frac{\{-0.0645 + (0.0583 - 0.0198 - 0.02)(0.5)\}}{0.1414} \\ &= -0.391 \\ d_2(S_T^2) &= \frac{\left\{ \ln\left(\frac{1500}{1700}\right) + (0.0583 - 0.0198 - 0.02)(0.5) \right\}}{0.1414} \\ &= \frac{\{(-0.1252 + 0.00925)\}}{0.1414} \\ &= -0.820 \\ q(S_T^1, S_T^2) &= e^{-\ln(1.06)(0.5)} \{N(-0.391) - N(-0.820)\} \\ &= 0.9713 \{0.3479 - 0.2061\} \\ &= 0.1381, \end{aligned}$$

or about \$0.14.

Suppose we wished to price an uncertain cash flow to be received one period from the present, where a period corresponds to a duration of time T . What do we do? Choose several ranges of the value of the market portfolio corresponding to the various states of nature that may occur—say three states: “recession,” “slow growth,” and “boom”—and estimate the cash flow in each of these states (see Figure 11.4). It would be unusual to have a large number of states, as the requirement of having to estimate the cash flows in each of those states is likely to exceed our forecasting abilities.

Suppose the cash-flow estimates are, respectively, CF_B , CF_{SG} , and CF_R , where the subscripts denote, respectively, “boom,” “slow growth,” and “recession.” Then,

$$\text{Value of the } CF = V_{CF} = q(S_T^3, S_T^4)CF_B + q(S_T^2, S_T^3)CF_{SG} + q(S_T^1, S_T^2)CF_R,$$

where $S_T^1 < S_T^2 < S_T^3 < S_T^4$, and the Arrow–Debreu prices are estimated from option prices or option pricing formulas according to Eq. (11.4).

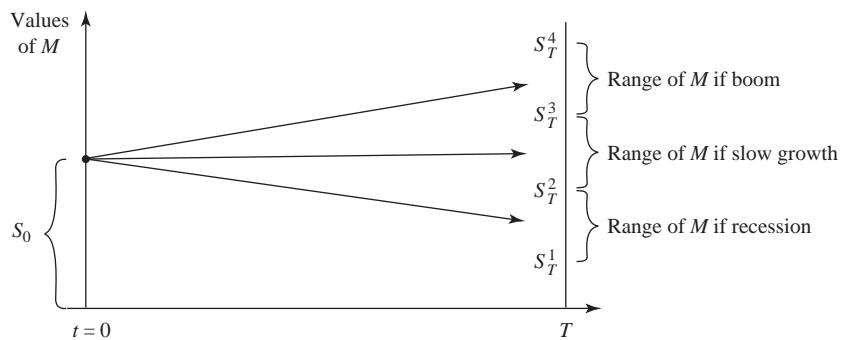


Figure 11.4
Constructing “states” as ranges of the future value of M .

We can go one (final) step further if we assume for a moment that the cash flow we wish to value can be described by a continuous function of the value of the market portfolio.

In principle, for a very fine partition of the range of possible values of the market portfolio, say $\{S_1, \dots, S_N\}$, where $S_i < S_{i+1}$, $S_N = \max S_T$, and $S_1 = \min S_T$, we can price the Arrow–Debreu securities that pay off in each of these $N - 1$ states defined by the partition:

$$\begin{aligned} q(S_1, S_2) &= C_2(S, S_2) - C_2(S, S_1) \\ q(S_2, S_3) &= C_2(S, S_3) - C_2(S, S_2), \dots, \text{etc.} \end{aligned}$$

Simultaneously, we could approximate a cash-flow function $CF(S_T)$ by a function that is constant in each of these ranges of S_T (a so-called step function); in other words, $\hat{CF}(S_T) = CF_i$, for $S_{i-1} \leq S_T \leq S_i$. For example,

$$\hat{CF}(S_T) = CF_i = \frac{CF(S, S_T = S_i) + CF(S, S_T = S_{i-1})}{2} \text{ for } S_{i-1} \leq S_T \leq S_i$$

This particular approximation is represented in Figure 11.5. The value of the approximate cash flow would then be

$$\begin{aligned} V_{CF} &= \sum_{i=1}^N \hat{CF}_i \cdot q(S_{i-1}, S_i) \\ &= \sum_{i=1}^N \hat{CF}_i [C_2(S, S_T = S_i) - C_2(S, S_T = S_{i-1})] \end{aligned} \tag{11.5}$$

Our approach is now clear. The precise value of the uncertain cash flow will be the sum of the approximate cash flows evaluated at the Arrow–Debreu prices as the norm of the

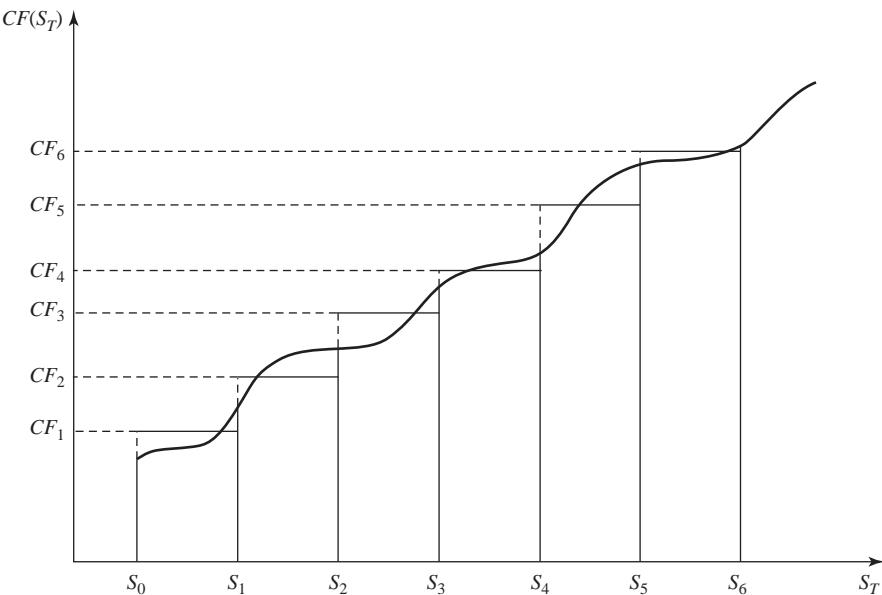


Figure 11.5
A discrete approximation to a continuous cash-flow function.

partition (the size of the interval $S_i - S_{i-1}$) tends to zero. It can be shown (and it is intuitively plausible) that the limit of Eq. (11.5) as $\max_i |S_{i+1} - S_i| \rightarrow 0$ is the integral of the cash-flow function multiplied by the second derivative of the call's price with respect to the exercise price. The latter is the continuum counterpart to the difference in the first derivatives of the call prices entering in Eq. (11.4).

$$\lim_{\substack{\max_i |S_{i+1} - S_i| \rightarrow 0 \\ i}} \sum_{i=1}^N \hat{C}F_i [C_2(S, S_T = S_{i+1}) - C_2(S, S_T = S_i)] = \int CF(S_T) C_{22}(S, S_T) dS_T. \quad (11.6)$$

As a particular case of a constant cash-flow stream, a risk-free bond paying \$1 in every state is then priced as per

$$q^{rf} = \frac{1}{(1 + r_f)} = \int_0^\infty C_{22}(S, S_T) dS_T. \quad (11.7)$$

See Box 11.2 for a numerical illustration of these ideas (Table 11.8).

**BOX 11.2 Extracting Arrow–Debreu Prices from Option Prices:
A Numerical Illustration**

Let us now illustrate the power of the approach adopted in this and the previous section. For that purpose, [Table 11.8](#) (adapted from [Pirkner et al. \(1999\)](#)) starts by recording call prices, obtained from the Black–Scholes formula for a call option, on an underlying index portfolio, currently valued at $S = 10$, for a range of strike prices going from $K = 7$ to $K = 13$ (columns 1 and 2). Column 3 computes the value of portfolio P of [Section 11.6](#). Given that the difference between the exercise prices is always 1 (i.e., $\delta = 1$), holding exactly one unit of this portfolio replicates the \$1 payoff of the Arrow–Debreu security associated with $K = 10$. This is shown on the bottom line of column 7, which corresponds to $S = 10$. From column 3, we learn that the price of this Arrow–Debreu security, which must be equal to the value of the replicating portfolio, is \$0.184. Finally, the last two columns approximate the first and second derivatives of the call price with respect to the exercise price. In the current context, this is naturally done by computing the first and second differences (the price increments and the increments of the increments as the exercise price varies) from the price data given in column 2. This is a literal application of [Eq. \(11.4\)](#). One thus obtains the full series of Arrow–Debreu prices for states of nature identified with values of the underlying market portfolios ranging from 8 to 12, confirming that the \$0.184 price occurs when the state of nature is identified as $S = 10$ (or $9.5 < S < 10.5$).

Table 11.8: Pricing an Arrow–Debreu state claim

K	$C(S, K)$	Cost of Position	Payoff if $S_T =$							ΔC	$\Delta(\Delta C) = q_\theta$
			7	8	9	10	11	12	13		
7	3.354									-0.895	
8	2.459									-0.789	0.106
9	1.670	+1.670	0	0	0	1	2	3	4	-0.625	0.164
10	1.045	-2.090	0	0	0	0	-2	-4	-6	-0.441	0.184
11	0.604	+0.604	0	0	0	0	0	1	2	-0.279	0.162
12	0.325									-0.161	0.118
13	0.164	0.184	0	0	0	1	0	0	0		

11.8 Arrow–Debreu Pricing in a Multiperiod Setting

The fact that the Arrow–Debreu pricing approach is static makes it ideal for the pricing of one-period cash flows, and it is, quite naturally, in this context that most of our discussion has been framed. But as we have emphasized previously, it is formally equally

appropriate for pricing multiperiod cash flows. The estimation (for instance, via option pricing formulas and the methodology introduced in the last two sections) of Arrow–Debreu prices for several periods ahead is inherently more difficult, however, and relies on more perilous assumptions than in the case of one period ahead prices. (This parallels the fact that the assumptions necessary to develop closed-form option pricing formulas are more questionable when they are used in the context of pricing long-term options.) Pricing long-term assets, whatever the approach adopted, requires making hypotheses to the effect that the recent past tells us something about the future, which, in ways to be defined and which vary from one model to the next, translates into hypotheses that some form of stationarity prevails. Completing the Arrow–Debreu pricing approach with an additional stationarity hypothesis provides an interesting perspective on the pricing of multiperiod cash flows. This is the purpose of the present section.

For notational simplicity, let us first assume that the same two states of nature (ranges of value of M) can be realized in each period and that all future state-contingent cash flows have been estimated. The structure of the cash flow is found in [Figure 11.6](#).

Suppose also that we have estimated, using our formulas derived earlier, the values of the one-period state-contingent claims as follows:

$$\begin{array}{c} \text{Tomorrow} \\ \begin{array}{cc} 1 & 2 \end{array} \\ \text{Today} \quad \begin{array}{c} 1 \\ 2 \end{array} \quad \left[\begin{array}{cc} 0.54 & 0.42 \\ 0.46 & 0.53 \end{array} \right] = \mathbf{q} \end{array}$$

where q_{11} ($=0.54$) is the price today of an Arrow–Debreu claim paying \$1 if state 1 (a boom) occurs tomorrow, given that we are in state 1 (boom) today. Similarly, q_{12} is the

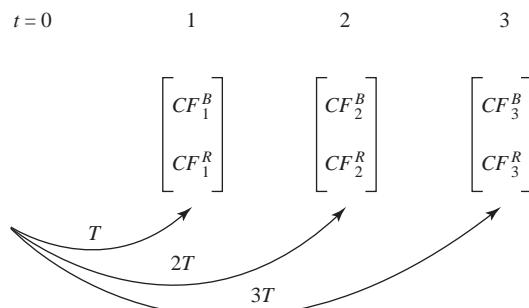


Figure 11.6

A multiperiod cash flow: two states of nature every period.

price today of an Arrow–Debreu claim paying \$1 if state 2 (recession) occurs tomorrow given that we are in state 1 today. Note that these prices differ because the distribution of the value of M tomorrow differs depending on the state today.

Now let us introduce our stationarity hypothesis: Suppose that \mathbf{q} , the matrix of values, is invariant through time.⁹ That is, the same two states of nature describe the possible futures at all future dates, and the contingent one-period prices remain the same. This allows us to interpret powers of the \mathbf{q} matrix, $\mathbf{q}^2, \mathbf{q}^3, \dots$ in a particularly useful way. Consider \mathbf{q}^2 (see also Figure 11.7):

$$\begin{aligned}\mathbf{q}^2 &= \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix} \cdot \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix} \\ &= \begin{bmatrix} (0.54)(0.54) + (0.42)(0.46) & (0.54)(0.42) + (0.42)(0.53) \\ (0.46)(0.54) + (0.53)(0.46) & (0.46)(0.42) + (0.53)(0.53) \end{bmatrix}\end{aligned}$$

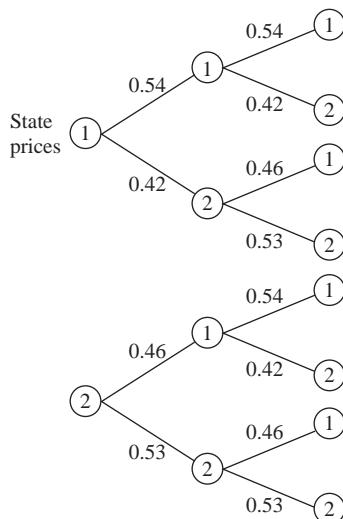


Figure 11.7
The evolution of state prices through time.

⁹ If this were not the case, the approach in Figure 11.7 would carry on provided we would be able to compute forward Arrow–Debreu prices. In other words, the Arrow–Debreu matrix would change from date to date, and it would have to be time-indexed. Mathematically, the procedure described would carry over, but the information requirement would, of course, be substantially larger.

Table 11.9: State-contingent cash flows

$t = 0$	1	2	3
state 1	$\begin{bmatrix} 42 \\ 65 \end{bmatrix}$	$\begin{bmatrix} 48 \\ 73 \end{bmatrix}$	$\begin{bmatrix} 60 \\ 58 \end{bmatrix}$
state 2			

Note that there are two ways to be in state 1 two periods from now, given that we are in state 1 today. Therefore, the price today of \$1, if state 1 occurs in two periods, given we are in state 1 today is:

$$\underbrace{(0.54)(0.54)}_{\text{value of } \$1 \text{ in 2 periods if state 1 occurs and the intermediate state is 1}} + \underbrace{(0.42)(0.46)}_{\text{value of } \$1 \text{ in 2 periods if state 1 occurs and the intermediate state is 2}}$$

Similarly, $q_{22}^2 = (0.46)(0.42) + (0.53)(0.53)$ is the price today, if today's state is 2, of \$1 contingent on state 2 occurring in two periods. In general, for powers N of the matrix \mathbf{q} , we have the following interpretation for q_{ij}^N : given that we are in state i today, it gives the price today of \$1, contingent on state j occurring in N periods. Of course, if we hypothesized three states, then the Arrow–Debreu matrices would be 3×3 and so forth.

How can this information be used in a “capital budgeting” problem? First we must estimate the cash flows. Suppose they are as outlined in [Table 11.9](#).

Then the present value (PV) of the cash flows, contingent on state 1 or state 2, are given by:

$$\begin{aligned} PV &= \begin{bmatrix} PV_1 \\ PV_2 \end{bmatrix} \\ &= \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix} \begin{bmatrix} 42 \\ 65 \end{bmatrix} + \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix}^2 \begin{bmatrix} 48 \\ 73 \end{bmatrix} + \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix}^3 \begin{bmatrix} 60 \\ 58 \end{bmatrix} \\ &= \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix} \begin{bmatrix} 42 \\ 65 \end{bmatrix} + \begin{bmatrix} 0.4848 & 0.4494 \\ 0.4922 & 0.4741 \end{bmatrix} \begin{bmatrix} 48 \\ 73 \end{bmatrix} + \begin{bmatrix} 0.4685 & 0.4418 \\ 0.4839 & 0.4580 \end{bmatrix} \begin{bmatrix} 60 \\ 58 \end{bmatrix} \\ &= \begin{bmatrix} 49.98 \\ 53.77 \end{bmatrix} + \begin{bmatrix} 56.07 \\ 58.23 \end{bmatrix} + \begin{bmatrix} 53.74 \\ 55.59 \end{bmatrix} = \begin{bmatrix} 159.79 \\ 167.59 \end{bmatrix}. \end{aligned}$$

This procedure can be expanded to include as many states of nature as one may wish to define. This amounts to choosing as fine a partition of the range of possible values of M as one wishes to choose. It makes no sense to construct a finer partition, however, if we have no real basis for estimating different cash flows in those states. For most practical problems,

three or four states are probably sufficient. But an advantage of this method is that it forces one to think carefully about what a project cash flow will be in each state, and what the relevant states, in fact, are.

One may wonder whether this methodology implicitly assumes that the states are equally probable. That is not the case. Although the probabilities, which would reflect the likelihood of the value of the market portfolio M lying in the various intervals, are not explicit, they are built into the prices of the state-contingent claims.

We close this chapter by suggesting a way to tie the approach proposed here with our previous work in this chapter. Risk-free cash flows are special (degenerate) examples of risky cash flows. It is thus easy to use the method of this section to price risk-free flows. The comparison with the results obtained with the method of [Section 11.3](#) then provides a useful check of the appropriateness of the assumptions made in the present context.

Consider our earlier example with Arrow–Debreu prices given by:

$$\begin{array}{cc} & \begin{matrix} 1 & 2 \end{matrix} \\ \text{State 1} & \left[\begin{matrix} 0.54 & 0.42 \end{matrix} \right] \\ \text{State 2} & \left[\begin{matrix} 0.46 & 0.53 \end{matrix} \right] \end{array}$$

If we are in state 1 today, the price of \$1 in each state tomorrow (i.e., a risk-free cash flow tomorrow of \$1) is $0.54 + 0.42 = 0.96$. This implies a risk-free rate of:

$$(1 + r_f^1) = \frac{1.00}{0.96} = 1.0416 \text{ or } 4.16\%.$$

To put it differently, $0.54 + 0.42 = 0.96$ is the price of a one-period discount bond paying \$1 in one period, given that we are in state 1 today. More generally, we would evaluate the following risk-free cash flow as:

$$t = 0 \quad \begin{matrix} 1 & 2 & 3 \\ 100 & 100 & 100 \end{matrix}$$

$$\begin{aligned} PV &= \begin{bmatrix} PV_1 \\ PV_2 \end{bmatrix} \\ &= \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \end{bmatrix} + \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix}^2 \begin{bmatrix} 100 \\ 100 \end{bmatrix} + \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix}^3 \begin{bmatrix} 100 \\ 100 \end{bmatrix} \\ &= \begin{bmatrix} 0.54 & 0.42 \\ 0.46 & 0.53 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \end{bmatrix} + \begin{bmatrix} 0.4848 & 0.4494 \\ 0.4922 & 0.4741 \end{bmatrix}^2 \begin{bmatrix} 100 \\ 100 \end{bmatrix} + \begin{bmatrix} 0.4685 & 0.4418 \\ 0.4839 & 0.4580 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \end{bmatrix} \end{aligned}$$

So

$$\begin{aligned}
 PV_1 &= [0.54 + 0.42]100 + [0.4848 + 0.4494]100 + [0.4685 + 0.4418]100 \\
 &= [0.96]100 + [0.9342]100 + [0.9103]100 \\
 &= 280.45
 \end{aligned}$$

where $[0.96]$ = price of a one-period discount bond given state 1 today, $[0.9342]$ = price of a two-period discount bond given state 1 today, $[0.9103]$ = price of a three-period discount bond given state 1 today. The PV given state 2 is computed analogously. Now this provides us with a **verification test**: if the price of a discount bond using this method does not coincide with the prices using the approach developed in [Section 11.3](#) (which relies on quoted coupon bond prices), then this must mean that our states are not well defined or numerous enough or that the assumptions of the option pricing formulas used to compute Arrow–Debreu prices are inadequate.

11.9 Conclusions

This chapter has served two main purposes. First, it has provided us with a platform to think more in depth about the all-important notion of market completeness.

Our demonstration that, in principle, a portfolio of simple calls and puts written on the market portfolio might suffice to reach a complete market structure suggests that the “Holy Grail” may not be totally out of reach. Caution must be exercised, however, in interpreting the necessary assumptions. Can we indeed assume that the market portfolio—and what do we mean by the latter—is an adequate reflection of all the economically relevant states of nature? And the time dimension of market completeness should not be forgotten. The most relevant state of nature for a Swiss resident of 40 years of age may be the possibility of a period of prolonged depression with high unemployment in Switzerland 25 years from now (when he nears retirement). Now extreme aggregate economic conditions would certainly be reflected in the Swiss Market Index (SMI), but options with 20-year maturities are not customarily traded. Is it because of a lack of demand (possibly meaning that our assumption as to the most relevant state is not borne out), or because the structure of the financial industry is such that the supply of securities for long horizons is deficient?¹⁰

The second part of the chapter discussed how Arrow–Debreu prices can be extracted from option prices (in the case where the relevant option is actively traded) or option pricing

¹⁰ A forceful statement in support of a similar claim is found in [Shiller \(1993\)](#) (see also the conclusions to Chapter 1). For the particular example discussed here, it may be argued that shorting the SMI (Swiss Market Index) would provide the appropriate hedge. Is it conceivable to take a short SMI position with a 20-year horizon?

formulas (in the case where they are not). This discussion helps make Arrow–Debreu securities a less abstract concept. In fact, in specific cases the detailed procedure is fully operational and may indeed be the wiser route to evaluating risky cash flows. The key hypotheses are similar to those we have just discussed: the relevant states of nature are adequately distinguished by the market portfolio, a hypothesis that may be deemed appropriate if the context is limited to the valuation of risky cash flows. Moreover, in the case where options are not traded, the quality of the extracted Arrow–Debreu prices depends on the appropriateness of the various hypotheses embedded in the option pricing formulas to which one has recourse. This issue has been abundantly discussed in the relevant literature.

References

- Banz, R., Miller, M., 1978. Prices for state-contingent claims: some estimates and applications. *J. Bus.* 51, 653–672.
- Breeden, D., Litzenberger, R.H., 1978. Prices of state-contingent claims implicit in option prices. *J. Bus.* 51, 621–651.
- Pirkner, C.D., Weigend, A.S., Zimmermann, H., 1999. Extracting Risk-Neutral Densities from Option Prices Using Mixture Binomial Trees. University of St-Gallen (Mimeoedged).
- Ross, S., 1976. Options and efficiency. *Q. J. Econ.* 90, 75–89.
- Shiller, R.J., 1993. Macro Markets-Creating Institutions for Managing Society's Largest Economic Risks. Clarendon Press, Oxford.

Appendix 11.1: Forward Prices and Forward Rates

Forward prices and forward rates correspond to the prices of (rates of return earned by) securities to be issued in the future.

Let $_k f_\tau$ denote the (compounded) rate of return on a risk-free discount bond to be issued at a future date k and maturing at date $k + \tau$. These forward rates are defined by the equations:

$$\begin{aligned}(1 + r_1)(1 + _1 f_1) &= (1 + r_2)^2 \\(1 + r_1)(1 + _1 f_2)^2 &= (1 + r_3)^3 \\(1 + r_2)^2(1 + _2 f_1) &= (1 + r_3)^3, \text{ etc.}\end{aligned}$$

We emphasize that the forward rates are *implied* forward rates, in the sense that the corresponding contracts are typically not traded. However, it is feasible to *lock in* these forward rates—that is, to guarantee their availability in the future. Suppose we wished to lock in the 1-year forward rate 1 year from now. This amounts to creating a new security “synthetically” as a portfolio of existing securities, and it is accomplished by simply

Table 11.10: Locking in a forward rate

$t =$	0	1	2
Buy a 2-yr bond	−1000	65	1065
Sell short a 1-yr bond	+1000	−1060	−995
	0	−995	1065

undertaking a series of *long* and *short* transactions today. For example, take as given the implied discount bond prices of [Table 11.5](#) and consider the transactions in [Table 11.10](#).

The portfolio we have constructed has a zero cash flow at date 0, requires an investment of \$995 at date 1, and pays \$1065 at date 2. The gross return on the date 1 investment is

$$\frac{1065}{995} = 1.07035.$$

That this is exactly equal to the corresponding forward rate can be seen from the forward rate definition:

$$1 + {}_1f_2 = \frac{(1+r_2)^2}{(1+r_1)} = \frac{(1.065163)^2}{1.06} = 1.07035.$$

Let us scale back the previous transactions to create a \$1000 payoff for the forward security. This amounts to multiplying all of the indicated transactions by $1000/1065 = 0.939$.

This price (\$934.34) is the no arbitrage price of this forward bond—no arbitrage in the sense that if there were any other contract calling for the delivery of such a bond at a price different from \$934.34, an arbitrage opportunity would exist (see [Table 11.11](#)).¹¹

Table 11.11: Creating a \$1000 payoff

$t =$	0	1	2
Buy 0.939 × 2-yr bonds	−939	61.0	1000
Sell short 0.939 × 1-yr bonds	+939	−995.34	
	0	−934.34	1000

¹¹ The approach of this section can, of course, be generalized to more distant forward rates.

The Martingale Measure: Part I

Chapter Outline

12.1 Introduction	361
12.2 The Setting and the Intuition	362
12.3 Notation, Definitions, and Basic Results	364
12.4 Uniqueness	369
12.5 Incompleteness	372
12.6 Equilibrium and No Arbitrage Opportunities	375
12.7 Application: Maximizing the Expected Utility of Terminal Wealth	377
12.7.1 Portfolio Investment and Risk-Neutral Probabilities	377
12.7.2 Solving the Portfolio Problem	380
12.7.3 A Numerical Example	381
12.8 Conclusions	383
References	384
Appendix 12.1 Finding the Stock and Bond Economy That Is Directly Analogous to the Arrow–Debreu Economy in Which Only State Claims Are Traded	384
Appendix 12.2 Proof of the Second Part of Proposition 12.6	386

12.1 Introduction

We have already introduced the concept of risk-neutral valuation within the specialized contexts of Arrow-Debreu pricing (Chapter 9) and the CCAPM (Chapter 10). In the present chapter we revisit it from the perspective of arbitrage pricing.

As we will shortly see, the theory of risk-neutral valuation - also referred to Martingale pricing - is actually a theory based on preference-free pure arbitrage principles.¹ That is, it is free of the structural assumptions on preferences, expectations and endowments that make the CAPM and the CCAPM so restrictive. In this respect the present chapter will illustrate how far one can go in pricing financial assets while abstracting from the usual structural assumptions.

¹ The theory of risk-neutral valuation was first developed by [Harrison and Kreps \(1979\)](#). [Pliska \(1997\)](#) provides an excellent review of the notion in discrete time. This chapter is based on his presentation. For historical reasons the risk-neutral probability measure is also referred to as the Martingale measure.

Recall that risk-neutral valuation offers a unique perspective on the asset valuation problem. Rather than modify the denominator—the discount factor—to take account of the risky nature of a cash flow to be valued, or the numerator, by transforming the expected cash flows into their certainty equivalent, risk-neutral valuation simply corrects the *probabilities* with respect to which the expectation of the future cash flows is taken. This is done in such a way that discounting at the risk-free rate is legitimate. It is thus a procedure by which an asset valuation problem is transformed into one in which the asset's expected cash flow, computed now with respect to a new set of risk-neutral probabilities, can be discounted at the risk-free rate. The risk-neutral valuation methodology thus places an arbitrary valuation problem into a context in which all fairly priced assets earn the risk-free rate.

Risk-neutral probability distributions naturally assume a variety of specialized forms when restricted to the specific settings of Chapters 9 and 10. Harking back to Chapter 9, we first illustrate them in the context of the well-understood, finite time Arrow-Debreu complete markets setting, arriving at the same results as Section 9.6 but from a different angle. This strategy serves to clarify the very tight relationship between Arrow-Debreu pricing and Martingale pricing despite the apparent differences in terminology and perspective. Chapter 13 focuses on a similar set of issues but within the dynamic context of the CCAPM.

12.2 The Setting and the Intuition

Our setting for preliminary discussion is the particularly simple one with which we are now long familiar. There are two dates, $t = 0$ and $t = 1$. At date $t = 1$, any one of $j = 1, 2, \dots, J$ possible states of nature can be realized; denote the j th state by θ_j and its objective probability by π_j . We assume $\pi_j > 0$ for all θ_j .

Securities are competitively traded in this economy. There is a risk-free security that pays a fixed return r_f ; its period t price is denoted by $q^b(t)$. We sometimes assume $q^b(0) = 1$, with its price at date 1 given by $q^b(1) \equiv q^b(\theta_j, 1) = (1 + r_f)$, for all states θ_j . Since the date 1 price of the security is $(1 + r_f)$ in any state, we can as well drop the first argument in the pricing function indicating the state in which the security is valued.²

Also traded are N fundamental risky securities, indexed $i = 1, 2, \dots, N$, which we think of as stocks.³ The period $t = 0$ price of the i th such security is represented as $q_i^e(0)$. In period $t = 1$ its contingent payoff, given that state θ_j is realized, is given by $q_i^e(\theta_j, 1)$.⁴ It is also

² In this chapter, it will be useful for the clarity of exposition to alter some of our previous notational conventions. One of the reasons is that we will want, symmetrically for all assets, to distinguish between their price at date 0 and their price at date 1 under any given state θ_j .

³ Fundamental securities are linearly independent.

⁴ In the notation of Chapter 9, $q_i^e(\theta_j, 1)$ is the cash flow associated with security i if state θ_j is realized, $CF^i(\theta_j)$.

assumed that investors may hold any linear combination of the fundamental risk-free and risky securities. No assumption is made, however, regarding the number of linearly independent securities vis-à-vis the number of states of nature: The securities market may or may not be complete. Neither is there any mention of agents' preferences. Otherwise the setting is standard Arrow–Debreu. Let \mathbf{S} denote the set of all fundamental securities, the stocks and the bond, and linear combinations thereof.

For this setting, the existence of a set of risk-neutral probabilities or, in more customary usage, a risk-neutral probability measure, effectively means the existence of a set of state probabilities, $\pi_j^{RN} > 0$, $j = 1, 2, \dots, J$ such that for each and every fundamental security $i = 1, 2, \dots, N$

$$q_i^e(0) = \frac{1}{(1 + r_f)} E\pi^{RN}(q_i^e(\theta, 1)) = \frac{1}{(1 + r_f)} \sum_{j=1}^J \pi_j^{RN} q_i^e(\theta_j, 1) \quad (12.1)$$

where the analogous relationship automatically holds for the risk-free security.

To gain some intuition as to what might be necessary, at a minimum, to guarantee the existence of such probabilities, first observe that in our setting the π_j^{RN} represent strictly positive numbers that must satisfy a large system of equations of the form

$$q_i^e(0) = \pi_1^{RN} \left(\frac{q_i^e(\theta_1, 1)}{1 + r_f} \right) + \dots + \pi_j^{RN} \left(\frac{q_i^e(\theta_J, 1)}{1 + r_f} \right), \quad i = 1, 2, \dots, N, \quad (12.2)$$

together with the requirement that $\pi_j^{RN} > 0$ for all j and $\sum_{j=1}^J \pi_j^{RN} = 1$.⁵

Such a system most certainly will not have a solution if there exist two fundamental securities, s and k , with the same $t = 0$ price, $q_s^e(0) = q_k^e(0)$, for which one of them, say k , pays as much as s in every state, and strictly more in at least one state; in other words,

$$q_k^e(\theta_j, 1) \geq q_s^e(\theta_j, 1) \quad \text{for all } j, \quad \text{and} \quad q_k^e(\theta_{\hat{j}}, 1) > q_s^e(\theta_{\hat{j}}, 1) \quad (12.3)$$

for at least one $j = \hat{j}$. Eq. (12.2) corresponding to securities s and k would, for any set $\{\pi_j^{RN}; j = 1, 2, \dots, N\}$, have the same left-hand sides, yet different right-hand sides, implying no solution to the system. But two such securities cannot themselves be consistently priced because, together, they constitute an *arbitrage opportunity*: Short one unit of security s , long one unit of security k , and pocket the difference $q_k^e(\theta_{\hat{j}}, 1) - q_s^e(\theta_{\hat{j}}, 1) > 0$ if state \hat{j} occurs; replicate the transaction many times over. These remarks suggest, therefore, that the existence of a risk-neutral measure is, in some intimate way, related to the absence of arbitrage opportunities in the financial markets. This is, in fact, the case, but first some notation, definitions, and examples are in order.

⁵ Compare this system of equations with those considered in Section 11.2 when extracting Arrow–Debreu prices from a complete set of prices for complex securities.

12.3 Notation, Definitions, and Basic Results

Consider a portfolio, P , composed of n_p^b risk-free bonds and n_p^i units of risky security i , $i = 1, 2, \dots, N$. No restrictions will be placed on n_p^b, n_p^i : Short sales of these assets are permitted; they can, therefore, take negative values, and fractional share holdings are acceptable. The value of this portfolio at $t = 0$, $V_P(0)$, is given by

$$V_P(0) = n_p^b q^b(0) + \sum_{i=1}^N n_p^i q_i^e(0), \quad (12.4)$$

while its value at $t = 1$, given that state θ_j , is realized is

$$V_P(\theta_j, 1) = n_p^b q^b(1) + \sum_{i=1}^N n_p^i q_i^e(\theta_j, 1). \quad (12.5)$$

With this notation we are now in a position to define our basic concepts.

Definition 12.1 A portfolio P in \mathbf{S} constitutes an arbitrage opportunity provided the following conditions are satisfied:

- (i) $V_P(0) = 0$,
- (ii) $V_P(\theta_j, 1) \geq 0$, for all $j \in \{1, 2, \dots, J\}$,
- (iii) $V_P(\theta_j, 1) > 0$, for at least one $\hat{J} \in \{1, 2, \dots, J\}$.

This is the standard sense of an arbitrage opportunity: With no initial investment and no possible losses (thus no risk), a strictly positive profit can be made in at least one state. Our second crucial definition is [Definition 12.2](#).

Definition 12.2 A probability measure $\{\pi_j^{RN}\}_{j=1}^J$ defined on the set of states $\theta_j, j = 1, 2, \dots, J$, is said to be a risk-neutral probability measure if

- (i) $\pi_j^{RN} > 0$, for all $j = 1, 2, \dots, J$, and
- (ii) $q_i^e(0) = E_{\pi^{RN}} \left\{ \frac{\tilde{q}_i^e(\theta, 1)}{1 + r_f} \right\},$

for all fundamental risky securities $i = 1, 2, \dots, N$ in \mathbf{S} .

Table 12.1: Fundamental securities for example 12.1

Period $t = 0$ Prices	Period $t = 1$ Payoffs		
		θ_1	θ_2
$q^b(0): 1$	$q^b(1):$	1.1	1.1
$q^e(0): 4$	$q^e(\theta_j, 1):$	3	7

Table 12.2: Fundamental securities for example 12.2

Period $t = 0$ Prices	Period $t = 1$ Payoffs			
		θ_1	θ_2	θ_3
$q^b(0): 1$	$q^b(1):$	1.1	1.1	1.1
$q_1^e(0): 2$	$q_1^e(\theta_j, 1):$	3	2	1
$q_2^e(0): 3$	$q_2^e(\theta_j, 1):$	1	4	6

Both elements of this definition are crucial. Not only must each individual security be priced equal to the present value of its expected payoff, the latter computed using the risk-neutral probabilities (and thus it must also be true of portfolios of them), but these probabilities must also be strictly positive. To find them, if they exist, it is necessary only to solve the system of equations implied by part (ii) of Eq. (12.7) of the risk-neutral probability definition. We illustrate this idea in the Examples 12.1 through 12.4.

Example 12.1 There are two periods and two fundamental securities, a stock and a bond, with prices and payoffs presented in Table 12.1.

By the definition of a risk-neutral probability measure, it must be the case that simultaneously

$$\begin{aligned} 4 &= \pi_1^{RN} \left(\frac{3}{1.1} \right) + \pi_2^{RN} \left(\frac{7}{1.1} \right) \\ 1 &= \pi_1^{RN} + \pi_2^{RN} \end{aligned}$$

Solving this system of equations, we obtain $\pi_1^{RN} = 0.65$, $\pi_2^{RN} = 0.35$.

For future reference note that the fundamental securities in this example define a complete set of financial markets for this economy, and that there are clearly no arbitrage opportunities among them.

Example 12.2 Consider next an analogous economy with three possible states of nature and three securities, as found in Table 12.2.

The relevant system of equations is now

$$\begin{aligned} 2 &= \pi_1^{RN} \left(\frac{3}{1.1} \right) + \pi_2^{RN} \left(\frac{2}{1.1} \right) + \pi_3^{RN} \left(\frac{1}{1.1} \right) \\ 3 &= \pi_1^{RN} \left(\frac{1}{1.1} \right) + \pi_2^{RN} \left(\frac{4}{1.1} \right) + \pi_3^{RN} \left(\frac{6}{1.1} \right) \\ 1 &= \pi_1^{RN} + \pi_2^{RN} + \pi_3^{RN}. \end{aligned}$$

The solution to this set of equations,

$$\pi_1^{RN} = 0.3, \quad \pi_2^{RN} = 0.6, \quad \pi_3^{RN} = 0.1,$$

satisfies the requirements of a risk-neutral measure. By inspection, we again observe that this financial market is complete and that there are no arbitrage opportunities among the three securities.

Example 12.3 To see what happens when the financial markets are incomplete, consider the securities in [Table 12.3](#).

For this example the relevant system is

$$\begin{aligned} 2 &= \pi_1^{RN} \left(\frac{1}{1.1} \right) + \pi_2^{RN} \left(\frac{2}{1.1} \right) + \pi_3^{RN} \left(\frac{3}{1.1} \right) \\ 1 &= \pi_1^{RN} + \pi_2^{RN} + \pi_3^{RN} \end{aligned}$$

Because this system is underdetermined, there will be many solutions. Without loss of generality, first solve for π_2^{RN} and π_3^{RN} in terms of π_1^{RN} :

$$\begin{aligned} 2.2 - \pi_1^{RN} &= 2\pi_2^{RN} + 3\pi_3^{RN} \\ 1 = \pi_1^{RN} &= \pi_2^{RN} + \pi_3^{RN}, \end{aligned}$$

which yields the solution $\pi_3^{RN} = 0.2 + \pi_1^{RN}$ and $\pi_2^{RN} = 0.8 - 2\pi_1^{RN}$.

Table 12.3: Fundamental securities for example 12.3

Period $t = 0$ Prices	Period $t = 1$ Payoffs			
		θ_1	θ_2	θ_3
$q^b(0): 1$	$q^b(1):$	1.1	1.1	1.1
$q^e_1(0): 2$	$q^e_1(\theta_j, 1):$	1	2	3

In order for a triple $(\pi_1^{RN}, \pi_2^{RN}, \pi_3^{RN})$ to simultaneously solve this system of equations, while also satisfying the strict positivity requirement of risk-neutral probabilities, the following inequalities must hold:

$$\begin{aligned}\pi_1^{RN} &> 0 \\ \pi_2^{RN} &= 0.8 - 2\pi_1^{RN} > 0 \\ \pi_3^{RN} &= 0.2 + \pi_1^{RN} > 0\end{aligned}$$

By the second inequality $\pi_1^{RN} < 0.4$, and by the third $\pi_1^{RN} > -0.2$. In order that all probabilities be strictly positive, it must, therefore, be the case that

$$0 < \pi_1^{RN} < 0.4,$$

with π_2^{RN} and π_3^{RN} given by the indicated equalities.

In an incomplete market, therefore, there appear to be many risk-neutral probability sets: any triple $(\pi_1^{RN}, \pi_2^{RN}, \pi_3^{RN})$ where

$$(\pi_1^{RN}, \pi_2^{RN}, \pi_3^{RN}) \in \{(\lambda, 8 - 2\lambda, 0.2 + \lambda) : 0 < \lambda < 0.4\}$$

serves as a risk-neutral probability measure for this economy.

Example 12.4 Lastly, we may as well see what happens if the set of fundamental securities contains an arbitrage opportunity (see [Table 12.4](#)).

Any attempt to solve the system of equations defining the risk-neutral probabilities fails in this case. There is no solution. Notice also the implicit arbitrage opportunity: risky security 2 dominates a portfolio of one unit of the risk-free security and one unit of risky security 1, yet it costs less.

It is also possible to have a solution in the presence of arbitrage. In this case, however, at least one of the solution probabilities will be zero, disqualifying the set for the risk-neutral designation.

Table 12.4: Fundamental securities for example 12.4

Period $t = 0$ Prices	Period $t = 1$ Payoffs			
		θ_1	θ_2	θ_3
$q^b(0): 1$	$q^b(1):$	1.1	1.1	1.1
$q_1^e(0): 2$	$q_1^e(\theta_1, 1):$	2	3	1
$q_2^e(0): 2.5$	$q_2^e(\theta_1, 1):$	4	5	3

Together with our original intuition, these examples suggest that arbitrage opportunities are incompatible with the existence of a risk-neutral probability measure. This is the substance of the first main result.

Proposition 12.1 Consider the two-period setting described earlier in this chapter. Then there exists a risk-neutral probability measure on \mathbf{S} , if and only if there are no arbitrage opportunities among the fundamental securities.

Proposition 12.1 tells us that, provided financial markets are characterized by the absence of arbitrage opportunities, our ambition to use distorted, risk-neutral probabilities to compute expected cash flows and discount at the risk-free rate has some legitimacy! Note, however, that the proposition admits the possibility that there may be many such measures, as in [Example 12.3](#).

Proposition 12.1 also provides us, in principle, with a method for testing whether a set of fundamental securities contains an arbitrage opportunity. If the system of [Eq. \(12.7.ii\)](#) has no solution probability vector where all the terms are strictly positive, an arbitrage opportunity is present. Unless we are highly confident of the actual states of nature and the payoffs to the various fundamental securities in those states, however, this observation is of limited use. But even for a very large number of securities it is easy to check the solution vector computationally.

Although we have calculated the risk-neutral probabilities with respect to the prices and payoff of the fundamental securities only, the analogous relationship must hold for arbitrary portfolios in \mathbf{S} —all linear combinations of the fundamental securities—in the absence of arbitrage opportunities. This result is formalized in [Proposition 12.2](#).

Proposition 12.2 Suppose the set of securities \mathbf{S} is free of arbitrage opportunities. Then for any portfolio \hat{P} in \mathbf{S}

$$V_{\hat{P}}(0) = \frac{1}{(1 + r_f)} E_{\pi^{RN}} \tilde{V}_{\hat{P}}(\theta, 1) \quad (12.8)$$

for any risk-neutral probability measure π^{RN} on \mathbf{S} .

Proof Let \hat{P} be an arbitrary portfolio in \mathbf{S} , and let it be composed of $n_{\hat{P}}^b$ bonds and $n_{\hat{P}}^i$ shares of fundamental risky asset i . In the absence of arbitrage, \hat{P} must be priced equal to the value of its constituent securities. In other words,

$$V_{\hat{P}}(0) = n_{\hat{P}}^b q^b(0) + \sum_{i=1}^N n_{\hat{P}}^i q_i^e(0) = n_{\hat{P}}^b E_{\pi^{RN}} \left(\frac{q^b(1)}{1 + r_f} \right) + \sum_{i=1}^N n_{\hat{P}}^i E_{\pi^{RN}} \left(\frac{\tilde{q}_i^e(\theta, 1)}{1 + r_f} \right),$$

for any risk-neutral probability measure π^{RN} ,

$$= E_{\pi^{RN}} \left\{ \frac{n_P^b q^b(1) + \sum_{i=1}^N n_P^i \tilde{q}_i^e(\theta, 1)}{1 + r_f} \right\} = \frac{1}{(1 + r_f)} E_{\pi^{RN}}(\tilde{V}_{\hat{P}}(\theta, 1)).$$

Proposition 12.2 is merely a formalization of the obvious fact that if every security in the portfolio is priced equal to the present value, discounted at r_f , of its expected payoffs computed with respect to the risk-neutral probabilities, the same must be true of the portfolio itself. This follows from the linearity of the expectations operator and the fact that the portfolio is valued as the sum total of its constituent securities, which must be the case in the absence of arbitrage opportunities.

A multiplicity of risk-neutral measures on \mathbf{S} does not compromise this conclusion in any way, because each of them assigns the same value to the fundamental securities and thus to the portfolio itself via Eq. (12.8). For completeness, we note that a form of a converse to Proposition 12.2 is also valid.

Proposition 12.3 Consider an arbitrary period $t = 1$ payoff $\tilde{x}(\theta, 1)$ and let M represent the set of all risk-neutral probability measures on the set \mathbf{S} . Assume \mathbf{S} contains no arbitrage opportunities. If

$$\frac{1}{(1 + r_f)} E_{\pi^{RN}} \tilde{x}(\theta, 1) = \frac{1}{(1 + r_f)} E_{\hat{\pi}^{RN}} \tilde{x}(\theta, 1) \quad \text{for any } \pi^{RN}, \hat{\pi}^{RN} \in M,$$

then there exists a portfolio in \mathbf{S} with the same $t = 1$ payoff as $\tilde{x}(\theta, 1)$.

It would be good to be able to dispense with the complications attendant to multiple risk-neutral probability measures on \mathbf{S} . When this is possible it is the subject of Section 12.4.

12.4 Uniqueness

Examples 12.1 and 12.2 both possessed unique risk-neutral probability measures. They were also complete markets models. This illustrates an important general proposition.

Proposition 12.4 Consider a set of securities \mathbf{S} without arbitrage opportunities. Then \mathbf{S} is complete if and only if there exists exactly one risk-neutral probability measure.

Proof Let us prove one side of the proposition, as it is particularly revealing. Suppose \mathbf{S} is complete and there were two risk-neutral probability measures, $\{\pi_j^{RN}: j = 1, 2, \dots, J\}$ and $\{\vec{\pi}_j^{RN}: j = 1, 2, \dots, J\}$. Then there must be at least one state \hat{J} for which $\pi_{\hat{J}}^{RN} \neq \vec{\pi}_{\hat{J}}^{RN}$. Since the market is complete, one must be able to construct a portfolio P in \mathbf{S} such that

$$V_P(0) > 0, \quad \text{and} \quad \begin{cases} V_P(\theta_j, 1) = 0 & j \neq \hat{J} \\ V_P(\theta_{\hat{J}}, 1) = 1 & j = \hat{J} \end{cases}$$

This is simply the statement of the existence of an Arrow–Debreu security associated with $\theta_{\hat{J}}$.

But then $\{\pi_j^{RN}: j = 1, 2, \dots, J\}$ and $\{\vec{\pi}_j^{RN}: j = 1, 2, \dots, J\}$ cannot both be risk-neutral measures since, by [Proposition 12.2](#),

$$\begin{aligned} V_P(0) &= \frac{1}{(1 + r_f)} E_{\pi^{RN}} \tilde{V}_P(\theta, 1) = \frac{x_{\hat{J}}^{RN}}{1 + r_f} \\ &\neq \frac{\vec{\pi}_{\hat{J}}^{RN}}{(1 + r_f)} = \frac{1}{(1 + r_f)} E_{\vec{\pi}^{RN}} \tilde{V}_P(\theta, 1) \\ &= V_P(0), \quad \text{a contradiction.} \end{aligned}$$

Thus, there cannot be more than one risk-neutral probability measure in a complete market economy.

We omit a formal proof of the other side of the proposition. Informally, the idea is as follows: if the market is not complete, then the fundamental securities do not span the space. Hence, the system of [Eq. \(12.6\)](#) contains more unknowns than equations, yet they are all linearly independent (no arbitrage). There must be a multiplicity of solutions and hence a multiplicity of risk-neutral probability measures.

Concealed in the proof of [Proposition 12.4](#) is an important observation: the price of an Arrow–Debreu security that pays one unit of payoff if event $\theta_{\hat{J}}$ is realized and nothing otherwise must be $\frac{\pi_{\hat{J}}^{RN}}{1 + r_f}$, the present value of the corresponding risk-neutral probability.

In general,

$$q_j(0) = \frac{\pi_j^{RN}}{(1 + r_f)}$$

where $q_j(0)$ is the $t = 0$ price of a state claim paying 1 if and only if state θ_j was realized. Provided the financial market is complete, risk-neutral valuation is nothing more than

valuing an uncertain payoff in terms of the value of a replicating portfolio of Arrow–Debreu claims. Notice, however, that we thus identify the all-important Arrow–Debreu prices without having to impose any of the economic structure of Chapter 9; in particular, knowledge of the agents’ preferences is not required. This approach can be likened to describing the Arrow–Debreu pricing theory from the perspective of Proposition 11.2. It is possible, and less restrictive, to limit our inquiry to extracting Arrow–Debreu prices from the prices of a (complete) set of complex securities and proceed from there to price arbitrary cash flows. In the absence of further structure, nothing can be said, however, about the determinants of Arrow–Debreu prices (or risk-neutral probabilities).

Let us illustrate with the data of our second example. There we identified the unique risk-neutral measure to be:

$$\pi_1^{RN} = 0.3, \quad \pi_2^{RN} = 0.6, \quad \pi_3^{RN} = 0.1,$$

Together with $r_f = 0.1$, these values imply that the Arrow–Debreu security prices must be

$$q_1(0) = \frac{0.3}{1.1} = 0.27; \quad q_2(0) = \frac{0.6}{1.1} = 0.55; \quad q_3(0) = \frac{0.1}{1.1} = 0.09.$$

Conversely, given a set of Arrow–Debreu claims with strictly positive prices, we can generate the corresponding risk-neutral probabilities and the risk-free rate. As noted in earlier chapters, the period zero price of a risk-free security (one that pays one unit of the numeraire in every date $t = 1$ state) in this setting is given by

$$q_{r_f} = \sum_{j=1}^J q_j(0),$$

and thus

$$(1 + r_f) = \frac{1}{q_{r_f}} = \frac{1}{\sum_{j=1}^J q_j(0)}$$

We define the risk-neutral probabilities $\{\pi^{RN}(\theta)\}$ according to

$$\pi_j^{RN} = \frac{q_j(0)}{\sum_{j=1}^J q_j(0)} \tag{12.9}$$

Clearly, $\pi_j^{RN} > 0$ for each state j (since $q_j(0) > 0$ for every state) and, by construction $\sum_{j=1}^J \pi_j^{RN} = 1$. As a result, the set $\{\pi_j^{RN}\}$ qualifies as a risk-neutral probability measure.⁶

Referring now to the example developed in Section 9.3, let us recall that we had found a complete set of Arrow–Debreu prices to be $q_1(0) = 0.24$; $q_2(0) = 0.3$; this means, in turn, that the unique risk-neutral measure for the economy there described is

$$\pi_1^{RN} = \frac{0.24}{0.54} = 0.444, \quad \pi_2^{RN} = \frac{0.3}{0.54} = 0.556.$$

For complete markets we see that the relationship between strictly positively priced state claims and the risk-neutral probability measure is indeed an intimate one: each implies the other. Since, in the absence of arbitrage possibilities, there can exist only one set of state claims prices, and thus only one risk-neutral probability measure, [Proposition 12.4](#) is reconfirmed.

12.5 Incompleteness

What about the case in which \mathbf{S} is an incomplete set of securities? By [Proposition 12.4](#) there will be a multiplicity of risk-neutral probabilities, but these will all give the same valuation to elements of \mathbf{S} ([Proposition 12.2](#)). Consider, however, a time $t = 1$ bounded state-contingent payoff vector $\tilde{x}(\theta, 1)$ that does *not* coincide with the payoff to any portfolio in \mathbf{S} . By [Proposition 12.4](#), different risk-neutral probability measures will assign different values to this payoff: essentially, its price is not well defined. It is possible, however, to establish *arbitrage bounds* on the value of this claim. For any risk-neutral probability π^{RN} , defined on \mathbf{S} , consider the following quantities:

$$H_x = \inf \left\{ E_{\pi^{RN}} \left[\frac{\tilde{V}_P(\theta, 1)}{1 + r_f} \right] : V_P(\theta_j, 1) \geq x(\theta_j, 1), \forall j = 1, 2, \dots, J \text{ and } P \in \mathbf{S} \right\} \quad (12.10)$$

$$L_x = \sup \left\{ E_{\pi^{RN}} \left[\frac{\tilde{V}_P(\theta, 1)}{1 + r_f} \right] : V_P(\theta_j, 1) \leq x(\theta_j, 1), \forall j = 1, 2, \dots, J \text{ and } P \in \mathbf{S} \right\}$$

In these evaluations we don't care what risk-neutral measure is used because any one of them gives identical valuations for all portfolios in \mathbf{S} . Since, for some γ , $\gamma q^b(1) > x(\theta, 1)$,

⁶ Note that we also assumed at the same expression as (12.9) back in Chapter 9.

for all j , H_x is bounded above by $\gamma q^b(0)$, and hence is well defined (an analogous comment applies to L_x). The claim is that the no arbitrage price of x , $q^x(0)$ lies in the range

$$L_x \leq q^x(0) \leq H_x$$

To see why this must be so, suppose that $q^x(0) > H_x$ and let P^* be any portfolio in \mathbf{S} for which

$$\begin{aligned} q^x(0) &> V_{P^*}(0) > H_x, \text{ and} \\ V_{P^*}(\theta_j, 1) &\geq x(\theta_j, 1), \quad \text{for all } \theta_j, j = 1, 2, \dots, N. \end{aligned} \tag{12.11}$$

We know that such a P^* exists because the set

$$\mathbf{S}_x = \{P : P \in \mathbf{S}, V_P(\theta_j, 1) \geq x(\theta_j, 1), \text{ for all } j = 1, 2, \dots, J\}$$

is closed. Hence there is a \hat{P} in \mathbf{S}_x such that $E_{\pi^{RN}} \frac{\tilde{V}_{\hat{P}}(\theta, 1)}{(1 + r_f)} = H_x$. By the continuity of the expectations operator, we can find a $\lambda > 1$ such that $\lambda \hat{P}$ is in \mathbf{S}_x and⁷

$$q^x(0) > \frac{1}{1 + r_f} E_{\pi^{RN}} \tilde{V}_{\lambda \hat{P}}(\theta, 1) = \lambda \frac{1}{1 + r_f} E_{\pi^{RN}} \tilde{V}_{\hat{P}}(\theta, 1) = \lambda H_x > H_x.$$

Since $\lambda > 1$, for all j , $V_{\lambda \hat{P}}(\theta_j, 1) > V_{\hat{P}}(\theta_j, 1) \geq x(\theta_j, 1)$; let $P^* = \lambda \hat{P}$. Now the arbitrage argument: sell the security with title to the cash flow $x(\theta_j, 1)$, and buy the portfolio P^* . At time $t = 0$, you receive, $q^x(0) - V_{P^*}(0) > 0$, while at time $t = 1$ the cash flow from the portfolio, by Eq. (12.11), fully covers the obligation under the short sale in every state. In other words, there is an arbitrage opportunity. An analogous argument demonstrates that $L_x \leq q^x(0)$.

In some cases it is readily possible to solve for these bounds.

Example 12.5 Revisit, for example, our earlier Example 12.3, and consider the payoff

$$\overline{x(\theta_j, 1)} : \left| \begin{array}{c|c|c} \theta_1 & \theta_2 & \theta_3 \\ \hline 0 & 0 & 1 \end{array} \right|$$

This security is most surely not in the span of the securities $(1.1, 1.1, 1.1)$ and $(1, 2, 3)$, a fact that can be confirmed by observing that the system of equations implied by equating

$$(0, 0, 1) = a(1.1, 1.1, 1.1) + b(1, 2, 3),$$

⁷ By $\lambda \hat{P}$ we mean a portfolio with constituent bonds and stocks in the proportions $(\lambda n_{\hat{P}}^b, \lambda n_{\hat{P}}^i)$.

in other words, the system:

$$\begin{aligned} 0 &= 1.1a + b \\ 0 &= 1.1a + 2b \\ 1 &= 1.1a + 3b \end{aligned}$$

has no solution. But any portfolio in \mathbf{S} can be expressed as a linear combination of $(1, 1, 1)$, $(1, 2, 3)$ and thus must be of the form

$$a(1, 1, 1) + b(1, 2, 3) = (a(1), a(1) + 2b, a(1) + 3b)$$

for some a, b real numbers.

We also know that in computing H_x, L_x , any risk-neutral measure can be employed. Recall that we had identified the solution of [Example 12.3](#) to be

$$(\pi_1^{RN}, \pi_2^{RN}, \pi_3^{RN}) \in \{(\lambda, 0.8 - 2\lambda, 0.2 + \lambda) : 0 < \lambda < 0.4\}$$

Without loss of generality, choose $\lambda = 0.2$; thus

$$(\pi_1^{RN}, \pi_2^{RN}, \pi_3^{RN}) = (0.2, 0.4, 0.4).$$

For any choice of a, b (thereby defining a $\tilde{V}_P(\theta; 1)$)

$$\begin{aligned} E_{\pi^{RN}} \left[\frac{\tilde{V}_P(\theta; 1)}{(1 + r_f)} \right] &= \frac{0.2\{(1.1)a + b\} + 0.4\{(1.1)a + 2b\} + 0.4\{(1.1)a + 3b\}}{1.1} \\ &= \frac{(1.1)a + (2.2)b}{1.1} = a + 2b. \end{aligned}$$

Thus,

$$H_x = \inf_{a, b \in R} \{(a + 2b) : a(1.1) + b \geq 0, a(1.1) + 2b \geq 0, \text{ and } a(1.1) + 3b \geq 1\}$$

Similarly,

$$L_x = \sup_{a, b \in R} \{(a + 2b) : a(1.1) + b \leq 0, a(1.1) + 2b \leq 0, a(1.1) + 3b \leq 1\}$$

Because the respective sets of admissible pairs are closed in R^2 , we can replace inf and sup by, respectively, min and max.

Solving for H_x, L_x thus amounts to solving small linear programs. The solutions, obtained via MATLAB, are detailed in [Table 12.5](#).

Table 12.5: Solutions for H_x and L_x

	H_x	L_x
a^*	-0.4545	-1.8182
b^*	0.5	1
H_x	0.5455	
L_x		0.1818

Table 12.6: The exchange economy of section 9.3—endowments and preferences

	Endowments			Preferences
	$t = 0$	$t = 1$		
Agent 1	10	1	2	$U^1(c_0, c_1) = \frac{1}{2}c_0^1 + 0.9(\frac{1}{3}\ln(c_1^1) + \frac{2}{3}\ln(c_2^1))$
Agent 2	5	4	6	$U^2(c_0, c_1) = \frac{1}{2}c_0^2 + 0.9(\frac{1}{3}\ln(c_1^2) + \frac{2}{3}\ln(c_2^2))$

The value of the security (state claim), we may conclude, lies in the interval (0.1818, 0.5455).

Before turning to the applications, there is one additional point of clarification.

12.6 Equilibrium and No Arbitrage Opportunities

Thus far we have made no reference to financial equilibrium, in the sense discussed in earlier chapters. Clearly, equilibrium implies no arbitrage opportunities: The presence of an arbitrage opportunity will induce investors to assume arbitrarily large short and long positions, which is inconsistent with the existence of equilibrium. The converse is also clearly not true. It could well be, in some specific market, that supply exceeds demand or, conversely, without this situation opening up an arbitrage opportunity in the strict sense understood in this chapter. In what follows the attempt is made to convey the sense of risk-neutral valuation as an equilibrium phenomenon.

To illustrate, let us return to the first example in Chapter 9. The basic data of that Arrow–Debreu equilibrium is provided in [Table 12.6](#) and the $t = 0$ corresponding equilibrium state prices are $q_1(0) = 0.24$ and $q_2(0) = 0.30$. In this case the risk-neutral probabilities are

$$\pi_1^{RN} = \frac{0.24}{0.54}, \quad \text{and} \quad \pi_2^{RN} = \frac{0.30}{0.54}.$$

Suppose a stock were traded where $q^e(\theta_1, 1) = 1$, and $q^e(\theta_2, 1) = 3$. By risk-neutral valuation (or equivalently, using Arrow–Debreu prices), its period $t = 0$ price must be

$$q^e(0) = 0.54 \left[\frac{0.24}{0.54}(1) + \frac{0.30}{0.54}(3) \right] = 1.14;$$

the price of the risk-free security is $q^b(0) = 0.54$.

Verifying this calculation is a bit tricky because, in the original equilibrium, this stock was not traded. Introducing such assets requires us to decide what the original endowments must be, that is, who owns what in period 0. We cannot just add the stock arbitrarily, as the wealth levels of the agents would change as a result, and, in general, this would alter the state prices, risk-neutral probabilities, and all subsequent valuations. The solution of this problem is to compute the equilibrium for a similar economy in which the two agents have the same preferences and in which the only traded assets are this stock and a bond. Furthermore, the initial endowments of these instruments must be such as to guarantee the same period $t = 0$ and $t = 1$ net endowment allocations as in the first equilibrium.

Let \hat{n}_e^i, \hat{n}_b^i denote, respectively, the initial endowments of the equity and debt securities of agent i , $i = 1, 2$. The equivalence noted previously is accomplished as outlined in [Table 12.7](#) (see [Appendix 12.1](#)).

A straightforward computation of the equilibrium prices yields the same $q^e(0) = 1.14$, and $q^b(0) = 0.54$ as predicted by risk-neutral valuation.

We conclude this section with one additional remark. Suppose one of the two agents were risk neutral; without loss of generality let this be agent 1. Under the original endowment scheme, his problem becomes:

$$\begin{aligned} & \max(10 + 1q_1(0) + 2q_2(0) - c_1^1 q_1(0) - c_2^1 q_2(0)) + 0.9 \left(\frac{1}{3} c_1^1 + \frac{2}{3} c_2^1 \right) \\ \text{s.t. } & c_1^1 q_1(0) + c_2^1 q_2(0) \leq 10 + q_1(0) + 2q_2(0) \end{aligned}$$

Table 12.7: Initial holdings of equity and debt achieving equivalence with Arrow–Debreu equilibrium endowments

	$t = 0$		
	Consumption	\hat{n}_e^i	\hat{n}_b^i
Agent 1	10	$\frac{1}{2}$	$\frac{1}{2}$
Agent 2	5	1	3

The first-order conditions are

$$c_1^1: \quad q_1(0) = \frac{1}{3} \cdot 0.9$$

$$c_2^1: \quad q_2(0) = \frac{2}{3} \cdot 0.9$$

from which it follows that $\pi_1^{RN} = \frac{\frac{1}{3} \cdot 0.9}{0.9} = \frac{1}{3}$ while $\pi_2^{RN} = \frac{\frac{2}{3} \cdot 0.9}{0.9} = \frac{2}{3}$; that is, in equilibrium, the risk-neutral probabilities coincide with the true probabilities. This is the source of the term *risk-neutral probabilities*: if at least one agent is risk neutral, the risk-neutral probabilities and the true probabilities coincide.

We conclude from this example that risk-neutral valuation holds in equilibrium, as it must because equilibrium implies no arbitrage. The risk-neutral probabilities thus obtained, however, are to be uniquely identified with that equilibrium, and it is meaningful to use them only for valuing securities that are elements of the participants' original endowments.

12.7 Application: Maximizing the Expected Utility of Terminal Wealth

12.7.1 Portfolio Investment and Risk-Neutral Probabilities

Risk-neutral probabilities are intimately related to the *basis* or the set of fundamental securities in an economy. Under no arbitrage, given the prices of fundamental securities, we obtain a risk-neutral probability measure, and vice versa. This raises the possibility that it may be possible to formulate any problem in wealth allocation, for example, the classic consumption-savings problem, in the setting of risk-neutral valuation. In this section we consider a number of these connections.

The simplest portfolio allocation problem with which we have dealt involves an investor choosing a portfolio so as to maximize the expected utility of his period $t = 1$ (terminal) wealth (we retain, without loss of generality, the two-period framework). In our current notation, this problem takes the form: choose portfolio P , among all feasible portfolios (i.e., P must be composed of securities in \mathbf{S} and the date-0 value of this portfolio (its acquisition price) cannot exceed initial wealth) so as to maximize expected utility of terminal wealth, which corresponds to the date-1 value of P :

$$\max_{\{n_p^b, n_p^i, i = 1, 2, \dots, N\}} EU(\tilde{V}_P(\theta, 1)) \quad (12.12)$$

$$\text{s.t. } V_P(0) = V_0, P \in \mathbf{S},$$

where V_0 is the investor's initial wealth, $U(\cdot)$ is her period utility function, assumed to have the standard properties, and n_p^b, n_p^i , are the positions (not proportions, but units of indicated assets) in the risk-free asset and the risky asset $i = 1, 2, \dots, N$, respectively, defining portfolio P . It is not obvious that there should be a relationship between the solvability of this problem and the existence of a risk-neutral measure, but this is the case.

Proposition 12.5 If Eq. (12.12) has a solution, then there are no arbitrage opportunities in \mathbf{S} . Hence there exists a risk-neutral measure on \mathbf{S} .

Proof The idea is that an arbitrage opportunity is a costless way to endlessly improve upon the (presumed) optimum. So no optimum can exist. More formally, we prove the proposition by contradiction. Let $\hat{P} \in \mathbf{S}$ be a solution to Eq. (12.12), and let \hat{P} have the structure $\{n_{\hat{P}}^b, n_{\hat{P}}^i : i = 1, 2, \dots, N\}$. Assume also that there exists an arbitrage opportunity, in other words, a portfolio \tilde{P} , with structure $\{n_{\tilde{P}}^b, n_{\tilde{P}}^i : i = 1, 2, \dots, N\}$, such that $V_{\tilde{P}}(0) = 0$ and $E\tilde{V}_{\tilde{P}}(\theta, 1) > 0$. Consider the portfolio P^* with structure

$$\begin{aligned} &\{n_{P^*}^b, n_{P^*}^i : i = 1, 2, \dots, N\} \\ &n_{P^*}^b, n_{\hat{P}}^b + n_{\tilde{P}}^b \text{ and } n_{P^*}^i = n_{\hat{P}}^i + n_{\tilde{P}}^i, \quad i = 1, 2, \dots, N. \end{aligned}$$

P^* is still feasible for the agent, and it provides strictly more wealth in at least one state. Since $U(\cdot)$ is strictly increasing,

$$EU(\tilde{V}_{P^*}(\theta, 1)) > EU(\tilde{V}_{\hat{P}}(\theta, 1)).$$

This contradicts \hat{P} as a solution to Eq. (12.12). We conclude that there cannot exist any arbitrage opportunities and thus, by Proposition 12.1, a risk-neutral probability measure on \mathbf{S} must exist.

Proposition 12.5 informs us that arbitrage opportunities are incompatible with an optimal allocation—the allocation can always be improved upon by incorporating units of the arbitrage portfolio. More can be said. The solution to the agents' problem can, in fact, be used to identify the risk-neutral probabilities. To see this, let us first rewrite the objective function in Eq. (12.12) as follows:

$$\begin{aligned} &\max_{\{n_p^i : i = 1, 2, \dots, N\}} EU \left((1 + r_f) \left\{ V_0 - \sum_{i=1}^N n_p^i q_i^e(0) \right\} + \sum_{i=1}^N n_p^i q_i^e(\theta, 1) \right) \\ &= \max_{\{n_p^i : i = 1, 2, \dots, N\}} \sum_{j=1}^J \pi_j U \left((1 + r_f) \left\{ V_0 + \sum_{i=1}^N n_p^i \frac{q_i^e(\theta_j, 1)}{1 + r_f} - \sum_{i=1}^N n_p^i q_i^e(0) \right\} \right) \quad (12.13) \\ &= \max_{\{n_p^i : i = 1, 2, \dots, N\}} \sum_{j=1}^J \pi_j U \left((1 + r_f) \left\{ V_0 + \sum_{i=1}^N n_p^i \left(\frac{q_i^e(\theta_j, 1)}{1 + r_f} - q_i^e(0) \right) \right\} \right) \end{aligned}$$

The necessary and sufficient first-order conditions for this problem are of the form:

$$0 = \sum_{j=1}^J \pi_j U_1 \left((1 + r_f) \left\{ V_0 + \sum_{i=1}^N n_p^i \left(\frac{q_i^e(\theta_j, 1)}{1 + r_f} - q_i^e(0) \right) \right\} \right) (1 + r_f) \left[\frac{q_i^e(\theta_j, 1)}{1 + r_f} - q_i^e(0) \right] \quad (12.14)$$

Note that the quantity $\pi_j U_1(V_P^1(\theta_j, 1))(1 + r_f)$ is strictly positive because $\pi_j > 0$ and $U(\cdot)$ is strictly increasing. If we normalize these quantities, we can convert them into probabilities. Let us define

$$\check{\pi}_j = \frac{\pi_j U_1(V_P(\theta_j, 1))(1 + r_f)}{\sum_{j=1}^J \pi_j U_1(V_P(\theta_j, 1))(1 + r_f)} = \frac{\pi_j U_1(V_P(\theta_j, 1))}{\sum_{j=1}^J \pi_j U_1(V_P(\theta_j, 1))}, \quad j = 1, 2, \dots, J.$$

Since $\check{\pi}_j > 0, j = 1, 2, \dots, J$, $\sum_{j=1}^J \check{\pi}_j = 1$ and by (12.14)

$$q_i^e(0) = \sum_{j=1}^J \check{\pi}_j \frac{q_i^e(\theta_j, 1)}{1 + r_f};$$

these three properties establish the set $\{\check{\pi}_j : j = 1, 2, \dots, N\}$ as a set of risk-neutral probabilities.

We have just proved one-half of the following proposition:

Proposition 12.6 Let $\{n_{p^*}^b, n_{p^*}^i : i = 1, 2, \dots, N\}$ be the solution to the optimal portfolio problem (12.12). Then the set $\{n_j^* : j = 1, 2, \dots, J\}$, defined by

$$\pi_j^* = \frac{\pi_j U_1(V_{P^*}(\theta_j, 1))}{\sum_{j=1}^J \pi_j U_1(V_{P^*}(\theta_j, 1))}, \quad (12.15)$$

constitutes a risk-neutral probability measure on \mathbf{S} . Conversely, if there exists a risk-neutral probability measure $\{\pi_j^{RN} : j = 1, 2, \dots, J\}$ on \mathbf{S} , there must exist a concave, strictly increasing, differentiable utility function $U(\cdot)$ and an initial wealth V_0 for which Eq. (12.12) has a solution.

Proof We have proved the first part. The proof of the less important converse proposition is relegated to [Appendix 12.2](#).

12.7.2 Solving the Portfolio Problem

Now we can turn to solving Eq. (12.12). Since there is as much information in the risk-neutral probabilities as in the security prices, it should be possible to fashion a solution to Eq. (12.12) using that latter construct. Here we will choose to restrict our attention to the case in which the financial markets are complete.

In this case there exists exactly one risk-neutral measure, which we denote by $\{\pi_j^{RN}: j = 1, 2, \dots, N\}$. Since the solution to Eq. (12.12) will be a portfolio in \mathbf{S} that maximizes the date $t = 1$ expected utility of wealth, the solution procedure can be decomposed into a two-step process:

Step 1 Solve

$$\begin{aligned} & \max EU(\tilde{x}(\theta, 1)) \\ \text{s.t. } & E_{\pi^{RN}} \left(\frac{\tilde{x}(\theta, 1)}{1 + r_f} \right) = V_0 \end{aligned} \tag{12.16}$$

The solution to this problem identifies the feasible uncertain payoff that maximizes the agent's expected utility. But why is the constraint a perfect summary of feasibility? The constraint makes sense first because, under complete markets, every uncertain payoff lies in \mathbf{S} . Furthermore, in the absence of arbitrage opportunities, every payoff is valued at the present value of its expected payoff computed using the unique risk-neutral probability measure. The essence of the budget constraint is that a feasible payoff be affordable: that its price equals V_0 , the agent's initial wealth.

Step 2 Find the portfolio P in \mathbf{S} such that

$$V_P(\theta_j, 1) = x(\theta_j, 1), \quad j = 1, 2, \dots, J.$$

In step 2 we simply find the precise portfolio allocations of fundamental securities that give rise to the optimal uncertain payoff identified in step 1. The theory is all in step 1; in fact, we have used all of our major results thus far to write the constraint in the indicated form.

Now let us work out a problem, first abstractly and then by a numerical example.

Equation (12.16) of step 1 can be written as

$$\max E_{\pi} U(\tilde{x}(\theta, 1)) - \lambda \left[E_{\pi^{RN}} \left(\frac{\tilde{x}(\theta, 1)}{1 + r_f} \right) - V_0 \right] \tag{12.17}$$

where λ denotes the Lagrange multiplier and where we have made explicit the probability distributions with respect to which each of the expectations is being taken.

Equation (12.17) can be rewritten as

$$\max_x \sum_{j=1}^J \pi_j \left[U(x(\theta_j, 1)) - \lambda \frac{\pi_j^{RN}}{\pi_j} \frac{x(\theta_j, 1)}{(1 + r_f)} \right] + \lambda V_0. \quad (12.18)$$

The necessary first-order conditions, one equation for each state θ_j , are thus

$$U_1(x(\theta_j, 1)) = \frac{\lambda \pi_j^{RN}}{\pi_j(1 + r_f)}, \quad j = 1, 2, \dots, J. \quad (12.19)$$

from which the optimal asset payoffs may be obtained as per

$$x(\theta_j, 1) = U_1^{-1} \left(\frac{\lambda \pi_j^{RN}}{\pi_j(1 + r_f)} \right), \quad j = 1, 2, \dots, J. \quad (12.20)$$

with U_1^{-1} representing the inverse of the MU function.

The Lagrange multiplier λ is the remaining unknown. It must satisfy the budget constraint when Eq. (12.20) is substituted for the solution; that is, λ must satisfy

$$E_{\pi^{RN}} \left(\frac{1}{(1 + r_f)} U_1^{-1} \left(\frac{\lambda \pi_j^{RN}}{\pi_j(1 + r_f)} \right) \right) = V_0. \quad (12.21)$$

A value for λ that satisfies Eq. (12.21) may not exist. For all the standard utility functions that we have dealt with, $U(x) = \ln x$ or $\frac{x^{1-\gamma}}{1-\gamma}$ or e^{-vx} , however, it can be shown that such a λ will exist. Let $\hat{\lambda}$ solve Eq. (12.21); the optimal feasible contingent payoff is thus given by

$$x(\theta_j, 1) = U_1^{-1} \left(\frac{\hat{\lambda} \pi_j^{RN}}{\pi_j(1 + r_f)} \right) \quad (12.22)$$

(from Eq. (12.21)). Given this payoff, step 2 involves finding the portfolio of fundamental securities that will give rise to it. This is accomplished by solving the customary system of linear equations.

12.7.3 A Numerical Example

Let us choose a utility function from the familiar CRRA class, $U(x) = \frac{x^{1-\gamma}}{1-\gamma}$, and consider the market structure of Example 12.2. Markets are complete, and the unique risk-neutral probability measure is as noted.

Since $U_1(x) = x^{-\gamma}$, $U_1^{-1}(y) = y^{-\frac{1}{\gamma}}$, Eq. (12.20) reduces to

$$x(\theta_j, 1) = \left(\frac{\lambda \pi_j^{RN}}{\pi_j(1+r_f)} \right)^{-\frac{1}{\gamma}} \quad (12.23)$$

from which follows the counterpart to Eq. (12.21):

$$\sum_{j=1}^J \pi_j^{RN} \left(\frac{1}{(1+r_f)} \left(\frac{\lambda \pi_j^{RN}}{\pi_j(1+r_f)} \right)^{-\frac{1}{\gamma}} \right) = V_0.$$

Isolating λ gives

$$\hat{\lambda} = \left\{ \sum_{j=1}^J \pi_j^{RN} \left(\frac{1}{(1+r_f)} \left(\frac{\pi_j^{RN}}{\pi_j(1+r_f)} \right)^{-\frac{1}{\gamma}} \right) \right\}^\gamma V_0^{-\gamma}. \quad (12.24)$$

Let us consider some numbers: Assume $\gamma = 3$, $V_0 = 10$, and that (π_1, π_2, π_3) , the true probability distribution, takes on the value $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Refer to Example 12.2 where the risk-neutral probability distribution was found to be $(\pi_1^{RN}, \pi_2^{RN}, \pi_3^{RN}) = (0.3, 0.6, 0.1)$. Accordingly, from Eq. (12.24)

$$\begin{aligned} \hat{\lambda} &= 10^{-3} \left\{ 0.3 \left(\frac{1}{(1.1)} \left(\frac{0.3}{\left(\frac{1}{3}\right)(1.1)} \right)^{-\frac{1}{3}} \right) \right. \\ &\quad \left. + 0.6 \left(\frac{1}{(1.1)} \left(\frac{0.6}{\left(\frac{1}{3}\right)(1.1)} \right)^{-\frac{1}{3}} \right) + 0.1 \left(\frac{1}{(1.1)} \left(\frac{0.1}{\left(\frac{1}{3}\right)(1.1)} \right)^{-\frac{1}{3}} \right) \right\}^3 \\ \hat{\lambda} &= \left(\frac{1}{1000} \right) \{0.2916 + 0.4629 + 0.14018\}^3 = 0.0007161. \end{aligned}$$

The distribution of the state-contingent payoffs follows from Eq. (12.23):

$$x(\theta_j, 1) = \left(\frac{0.0007161 \pi_j^{RN}}{\pi_j(1+r_f)} \right)^{-\frac{1}{\gamma}} = \begin{cases} 11.951 & j = 1 \\ 9.485 & j = 2 \\ 17.236 & j = 3. \end{cases} \quad (12.25)$$

The final step is to convert this payoff to a portfolio structure via the identification:

$$(11.951, 11.485, 17.236) = n_p^b(1.1, 1.1, 1.1) + n_p^1(3, 2, 1) + n_p^2(1, 4, 6) \text{ or}$$

$$11.951 = 1.1n_p^b + 3n_p^1 + n_p^2$$

$$11.485 = 1.1n_p^b + 2n_p^1 + 4n_p^2$$

$$17.236 = 1.1n_p^b + n_p^1 + bn_p^2$$

The solution to this system of equations is

$$n_p^b = 97.08 \text{(invest a lot in the risk-free asset)}$$

$$n_p^1 = -28.192 \text{(short the first stock)}$$

$$n_p^2 = -10.225 \text{(also short the second stock)}$$

Lastly, we confirm that this portfolio is feasible:

Cost of portfolio = $97.08 + 2(-28.192) + 3(-10.225) = 10 = V_0$, the agent's initial wealth, as required.

Note the computational simplicity of this method: we need only solve a linear system of equations. Using more standard methods would result in a system of three nonlinear equations to solve. Analogous methods are also available to provide bounds in the case of market incompleteness.

12.8 Conclusions

Under the procedure of risk-neutral valuation, we construct a new probability distribution—the risk-neutral probabilities—under which all assets may be valued at their expected payoff discounted at the risk-free rate. More formally it would be said that we undertake a *transformation of measure* by which all assets are then expected to earn the risk-free rate. The key to our ability to find such a measure is that the financial markets exhibit no arbitrage opportunities.

Our setting was the standard Arrow–Debreu two-period equilibrium and we observed the intimate relationship between the risk-neutral probabilities and the relative prices of state claims. Here the practical applicability of the idea is limited. Applying these ideas to the real-world would, after all, require a denumeration of all future states of nature and the contingent payoffs to all securities in order to compute the relevant risk-neutral probabilities, something for which there would be no general agreement.

Even so, this particular way of approaching the optimal portfolio problem was shown to be a source of useful insights. In more restrictive settings, it is also practically powerful and, as noted in Chapter 11, lies behind all modern derivatives pricing.

References

- Harrison, M., Kreps, D., 1979. Martingales and multi-period securities market. *J. Econ. Theor.* 20, 381–408.
 Pliska, S.R., 1997. *Introduction to Mathematical Finance: Discrete Time Models*. Basil Blackwell, Malden, MA.

Appendix 12.1 Finding the Stock and Bond Economy That Is Directly Analogous to the Arrow–Debreu Economy in Which Only State Claims Are Traded

The Arrow–Debreu economy is summarized in [Table 12.6](#). We wish to price the stock and bond with the payoff structures in [Table A12.1](#).

In order for the economy in which the stock and bond are traded to be equivalent to the Arrow–Debreu economy where state claims are traded, we need the former to imply the same effective endowment structure. This is accomplished as shown on the next page

Agent 1: Let his endowments of the stock and bond be denoted by \hat{z}_1^e and \hat{z}_1^b , then,

$$\text{In state } \theta_1: \hat{z}_1^b + \hat{z}_1^e = 1$$

$$\text{In state } \theta_2: \hat{z}_1^b + 3\hat{z}_1^e = 2$$

$$\text{Solution: } \hat{z}_1^e + \hat{z}_1^b = \frac{1}{2} \text{ (half a share and half a bond)}$$

Agent 2: Let his endowments of the stock and bond be denoted by \hat{z}_2^e and \hat{z}_2^b , then,

$$\text{In state } \theta_1: \hat{z}_2^b + \hat{z}_2^e = 4$$

$$\text{In state } \theta_2: \hat{z}_2^b + 3\hat{z}_2^e = 6$$

$$\text{Solution: } \hat{z}_2^e = \hat{z}_2^b = 3$$

Table A12.1: Payoff structure

$t = 0$	$t = 1$	
	θ_1	θ_2
$-q_e(0)$	1	3
$-q_b(0)$	1	1

With these endowments the decision problems of the agent become:

Agent 1:

$$\max_{z_1^e + z_1^b} \frac{1}{2} \left(10 + \frac{1}{2} q^e + \frac{1}{2} q^b - z_1^b q^b \right) + 0.9 \left(\frac{1}{3} \ln(z_1^e + z_1^b) + \frac{2}{3} \ln(3z_1^e + z_1^b) \right)$$

Agent 2:

$$\max_{z_2^e + z_2^b} \frac{1}{2} (5 + q^e + 3q^b - z_2^e q^e - z_2^b q^b) + 0.9 \left(\frac{1}{3} \ln(z_2^e + z_2^b) + \frac{2}{3} \ln(3z_2^e + z_2^b) \right)$$

The FOCs are

$$z_1^e: \frac{1}{2} q^e = 0.9 \left(\frac{1}{3} \left(\frac{1}{z_1^e + z_1^b} \right) + \frac{2}{3} \left(\frac{1}{3z_1^e + z_1^b} \right) (3) \right)$$

$$z_1^b: \frac{1}{2} q^b = 0.9 \left(\frac{1}{3} \left(\frac{1}{z_1^e + z_1^b} \right) + \frac{2}{3} \left(\frac{1}{3z_1^e + z_1^b} \right) \right)$$

$$z_2^e: \frac{1}{2} q^e = 0.9 \left(\frac{1}{3} \left(\frac{1}{z_2^e + z_2^b} \right) + \frac{2}{3} \left(\frac{1}{3z_2^e + z_2^b} \right) (3) \right)$$

$$z_2^b: \frac{1}{2} q^b = 0.9 \left(\frac{1}{3} \left(\frac{1}{z_2^e + z_2^b} \right) + \frac{2}{3} \left(\frac{1}{3z_2^e + z_2^b} \right) \right)$$

Since these securities span the space and since the period 1 and period 2 endowments are the same, the real consumption allocations must be the same as in the Arrow–Debreu economy:

$$c_1^1 = c_1^2 = 2.5$$

$$c_2^1 = c_2^2 = 4$$

Thus,

$$q^e = 2(0.9) \left\{ \frac{1}{3} \left(\frac{1}{2.5} \right) + \frac{2}{3} \left(\frac{1}{4} \right) 3 \right\} = 1.14$$

$$q^b = 2(0.9) \left\{ \frac{1}{3} \left(\frac{1}{2.5} \right) + \frac{2}{3} \left(\frac{1}{4} \right) \right\} = 0.54,$$

as computed previously.

To compute the corresponding security holding, observe that:

Agent 1:

$$\begin{aligned} z_1^e + z_1^b &= 2.5 \Rightarrow z_1^e = 0.75 \\ 3z_1^e + z_1^b &= 4 \Rightarrow z_1^b = 1.75 \end{aligned}$$

Agent 2: (same holdings)

$$\begin{aligned} z_2^e &= 0.75 \\ z_2^b &= 1.75 \end{aligned}$$

Supply must equal demand in equilibrium:

$$\hat{z}_1^e + \hat{z}_2^e = \frac{1}{2} + 1 = 1.5 = z_1^e + z_2^e$$

$$\hat{z}_1^b + \hat{z}_2^b = \frac{1}{2} + 3 = 3.5 = z_1^b + z_2^b$$

The period zero consumptions are identical to the earlier calculation as well.

Appendix 12.2 Proof of the Second Part of Proposition 12.6

Define $\hat{U}(x, \theta_j) = x \left\{ \frac{\pi_j^{RN}}{\pi_j(1+r_f)} \right\}$, where $\{\pi_j: j = 1, 2, \dots, J\}$ are the true objective state probabilities. This is a state-dependent utility function that is linear in wealth. We will show that for this function, Eq. (12.13), indeed, has a solution. Consider an arbitrary allocation of wealth to the various fundamental assets $\{n_P^i: i = 1, 2, \dots, N\}$ and let P denote that portfolio. Fix the wealth at any level V_0 , arbitrary. We next compute the expected utility associated with this portfolio, taking advantage of representation (12.14):

$$\begin{aligned} E\hat{U}(\tilde{V}_P(\theta, 1)) &= E\hat{U}\left\{(1+r_f)\left[V_0 + \sum_{i=1}^N n_P^i \left(\frac{\tilde{q}_i^e(\theta, 1)}{(1+r_f)} - q_i^e(0)\right)\right]\right\} \\ &= \sum_{j=1}^J \pi_j (1+r_f) \left\{ V_0 + \sum_{i=1}^N n_P^i \left(\frac{q_i^e(\theta_j, 1)}{(1+r_f)} - q_i^e(0)\right) \right\} \frac{\pi_j^{RN}}{\pi_j(1+r_f)} \\ &= \sum_{j=1}^J \pi_j^{RN} \left\{ V_0 + \sum_{i=1}^N n_P^i \left(\frac{q_i^e(\theta_j, 1)}{(1+r_f)} - q_i^e(0)\right) \right\} \\ &= \sum_{j=1}^J \pi_j^{RN} V_0 + \sum_{j=1}^J \pi_j^{RN} \sum_{i=1}^N n_P^i \left(\frac{q_i^e(\theta_j, 1)}{(1+r_f)} - q_i^e(0)\right) \\ &= V_0 + \sum_{i=1}^N n_P^i \left(\sum_{j=1}^J \pi_j^{RN} \left(\frac{q_i^e(\theta_j, 1)}{(1+r_f)} - q_i^e(0)\right) \right) \\ &= V_0 \end{aligned}$$

in other words, with this utility function, every trading strategy has the same value. Thus, problem (12.12) has, trivially, a solution.

The Martingale Measure: Part II

Chapter Outline

13.1 Introduction	387
13.2 Discrete Time Infinite Horizon Economies: A CCAPM Setting	388
13.3 Risk-Neutral Pricing in the CCAPM	390
13.4 The Binomial Model of Derivatives Valuation	397
13.5 Continuous Time: An Introduction to the Black–Scholes Formula	407
13.6 Dybvig's Evaluation of Dynamic Trading Strategies	410
13.7 Conclusions	414
References	414
Appendix 13.1: Risk-Neutral Valuation When Discounting at the Term Structure of Multiperiod Discount Bond	414

13.1 Introduction

We continue our study of risk-neutral valuation, by extending the settings to one with many time periods. This will be accomplished in two very different ways. First, we revisit the concept in the CCAPM setting. Recall that this is a discrete time, general equilibrium framework: preferences and endowment processes must be specified, and no-trade prices computed. We will demonstrate that, here as well, assets may be priced equal to the present value, discounted at the risk-free rate of interest, of their expected payoffs when expectations are computed using the set of risk-neutral probabilities. We would expect this to be possible. The CCAPM is an equilibrium model (hence there are no-arbitrage opportunities and thus a set of risk-neutral probabilities must exist) with complete markets (hence this set is unique).

Second, we extend the idea to the partial equilibrium setting of equity derivatives (e.g., equity options) valuation. The key to derivatives pricing is to have an accurate model of the underlying price process. We hypothesize such a process (it is not derived from underlying fundamentals—preferences, endowments, etc.; rather, it is a pure statistical model) and demonstrates that, in the presence of *local* market completeness and *local* no-arbitrage situations, there exists a transformation of measure by which all derivatives

written on that asset may be priced equal to the present value, discounted at the risk-free rate, of their expected payoffs computed using this transformed measure.¹ The Black–Scholes formula, for example, may be derived in this way.

13.2 Discrete Time Infinite Horizon Economies: A CCAPM Setting

As in the previous chapter, time evolves according to $t = 0, 1, \dots, T, T + 1, \dots$. We retain the context of a single good endowment economy and presume the existence of a complete markets Arrow–Debreu financial structure. In period t , any one of N_t possible states, indexed by θ_t , may be realized.

We will assume that a period t event is characterized by two quantities:

- i. The actually occurring period t event as characterized by θ_t .
- ii. The unique history of events $(\theta_1, \theta_2, \dots, \theta_{t-1})$ that precedes it.

Requirement (ii), in particular, suggests an evolution of uncertainty similar to that of a tree structure in which the branches never join (two events always have distinct prior histories). Although this is a stronger assumption than what underlies the CCAPM, it will allow us to avoid certain notational ambiguities; subsequently, assumption (ii) will be dropped. We are interested more in the idea than in any broad application, so generality is not an important consideration.

Let $\pi(\theta_t, \theta_{t+1})$ represent the probability of state θ_{t+1} being realized in period $t + 1$, given that θ_t is realized in period t . The financial market is assumed to be complete in the following sense: at every date t , and for every state θ_t , there exists a short-term contingent claim that pays one unit of consumption if state θ_{t+1} is realized in period $t + 1$ (and nothing otherwise). We denote the period t , state θ_t price of such a claim by $q(\theta_t, \theta_{t+1})$.

Arrow–Debreu long-term claims (relative to $t = 0$) are not formally traded in this economy. Nevertheless, they can be synthetically created by dynamically trading short-term claims. (In general, more frequent trading can substitute for fewer claims.) To illustrate, let $q(\theta_0, \theta_{t+1})$ represent the period $t = 0$ price of a claim to one unit of the numeraire, if and only if, event θ_{t+1} is realized in period $t + 1$. It must be the case that

$$q(\theta_0, \theta_{t+1}) = \prod_{s=0}^t q(\theta_s, \theta_{s+1}), \quad (13.1)$$

where $(\theta_0, \dots, \theta_t)$ is the unique prior history of θ_{t+1} . By the uniqueness of the path to θ_{t+1} , $q(\theta_0, \theta_{t+1})$ is well defined. By no-arbitrage arguments, if the long-term Arrow–Debreu

¹ By *local* we mean that valuation is considered only in the context of the derivative, the underlying asset (a stock), and a risk-free bond.

security were also traded, its price would conform to Eq. (13.1). Arrow–Debreu securities can thus be effectively created via dynamic (recursive) trading, and the resulting financial market structure is said to be dynamically complete.² By analogy, the price in period t , state θ_t , of a security that pays one unit of consumption if state θ_{t+J} is observed in period $t + J$, $q(\theta_t, \theta_{t+J})$, is given by

$$q(\theta_t, \theta_{t+J}) = \prod_{s=t}^{t+J-1} q(\theta_s, \theta_{s+1}).$$

It is understood that θ_{t+J} is feasible from θ_t ; that is, given that we are in state θ_t in period t , there is some positive probability for the economy to find itself in state θ_{t+J} in period $t + J$. Otherwise the claims price must be zero.

Since our current objective is to develop risk-neutral pricing representations, a natural next step is to define risk-free bond prices and associated risk-free rates. Given that the current date state is (θ_t, t) , the price, $q^b(\theta_t, t)$, of a risk-free one-period (short-term) bond is given by (no arbitrage)

$$q^b(\theta_t, t) = \sum_{\theta_{t+1}=1}^{N_{t+1}} q(\theta_t, \theta_{t+1}); \quad (13.2)$$

Note here that the summation sign applies across all N_{t+1} future states of nature. The corresponding risk-free rate must satisfy

$$(1 + r_f(\theta_t)) = \{q^b(\theta_t, t)\}^{-1}$$

Pricing a k -period risk-free bond is similar:

$$q^b(\theta_t, t+k) = \sum_{\theta_{t+k}=1}^{N_{t+k}} q(\theta_t, \theta_{t+k}). \quad (13.3)$$

The final notion is that of an *accumulation factor*, denoted by $g(\theta_t, \theta_{t+k})$, and defined for a specific path $(\theta_t, \theta_{t+1}, \dots, \theta_{t+k})$ as follows:

$$g(\theta_t, \theta_{t+k}) = \prod_{s=t}^{t+k-1} q^b(\theta_s, s+1). \quad (13.4)$$

² This fact suggests that financial markets may need to be “very incomplete” if incompleteness *per se* is to have a substantial effect on equilibrium asset prices and, for example, have a chance to resolve some of the puzzles uncovered in Chapter 10. See Telmer (1993).

The idea being captured by the accumulation factor is this: An investor who invests one unit of consumption in short-term risk-free bonds from date t to $t + k$, continually rolling over his investment, will accumulate $[g(\theta_t, \theta_{t+k})]^{-1}$ units of consumption by date $t + k$, if events $\theta_{t+1}, \dots, \theta_{t+k}$ are realized. Alternatively,

$$[g(\theta_t, \theta_{t+k})]^{-1} = \prod_{s=0}^{k-1} (1 + r_f(\theta_{t+s})). \quad (13.5)$$

Note that from the perspective of date t , state θ_t , the factor $[g(\theta_t, \theta_{t+k})]^{-1}$ is an uncertain quantity as the actual state realizations in the succeeding time periods are not known at period t . From the $t = 0$ perspective, $[g(\theta_t, \theta_{t+k})]^{-1}$ is in the spirit of a (conditional) forward rate.

Let us illustrate with the two-date forward accumulation factor. We take the perspective of the investor investing one unit of the numeraire in a short-term risk-free bond from date t to $t + 2$. His first investment is certain since the current state θ_t is known and it returns $(1 + r_f(\theta_t))$. At date $t + 1$, this sum will be invested again in a one-period risk-free bond with return $(1 + r_f(\theta_{t+1}))$ contracted at $t + 1$ and received at $t + 2$. From the perspective of date t , this is indeed an uncertain quantity. The compounded return on the investment is: $(1 + r_f(\theta_t))(1 + r_f(\theta_{t+1}))$. This is the inverse of the accumulation factor $g(\theta_t, \theta_{t+1})$ as spelled out in Eq. (13.5).

Let us next translate these ideas directly into the CCAPM settings.

13.3 Risk-Neutral Pricing in the CCAPM

We make two additional assumptions in order to restrict our current setting to the context of the CCAPM.

A13.1 There is one agent in the economy with time-separable VNM preferences represented by

$$U(\tilde{c}) = E_0 \left(\sum_{t=0}^{\infty} U(\tilde{c}_t, t) \right),$$

where $U(\tilde{c}_t, t)$ is a family of, strictly increasing, concave, differentiable period utility functions, with $U_1(c_t, t) > 0$ for all t , $\tilde{c}_t = c(\theta_t)$ is the uncertain period t consumption, and E_0 the expectations operator conditional on date $t = 0$ information.

This treatment of the agent's preferences is quite general. For example, $U(c_t, t)$ could be of the form $\delta^t U(c_t)$ as in earlier chapters. Alternatively, the period utility function could itself be changing through time in deterministic fashion, or some type of habit formation could be

postulated. In all cases, it is understood that the set of feasible consumption sequences will be such that the sum exists (is finite).

A13.2 Output in this economy, $\tilde{Y}_t = Y_t(\theta_t)$ is exogenously given and, by construction, represents the consumer's income. In equilibrium it represents his consumption as well.

Recall that equilibrium-contingent claims prices in the CCAPM economy are no-trade prices, supporting the consumption sequences $\{\tilde{c}_t\}$ in the sense that at these prices, the representative agent does not want to purchase any claims; that is, at the prevailing contingent-claims prices, his existing consumption sequence is optimal. The loss in period t utility experienced by purchasing a contingent claim $q(\theta_t, \theta_{t+1})$ is exactly equal to the resultant increase in expected utility in period $t + 1$. There is no benefit to further trade. More formally,

$$U_1(c(\theta_t), t)q(\theta_t, \theta_{t+1}) = \pi(\theta_t, \theta_{t+1})U_1(c(\theta_{t+1}), t+1), \text{ or} \\ q(\theta_t, \theta_{t+1}) = \pi(\theta_t, \theta_{t+1}) \left\{ \frac{U_1(c(\theta_{t+1}), t+1)}{U_1(c(\theta_t), t)} \right\}. \quad (13.6)$$

Equation (13.7) corresponds to the state claim pricing in Section 10.4. State probabilities and intertemporal rates of substitution appear once again as the determinants of equilibrium Arrow–Debreu prices. Note that the more general utility specification adopted in this chapter does not permit bringing out explicitly the element of time discounting embedded in the intertemporal marginal rates of substitution. A short-term risk-free bond is thus priced according to

$$q^b(\theta_t, t+1) = \sum_{\theta_{t+1}=1}^{N_{t+1}} q(\theta_t, \theta_{t+1}) = \frac{1}{U_1(c(\theta_t), t)} E_t \{ U_1(c(\theta_{t+1}), t+1) \}. \quad (13.7)$$

Risk-neutral valuation is in the spirit of discounting at the risk-free rate. Accordingly, we may ask: at what probabilities must we compute the expected payoff to a security in order to obtain its price by discounting that payoff at the risk-free rate? But which risk-free rates are we speaking about? In a multiperiod context, there are two possibilities, and the alternative we choose will govern the precise form of the probabilities themselves.

The spirit of the dilemma is portrayed in Figure 13.1, which illustrates the case of a $t = 3$ period cash flow.

Under the first alternative, the cash flow is discounted at a series of consecutive short (one-period) rates, while in the second we discount back at the term structure of



Figure 13.1
Two possibilities of discounting a $t = 3$ period cash flow.

multiperiod discount bonds. These methods provide the same price, although the form of the risk-neutral probabilities will differ substantially. Here we offer a discussion of alternative 1; alternative 2 is considered in [Appendix 13.1](#).

Since the one-period state claims are the simplest securities, we will first ask what the risk-neutral probabilities must be in order that they be priced equal to the present value of their expected payoff, discounted at the risk-free rate.³ As before, let these numbers be denoted by $\pi^{RN}(\theta_t, \theta_{t+1})$. They are defined by:

$$q(\theta_t, \theta_{t+1}) = \pi(\theta_t, \theta_{t+1}) \left\{ \frac{U_1(c(\theta_{t+1}), t+1)}{U_1(c(\theta_t), t)} \right\} = q^b(\theta_t, t+1) [\pi^{RN}(\theta_t, \theta_{t+1})].$$

The second equality reiterates the tight relationship found in Chapter 12 between Arrow–Debreu prices and risk-neutral probabilities. Substituting [Eq. \(13.7\)](#) for $q^b(\theta_t, t+1)$ and rearranging terms, one obtains:

$$\begin{aligned} \pi^{RN}(\theta_t, \theta_{t+1}) &= \pi(\theta_t, \theta_{t+1}) \left\{ \frac{U_1(c(\theta_{t+1}), t+1)}{U_1(c(\theta_t), t)} \right\} \frac{U_1(c(\theta_t), t)}{E_t \{ U_1(c(\theta_{t+1}), t+1) \}} \\ &= \pi(\theta_t, \theta_{t+1}) \left\{ \frac{U_1(c(\theta_{t+1}), t+1)}{E_t U_1(c(\theta_{t+1}), t+1)} \right\}. \end{aligned} \tag{13.8}$$

Since $U(c(\theta_t), t)$ is assumed to be strictly increasing, $U_1 > 0$ and $\pi^{RN}(\theta_t, \theta_{t+1}) > 0$ (without loss of generality we may assume $\pi(\theta_t, \theta_{t+1}) > 0$). Furthermore, by construction, $\sum_{\theta_{t+1}=1}^{N_{t+1}} \pi^{RN}(\theta_t, \theta_{t+1}) = 1$. The set $\{\pi^{RN}(\theta_t, \theta_{t+1})\}$ thus defines a set of conditional (on θ_t) risk-neutral transition probabilities. As in our earlier more general setting, if the representative agent is risk neutral, $U_1(c(\theta_t), t) \equiv \text{constant}$ for all t , and $\pi^{RN}(\theta_t, \theta_{t+1})$

³ Recall that since all securities can be expressed as portfolios of state claims, we can use the state claims alone to construct the risk-neutral probabilities.

coincides with $\pi(\theta_t, \theta_{t+1})$, the true probability. Using these transition probabilities, makes it possible to discount expected future consumption flows at the intervening risk-free rates. Notice how the risk-neutral probabilities are related to the true probabilities: They represent the true probabilities scaled up or down by the relative consumption scarcities in the different states. For example, if, for some state θ_{t+1} , the representative agent's consumption is usually low, his marginal utility of consumption in that state will be much higher than average marginal utility and thus

$$\pi^{RN}(\theta_t, \theta_{t+1}) = \pi(\theta_t, \theta_{t+1}) \left\{ \frac{U_1(c(\theta_{t+1}), t+1)}{E_t U_1(c(\theta_{t+1}), t+1)} \right\} > \pi(\theta_t, \theta_{t+1}).$$

The opposite will be true if a state has a relative abundance of consumption. When we compute expected payoffs to assets using risk-neutral probabilities, we are thus implicitly taking into account both the (no-trade) relative equilibrium scarcities (prices) of their payoffs and their objective relative scarcities. This allows discounting at the risk-free rate: No further risk adjustment need be made to the discount rate as all such adjustments have been implicitly undertaken in the expected payoff calculation.

To gain a better understanding of this notion, let us go through a few examples.

Example 13.1 Denote a stock's associated dividend stream by $\{d(\theta_t)\}$. Under the basic state-claim valuation perspective of Section 11.2, its ex-dividend price at date t , given that θ_t has been realized, is:

$$q^e(\theta_t, t) = \sum_{s=t+1}^{\infty} \sum_{j=1}^{N_s} q(\theta_t, \theta_s(j)) d(\theta_s(j)), \quad (13.9)$$

or, with a recursive representation,

$$q^e(\theta_t, t) = \sum_{\theta_{t+1}} q(\theta_t, \theta_{t+1}) \{q^e(\theta_{t+1}, t+1) + d(\theta_{t+1})\} \quad (13.10)$$

Equation (13.10) may also be expressed as

$$q^e(\theta_t, t) = q^b(\theta_t, t+1) E_t^{RN} \{q^e(\tilde{\theta}_{t+1}, t+1) + d(\tilde{\theta}_{t+1})\}, \quad (13.11)$$

where E_t^{RN} denotes the expectation taken with respect to the relevant risk-neutral transition probabilities; equivalently,

$$q^e(\theta_t, t) = \frac{1}{1 + r_f(\theta_t)} E_t^{RN} \{q^e(\tilde{\theta}_{t+1}, t+1) + d(\tilde{\theta}_{t+1})\}.$$

Returning again to the present value expression, Eq. (13.9), we have

$$\begin{aligned} q^e(\theta_t, t) &= \sum_{s=t+1}^{\infty} E_t^{RN} \{g(\theta_t, \tilde{\theta}_s)d(\tilde{\theta}_s)\} \\ &= \sum_{s=t+1}^{\infty} E_t^{RN} \frac{d(\tilde{\theta}_s)}{\prod_{j=0}^{s-1} (1 + r_f(\tilde{\theta}_{t+j}))}. \end{aligned} \quad (13.12)$$

What does Eq. (13.12) mean? Any state $\hat{\theta}_s$ in period $s \geq t + 1$ has a unique sequence of states preceding it. The product of the risk-neutral transition probabilities associated with the states along the path defines the (conditional) risk-neutral probability of $\hat{\theta}_s$ itself. The product of this probability and the payment as $d(\hat{\theta}_s)$ is then discounted at the associated accumulation factor—the present value factor corresponding to the risk-free rates identified with the succession of states preceding $\hat{\theta}_s$. For each $s \geq t + 1$, the expectation represents the sum of all these terms, one for each θ_s feasible from θ_t .

Since the notational intensity tends to obscure what is basically a very straightforward idea, let us turn to a small numerical example.

Example 13.2 Let us value a two-period equity security, where $U(c_t, t) \equiv U(c_t) = \ln c_t$ for the representative agent (no discounting). The evolution of uncertainty is given by Figure 13.2 where

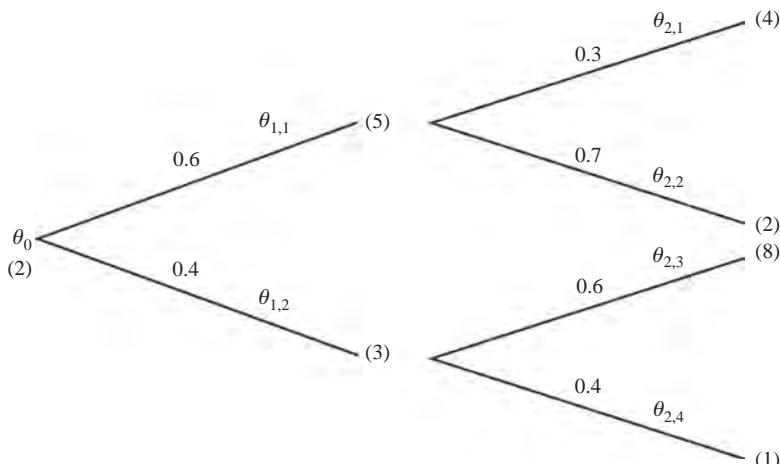


Figure 13.2

The structure of security payoffs—two periods—two states at each node.

$$\begin{aligned}
 \pi(\theta_0, \theta_{1,1}) &= 0.6 & \pi(\theta_{1,1}, \theta_{2,1}) &= 0.3 \\
 \pi(\theta_0, \theta_{1,2}) &= 0.4 & \pi(\theta_{1,1}, \theta_{2,2}) &= 0.7 \\
 && \pi(\theta_{1,2}, \theta_{2,3}) &= 0.6 \\
 && \pi(\theta_{1,2}, \theta_{2,4}) &= 0.4
 \end{aligned}$$

The consumption at each node, which equals the dividend, is represented as the quantity in parentheses. To value this asset risk neutrally, we consider three stages.

1. Compute the (conditional) risk-neutral probabilities at each node.

$$\pi^{RN}(\theta_0, \theta_{1,1}) = \pi(\theta_0, \theta_{1,1}) \left\{ \frac{U_1(c(\theta_{1,1}))}{E_0\{U_1(c_1(\tilde{\theta}_1))\}} \right\} = \frac{\binom{1}{5}}{\left(0.6\binom{1}{5} + 0.4\binom{1}{3}\right)} = 0.4737$$

$$\pi^{RN}(\theta_0, \theta_{1,2}) = 1 - \pi^{RN}(\theta_0, \theta_{1,1}) = 0.5263$$

$$\pi^{RN}(\theta_{1,1}, \theta_{2,1}) = \pi(\theta_{1,1}, \theta_{2,1}) \left\{ \frac{\frac{1}{4}}{\left(0.3\binom{1}{4} + 0.7\binom{1}{2}\right)} \right\} = 0.1765$$

$$\pi^{RN}(\theta_{1,1}, \theta_{2,2}) = 1 - \pi^{RN}(\theta_{1,1}, \theta_{2,1}) = 0.8235$$

$$\pi^{RN}(\theta_{1,2}, \theta_{2,4}) = \pi(\theta_{1,2}, \theta_{2,4}) \left\{ \frac{1}{\left(0.6\binom{1}{8} + 0.4(1)\right)} \right\} = 0.8421$$

$$\pi^{RN}(\theta_{1,2}, \theta_{2,3}) = 0.1579$$

2. Compute the conditional bond prices.

$$q^b(\theta_0, 1) = \frac{1}{U_1(c_0)} E_0\{U_1(c_1(\tilde{\theta}))\} = \frac{1}{\binom{1}{2}} \left\{ 0.6\binom{1}{5} + 0.4\binom{1}{3} \right\} = 0.5066$$

$$q^b(\theta_{1,1}, 2) = \frac{1}{\binom{1}{5}} \left\{ 0.3\binom{1}{4} + 0.7\binom{1}{2} \right\} = 2.125$$

$$q^b(\theta_{1,1}, 2) = \frac{1}{\binom{1}{3}} \left\{ 0.6\binom{1}{8} + 0.4\binom{1}{1} \right\} = 1.425$$

3. Value the asset.

$$\begin{aligned}
 q^e(\theta_0, 0) &= \sum_{s=1}^2 E_0^{RN} \{g(\theta_0, \tilde{\theta}_s) d_s(\tilde{\theta}_s)\} \\
 &= q^b(\theta_0, 1) \{\pi^{RN}(\theta_0, \theta_{1,1})(5) + \pi^{RN}(\theta_0, \theta_{1,2})(3)\} \\
 &\quad + q^b(\theta_0, 1) q^b(\theta_{1,1}, 2) \{\pi^{RN}(\theta_0, \theta_{1,1}) \pi^{RN}(\theta_{1,1}, \theta_{2,1})(4) \\
 &\quad + \pi^{RN}(\theta_0, \theta_{1,1}) \pi^{RN}(\theta_{1,1}, \theta_{2,2})(2)\} \\
 &\quad + q^b(\theta_0, 1) q^b(\theta_{1,2}, 2) \{\pi^{RN}(\theta_0, \theta_{1,2}) \{\pi^{RN}(\theta_{1,2}, \theta_{2,3})(8) \\
 &\quad + \pi^{RN}(\theta_0, \theta_{1,2}) \pi^{RN}(\theta_{1,2}, \theta_{2,4})(1)\} \\
 &= 4.00
 \end{aligned}$$

At a practical level this appears to be a messy calculation at best, but it is not obvious how we might compute the no-trade equilibrium asset prices more easily. The Lucas tree methodologies, for example, do not apply here as the setting is not infinitely recursive. This leaves us to solve for the equilibrium prices by working back through the tree and solving for the no-trade prices at each node. It is not clear that this will be any less involved.

Sometimes, however, the risk-neutral valuation procedure does allow for a very succinct, convenient representation of specific asset prices or price interrelationship. A case in point is that of a long-term discount bond.

Example 13.3 To price at time t , state θ_t , a long-term discount bond maturing in date $t+k$, observe that the corresponding dividend $d_{t+k}(\theta_{t+k}) \equiv 1$ for every θ_{t+k} feasible from state θ_t . Applying Eq. (13.12) yields

$$q^b(\theta_t, t+k) = E_t^{RN} g(\theta_t, \tilde{\theta}_{t+k}), \text{ or}$$

$$\frac{1}{(1+r_f(\theta_t, t+k))^k} = E_t^{RN} \left\{ \frac{1}{\prod_{s=t}^{t+k-1} (1+r_f(\theta_s, s+1))} \right\} \quad (13.13)$$

Equation (13.13), in either of its forms, informs us that the long-term rate is the expectation of the short rates taken with respect to the risk-neutral transition probabilities. This is generally not true if the expectation is taken with the ordinary or true probabilities.

At this point we draw this formal discussion to a close. We now have an idea what risk-neutral valuation might mean in a CCAPM context. Appendix 13.1 briefly discusses the second valuation procedure and illustrates it with the pricing of call and put options.

We thus see that the notion of risk-neutral valuation carries over easily to a CCAPM context. This is not surprising: The key to the existence of a set of risk-neutral probabilities is the presence of a complete set of securities markets, which is the case with the CCAPM. In fact, the somewhat weaker notion of dynamic completeness was sufficient.

We next turn our attention to equity derivatives pricing. The setting is much more specialized and not one of general equilibrium (though not inconsistent with it). One instance of this specialization is that the underlying stock's price is presumed to follow a specialized stochastic process. The term structure is also presumed to be flat. These assumptions, taken together, are sufficient to generate the existence of a unique risk-neutral probability measure, which can be used to value any derivative security written on the stock. That these probabilities are uniquely identified with the specific underlying stock has led us to dub them *local*.

13.4 The Binomial Model of Derivatives Valuation

Under the binomial abstraction we imagine a many-period world in which, at every date-state node only a stock and a bond are traded. With only two securities to trade, dynamic completeness requires that at each node there be only two possible succeeding states. For simplicity, we will also assume that the stock pays no dividend, in other words, that $d(\theta_t) \equiv 0$ for all $t \leq T$. Lastly, in order to avoid any ambiguity in the risk-free discount factors, it is customary to require that the risk-free rate be constant across all dates and states. We formalize these assumptions as follows:

A13.3 The risk-free rate is constant;

$$q^b(\theta_t, t+1) = \frac{1}{1+r_f} \quad \text{for all } t \leq T.$$

A13.4 The stock pays no dividends: $d(\theta_t) \equiv 0$ for all $t \leq T$.

A13.5 The rate of return to stock ownership follows an i.i.d. process of the form:

$$q^e(\theta_{t+1}, t+1) = \begin{cases} uq^e(\theta_t, t), & \text{with probability } \pi \\ dq^e(\theta_t, t), & \text{with probability } 1 - \pi, \end{cases}$$

where u (up) and d (down) represent gross rates of return. In order to preclude the existence of an arbitrage opportunity, it must be the case that

$$u > R_f > d,$$

where, in this context, $R_f = 1 + r_f$.

There are effectively only two possible future states in this model ($\theta_t \in \{\theta_1, \theta_2\}$ where θ_1 is identified with u and θ_2 identified with d). Thus, the evolution of the stock's price can be represented by a simple tree structure as seen in [Figure 13.3](#).

Why such a simple setting should be of use is not presently clear, but it will become so shortly.

In this context, the risk-neutral probabilities can be easily computed from [Eq. \(13.11\)](#), specialized to accommodate $d(\theta_t) \equiv 0$:

$$\begin{aligned} q^e(\theta_t, t) &= q^b(\theta_t, t+1) E_t^{RN} \{q^e(\tilde{\theta}_{t+1}, t+1)\} \\ &= q^b(\theta_t, t+1) \{\pi^{RN} u q^e(\theta_t, t) + (1 - \pi^{RN}) d q^e(\theta_t, t)\} \end{aligned} \quad (13.14)$$

This implies

$$\begin{aligned} R_f &= \pi^{RN} u + (1 - \pi^{RN}) d, \text{ or} \\ \pi^{RN} &= \frac{R_f - d}{u - d}. \end{aligned} \quad (13.15)$$

The power of this simple context is made clear when comparing [Eq. \(13.15\)](#) with [Eq. \(13.8\)](#). Here risk-neutral probabilities can be expressed without reference to marginal rates of substitution, that is, to agents' preferences.⁴ This provides an immense

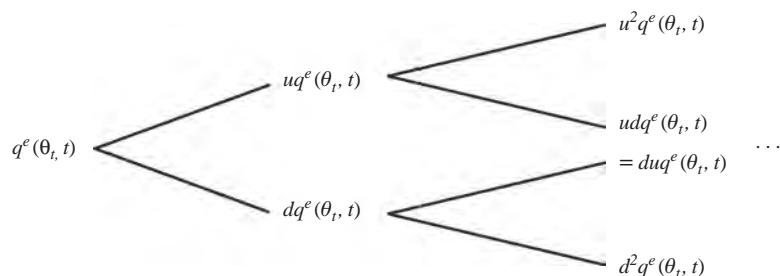


Figure 13.3
A binomial tree structure.

⁴ Notice that the risk-neutral probability distribution is i.i.d. as well.

simplification, which all derivative pricing will exploit in one way or another. Of course, the same is true for one-period Arrow–Debreu securities since they are priced equal to the present value of their respective risk-neutral probabilities:

$$q(\theta_t, \theta_{t+1} = u) = \left(\frac{1}{R_f} \right) \left(\frac{R_f - d}{u - d} \right), \text{ and}$$

$$q(\theta_t, \theta_{t+1} = d) = \left(\frac{1}{R_f} \right) \left(1 - \frac{R_f - d}{u - d} \right) = \left(\frac{1}{R_f} \right) \left(\frac{u - R_f}{u - d} \right).$$

Furthermore, since the risk-free rate is assumed constant in every period, the price of a claim to one unit of the numeraire to be received $T - t > 1$ periods from now if state θ_T is realized is given by

$$q(\theta_t, \theta_T) = \frac{1}{(1 + r_f(\theta_t, T))^{T-t}} \sum_{\{\theta_t, \dots, \theta_{T-1}\} \in \Omega} \prod_{s=t}^{T-1} \pi^{RN}(\theta_s, \theta_{s+1}),$$

where Ω represents the set of all time paths $\{\theta_t, \theta_{t+1}, \dots, \theta_{T-1}\}$ leading to θ_T . In the binomial setting this becomes

$$q(\theta_t, \theta_T) = \frac{1}{(R_f)^{T-t}} \binom{T-t}{s} (\pi^{RN})^s (1 - \pi^{RN})^{T-t-s}, \quad (13.16)$$

where s is the number of intervening periods in which the u state is observed on any path from θ_t to θ_T . The expression $\binom{T-t}{s}$ represents the number of ways s successes (u moves) can occur in $T - t$ trials. A standard result states

$$\binom{T-t}{s} = \frac{(T-t)!}{s!(T-t-s)!}.$$

The explanation is as follows. Any possible period T price of the underlying stock will be identified with a unique number of u and d realizations. Suppose, for example, that s_1

u -realizations are required. There are then $\binom{T-t}{s_1}$ possible paths, each of which has

exactly s_1 u and $(T - t - s_1)$ d states, leading to the prespecified period T price. Each path has the common risk-neutral probability $(\pi^{RN})^{s_1} (1 - \pi^{RN})^{T-t-s_1}$. As an example, suppose $T - t = 3$, and the particular final price is the result of two up moves and one down move. Then, there are $3 = \frac{3!}{2!1!} = \frac{3 \cdot 2 \cdot 1}{(2 \cdot 1)(1)}$ possible paths leading to that final state: uud , udu , and duu .

To illustrate the simplicity of this setting we again consider several examples.

Example 13.4 A European call option revisited: let the option expire at $T > t$; the price of a European equity call with exercise price K , given the current date-state (θ_t, t) , $C_e(\theta_t, t)$ is

$$\begin{aligned} C_e(\theta_t, t) &= \left(\frac{1}{R_f} \right)^{T-t} E_t^{RN} (\max\{q^e(\theta_T, T) - K, 0\}) \\ &= \left(\frac{1}{R_f} \right)^{T-t} \sum_{s=0}^{T-t} \binom{T-t}{s} (\pi^{RN})^s (1-\pi^{RN})^{T-t-s} (\max\{q^e(\theta_t, t) u^s d^{T-t-s} - K, 0\}) \end{aligned}$$

When taking the expectation, we sum over all possible values of $s \leq T-t$, thus weighting each possible option payoff by the risk-neutral probability of attaining it.

Define the quantity \hat{s} as the minimum number of intervening up states necessary for the underlying asset, the stock, to achieve a price in excess of K . The prior expression can then be simplified to:

$$C_e(\theta_t, t) = \frac{1}{(R_f)^{T-t}} \sum_{s=\hat{s}}^{T-t} \binom{T-t}{s} (\pi^{RN})^s (1-\pi^{RN})^{T-t-s} [q^e(\theta_t, t) u^s d^{T-t-s} - K], \quad (13.17)$$

or

$$\begin{aligned} C_e(\theta_t, t) &= \frac{1}{(R_f)^{T-t}} \sum_{s=\hat{s}}^{T-t} \binom{T-t}{s} (\pi^{RN})^s (1-\pi^{RN})^{T-t-s} q^e(\theta_t, t) u^s d^{T-t-s} \\ &\quad - \sum_{s=\hat{s}}^{T-t} \binom{T-t}{s} (\pi^{RN})^s (1-\pi^{RN})^{T-t-s} K \end{aligned} \quad (13.18)$$

The first term within the braces of Eq. (13.18) is the risk-neutral expected value at expiration of the acquired asset if the option is exercised, whereas the second term is the risk-neutral expected cost of acquiring it. The difference is the risk-neutral expected value of the call's payoff (value) at expiration.⁵ To value the call today, this quantity is then put on a present value basis by discounting at the risk-free rate R_f .

This same valuation can also be obtained by working backward, recursively, through the tree. Since markets are complete, in the absence of arbitrage opportunities any asset—the call included—is priced equal to its expected value in the succeeding time period discounted at R_f . This implies

$$C_e(\theta_t, t) = q^b(\theta_t, t) E^{RN} C_e(\tilde{\theta}_{t+1}, t+1). \quad (13.19)$$

Let us next illustrate how this fact may be used to compute the call's value in a simple three-period example.

⁵ Recall that there is no actual transfer of the security. Rather, this difference $q^e(\theta_T, T) - K$ represents the amount of money the writer (seller) of the call must transfer to the buyer at the expiration date if the option is exercised.

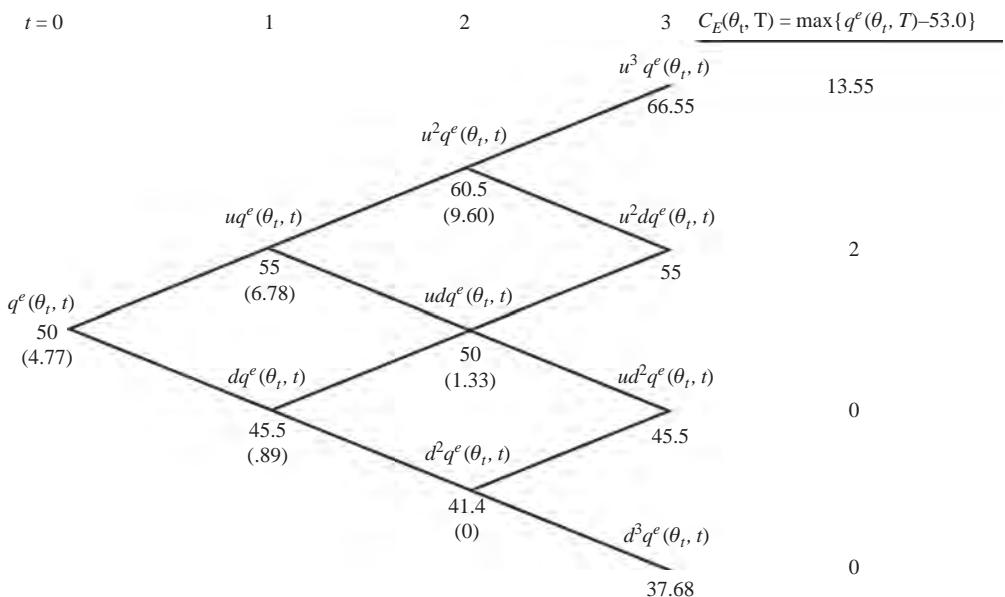


Figure 13.4
The binomial tree of Example 13.5.

Example 13.5 Let $u = 1.1$, $d = \frac{1}{u} = 0.91$, $q^e(\theta_t, t) = \$50$, $K = \$53$, $R_f = 1.05$, $T - t = 3$.

$$\pi^{RN} = \frac{R_f - d}{u - d} = \frac{1.05 - 0.91}{1.1 - 0.91} = 0.70$$

The numbers in parentheses in Figure 13.4 are the recursive values of the call, working backward in the manner of Eq. (13.19). These are obtained as follows:

$$C_e(u^2, t+2) = \frac{1}{1.05} \{0.70(13.55) + 0.30(2)\} = 9.60$$

$$C_e(ud, t+2) = \frac{1}{1.05} \{0.70(2) + 0.30(0)\} = 1.33$$

$$C_e(u, t+1) = \frac{1}{1.05} \{0.70(9.60) + 0.30(1.33)\} = 6.78$$

$$C_e(d, t+1) = \frac{1}{1.05} \{0.70(1.33) + 0.30(0)\} = 0.89$$

$$C_e(\theta_t, t) = \frac{1}{1.05} \{0.70(6.78) + 0.30(0.89)\} = 4.77$$

Table 13.1: Payoff pattern—Asian option

t	$t + 1$	$t + 2$...	$T - 1$	T
0	0	0		0	$\max\{q_{AVG}^e(\theta_T, T) - K, 0\}$

For a simple call, its payoff at expiration is dependent only upon the value of the underlying asset (relative to K) at that time, regardless of its price history. For example, the value of the call when $q^e(\theta_T, T) = 55$ is the same if the price history is (50,55,50,55) or (50,45.5,50,55).

For other derivatives, however, this is not the case; they are *path dependent*. An Asian (path-dependent) option is a case in point. Nevertheless, the same valuation methods apply: its expected payoff is computed using the risk-neutral probabilities, and then discounted at the risk-free rate.

Example 13.6 A path-dependent option: we consider an Asian option for which the payoff pattern assumes the form outlined in [Table 13.1](#).

Where $q_{AVG}^e(\theta_T, T)$ is the average price of the stock along the path from $q^e(\theta_t, t)$ to, and including, $q^e(\theta_T, T)$. We may express the period t value of such an option as

$$C_A(\theta_t, t) = \frac{1}{(R_f)^{T-t}} E_t^{RN} \max\{q_{AVG}^e(\theta_T, T) - K, 0\}$$

A simple numerical example with $T - t = 2$ follows. Let $q^e(\theta_t, t) = 100$, $K = 100$, $u = 1.05$, $d = \frac{1}{u} = 0.95$, and $R_f = 1.005$. The corresponding risk-neutral probabilities are

$$\pi^{RN} = \frac{R_f - d}{u - d} = \frac{1.005 - 0.95}{1.05 - 0.95} = 0.55; 1 - \pi^{RN} = 0.45$$

With two periods remaining, the possible evolutions of the stock's price and corresponding option payoffs are those found in [Figure 13.5](#).

Thus,

$$C_A(\theta_t, t) = \frac{1}{(1.005)^2} \{(0.55)^2(5.083) + (0.55)(0.45)(1.67)\} = \$1.932$$

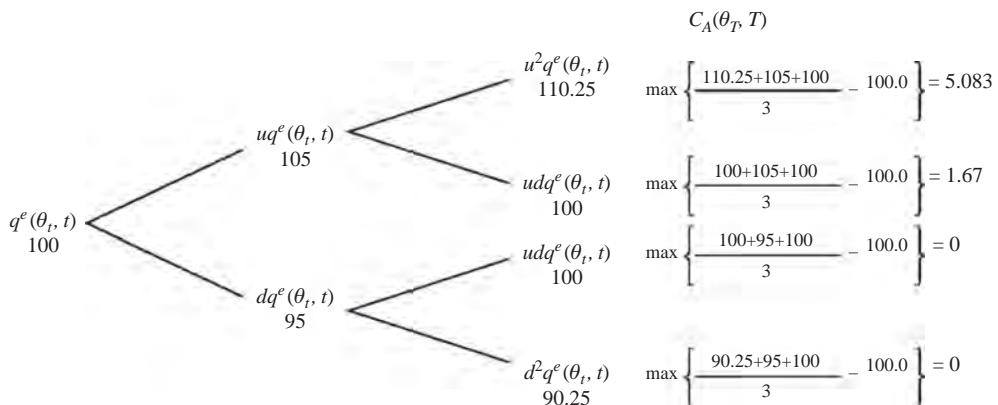


Figure 13.5
Evolution of the stock's price and the Asian option payoffs.

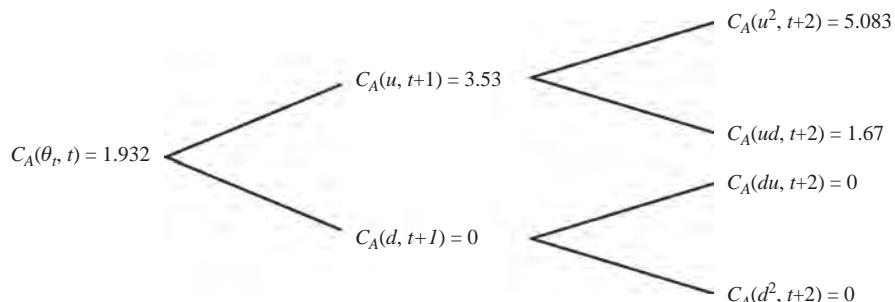


Figure 13.6
Computing recursively the value of the Asian option.

Note that we may as well work backward, recursively, in the price/payoff tree as shown in Figure 13.6,

where $C_A(\theta_{t+1} = u, t + 1) = 3.53 = \frac{1}{(1.005)} \{0.55(5.083) + 0.45(1.67)\}$, and

$$C_A(\theta_t, t) = \frac{1}{(1.005)} \{0.55(3.53) + 0.45(0)\} = \$1.932.$$

A number of fairly detailed comments are presently in order. Note that with a path-dependent option it is not possible to apply, naively, a variation on Eq. (13.18). Unlike with straightforward calls, the value of this type of option is not the same for all paths leading to the same final-period asset price.

Who might be interested in purchasing such an option? For one thing, they have payoff patterns similar in spirit to an ordinary call, but are generally less expensive (there is less upward potential in the average than in the price itself). This feature has contributed to the usefulness of path-dependent options in foreign exchange trading. Consider a firm that needs to provide a stream of payments (say, perhaps, for factory construction) in a foreign currency. It would want protection against a rise in the value of the foreign currency relative to its own because such a rise would increase the cost of the payment stream in terms of the firm's own currency. Since many payments are to be made, what is of concern is the average price of the foreign currency rather than its price at any specific date. By purchasing the correct number of Asian calls on the foreign currency, the firm can create a payment for itself if, on average, the foreign currency's value exceeds the strike price—the level above which the firm would like to be insured. By analogous reasoning, if the firm wished to protect the average value of a stream of payments, it was receiving in a foreign currency, the purchase of Asian puts would be one alternative.

We do not want to lose sight of the fact that risk-neutral valuation is a direct consequence of the dynamic completeness (at each node there are two possible future states and two securities available for trade) and the no-arbitrage assumption, a connection that is especially apparent in the binomial setting. Consider a call option with expiration one period from the present. Over this period the stock's price behavior and the corresponding payoffs to the call option are as found in [Figure 13.7](#).

By the assumed dynamic completeness we know that the payoff to the option can be replicated on a state-by-state basis by a position in the stock and the bond. Let this position be characterized by a portfolio of Δ shares and a bond investment of value B

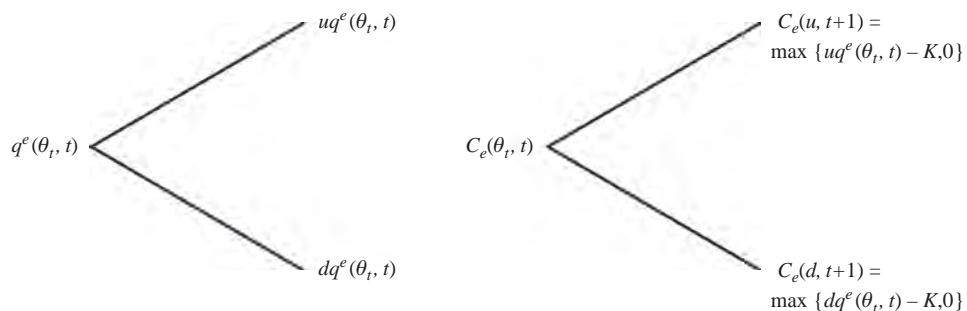


Figure 13.7

A call option one period to expiration and the underlying stock's price.

(for simplicity of notation we suppress the dependence of these latter quantities on the current state and date). Replication requires

$$uq^e(\theta_t, t)\Delta + R_f B = Ce(u, t+1), \text{ and}$$

$$dq^e(\theta_t, t)\Delta + R_f B = Ce(d, t+1),$$

from which follows

$$\Delta = \frac{Ce(u, t+1) - Ce(d, t+1)}{(u-d)q^e(\theta_t, t)}, \text{ and}$$

$$B = \frac{uC_e(d, t+1) - dC_e(u, t+1)}{(u-d)R_f},$$

By the no-arbitrage assumption:

$$\begin{aligned} C_e(\theta_t, t) &= \Delta q^e(\theta_t, t) + B \\ &= \left\{ \frac{Ce(u, t+1) - Ce(d, t+1)}{(u-d)q^e(\theta_t, t)} q^e(\theta_t, t) \frac{uC_e(d, t+1) - dC_e(u, t+1)}{(u-d)R_f} \right\} \\ &= \left(\frac{1}{R_f} \right) \left\{ \left(\frac{R_f - d}{u - d} \right) C_e(u, t+1) + \left(\frac{u - R_f}{u - d} \right) C_e(d, t+1) \right\} \\ &= \left(\frac{1}{R_f} \right) \{ \pi^{RN} C_e(u, t+1) + (1 - \pi^{RN}) C_e(d, t+1) \}, \end{aligned}$$

which is just a specialized case of [Eq. \(13.18\)](#).

Valuing an option (or other derivative) using risk-neutral valuation is thus equivalent to pricing its replicating portfolio of stock and debt. Working backward in the tree corresponds to recomputing the portfolio of stock and debt that replicates the derivative's payoffs at each of the succeeding nodes. In the earlier example of the Asian option, the value 3.53 at the intermediate u node represents the value of the portfolio of stocks and bonds necessary to replicate the option's values in the second-period nodes leading from it (5.083 in the u state, 1.67 in the d state).

Let us see how the replicated portfolio evolves in the case of the Asian option written on a stock.

$$\Delta_u = \frac{C_A(u^2, t+2) - C_A(ud, t+2)}{(u-d)q^e(\theta_t, t)} = \frac{5.083 - 1.67}{(1.05 - 0.95)(105)} = 0.325$$

$$B_u = \frac{uC_A(ud, t+2) - dC_A(u^2, t+2)}{(u-d)R_f} = \frac{(1.05)(1.67) - (0.95)(5.083)}{(1.05 - 0.95)(1.005)} = -30.60$$

$\Delta_d = 0$
 $B_d = 0$ (all branches leading from the “d” node result in zero option value)

$$\Delta = \frac{C_A(u, t+1) - C_A(d, t+1)}{(u-d)q^e(\theta_t, t)} = \frac{3.53 - 0}{(1.05 - 0.95)(100)} = 0.353$$

$$B = \frac{uC_A(d, t+2) - dC_A(u, t+1)}{(u-d)R_f} = \frac{(1.05)(0) - (0.95)(3.53)}{(1.05 - 0.95)(1.005)} = -33.33$$

We interpret these numbers as follows. In order to replicate the value of the Asian option, regardless of whether the underlying stock’s price rises to \$105 or falls to \$95.20, it is necessary to construct a portfolio composed of a loan of \$33.33 at R_f in conjunction with a long position of 0.353 share. The net cost is

$$0.353(100) - 33.33 = \$1.97,$$

the cost of the call, except for rounding errors. To express this idea slightly differently, if you want to replicate, at each node, the value of the Asian option, borrow \$33.33 (at R_f) and, together with your own capital contribution of \$1.97, take this money and purchase 0.353 share of the underlying stock.

As the underlying stock’s value evolves through time, this portfolio’s value will evolve so that at any node it matches exactly the call’s value. At the first u node, for example, the portfolio will be worth \$3.53. Together with a loan of \$30.60, this latter sum will allow the purchase of 0.325 share, with no additional capital contribution required. Once assembled, the portfolio is entirely self-financing, no additional capital need be added, and none may be withdrawn (until expiration).

This discussion suggests that Asian options represent a levered position in the underlying stock. To see this, note that at the initial node the replicating portfolio consists of a \$1.97 equity contribution by the purchaser in conjunction with a loan of \$30.60. This implies a debt/equity ratio of $\frac{\$30.60}{\$1.97} \cong 15.5!$ For the analogous straight call, with the same exercise price as the Asian and the same underlying price process, the analogous quantities are, respectively, \$3.07 and \$54.47, giving a debt/equity ratio of approximately 18. Call-related securities are thus attractive instruments for speculation! For a relatively small cash outlay, a stock’s entire upward potential (within a limited span of time) can be purchased.

Under this pricing perspective, there are no-arbitrage opportunities within the universe of the underlying asset, the bond, or any derivative asset written on the underlying asset.

We were reminded of this fact in the prior discussion! The price of the call at all times equals the value of the replicating portfolio. It does not, however, preclude the existence of such opportunities among different stocks or among derivatives written on different stocks.

These discussions make apparent the fact that binomial risk-neutral valuation views derivative securities, and call options in particular, as redundant assets, redundant in the sense that their payoffs can be replicated with a portfolio of preexisting securities.

The presence or absence of these derivatives is deemed not to affect the price of the underlying asset (the stock) on which they are written. This is in direct contrast to our earlier motivation for the existence of options: their desirable property in assisting in the completion of the market. In principle, the introduction of an option has the potential of changing all asset values if it makes the market more complete.

This issue has been examined fairly extensively in the literature. From a theoretical perspective, [Detemple and Selden \(1991\)](#) construct a mean variance example where there is one risky asset, one risk-free asset, and an incomplete market. There the introduction of a call option is shown to increase the equilibrium price of the risky asset. In light of our earlier discussions, this is not entirely surprising: The introduction of the option enhances opportunities for risk sharing, thereby increasing demand and consequently the price of the risky asset. This result can be shown not to be fully applicable to all contexts, however. On the empirical side, [Detemple and Jorion \(1990\)](#) examine a large sample of options introductions over the period 1973 to 1986 and find that, on average, the underlying stock's price rises 3% as a result and its volatility diminishes.

13.5 Continuous Time: An Introduction to the Black–Scholes Formula

Although the binomial model presents a transparent application of risk-neutral valuation, it is not clear that it represents the accurate description of the price evolution of any known security. We deal with this issue presently.

Fat tails aside, there is ample evidence to suggest that stock prices may be modeled as being lognormally distributed; more formally,

$$\ln q^e(\theta_T, T) \sim N(\ln q^e(\theta_t, t) + \mu(T - t), \sigma\sqrt{T - t}),$$

where μ and σ denote, respectively, the mean and standard deviation of the continuously compounded rate of return over the reference period, typically 1 year. Regarding t as the present time, this expression describes the distribution of stock prices at some time T in the future given the current price $q^e(\theta_t, t)$. The length of the time horizon $T - t$ is measured in years.

The key result is this: properly parameterized, the distribution of final prices generated by the binomial distribution can arbitrarily well approximate the prior lognormal distribution when the number of branches becomes very large. More precisely, we may imagine a binomial model in which we divide the period $T - t$ into n subintervals of equal length $\Delta T(n) = (T - t)/n$. If we adjust u , d , p (the true probability of a u price move), and R_f appropriately, then as $n \rightarrow \infty$, the distribution of period T prices generated by the binomial model will converge in probability to the hypothesized lognormal distribution. The adjustment requires that

$$\begin{aligned} u(n) &= e^{\sigma\sqrt{\Delta t(n)}}, & d(n) &= \frac{1}{u(n)}, & p &= \frac{e^{\mu\Delta t(n)} - d(n)}{u(n) - d(n)}, \quad \text{and} \\ R_f(n) &= (R_f)^{\frac{1}{n}} \end{aligned} \tag{13.20}$$

For this identification, the binomial valuation formula for a call option, Eq. (13.18), converges to the Black–Scholes formula for a European call option written on a nondividend paying stock:

$$C_e(\theta_t, t) = q^e(\theta_t, T)N(d_1) - Ke^{-\hat{r}_f(t-t)}N(d_2) \tag{13.21}$$

where $N(\cdot)$ is the cumulative normal probability distribution function,

$$\begin{aligned} \hat{r}_f &= \ell n(R_f) \\ d_1 &= \frac{\ell n\left(\frac{q_e(\theta_t, t)}{K}\right) + (T - t)\left(\hat{r}_f + \frac{\sigma^2}{2}\right)}{\sigma\sqrt{T - t}} \\ d_2 &= d_1 - \sigma\sqrt{T - t} \end{aligned}$$

Cox and Rubinstein (1979) provide a detailed development and proof of this equivalence, but we can see the rudiments of its origin in Eq. (13.18), which we now present, modified to make apparent its dependence on the number of subintervals n :

$$\begin{aligned} C_e(\theta_t, t; n) &= \frac{1}{(R_f(n))^n} \left\{ \sum_{s=a(n)}^n \binom{n}{s} (\pi(n)^{RN})^s (1 - \pi(n)^{RN})^{n-s} q^e(\theta_t, t) \right. \\ &\quad \left. - K \sum_{s=a(n)}^n \binom{n}{s} (\pi(n)^{RN})^s (1 - \pi(n)^{RN})^{n-s} \right\} \end{aligned} \tag{13.22}$$

where $\pi(n)^{RN} = (R_f(n) - d(n))/(u(n) - d(n))$ g and $a(n)$ is the minimum number of up moves for the option to have positive payoff.

Rearranging terms yields

$$C_e(\theta_t, t; n) = q^e(\theta_t, t) \sum_{s=a(n)}^n \binom{n}{s} \left(\frac{\pi(n)^{RN}}{R_f(n)} \right)^s \left(\frac{1-\pi(n)^{RN}}{R_f(n)} \right)^{n-s} - K \left(\frac{1}{R_f(n)} \right)^n \sum_{s=a}^n \binom{n}{s} (\pi(n)^{RN})^s (1-\pi(n)^{RN})^{n-s}, \quad (13.23)$$

which is of the general form

$$C_e(\theta_t, t; n) = q_e(\theta_t, t) \times \text{Probability} - (\text{present value factor}) \times K \times \text{Probability},$$

as per the Black–Scholes formula. Since, at each step of the limiting process (i.e., for each n , as $n \mapsto \infty$), the call valuation formula is fundamentally an expression of risk-neutral valuation, the same must be true of its limit. As such, the Black–Scholes formula represents the first hint at the translation of risk-neutral methods to the case of continuous time.

Let us conclude this section with a few more observations. The first concerns the relationship of the Black–Scholes formula to the replicating portfolio idea. Since at each step of the limiting process the call's value is identical to that of the replicating portfolio, this notion must carry over to the continuous time setting. This is indeed the case: in a context when investors may continuously and costlessly adjust the composition of the replicating portfolio, the initial position to assume (at time t) is one of $N(d_1)$ shares, financed in part by a risk-free loan of $Ke^{-R_f T}N(d_2)$. The net cost of assembling the portfolio is the Black–Scholes value of the call.

Notice also that neither the mean return on the underlying asset nor the true probabilities explicitly enter anywhere in the discussion.⁶ None of this is surprising. The short explanation is simply that risk-neutral valuation abandons the true probabilities in favor of the risk-neutral ones, and, in doing so, all assets are determined to earn the risk-free rate. The underlying assets' mean return still matters, but it is now R_f . More intuitively, risk-neutral valuation is essentially no-arbitrage pricing. In a world with full information and without transaction costs, investors will eliminate all arbitrage opportunities regardless of their objective likelihood or of the mean returns of the assets involved.

It is sometimes remarked that to purchase a call option is to buy volatility, and we need to understand what this expression is intended to convey. Returning to the binomial approximation (in conjunction with Eq. (13.18)), we observe first that a larger σ implies the possibility of a higher underlying asset price at expiration, with the attendant higher call payoff. More formally, σ is the only statistical characteristic of the underlying stock's price

⁶ They are implicitly present in the equilibrium price of the underlying asset.

process to appear in the Black–Scholes formula. Given r_f , K , and $q^e(\theta_t, t)$, there is a unique identification between the call's value and σ . For this reason, estimates of an asset's volatility are frequently obtained from its corresponding call price by inverting the Black–Scholes formula. This is referred to as an *implied volatility estimate*.

The use of risk-neutral methods for the valuation of options is probably the area in which asset pricing theory has made the most progress. Indeed, Merton and Scholes were awarded the Nobel Prize for their work (Fischer Black had died). So much progress has, in fact, been made that the finance profession has largely turned away from conceptual issues in derivatives valuation to focus on the development of fast computer valuation algorithms that mimic the risk-neutral methods. This, in turn, has allowed the use of derivatives, especially for hedging purposes, to increase so enormously over the past 30 years.

13.6 Dybvig's Evaluation of Dynamic Trading Strategies

Let us next turn to a final application of these methods: the evaluation of dynamic trading strategies. To do so, we retain the partial equilibrium setting of the binomial model, but invite agents to have preferences over the various outcomes. Note that under the pure pricing perspective of [Section 13.4](#), preferences were irrelevant. All investors would agree on the prices of call and put options (and all other derivatives) regardless of their degrees of risk aversion, or their subjective beliefs as to the true probability of an up or down state. This simply reflects the fact that any rational investor, whether highly risk averse or risk neutral, will seek to profit by an arbitrage opportunity, whatever the likelihood, and that in equilibrium, assets should thus be priced so that such opportunities are absent. In this section our goal is different, and preferences will have a role to play. We return to Assumption A13.1.

Consider the optimal consumption problem of an agent who takes security prices as given and who seeks to maximize the present value of time-separable utility (A13.1). His optimal consumption plan solves

$$\begin{aligned} \max E_0 & \left(\sum_{t=0}^{\infty} U(\tilde{c}_t, t) \right) \\ \text{s.t.} & \sum_{t=0}^{\infty} \sum_{s \in N_t} q(\theta_0, \theta_t(s)) c(\theta_t(s)) \leq Y_0, \end{aligned} \tag{13.24}$$

where Y_0 is his initial period 0 wealth and $q(\theta_0, \theta_t(s))$ is the period $t = 0$ price of an Arrow–Debreu security paying one unit of the numeraire if state s is observed at

time $t > 0$. Assuming a finite number of states and expanding the expectations operator to make explicit the state probabilities, we find that the Lagrangian for this problem is

$$L(\cdot) = \sum_{t=0}^{\infty} \sum_{s=1}^{N_t} \pi(\theta_0, \theta_t(s)) U(c(\theta_t(s), t)) \\ + \lambda \left(Y_0 - \sum_{t=0}^{\infty} \sum_{s=1}^{N_t} q(\theta_0, \theta_t(s)) c(\theta_t(s)) \right),$$

where $\pi(\theta_0, \theta_t(s))$ is the conditional probability of state s occurring, at time t and λ the Lagrange multiplier.

The first-order condition is

$$U_1(c(\theta_t(s)), t) \pi(\theta_0, \theta_t(s)) = \lambda q(\theta_0, \theta_t(s)).$$

By the concavity of $U(\cdot)$, if $\theta_t(1)$ and $\theta_t(2)$ are two states, then

$$\frac{q(\theta_0, \theta_t(1))}{\pi(\theta_0, \theta_t(1))} > \frac{q(\theta_0, \theta_t(2))}{\pi(\theta_0, \theta_t(2))}, \text{ if and only if } c(\theta_t(1), t) < c(\theta_t(2), t). \quad (13.25)$$

It follows that if

$$\frac{q(\theta_0, \theta_t(1))}{\pi(\theta_0, \theta_t(1))} > \frac{q(\theta_0, \theta_t(2))}{\pi(\theta_0, \theta_t(2))}, \text{ then } c(\theta_t(1)) = c(\theta_t(2)).$$

The $q(\theta_0, \theta_t(s))/\pi(\theta_0, \theta_t(s))$ ratio measures the relative scarcity of consumption in state $\theta_t(s)$: A high ratio in some state suggests that the price of consumption is very high relative to the likelihood of that state being observed. This suggests that consumption is scarce in the high $q(\theta_0, \theta_t(s))/\pi(\theta_0, \theta_t(s))$ states. A rational agent will consume less in these states and more in the relatively cheaper ones, as Eq. (13.25) suggests.

This observation is, in fact, quite general as Proposition 13.1 demonstrates.

Proposition 13.1 (Dybvig, 1988) Consider the consumption allocation problem described by Eq. (13.24). For any rational investor for which $U_{11}(c_t, t) < 0$, his optimal consumption plan is a decreasing function of $q(\theta_0, \theta_t(s))\pi(\theta_0, \theta_t(s))$. Furthermore, for any consumption plan with this monotonicity property, there exists a rational investor with concave period utility function $U(c_t, t)$ for which the consumption plan is optimal in the sense of solving Problem (13.24).

Dybvig (1988) illustrates the power of this result most effectively in the binomial context where the price-to-probability ratio assumes an especially simple form. Recall that in the binomial model the state at time t is completely characterized by the number of up states, u ,

preceding it. Consider a state $\theta_t(s)$ where s denotes the number of preceding up states. The true conditional probability of $\theta_t(s)$ is

$$\pi(\theta_0, \theta_t(s)) = \pi^s(1 - \pi)^{t-s},$$

while the corresponding state claim has price

$$q(\theta_0, \theta_t(s)) = (R_f)^{-t}(\pi^{RN})^s(1 - \pi^{RN})^{t-s}.$$

The price/probability ratio thus assumes the form

$$\frac{q(\theta_0, \theta_t(s))}{\pi(\theta_0, \theta_t(s))} = (R_f)^{-t} \left(\frac{\pi^{RN}}{\pi} \right)^s \left(\frac{1 - \pi^{RN}}{1 - \pi} \right)^{t-s} = (R_f)^{-t} \left(\frac{\pi^{RN}(1 - \pi)}{(1 - \pi^{RN})\pi} \right)^s \left(\frac{1 - \pi^{RN}}{1 - \pi} \right)^t.$$

We now specialize the binomial process by further requiring the condition in Assumption A13.6.

A13.6 $\pi u + (1 - \pi)d > R_f$, in other words, the expected return on the stock exceeds the risk-free rate.

Assumption A13.6 implies that

$$\pi > \frac{R_f - d}{u - d} = \pi^{RN},$$

so that

$$\frac{\pi^{RN}(1 - \pi)}{(1 - \pi^{RN})} < 1,$$

for any time t , and the price probability ratio $q(\theta_0, \theta_t(s))/\pi(\theta_0, \theta_t(s))$ is a decreasing function of the number of preceding up moves, s . By Proposition 13.1 the period t level of optimal, planned consumption across states $\theta_t(s)$ is thus an *increasing function of the number of up moves, s , preceding it*.

Let us now specialize our agent's preferences to assume that he is only concerned with his consumption at some terminal date T , at which time he consumes his wealth.

Problem (13.24) easily specializes to this case:

$$\begin{aligned} & \max \sum_{s \in N_T} \pi(\theta_0, \theta_T(s)) U(c(\theta_T(s))) \\ & \text{s.t. } \sum_{s \in N_T} q(\theta_0, \theta_T(s)) c(\theta_T(s)) \leq Y_0 \end{aligned} \tag{13.26}$$

$$q_e(\theta_0, 0) = 16, \quad u = 2, \quad d = \frac{1}{2}, \quad R_f = 1$$

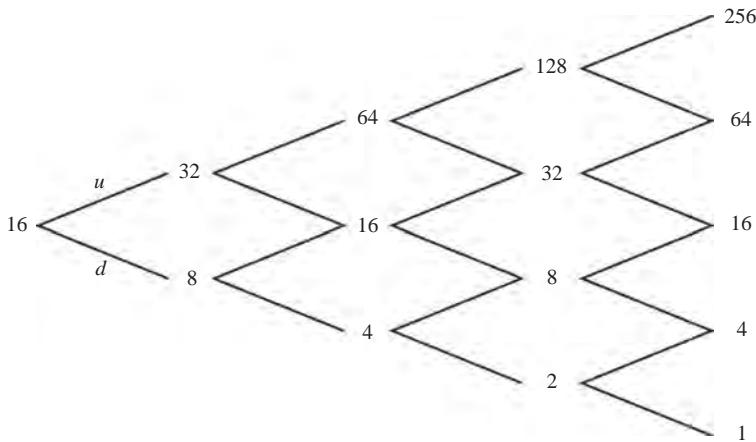


Figure 13.8
Binomial evolution of a stock's price over four periods.

In effect, we set $U(c_t, t) \equiv 0$ for $t \leq T$.

Remember also that a stock, from the perspective of an agent who is concerned only with terminal wealth, can be viewed as a portfolio of period t state claims. The results of [Proposition 13.1](#) thus apply to this security as well.

[Dybvig \(1988\)](#) shows how these latter observations can be used to assess the optimality of many commonly used trading strategies. The context of his discussion is illustrated with the example in [Figure 13.8](#) where the investor is presumed to consume his wealth at the end of the trading period.

For this particular setup, $\pi^{RN} = \frac{1}{3}$. He considers the following frequently cited equity trading strategies:

1. Technical analysis: buy the stock and sell it after an up move; buy it back after a down move; invest at R_f (zero in this example) when out of the market. But under this strategy

$$c_4(\theta_t(s)|uuuu) = \$32, \text{ yet}$$

$$c_4(\theta_t(s)|udud) = \$48; \text{ in other words,}$$

the investor consumes more in the state with the fewer preceding up moves, which violates the optimality condition. This cannot be an optimal strategy.

2. Stop-loss strategy: buy and hold the stock, sell only if the price drops to \$8, and stay out of the market. Consider, again, two possible evolutions of the stock's price:

$$c_4(\theta_t(s)|duuu) = \$8$$

$$c_4(\theta_t(s)|udud) = \$16.$$

Once again, consumption is not an increasing function of the number of up states under this trading strategy, which must, therefore, be suboptimal.

13.7 Conclusions

We have extended the notion of risk-neutral valuation to two important contexts: the dynamic setting of the general equilibrium consumption CAPM and the partial equilibrium binomial model. The return on our investment is particularly apparent in the latter framework. The reasons are clear: in the binomial context, which provides the conceptual foundations for an important part of continuous time finance, the risk-neutral probabilities can be identified independently from agents' preferences. Knowledge of the relevant intertemporal marginal rates of substitution, in particular, is superfluous. This is the huge dividend of the twin modeling choices of binomial framework and arbitrage pricing. It has paved the way for routine pricing of complex derivative-based financial products and for their attendant use in a wide range of modern financial contracts.

References

- Cox, J., Rubinstein, M., 1979. Option Markets. Prentice-Hall, Upper Saddle River, NJ.
- Detemple, J., Jorion, P., 1990. Option listing and stock returns: an empirical analysis. *J. Bank. Financ.* 14, 781–801.
- Detemple, J., Selden, L., 1991. A general equilibrium analysis of option and stock market interactions. *Int. Econ. Rev.* 32, 279–303.
- Dybvig, P.H., 1988. Inefficient dynamic portfolio strategies or how to throw away a million dollars in the stock market. *Rev. Financ. Stud.* 1, 67–88.
- Telmer, C., 1993. Asset pricing puzzles and incomplete markets. *J. Financ.* 48, 1803–1832.

Appendix 13.1: Risk-Neutral Valuation When Discounting at the Term Structure of Multiperiod Discount Bond

Here we seek a valuation formula where we discount not at the succession of one-period rates but at the term structure. This necessitates a different set of risk-neutral probabilities with respect to which the expectation is taken.

Define the k -period, time-adjusted risk-neutral transition probabilities as:

$$\hat{\pi}^{RN}(\theta_t, \theta_{t+k}) = \left\{ \frac{\pi^{RN}(\theta_t, \theta_{t+k})g(\theta_t, \theta_{t+k})}{q^b(\theta_t, \theta_{t+k})} \right\}$$

where $\pi^{RN}(\theta_t, \theta_{t+k}) = \prod_{s=t}^{t+k-1} \pi^{RN}(\theta_t, \theta_{s+1})$, and $\{\theta_t, \dots, \theta_{t+k-1}\}$ is the path of states preceding θ_{t+k} . Clearly, the $\hat{\pi}^{RN}()$ are positive since $\pi^{RN}() \geq 0$, $g(\theta_t, \theta_{t+k}) > 0$ and $q^b(\theta_t, \theta_{t+k}) > 0$. Furthermore, by Eq. (13.13),

$$\begin{aligned} \sum_{\theta_{t+k}} \hat{\pi}^{RN}(\theta_t, \theta_{t+k}) &= \left(\frac{1}{q^b(\theta_t, \theta_{t+k})} \right) \sum_{\theta_{t+k}} \pi^{RN}(\theta_t, \theta_{t+k}) g(\theta_t, \theta_{t+k}) \\ &= \frac{q^b(\theta_t, \theta_{t+k})}{q^b(\theta_t, \theta_{t+k})} = 1. \end{aligned}$$

Let us now use this approach to price European call and put options. A European call option contract represents the right (but not the obligation) to buy some underlying asset at some prespecified price (referred to as the exercise or strike price) at some prespecified future date (date of contract expiration). Since such a contract represents a right, its payoff is as shown in Table A13.1, where T represents the time of expiration and K the exercise price.

Table A13.1: Payoff pattern—European call option

t	$t + 1$	$t + 2$...	$T - 1$	T
0	0	0		0	$\max\{q^e(\theta_T, T) - K, 0\}$

Let $C_e(\theta_t, t)$ denote the period t , state θ_t price of the call option. Clearly,

$$\begin{aligned} C_e(\theta_t, t) &= E_t^{RN}\{g(\theta_t, \tilde{\theta}_T)(\max\{q^e(\tilde{\theta}_t, T) - K, 0\})\} \\ &= q^b(\theta_t, T)\hat{E}_t^{RN}\{\max\{q^e(\tilde{\theta}_t, T) - K, 0\}\}, \end{aligned}$$

where \hat{E}_t^{RN} denotes the expectations operator corresponding to the $\{\pi^{RN}\}$. A European put option is the right to sell some underlying asset at some prespecified price (exercise price K) at some prespecified future date T . It is similarly priced according to

$$\begin{aligned} P_e(\theta_t, t) &= E_t^{RN}\{g(\theta_t, \tilde{\theta}_T)(\max\{K - q^e(\tilde{\theta}_t, T), 0\})\} \\ &= q^b(\theta_t, T)\hat{E}_t^{RN}\{\max\{K - q^e(\tilde{\theta}_T, T), 0\}\} \end{aligned}$$

The Arbitrage Pricing Theory

Chapter Outline

14.1 Introduction	417
14.2 Factor Models: A First Illustration	419
14.2.1 Using the Market Model	420
14.3 A Second Illustration: Multifactor Models, and the CAPM	421
14.4 The APT: A Formal Statement	424
14.5 Macroeconomic Factor Models	426
14.6 Models with Factor-Mimicking Portfolios	428
14.6.1 The Size and Value Factors of Fama and French (1993)	428
14.6.2 Momentum Portfolios	434
14.7 Advantage of the APT for Stock or Portfolio Selection	436
14.8 Conclusions	437
References	437
Appendix A.14.1: A Graphical Interpretation of the APT	438
Statement and Proof of the APT	439
The CAPM and the APT	441
Appendix 14.2: Capital Budgeting	441

14.1 Introduction

We have already made two attempts (Chapters 11–13) at asset pricing from an arbitrage perspective, i.e., without specifying a complete equilibrium structure. Here we try again from a different, more empirical angle. Before doing so, let us first collect a few thoughts as to the differences between an arbitrage approach and equilibrium modeling.

In the context of general equilibrium theory, we make hypotheses about agents—consumers, producers, investors; in particular, we start with some form of rationality hypothesis leading to the specification of maximization problems under constraints. We also make hypotheses about markets: typically, we assume that supply equals demand in all markets under consideration, and that the markets are competitive.

We have repeatedly used the fact that in general equilibrium with fully informed optimizing agents, there can be no-arbitrage opportunities—in other words, no possibilities

to make money risklessly at zero cost. An arbitrage opportunity indeed implies that at least one agent can reach a higher level of utility without violating his or her budget constraint (since there is no extra cost to do so).

In particular, our assertion that one can price any asset (income stream) from the knowledge of Arrow–Debreu prices relied implicitly on a no-arbitrage hypothesis: with a complete set of Arrow–Debreu securities, some portfolio thereof will exactly replicate the asset’s cash-flow stream. If there are to be no-arbitrage opportunities, the market price of the asset and the value of the replicating portfolio of Arrow–Debreu securities must therefore be the same. An arbitrageur could otherwise make arbitrarily large profits by short selling large quantities of the more expensive of the two and buying the cheaper in equivalent amounts. Such an arbitrage portfolio would have zero cost and be riskless.

While general equilibrium implies the no-arbitrage condition, it is more restrictive in the sense of imposing a heavier structure on modeling. And the reverse implication is *not* true: no-arbitrage opportunities—the fact that all arbitrage opportunities have been exploited—does not imply that a general equilibrium in all markets has been obtained.¹ Nevertheless, or precisely for that reason, it is interesting to see how far one can go in exploiting the less restrictive hypothesis that no-arbitrage opportunities are left unexploited.

The underlying logic of the arbitrage pricing theory (APT), to be reviewed in this chapter is, in a sense, very similar to the fundamental logic of the Arrow–Debreu model, and it is very much in the spirit of a complete markets structure. It distinguishes itself in two major ways: first it replaces the underlying structure based on fundamental securities paying exclusively in a given state of nature with other fundamental securities exclusively remunerating some form of risk taking. More precisely, the APT abandons the analytically powerful, but empirically cumbersome, concept of states of nature as the basis for the definition of its primitive securities. It replaces it with the hypothesis that there exists a (stable) set of factors (random quantities) that are essential and exhaustive determinants of all asset returns. Factors may be macroeconomic in nature (e.g., the random growth rate of gross domestic product, GDP), or behavioral (e.g., the momentum factor to be considered shortly). Typically factors may be separated as the return distributions on specific financial assets. A primitive security, in a factor context, will then be defined as a security whose risk is exclusively determined by its association with one specific underlying risk factor and which is totally statistically independent of (unaffected by) any other risk factor. Its expected return premium thus represents compensation to investors for bearing the factor specific risk. The second difference from the Arrow–Debreu pricing of Chapter 9 is that the prices of the primitive securities are not derived from fundamentals—supply and

¹ An arbitrage portfolio is a self-financing (zero net-investment) portfolio. An arbitrage opportunity exists if an arbitrage portfolio exists that yields nonnegative cash flows in all states of nature and positive cash flows in some states (Chapter 11).

demand, themselves resulting from agents' endowments and preferences—but will be deduced empirically from observed asset returns without attempting to identify their underlying macroeconomic or behavioral determinants. Once the price of each fundamental security has been inferred from observed return distributions, the usual arbitrage argument applied to the pricing of complex securities will be made (in the spirit of Chapter 12).²

14.2 Factor Models: A First Illustration

The main building block of the APT is a factor model, the notion of which we introduced back in Chapter 2. As discussed previously, this is the structure that replaces the concept of states of nature. The motivation has been evoked before: states of nature are analytically powerful abstractions for valuation. In practice, however, they are difficult to work with and, moreover, often not verifiable, implying that contracts cannot necessarily be written contingent on a specific state of nature being realized. We discussed these shortcomings of the Arrow–Debreu pricing theory in Chapter 9. The temptation is thus irresistible to attack the asset pricing problem from the opposite angle and build the concept of primitive securities on an empirically more operational notion, at a cost of abstracting from its potential theoretical credentials. This structure is what factor models accomplish.

The simplest conceivable factor model is a one-factor market model, which asserts that *ex post* returns on individual assets can be entirely ascribed either to their own specific stochastic components or to their common association with a single factor. This simple factor model can thus be summarized by following the relationship:³

$$\tilde{r}_j = \alpha_j + \beta_j \tilde{f}_1 + \tilde{\varepsilon}_j \quad (14.1)$$

with $E\tilde{\varepsilon}_j = 0$, $\text{cov}(\tilde{f}_1, \tilde{\varepsilon}_j) = 0$, $\forall j$, and $\text{cov}(\tilde{\varepsilon}_j, \tilde{\varepsilon}_k) = 0$, $\forall j \neq k$.

This model states that there are three components in individual returns: (1) an asset-specific constant α_j ; (2) a common influence, in this case the unique factor, \tilde{f}_1 , a random quantity, which affects all assets in varying degrees, with β_j measuring the sensitivity of asset j 's return to fluctuations in the single factor; and (3) an asset-specific stochastic term $\tilde{\varepsilon}_j$ summarizing all other stochastic components of \tilde{r}_j that are unique to asset j . Equation (14.1) has no bite (such an equation can always be written) until one adds the hypothesis $\text{cov}(\tilde{\varepsilon}_j, \tilde{\varepsilon}_k) = 0$, $j \neq k$, which signifies that *all* return characteristics common to different assets are subsumed in their link with the factor \tilde{f}_1 .

² The arbitrage pricing theory was first developed by Ross (1976) and substantially interpreted by Huberman (1982) and Connor (1984) among others. For a presentation emphasizing practical applications, see Burmeister et al. (1994).

³ Factors are frequently measured as deviations from their mean. When this is the case, α_j becomes an estimate of the mean return on asset j .

From our study of the capital asset pricing model (CAPM), it is natural to identify \tilde{f}_1 with \tilde{r}_M , the return on the market portfolio. With this identification, Eq. (14.1) is referred to as the “market model.” If the market model were to be empirically verified, then the CAPM would be the undisputed endpoint of asset pricing. As the APT does not require the assumptions of the CAPM, it does not, *a priori*, identify \tilde{f}_1 with \tilde{r}_M .

From the empirical perspective, one may say that it is quite unlikely that a single-factor model will suffice.⁴ But the strength of the APT is that it is agnostic as to the number of underlying factors (and to their identity). As we increase the number of factors, hoping that this will not require a number too large to be operational, a generalization of Eq. (14.1) becomes more and more plausible. For pedagogical purposes, let us for the moment maintain the hypothesis of one common factor which we identify with (some proxy for) the market portfolio.⁵

14.2.1 Using the Market Model

Besides serving as a potential basis for the APT, the market model, despite all its weaknesses, is also of interest on two other grounds. First it produces estimates for the β 's that play a central role in the CAPM. Note, however, that estimating β 's from past data alone is useful only to the extent that some degree of stationarity in the relationship between asset returns and the return on the market is present. Empirical observations suggest a fair amount of stationarity is plausible at the level of portfolios, but not of individual assets. On the other hand, estimating the β 's does not require all the assumptions of the market model; in particular, a violation of the $\text{cov}(\tilde{\varepsilon}_i, \tilde{\varepsilon}_k) = 0, i \neq k$ hypothesis is not damaging.

The second source of interest in the market model, crucially dependent upon the latter hypothesis being approximately valid, is that it permits economizing on the computation of the matrix of variances and covariances of asset returns at the heart of the MPT. Indeed, under the market model hypothesis, one can write (you are invited to prove these statements):

$$\sigma_j^2 = \beta_j^2 \sigma_M^2 + \sigma_{\varepsilon_j}^2, \quad \forall j$$

$$\sigma_{ij} = \beta_i \beta_j \sigma_M^2$$

This effectively means that the information requirements for the implementation of MPT can be substantially weakened. Suppose there are N risky assets under consideration.

⁴ Recall the difficulty in constructing the empirical counterpart of M .

⁵ Fama (1973), however, demonstrates that in its form (14.1) the market model is inconsistent in the following sense: the fact that the market is, by definition, the collection of all individual assets implies an exact linear relationship between the disturbances ε_j ; in other words, when the single factor is interpreted to be the market, the hypothesis $\text{cov}(\tilde{\varepsilon}_j, \tilde{\varepsilon}_k) = 0, \forall j \neq k$ cannot be strictly valid. While we ignore this criticism in view of our purely pedagogical objective, it is a fact that if a single-factor model had a chance of being empirically verified (in the sense of all the assumptions in Eq. (14.1) being confirmed), the unique factor could not be the market.

In that case, the computation of the efficient frontier requires knowledge of N expected returns, N variances, and $(N^2 - N)/2$ covariance terms. (N^2 is the total number of entries in the matrix of variances and covariances; take away the N variance/diagonal terms, and divide by 2 since $\sigma_{ij} = \sigma_{ji}$, $\forall i, j$.)

Working via the market model, on the other hand, requires estimating Eq. (14.1) for the N risky returns, producing estimates for the $N \beta_j$'s and the $N \sigma_{\tilde{\varepsilon}_j}^2$ and estimating the variance of the market return. This represents $2N + 1$ information items, many fewer than under the previous case.

14.3 A Second Illustration: Multifactor Models, and the CAPM

The APT approach is generalizable to any number of factors. It does not, however, provide any clue as to what these factors should be or any particular indication as to how they should be selected. This is both its strength and its weakness. Suppose we can agree on a two-factor model:

$$\tilde{r}_{j,t} = \alpha_j + b_{j,1} \tilde{f}_t^1 + b_{j,2} \tilde{f}_t^2 + \tilde{\varepsilon}_{j,t} \quad (14.2)$$

with $E\tilde{\varepsilon}_j = 0$, $\text{cov}(\tilde{f}_t^1, \tilde{\varepsilon}_j) = \text{cov}(\tilde{f}_t^2, \tilde{\varepsilon}_j) = 0$, $\forall j$, $\text{cov}(\tilde{f}_t^1, \tilde{f}_t^2) = 0$ and $\text{cov}(\tilde{\varepsilon}_k, \tilde{\varepsilon}_j) = 0$, $\forall j \neq k$.

As was the case for Eq. (14.1), Eq. (14.2) implies that one cannot reject, empirically, the hypothesis that the *ex post* return on an asset j has two stochastic components: one specific, $(\tilde{\varepsilon}_j)$, and one systematic, $(b_{j,1} \tilde{f}_t^1 + b_{j,2} \tilde{f}_t^2)$. What is new is that the systematic component is not viewed as the result of a single common factor influencing all assets as in the market model. Common or systematic issues may now be traced to two fundamental factors affecting, in varying degrees, the returns on individual assets (and thus on portfolios as well). Without loss of generality, we may assume that these factors are uncorrelated. As before, an expression such as Eq. (14.2) is useful only to the extent that it describes a relationship that is relatively stable over time. The two factors \tilde{f}_t^1 and \tilde{f}_t^2 must really summarize *all* that is common in individual asset returns. What could these fundamental factors be?

Much of the remainder of this chapter is devoted to presenting the answer the literature provides to this question. For the present conceptual discussion, let us assume that \tilde{f}_t^1 is the price of energy and that it is perfectly proxied by the value-weighted return on a portfolio of international oil stocks:

$$\tilde{f}_t^1 = \tilde{f}_t^{P_1}$$

Let \tilde{f}_t^2 be a measure of the stock market's risk sensitivity and let us suppose that it is well proxied by a long position in a portfolio (defaultable) of Baa bonds financed by a short position in (default-free) 10-year US Treasury notes:

With these definitions, we can write Eq. (14.2) as

$$\tilde{r}_{j,t} = \alpha_j + b_{j,1}\tilde{r}_t^{P_1} + b_{j,2}\tilde{r}_t^{P_2} + \tilde{\epsilon}_{j,t} \quad (14.3)$$

Viewing Eq. (14.3) as a regression equation, the $b_{j,k}$, $k = 1, 2$ coefficients are referred to as factor loadings. Let us assume (falsely, as it turns out) that $\text{cov}(\tilde{r}_t^{P_1}, \tilde{r}_t^{P_2}) = \text{cov}(\tilde{r}_t^{P_i}, \tilde{\epsilon}_{j,t}) = \text{cov}(\tilde{\epsilon}_{j,t}, \tilde{\epsilon}_{i,t}) = 0$, $j \neq i, i, j \in \{1, 2\}$, and that a portfolio is also traded with zero sensitivity to either factor and zero asset-specific risk (i.e., a risk-free asset).

The APT then states that there exist scalars $\lambda_0, \lambda_1, \lambda_2$ such that

$$\bar{r}_j = \lambda_0 + \lambda_1 b_{j,1} + \lambda_2 b_{j,2} \quad (14.4)$$

That is, the expected return on an arbitrary asset j is perfectly and completely described by a linear function of asset j 's factor loadings $b_{j,1}, b_{j,2}$. This can appropriately be viewed as a (two-factor) generalization of the security market line (SML).

Furthermore, the coefficients of the linear function are

$$\begin{aligned}\lambda_0 &= r_f \\ \lambda_1 &= \bar{r}_{P_1} - r_f \\ \lambda_2 &= \bar{r}_{P_2} - r_f.\end{aligned}$$

The reader will note the resemblance of Eqs. (14.3) and (14.4) to the results of the Fama and MacBeth (1973) methodology (Chapter 8) and herein lies the manner by which a proposed set of factors can be tested as to their efficacy.

The APT agrees with the CAPM that the risk premium on an asset, $\bar{r}_j - \lambda_0$, is not a function of its specific or diversifiable risk. It potentially disagrees with the CAPM in the identification of the systematic risk. The APT decomposes the systematic risk into elements of risk associated with a particular asset's sensitivity to a few fundamental common factors.

Note also the parallels with the Arrow–Debreu pricing approach. In both contexts, every individual asset or portfolio can be viewed as a complex security, or a combination of primitive securities: Arrow–Debreu securities in one case, the pure-factor portfolios in the other. Once the prices of the primitive securities are known, it is a simple step to compose replicating portfolios and, by a no-arbitrage argument, price complex securities and arbitrary cash flows. The difference, of course, resides in the identification of the primitive security. Although the Arrow–Debreu approach sticks to the conceptually clear notion of states of nature, the APT takes the position that there exist a few common and stable sources of risk and that they can be empirically identified. Once the corresponding

risk premia are identified, by observing the market-determined premia on the primitive securities (the portfolios with unit sensitivity to a particular factor and zero sensitivity to all others), the pricing machinery can be put to work.

Let us illustrate. In our two-factor examples, a security j with, say, $b_{j,1} = 0.8$ and $b_{j,2} = 0.4$ is like a portfolio with proportions of 0.8 of the pure portfolio P_1 , 0.4 of pure portfolio P_2 , and consequently proportion -0.2 in the riskless asset. By our usual (no-arbitrage) argument, the expected rate of return on that security must be

$$\begin{aligned}\bar{r}_j &= -0.2r_f + 0.8\bar{r}_{P_1} + 0.4\bar{r}_{P_2} \\ &= -0.2r_f + 0.8r_E + 0.4r_f + 0.8(\bar{r}_{P_1} - r_f) + 0.4(\bar{r}_{P_2} - r_f) \\ &= r_f + 0.8(\bar{r}_{P_1} - r_f) + 0.4(\bar{r}_{P_2} - r_f) \\ &= \lambda_0 + b_{j,1}\lambda_1 + b_{j,2}\lambda_2\end{aligned}$$

The APT equation can thus be seen as the immediate consequence of the linkage between pure-factor portfolios and complex securities in an arbitrage-free context. The reasoning is directly analogous to our derivation of the value additivity theorem in Chapter 11 and leads to a similar result: Diversifiable risk is not priced in a complete (or quasi-complete) market world.

Though potentially more general, the APT does not necessarily contradict the CAPM. That is, it may simply provide another, more disaggregated, way of writing the expected return premium associated with systematic risk, and thus a decomposition of the latter in terms of its fundamental elements. Clearly, the two theories have the same implications if (keeping with our two-factor model, the generalization is trivial):

$$\beta_j(\bar{r}_m - r_f) = b_{j,1}(\bar{r}_{P_1} - r_f) + b_{j,2}(\bar{r}_{P_2} - r_f) \quad (14.5)$$

Let β_{P_1} be the (market) beta of the pure portfolio P_1 and similarly for β_{P_2} . Then if the CAPM is valid, not only is the LHS of Eq. (14.5) the expected risk premium on asset j , but we also have

$$\begin{aligned}\bar{r}_{P_1} - r_f &= \beta_{P_1}(\bar{r}_M - r_f) \\ \bar{r}_{P_2} - r_f &= \beta_{P_2}(\bar{r}_M - r_f)\end{aligned}$$

Thus, the APT expected risk premium may be written as

$$b_{j,1}[\beta_{P_1}(\bar{r}_M - r_f)] + b_{j,2}[\beta_{P_2}(\bar{r}_M - r_f)] = (\beta_{j,1}\beta_{P_1} + \beta_{j,2}\beta_{P_2})(\bar{r}_M - r_f)$$

which is the CAPM equation provided:

$$\beta_j = b_{j,1}\beta_{P_1} + b_{j,2}\beta_{P_2}$$

In other words, CAPM and APT have identical implications if the sensitivity of an arbitrary asset j with the market portfolio fully summarizes its relationship with the two underlying common factors. In that case, the CAPM would be another, more synthetic, way of writing the APT.⁶

In reality, of course, there are reasons to think that the APT with an arbitrary number of factors will always do at least as well in identifying the sources of systematic risk as the CAPM. Indeed, this will generally be the case. But first, a formal statement of the APT.

14.4 The APT: A Formal Statement

The APT assumes a model of the financial markets that is both frictionless and competitive and in which the returns to each traded asset i are governed by a K factor structure of the following form:

- i. $\tilde{R}_i = E\tilde{R}_i + \mathbf{b}_i^T \tilde{\mathbf{f}} + \tilde{\varepsilon}_i$ where (14.6i)

$\tilde{\mathbf{f}} = [\tilde{f}^1, \tilde{f}^2, \tilde{f}^3, \dots, \tilde{f}^K]$ is a $K \times 1$ vector of random factors, and \mathbf{b}_i^T is a $K \times 1$ vector of factor sensitivities (constants) for asset i where

- ii. $E\tilde{f} = 0, E\tilde{f}^i \tilde{f}^j = 0, \text{ for all } i, j,$ (14.6ii)

$$E\tilde{\varepsilon}_i \tilde{f} = 0 \text{ and } E\tilde{\varepsilon}_i = 0 \text{ for all } i.$$

The APT also assumes that a very large number of such assets is traded in the sense that for well-diversified portfolios with weights $\approx (1/N)$, where N is large, $\tilde{\varepsilon}_P = \sum_{i=1}^N (1/N) \tilde{\varepsilon}_i \approx 0$. Although not specifically required by the APT, we will further assume the financial markets are complete and that a risk-free asset is traded.

Given the above structure, [Ross \(1976\)](#) demonstrates that if there are no-arbitrage opportunities, then there exists a $K \times 1$ vector of factor risk premia λ_K such that for any asset i ,

$$E\tilde{R}_i - R_f \approx \mathbf{b}_i^T \boldsymbol{\lambda}_K \quad (14.7)$$

Expressions (14.6i, ii) and their consequence (14.7) constitute the APT.

Since the conclusion of the APT is an approximation, expression (14.7), it does not directly provide testable implications for asset returns.⁷ Within the context of a full equilibrium model, however, [Connor \(1984\)](#) details additional requirements such that Eq. (14.7) can be

⁶ The observation in footnote 5, however, suggests this could be true as an approximation only.

⁷ Note that Eqs. (14.6i, ii) and (14.7) suggest the Fama and MacBeth (1973) procedure of Section 8.9.1.

expressed as an equality (exact factor pricing).⁸ Alternatively, Ross (1976) demonstrates that the approximation (14.7) becomes increasingly more accurate as the number of assets increases, while Grinblatt and Titman (1985), arguing from a different angle, conclude that deviations from exact factor pricing are likely to be very small. Taken together, these studies suggest that empirical exercises based on Eq. (14.7), assumed to hold with equality, can be justified from several perspectives.

To interpret Eq. (14.7) more fully, consider a well-diversified portfolio i (well diversified in the sense of Connor (1984)), for which idiosyncratic risk has been entirely diversified away so that

$$\tilde{R}_i = E\tilde{R}_i + \mathbf{b}_i^T \tilde{f} \quad (14.8)$$

Since we have assumed the asset market is complete we know there exists a unique SDE, \tilde{m} , within the space of payoffs, that allows us to price all assets.⁹

Accordingly, Eq. (14.8) allows us to write

$$\tilde{m}\tilde{R}_i = \tilde{m}E\tilde{R}_i + \tilde{m}\mathbf{b}_i^T \tilde{f}, \text{ and, by Theorem 10.3:}$$

$$1 = E\tilde{m}\tilde{R}_i = E\tilde{R}_i E\tilde{m} + \mathbf{b}_i^T E\tilde{m}\tilde{f}. \text{ Thus,}$$

$$1 = \frac{E\tilde{R}_i}{R_f} + \mathbf{b}_i^T E\tilde{m}\tilde{f}, \text{ and}$$

$$E\tilde{R}_i - R_f = E\tilde{r}_i - r_f = \mathbf{b}_i^T [-R_f \pi(\tilde{f})]$$

where we define $-R_f \pi(\tilde{f}) = \lambda_K$, and interpret $\pi(\tilde{f})$ as defining the $K \times 1$ vector of factor risk premia.

As we have noted earlier, the APT is not really an economic model in the sense of our discussions in Chapters 3–10. Its only claim is that there exists a factor structure that captures all stock returns. Nothing is said as to the identity of the factors \tilde{f} nor even their number. In this sense, the APT is silent as to the fundamental underlying source of asset premia and return comovement. The conclusion to the theory does not even guarantee that all factor risk premia are strictly positive. Indeed, the APT is basically a generalization of the “market model” of Section 14.2.1 that allows a parsimonious structuring of the variance–covariance matrix of returns. Accordingly, without a theoretical discipline, there has arisen a large literature proposing a variety of factors for explaining returns.

⁸ Connor (1984) requires, in particular, that the universe of assets under consideration is very large and that no single asset accounts for more than a trivial proportion of the economy’s total wealth.

⁹ See Appendix 14.1 for a graphical interpretation of the APT.

Roughly speaking these models are of two types, those that employ macroeconomic, business-cycle-related factors (e.g., the growth rate of GDP), and those whose factors are returns on the so-called factor-mimicking portfolios. In the latter case, the portfolio return patterns are presumed to capture some important and variable underlying feature of the securities markets. It may be macroeconomic or behavioral. The connection here is generally far from transparent; the important thing, most profoundly in the eyes of practitioners, is that the factors “work,” meaning that they greatly assist in explaining, statistically, a wide class of portfolio returns. We explore macro factor models in [Section 14.5](#), while [Section 14.6](#) reviews models based on factor-mimicking portfolios. Most of the attention in the literature has focused on the latter approach because of the bountiful availability of financial return data: model performance can be assessed at monthly or even daily return frequencies. Macroeconomic data (e.g., GDP growth) is frequently available only at quarterly frequencies at best, and even so, is open to substantial subsequent revision. Of necessity, macro models therefore take a longer-term perspective.

Especially as regards the factor-mimicking portfolios, the reader should keep in mind that these portfolios’ return patterns reflect a specific type of risk for which investors should, in equilibrium, receive compensating average return risk premia. Accordingly, factor-mimicking portfolios should display positive average returns appropriate to the undiversifiable risks they represent.

14.5 Macroeconomic Factor Models

A few introductory remarks are in order here in order to set the stage. First, the focus of this literature is to explain the cross section of equity returns using the Fama and MacBeth (1973) methodology (see [Section 8.9.1](#)). More specifically, it is the cross section of equity *portfolio* returns, where some *a priori* rule is adopted that assigns each individual stock to one of a limited number of well-defined portfolios.¹⁰ Recently, the emphasis in the literature has been to explain the cross-sectional return patterns for the 25 [Fama and French \(1993\)](#) portfolios whose construction we detail in the next section, asking our readers’ forbearance until then. A pervasive aspect to these studies is also to assess whether the addition of the “market factor,” $\tilde{r}_M - r_f$, significantly improves the ability of the model under study to explain the cross section of returns beyond a corresponding model that is solely macro factor or factor-mimicking-portfolio based.

Lastly, it is unreasonable to presume that causality runs only from exogenous macroeconomic events to equity return patterns. As has become abundantly clear from the

¹⁰ Recall that under the Fama and MacBeth (1973) methodology, the first pass beta regression estimates are much more precisely estimated when portfolios are the objects of discussion.

events of the recent financial crisis, financial events can have monumental consequences for the macroeconomy. In particular, the loss in the US aggregate stock market valuation, in the 2008–2009 acute phase of the crisis, undoubtedly influenced the subsequent US GDP growth rate via its effects on consumption demand (via the wealth effect) and investment demand (via increased future cash-flow uncertainty).

Since asset prices and returns are jointly determined by discount factors and expected cash-flow streams (recall the discussion in Chapter 10), each macro factor must be plausibly related to at least one of these two quantities. [Chen et al. \(1986\)](#) represents the first widely studied macro factor model. Its focus is to assess the ability of a class of macro factors to explain the cross sections of returns to 20 equally weighted portfolios which represent a partition of all NYSE stocks according to market value ranking. After exploring the explanatory power of a wide class of factors, [Chen et al. \(1986\)](#) identify the following as most significant:

cash flow related:

1. $g_t^{\text{IP}} = \ln \text{IP}_t - \ln(\text{IP}_{t-1})$, where IP_t is an index of the level of industrial production in period t and g_t^{IP} is its (continuously compounded) growth rate,

discount factor related:

1. UPR_t = unanticipated risk premia in period t
 $= (\text{Baa and lower corporate bond portfolio return in period } t) - \text{LTG}_t$, where LTG_t is the period t return on a portfolio of (default-free) long-term US Treasury bonds.
2. TS_t = the slope of the term structure
 $= \text{LTG}_t - \text{TB}_{t-1}$, where LTG_t is measured as in (1) immediately above and TB_{t-1} is the Treasury bill (short rate) in period $t-1$.
3. a measure of unanticipated inflation,
 $\text{INF}_t = \text{INF}_t - E_{t-1}\text{INF}_t$,
where INF_t is *ex post* inflation from period $t-1$ to t and $E_{t-1}\text{INF}_t$ is expected inflation for period t conditional on information at the close of period $t-1$.

Although this latter quantity was somewhat indirectly measured in [Chen et al. \(1986\)](#), with the advent of Treasury Inflation Protected Securities (TIPS), it would be natural to measure $E_{t-1}\text{INF}_t$ as

$$E_{t-1}\text{INF}_t = r_{t-1}^{\text{1-year nominal Treasury}} - r_{t-1}^{\text{1-year TIPS bond}}$$

When added to this set of explanatory variables, [Chen et al. \(1996\)](#) found that the market index was statistically insignificant in its contribution to explaining the cross section of returns. In what is a somewhat surprising outcome vis-à-vis the consumption capital asset pricing model (CCAPM) theory of Chapter 10, the growth rate of real consumption per

capita was similarly insignificant. The reader is referred to the article itself for the numerous details associated with the construction of the aforementioned factors.

Other macro factors have been proposed in the recent asset pricing literature. Jagannathan and Wang (1996) employ the growth rate of labor income as a proxy for the return on human capital which they propose as an important systematic risk factor. They also employ a measure of the default premium on corporate debt as a measure of the stage of the business cycle. Santos and Veronesi (2006) explore the influence of the labor income/consumption ratio, while Lettau and Ludvigson (2001) detail the consequences of including the consumption/wealth ratio.

Total factor productivity (TFP) risk (see web-related chapter) and various demographic risk measures have also been suggested as macro factors. TFP shocks directly influence the return on capital, and, because of their well-known high intertemporal persistence, can be expected to influence this return over an extended number of years. Demographic risks are potentially related to the fraction of the population in retirement in contrast to the fraction which is saving for retirement. Since retired persons are typically selling financial assets to finance their retirement consumption, as their fraction of the population increases (as is presently the case in most developed countries), lower equity prices and returns may result. See Abel (2001) and Geanakoplos et al. (2004).

14.6 Models with Factor-Mimicking Portfolios

As noted previously, the APT is silent as to the identity of the factors underlying common stock returns. From the discussion in Chapter 8, however, we are aware that (BE/ME) and “size” (market value of all common equity) are two quantities with substantial ability to explain the cross section of returns, and it is thus natural to propose “factors” related to them. [Fama and French \(1993\)](#) were among the first to take up this agenda.

14.6.1 The Size and Value Factors of [Fama and French \(1993\)](#)

To construct these factors, [Fama and French \(1993\)](#) first sort the universe of NYSE, AMEX, and NASDAQ stocks into (first sorting) three (B/ME) value-weighted portfolios as ranked lowest to highest with (second sorting) these portfolios then subdivided into the half with the higher ME values, and the other half with the lower ME values ([Table 14.1](#)).¹¹

¹¹ This is the same collection of stocks (and their return histories) as is used in [Fama and French \(1992\)](#); see Section 8.9.1.

Table 14.1: Stock sort underlying the Fama and French (1993) factor construction

Biggest (B) ½ of stocks ranked by ME	BL	BM	BH
Smallest (S) ½ of stocks ranked by ME	SL	SM	SH
	Lowest (L) 30% of stocks as ranked by (BE/ME)	Middle (M) 40% of stocks as ranked by (BE/ME)	Highest (H) 30% of stocks as ranked by (BE/ME)

The Fama–French “size” and “value” factors are then constructed from long and short positions in these six portfolios as follows:

$$\begin{aligned} \text{The “size” factor} &= \text{SMB}_t =_{\text{def}} ((1/3)r_t^{\text{SH}} + (1/3)r_t^{\text{SM}} + (1/3)r_t^{\text{SL}}) \\ &\quad - ((1/3)r_t^{\text{BH}} + (1/3)r_t^{\text{BM}} + (1/3)r_t^{\text{BL}}) \\ \text{The “value” factor} &\equiv \text{HML}_t = \text{def}((1/2)r_t^{\text{SH}} + (1/2)r_t^{\text{BH}}) \\ &\quad - ((1/2)r_t^{\text{SL}} + (1/2)r_t^{\text{BL}}). \end{aligned}$$

In these definitions r_t^{ij} refers to the net return on portfolio ij , where $i \in \{S, B\}$, $j \in \{L, M, H\}$ and the “minus” sign denotes a long position financed by a short position. The SMB and HML factor titles are acronyms for “small minus (i.e., short) big” and “high (BE/ME) minus low,” respectively. More precisely, a \$100,000 SMB portfolio, for example, is composed of a \$100,000 long position in the portfolio defined by $w_{SB} = w_{SM} = w_{SL} = 1/3$, a \$100,000 short position in the portfolio defined by $w_{BH} = w_{BM} = w_{BL} = 1/3$, and \$100,000 invested in risk-free assets. The return $\text{SMB}_t - r_f$ thus represents the excess return on this portfolio; similarly for HML_t .¹² By forming long–short portfolios of these types, the small firm premium and value premium are strengthened and made more sensitive to business cycle variation as they are fundamentally highly leveraged entities. As constructed, SMB and HML are pure-factor-mimicking portfolios, although it is not clear at this point what fundamental economic determinants they represent.

Fama and French (1993) evaluate their three-factor model $\text{SMB}_t, \text{HML}_t, r_t^M - r_f$ (the market factor) as regards its ability to explain the cross-sectional pattern of returns across a universe of 25 specifically constructed, value-weighted portfolios. Since it has become commonplace in the literature to judge a model as regards its ability to explain the cross section of returns on these specific 25 portfolios, it behooves us to become acquainted with their construction.

The assignment protocol uses the same data set and is in the same spirit as their factor construction except that stocks are sorted at a higher level of refinement: at the start of a year every stock in the above universe is ranked by market value (“size”) from lowest to highest, and assigned to one of five portfolios on this basis. In particular, the first quintile portfolio contains the lowest fifth of all stocks ranked by size; the top quintile contains the highest

¹² In the asset pricing literature financial factors are always constructed in this way.

Table 14.2: The Fama-French portfolios⁽ⁱ⁾

Highest ME quintile					(5,5)
				(4,3)	
Lowest ME quintile					
	(1,1)		(3,1)		(1,5)
	Lowest (BE/ME) quintile				Highest (BE/ME) quintile

⁽ⁱ⁾entry (i, j) denotes the ith (BE/ME) quintile with the Jth quintile ME stocks

fifth, with all other stocks assigned to the remaining three quintile portfolios in an identical manner. The stocks are then reranked into five subquintiles based on their ascending B/ME (the ratio of the book value to the market equity value), with the sorting redone annually. The net result is to assign, each year, every stock to one of the so-called “25 Fama–French value-weighted portfolios.” Table 14.2 is intended to reflect this assignment.

In particular, stocks assigned to the value-weighted portfolio (1,1) are stocks with the lowest ME and the lowest (BE/ME). These firms are likely to be small and growing rapidly: most of their value is in the form of their future growth opportunities, not in their fixed assets. Small software companies come to mind. At the other end of the spectrum, stocks in the (5,5) portfolio are typically the stocks of large, well-established firms with large fixed assets, but more modest future growth opportunities. The other 23 portfolios basically reflect various degrees of these basic two distinctions.¹³

Why should this specific collection of portfolios be of interest? From one perspective, this sorting encompasses many if not most “investment strategies” that investors adopt. For example, young persons with extended future working lives may view themselves as best served by investing in a high-risk–high-reward small firm strategy somewhere in the range of (1,1)–(1,5). Alternatively, older persons who need income and wealth stability may prefer (3,5)–(5,5) portfolios. Second, the annual rebalancing of the portfolios reflects the notion of a “dynamic asset allocation” strategy typical of many hedge funds.

These factors were employed in the standard Fama and MacBeth (1973) two-step regression procedure, which we review for clarity:

First-pass regression: for all 25 Fama–French portfolios, regress

$$\tilde{r}_t^i - r_f = \hat{\alpha}_i + \hat{\beta}_M^i (\tilde{r}_t^M - r_f) + \hat{\beta}_{SMB}^i (\widetilde{SMB}_t - r_f) + \hat{\beta}_{HML}^i (\widetilde{HML}_t - r_f) + \varepsilon_t^i$$

¹³ The sorting into portfolios also allows more precise estimation of market betas. However, there is something suspect in using similarly constructed portfolios to provide both the factors themselves and the portfolios whose return patterns are to be explained. See Daniel and Titman (1997).

Table 14.3: Coefficient estimates: Fama and French (1993) three-factor model

Coefficient <i>t</i> -statistic	$\hat{\gamma}_0$ 0.031 (2.265)	$\hat{\gamma}_M$ −0.012 (−0.753)	$\hat{\gamma}_{SMB}$ 0.003 (0.566)	$\hat{\gamma}_{HML}$ 0.012 (2.034)
$R^2_{Adj.} = 0.65$				

Second-pass regression: using the $\{\vec{r}^i, \hat{\beta}_M^i, \hat{\beta}_{SMB}^i, \hat{\beta}_{HML}^i\}$, for $i = 1, 2, \dots, 25$, regress

$$\hat{r}^i - r_f = \gamma_0 + \hat{\gamma}_M \hat{\beta}_M^i + \hat{\gamma}_{SMB} \hat{\beta}_{SMB}^i + \hat{\gamma}_{HML} \hat{\beta}_{HML}^i + \tilde{u}_i$$

The results are listed in [Table 14.3](#), for one representative historical period.¹⁴

Unlike single-factor CAPM, the [Fama and French \(1992, 1993\)](#) multifactor model appears to explain the cross-sectional variation in returns to the 25 Fama–French portfolios reasonably well: all the coefficients are significant and the R^2 is reasonably high: 65% of the variation in average returns across the 25 portfolios is explained by variation in the three factors. Unfortunately, and contrary to theory, the estimate on the CAPM beta is negative and not significant.

There is no generally accepted story to explain why the SMB and HML factors work so successfully, but they must be capturing fundamental systematic macroeconomic risks or behavior biases. Systematic risk in the CAPM sense is business cycle risk: variation in the rate of growth of the underlying economy's GDP. This sort of risk must affect the profitability of all firms, though to differing degrees. Small firms, for example, are frequently credit constrained in cyclical downturns. This possibility is much less likely for large, well-established firms: established firms in slow growth industries may suffer very significant demand reductions in recessions, whereas the demand growth for small, high tech firms may not diminish at all. Unfortunately, we do not know what specific phenomena principally underlie the SMB and HML factors. It is a subject of on-going research.

In particular, [Liew and Vassalou \(2000\)](#) demonstrate that SMB and HML are leading indicators of future GDP growth. [Vassalou \(2003\)](#) goes on to show that a factor measuring news related to future GDP growth, in conjunction with the standard market factor, can explain equity returns as well as the SMB and HML factors with regard to returns on the 25 Fama and French portfolios. More recently, [Zhang \(2005\)](#) has proposed a model in which investment irreversibility, coupled with counter-cyclical risk aversion, serves as the underlying determinant of the HML factor. His results are derived within the framework of a dynamic general equilibrium representative agent model and, as such, his setting is much more highly structured (and therefore much more informative—recall the remarks of Lucas

¹⁴ See the web notes to this chapter where we interpret this regression and provide comparative results in regard to the CAPM versus the FF three factor model.

in Chapter 2) than the APT. The idea is that high book to market value firms are those with high fixed assets, while growth firm assets are largely in the form of “growth options” (see Web Chapter B). Because of irreversibility, value firms find it extremely costly to reduce their capital stocks to a more efficient level in bad times while low book to market growth firms are unaffected in the same way because their capital stock is in the form of the human capital of their employees which can be reduced by discharge. In good times value firms feel less pressure to increase their capital stock as they enter these times with already excess capital. Growth firms do expand in good times, but the cost of expanding their capital stocks is comparatively low. The net effect of this phenomenon is that value firms are fundamentally more exposed to business cycle risk and, as such, require a higher equilibrium return as the HML factor presumes.

Various behavioral theories of the value premium have also been proposed. The idea in Lakonishok et al. (1994) is that the value premium centers around investors’ overreaction to and over-extrapolation of recent news about firm returns. Unlike the rational economic story of [Zhang \(2005\)](#), value stocks are not fundamentally more risky than growth stocks, but are cheap and pay high returns because investors consistently underestimate their future prospects (with the opposite being true of growth stocks). Barberis and Huang (2001) are able to generate a value premium in the context of a model with loss aversion and mental accounting. Their idea is this: losses are very painful to loss-averse investors who, because of mental accounting (see Chapter 3), focus on losses to individual stocks rather than on their overall portfolio’s return. Accordingly, value stocks are ones that suffered losses in the past and, as a result of mental accounting, investors are individually focused on these losses and are pessimistic regarding their future prospects.

As we conclude this section on the [Fama and French \(1993\)](#) factors, it is of interest to get some idea as to the actual return properties of the HML and SMB portfolios. These are found in [Tables 14.4 and 14.5](#) for a prefinancial-crisis data period.

Table 14.4: International performance of the SMB factor for the period January 1997–December 2006; principal world stock markets, annualized returns

Country	Quarterly Rebalancing			Semiannual Rebalancing			Annual Rebalancing		
	Mean (%)	SD (%)	t-Value	Mean (%)	SD (%)	t-Value	Mean (%)	SD (%)	t-Value
Australia	6.21	15.88	1.38	2.79	16.06	0.61	5.88	19.15	1.06
Canada	4.85	10.71	2.01	6.02	10.79	2.46	5.16	10.15	2.21
France	5.22	11.70	1.66	5.46	11.42	1.79	5.40	10.49	1.92
Germany	2.07	9.69	0.68	0.82	9.63	0.27	0.46	9.94	0.14
Italy	2.19	10.50	0.67	2.92	10.32	0.89	0.59	10.29	0.18
Japan	6.78	15.27	1.81	6.92	15.37	1.82	6.57	14.05	1.87
Netherlands	2.00	12.98	0.67	1.82	12.73	0.62	2.40	11.92	0.88
Switzerland	-4.13	10.81	-1.26	-3.39	11.01	-1.02	-1.20	10.88	-0.37
UK	3.37	11.15	1.33	3.02	11.09	1.20	3.17	11.00	1.25
USA	10.73	13.65	3.48	11.46	13.93	3.62	6.45	10.57	2.65

The high returns evidenced in [Tables 14.4 and 14.5](#) invite investors to attempt to reap the excess returns so evident in the tables, and this has the implication that over time they will be reduced in equilibrium. This has been the case most profoundly for the SMB factor.

Dimson and Marsh (1999) and Fama and French (2012) both find little evidence of significant size premia internationally for recent data sets. This finding does not carry over to the HML factor, however, which seems to remain quite robust as regards the magnitude of its risk premium. There is one data set that would seem to constitute an exception to this rule, and that is found in [Daniel et al. \(2014\)](#). While not the focus of their study, these authors compute the average annualized returns (based on monthly rebalancing) to the market, SMB, and HML factors. The results are presented in [Table 14.6](#).

In both periods, the market factor returns are robustly significant, and in neither period is this the case for SMB or the HML factors.

Table 14.5: International performance of the HML factor for the period January 1997–December 2006; principal world stock markets, annualized returns

Country	Quarterly Rebalancing			Semiannual Rebalancing			Annual Rebalancing		
	Mean (%)	SD (%)	t-Value	Mean (%)	SD (%)	t-Value	Mean (%)	SD (%)	t-Value
Australia	9.30	14.53	2.26	9.14	14.06	2.29	5.93	13.74	1.48
Canada	7.44	11.06	2.98	8.56	10.69	3.53	8.16	10.46	3.41
France	12.51	9.09	5.13	12.05	9.26	4.85	10.32	9.90	3.90
Germany	5.55	6.42	2.75	3.14	5.96	1.66	4.56	5.98	2.40
Italy	7.29	9.77	2.38	7.47	9.24	2.55	7.43	9.18	2.54
Japan	8.75	10.12	3.53	6.85	9.77	2.84	7.71	9.32	3.29
Netherlands	0.75	11.50	0.28	0.96	11.60	0.36	0.68	11.17	0.26
Switzerland	8.66	10.34	2.77	7.62	10.57	2.38	8.48	9.87	2.83
UK	8.33	6.09	6.06	7.45	5.90	5.56	6.91	5.84	5.14
USA	7.99	12.24	2.89	7.98	12.12	2.90	6.74	8.64	3.39

Table 14.6: Excess returns, annualized, for the SMB, HML and Market factors

February 1990 – August 2013			
Average	Market 7.01	SMB 3.04	HML 3.75
February 1990 – August 2013			
Standard error	(2.62)	(1.70)	(1.91)
February 1990 – August 2013			
Average	7.20	2.51	2.95
Standard error	(3.37)	(2.25)	(2.60)

14.6.2 Momentum Portfolios

Momentum portfolios and the momentum factor to which they give rise are fundamentally distinct from the HML and SMB factor portfolios. The basic reference is Jegadeesh and Titman (1993), although the notion of a momentum strategy (for portfolio construction) preceded them (e.g., De Bont and Thaler, 1985, 1987). Using return data on all NYSE and AMEX stocks for the period January 1965–December 1989, Jegadeesh and Titman (1993) rank, at the start of each month, these stocks from lowest to highest on the basis of their historical returns in the $J = 3, 6, 9$, and 12-month prior periods. For each of these periods, all stocks are then assigned to one of ten portfolios again on a return ranked basis: the lowest decile portfolio contains the lowest 10% of all stocks ranked by returns, the second decile the next lowest 10%, etc. Each portfolio is equal weighted. The authors then construct a “buy–sell” portfolio where the lowest decile stock portfolio is sold short to finance an equal-value long position in the highest decile portfolio.¹⁵ The return on this long–short portfolio for $K = 3, 6, 9$, and 12-month forward horizons is then computed. Each month the portfolios are reassembled as per above, and the indicated returns computed. The cumulative results are presented in Table 14.7.

Note that the monthly returns going forward on the buy portfolios are generally much higher relative to the sell portfolios indicating some persistence at least as regards to relative returns. For buy–sell portfolio formation based on 12 months of historical data and held for 3 months, the excess return was 1.31% per month on average which is enormous, although it does tend to peter out for longer horizons. For delayed portfolio formation (Panel B), the corresponding figure is 1.49%, or roughly 18% on an annual basis. These results are in direct contradiction to market efficiency as it is commonly understood. Furthermore, the results are driven neither by significant systematic risk differentials nor by profound differences in market capitalizations (This is confirmed in Table II of Jegadeesh and Titman, 1993). Accordingly, they also contradict the implications of the standard CAPM model (something we address in Chapter 8). Subsequently, these same authors report that similar return patterns are observed in later historical periods (Jegadeesh and Titman, 2001). Rouwenhorst (1998) finds the same patterns in European stock market data, while Asness et al. (2013) find that momentum phenomena are pervasive, being present in currency markets, commodity markets, etc. The acronym for the momentum factor is UMD, signifying “up minus down.”

The momentum portfolio constructed as per above appears to represent another fundamental APT factor. Such a conclusion, however, is based on empirical evidence alone. Again,

¹⁵ Just as in the construction of the SMB and HML portfolios, a \$100,000 momentum portfolio is composed of a \$100,000 long position in the decile 10 portfolio, a \$100,000 short position in the decile 1 portfolio, and \$100,000 in risk-free assets. The excess return on the momentum portfolio is the return on the aforementioned portfolio less r_f .

Table 14.7: Excess returns on buy, sell, and buy–sell portfolios^a

Panel A					Panel B ^b			
J =	K = 3	6	9	12	K = 3	6	9	12
6 Sell	0.0087	0.0079	0.0072	0.0080	0.0066	0.0068	0.0067	0.0076
	(1.67)	(1.56)	(1.48)	(1.66)	(1.28)	(1.35)	(1.38)	(1.58)
6 Buy	0.0171	0.0174	0.0174	0.0166	0.0179	0.0178	0.0175	0.0166
	(4.28)	(4.33)	(4.31)	(4.13)	(4.47)	(4.41)	(4.32)	(4.13)
6 Buy-sell	0.0084	0.0095	0.0102	0.0086	0.0114	0.0110	0.0108	0.0090
	(2.44)	(3.07)	(3.76)	(3.36)	(3.37)	(3.61)	(4.01)	(3.54)
12 Sell	0.0060	0.0065	0.0075	0.0155	0.0048	0.0058	0.0070	0.0085
	(1.170)	(1.29)	(1.48)	(1.74)	(0.93)	(1.15)	(1.40)	(1.71)
12 Buy	0.0192	0.0179	0.0168	0.0155	0.0196	0.0179	0.0167	0.0154
	(4.63)	(4.36)	(4.10)	(3.81)	(4.73)	(4.36)	(4.09)	(3.79)
12 Buy-sell	0.0131 ^c	0.0114	0.0093	0.0068	0.0149	0.0121	0.0096	0.0069
	(3.74)	(3.40)	(2.95)	(2.25)	(4.28)	(3.65)	(3.09)	(2.31)

^aData from Jegadeesh and Titman (1993), Table 1. Returns are average monthly excess returns for the indicated horizons. T-statistics are in parentheses. Excess returns signify returns above the risk-free rate. Portfolios based on J month lagged returns and held for K months.

^bPanel B describes the exactly analogous results except that the buy–sell portfolio is (and its buy and sell constituents) constructed 1 week after the historical returns are available. Measured returns are thus delayed by 1 week.

^cThe number 1.31 is to be interpreted as follows: Consider a portfolio composed of a long position of \$100 in the buy portfolio, a \$100 short position in the short portfolio, and \$100 in risk-free assets. After subtracting the return on the risk-free assets, the portfolio nets \$1.31 per month on average.

to date, there is no consensus as to the underlying macro (or psychological) factor for which it serves as the “factor-mimicking portfolio”.

While the SMB, HML, and UMD factors have received the most attention in the literature, they are by no means the only ones to be proposed. The accrual factor, which is based on accounting data (Sloan, 1996), deserves mention. The idea here is based on the accounting distinction between the cash component of current earnings (cash actually received) and the accrual component (cash promised to the firm by customers, etc., and cash promised by the firm in the sense of delayed payments). The cash component of earnings, and its relative contribution to earnings, in particular, is viewed as more informative of future earnings potential than the accrual component. Accordingly it is possible to develop a factor formed by a long position in a portfolio of firms with low levels of accruals financed by a short position in firms’ responding high levels of accruals. Let us simply denote it as the ACR factor and eschew the details of its construction.

Daniel and Titman (2006) propose a composite share issuance variable, ISU_t, which measures “the amount of equity a firm issues (or retires) in exchange for cash or services. Thus seasoned issues, employee stock option plans and share based acquisitions increase the issuance measure while repurchases, dividends, and other actions that take cash out of the firm reduce the issuance measure” (Daniel and Titman, 2006, p. 1608). The economics

Table 14.8: Ex post Sharpe ratios: various combinations of the listed factor-mimicking portfolio^a

Portfolio Proportions in Percent						
Market 100	SMB	HML	UMD	ISU	ACR	Ex post Sharpe Ratio
75.07	24.93					0.31
28.19	14.63	57.18				0.32
21.13	10.16	41.92	26.79			0.80
18.82	15.33	13.67	9.55	42.44		1.18
17.35	14.47	12.32	8.18	36.44	11.04	1.55
						1.60

Data period: July 1963–December 2012 portfolios annually rebalanced.

^aFrom [Daniel \(2012\)](#).

behind this measure is the observation that firms tend to issue shares when management receives favorable information regarding investment opportunities not previously reported and tend to repurchase shares in the absence of such information.

Given these new factors, we conclude this section with a comment regarding the relative importance of the many factor-mimicking portfolios. This is provided by [Daniel \(2012\)](#) who, in the tradition of MPT, treats each of the factors, $\tilde{r}^M - r_f$ (the market), \widehat{SMB} , \widehat{HML} , \widehat{ISU} and \widehat{ACR} as a distinct portfolio, and computes the *ex post* maximum Sharpe ratio realizable as progressively more assets are added. The results of this exercise are presented in [Table 14.8](#).

Note that the largest proportional boosts come from the addition of the HML and UMD factors, while the SMB factor's contribution is slight. Note also the power of the ISU_t factor.

14.7 Advantage of the APT for Stock or Portfolio Selection

The APT helps to identify the sources of systematic risk, and to split systematic risk into its fundamental components. It can thus serve as a tool for helping the portfolio manager modulate his risk exposure. For example, studies show that, among US stocks, the stocks of chemical companies are much more sensitive to short-term inflation risk than stocks of electrical equipment companies. This would be compatible with both having the same exposure to variations in the market return (same beta). Such information can be useful in at least two ways. When managing the portfolio of an economic agent whose natural position is very sensitive to short-term inflation risk, chemical stocks may be a lot less attractive than electricals, all other things equal (even though they may both have the same market beta). Second, conditional expectations, or accurate predictions, on short-term inflation may be a lot easier to achieve than predictions of the market's return. Such a refining of the information requirements needed to take aggressive positions can, in that

context, be of great use. To summarize these thoughts more succinctly, the APT reminds investors that when they construct their preferred equity portfolios, they are in reality constructing portfolios of factors. Accordingly, the portfolio risk premia they earn will be determined not so much by the risk premia of the individual assets in their portfolios, but the individual risk premia of the assorted factors they have thereby implicitly elected to hold.

14.8 Conclusions

We have now completed our review of asset pricing theories. At this stage, it may be useful to draw a final distinction between the equilibrium theories covered in Chapters 8–10 and the theories based on arbitrage such as the Martingale pricing theory and the APT.

Equilibrium theories aim at providing a complete theory of value on the basis of *primitives*: preferences, technology, and market structure. They are inevitably *heavier*, but their weight is proportional to their ambition. By contrast, arbitrage-based theories can only provide a relative theory of value. With what may be viewed as a minimum of assumptions, they

- offer bounds on option values as a function of the price of the underlying asset, the stochastic behavior of the latter being taken as given (and unexplained);
- permit estimating the value of arbitrary cash flows or securities using risk-neutral measures extracted from the market prices of a set of fundamental securities, or in the same vein, using Arrow–Debreu prices extracted from a complete set of complex securities prices;
- explain expected returns on any asset or cash-flow stream once the price of risk associated with pure-factor portfolios has been estimated from market data on the basis of a postulated return-generating process.

Arbitrage-based theories currently have the upper hand in practitioner circles where their popularity far outstrips the degree of acceptance of equilibrium theories. This, possibly temporary, state of affairs may be interpreted as a measure of our ignorance and the resulting needs to restrain our ambitions.

References

- Abel, A., 2001. Will requests attenuate the predicated meltdown in stock prices when baby boomers retire? *Rev. Econ. Stat.* 83, 589–595.
- Asness, C., Moskowitz, T., Pedersen, L., 2013. Value and momentum everywhere. *J. Finan.* 68, 929–985.
- Barberis, N., Huang, M., 2001. Mental accounting, loss aversion, and individual stock returns. *J. Finan.* 56, 1247–1292.
- Burmeister, E., Roll, R., Ross, S.A., 1994. A practitioner's guide to arbitrage pricing theory. *A Practitioner's Guide to Factor Models*. Research Foundation of the Institute of Chartered Financial Analysts, Charlottesville, VA.
- Chen, N.F., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. *J. Bus.* 59 (3), 383–404.
- Connor, G., 1984. A unified beta pricing theory. *J. Econ. Theory*. 34 (1), 13–31.

- Daniel, K., 2012. Behavioral finance: a selective survey for owners of money. Presentation, JOIM Fall Conference.
- Daniel, K., Hodrick, R., Lu, Z., 2014. The carry trade: risks and drawdowns. Working Paper, Columbia Business School.
- Daniel, K., Titman, S., 1997. Evidence on the characteristics of cross sectional variation in stock returns. *J. Finan.* 52 (1), 1–33.
- Daniel, K., Titman, S., 2006. Market reactions to tangible and intangible information. *J. Finan.* 61 (4), 1605–1643.
- De Bont, W., Thaler, R., 1985. Does the stock market overreact? *J. Finan.* 40, 793–805.
- De Bont, W., Thaler, R., 1987. Further evidence on investor overreaction and stock market seasonality. *J. Finan.* 42, 557–581.
- Dimson, E., Marsh, P., 1999. Murphy's law and market anomalies. *J. Portf. Manage.* 25, 53–69.
- Fama, E.F., 1973. A note on the market model and the two-parameter model. *J. Finan.* 28 (5), 1181–1185.
- Fama, E., French, K., 1992. The cross section of expected stock returns. *J. Finan.* 47, 427–465.
- Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. *J. Finan. Econ.* 33, 3–56.
- Fama, E., French, K., 2012. Size, value and momentum in international stock returns. *J. Financ. Econ.* 105, 457–472.
- Fama, E., MacBeth, J., 1973. Risk, return, and equilibrium: empirical tests. *J. Polit. Econ.* 81, 607–636.
- Geonakoplos, J., Magill, Quinzii, M., M., 2004. Demography and the long – run predictability of the stock market. *Brookings Pap. Econ. Act.* 1, 241–307.
- Grinblatt, M., Titman, S., 1985. Approximate factor structures: interpretations and implications for empirical tests. *J. Finan.* 15 (5), 1367–1373.
- Huberman, G., 1982. A simple approach to arbitrage pricing. *J. Econ. Theory.* 28, 183–191.
- Jagannathan, R., Wang, Z., 1996. The conditional capm and the cross-section of expected returns. *J. Finan.* 51, 3–53.
- Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for stock market efficiency. *J. Finan.* 48, 65–91.
- Jegadeesh, N., Titman, S., 2001. Profitability of Momentum Strategies: An Evaluation of Alternative Explanations. *The Journal of Finance.* 56, 699–720.
- Lakoneshok, J., Schleefer, A., Visblny, R., 1994. Contrarian investment, extrapolation, and risk. *J. Finan.* 44, 154–1578.
- Lettav, M., Ludvigson, S., 2001. Consumption, aggregated wealth and expected stock returns. *J. Finan.* 55, 815–849.
- Liew, J., Vassalou, M., 2000. Can book to market, size and momentum be risk factors that predict economic growth? *J. Finan. Econ.* 57, 221–245.
- Ross, S.A., 1976. The arbitrage pricing theory. *J. Econ. Theory.* 1, 341–360.
- Rouwenhorst, G., 1998. International momentum strategies. *J. Finan.* 53, 267–284.
- Santos, T., Veronesi, P., 2006. Labor income and predictable stock returns. *Rev. Finan. Stud.* 19 (1), 1–44.
- Santos, T. Veronesi, P. Labor income and predictable stock returns. *Rev. Financ. Stud.* 19, 1–44.
- Sloan, R., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? *Account. Rev.* 71 (3), 289–315.
- Vassalou, M., 2003. News related to future GDP growth as a risk factor in equity returns. *J. Finan. Econ.* 68, 47–73.
- Zhang, L., 2005. The value premium. *J. Finan.* 60 (1), 67–103.

Appendix A.14.1: A Graphical Interpretation of the APT

For illustrative purposes, we assume one common factor. As noted in Connor (1984), the APT requires existence of a *rich* market structure with a large number of assets with different characteristics and a minimum number of trading restrictions. Such a market

structure, in particular, makes it possible to form a portfolio P with the following three properties:

Property 1 P has zero cost; in other words, it requires no investment. This is the first requirement of an arbitrage portfolio.

Let us denote w_i as the **value** of the position in the i th asset in portfolio P . Portfolio P is then fully described by the vector $\mathbf{w}^T = (w_1, w_2, \dots, w_N)$, and the zero cost condition becomes

$$\sum_{i=1}^N w_i = 0 = \mathbf{w}^T \cdot \mathbf{1}$$

with $\mathbf{1}$ the (column) vector of 1's. (Positive positions in some assets must be financed by short sales of others.)

Property 2 P has zero sensitivity (zero beta) to the common factor:¹⁶

$$\sum_i^N w_i \beta_i = 0 = \mathbf{w}^T \cdot \boldsymbol{\beta}$$

Property 3 P is a well-diversified portfolio. The specific risk of P is (almost) totally eliminated:

$$\sum_i^N w_i^2 \sigma_{\varepsilon_i}^2 \cong 0$$

The APT builds on the assumed existence of such a portfolio, which requires a rich market structure.

Statement and Proof of the APT

The APT relationship is the direct consequence of the factor structure hypothesis, the existence of a portfolio P satisfying these conditions, and the no-arbitrage assumption. Given that returns have the structure of Eq. (14.1), Properties 2 and 3 imply that P is riskless. The fact that P has zero cost (**Property 1**) then entails that an arbitrage opportunity will exist unless:

$$\bar{r}_P = 0 = \mathbf{w}^T \cdot \bar{r} \quad (\text{A.14.1})$$

¹⁶ Remember that the beta of a portfolio is the weighted sum of the betas of the assets in the portfolio.

The APT theorem states, as a consequence of this succession of statements, that there must exist scalars λ_0, λ_1 , such that

$$\begin{aligned}\bar{r} &= \lambda_0 \cdot \mathbf{1} + \lambda_1 \beta, \text{ or} \\ \bar{r}_i &= \lambda_0 + \lambda_1 \beta_i \text{ for all assets, } i\end{aligned}\tag{A.14.2}$$

which is the main equation of the APT.

Equation (A.14.2) and Properties 1 and 2 are statements about four vectors: w , β , $\mathbf{1}$, and \bar{r} .

Property 1 states that w is orthogonal to $\mathbf{1}$. **Property 2** asserts that w is orthogonal to β . Together these statements imply a geometric configuration that we can easily visualize if we fix the number of risky assets at $N = 3$, which implies that all the vectors have dimension 3. This is illustrated in [Figure 14.1](#).

Equation (A.14.2)—no arbitrage—implies that w and \bar{r} are orthogonal. But this means that the vector \bar{r} must lie in the plane formed by $\mathbf{1}$ and β , or that \bar{r} can be written as a linear combination of $\mathbf{1}$ and β , as Eq. (A.14.2) asserts.

More generally, one can deduce from the triplet

$$\sum_I w_i = \sum_I w_i \beta_i = \sum_I w_i \bar{r}_i = 0$$

that there exist scalars λ_0, λ_1 such that

$$\bar{r}_i = \lambda_0 + \lambda_1 \beta_i \text{ for all } i$$

This is a consequence of the orthonormal projection of the vector \bar{r}_i into the subspace spanned by the other two.

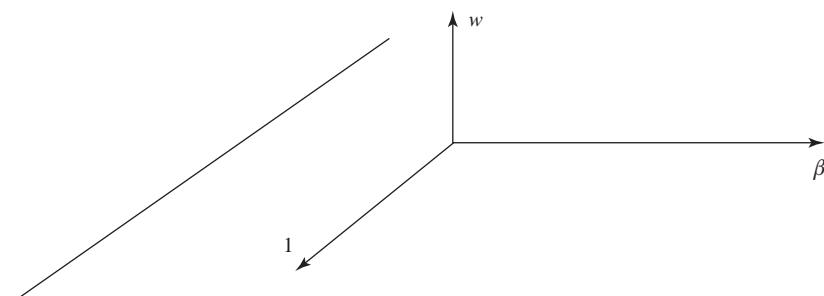


Figure 14.1
Geometric representation: w orthogonal to 1 and β .

The CAPM and the APT

Suppose that there exists a risk-free asset or, alternatively, that the sufficiently rich market structure hypothesis permits constructing a fully diversified portfolio with zero sensitivity to the common factor (but positive investment). Then

$$\bar{r}_f = r_f = \lambda_0$$

That is, λ_0 is the return on the risk-free asset or the risk-free portfolio.

Now let us compose a portfolio Q with unitary sensitivity to the common factor: $\beta = 1$. Then applying the APT relation, one gets

$$\bar{r}_Q = r_f + \lambda_1 \cdot \mathbf{1}$$

Thus, $\lambda_1 = \bar{r}_Q - r_f$, excess return on the pure-factor portfolio Q . It is now possible to rewrite Eq. (A.14.2) as

$$\bar{r}_i = r_f + \beta_i(\bar{r}_Q - r_f) \quad (\text{A.14.3})$$

If, as we have assumed, the unique common factor is the return on the market portfolio, in which case $Q = M$ and $\bar{r}_Q \equiv \bar{r}_M$, then Eq. (13.4) is simply the CAPM equation:

$$\bar{r}_i = r_f + \beta_i(\bar{r}_M - r_f)$$

Appendix 14.2: Capital Budgeting

To illustrate an example of the use of the [Fama and French \(1992, 1993\)](#) three factor model for a capital budgeting cost of capital exercise, let us estimate Microsoft Corporation's cost-of-capital. We contract this method with the CAPM. The estimates are derived from 5 years of monthly data from the period 1996.1 to 2000.12.

A. CAPM results

1. Standard regression

$$\tilde{r}_t^{\text{micro}} - r_f = \hat{\alpha}^{\text{micro}} + \hat{\beta}[\tilde{r}_t^M - r_f] + \tilde{\varepsilon}_t^{\text{micro}}$$

Estimate	0.994	1.71
Standard error	(1.41)	(0.289)
t-statistic	704	5.91

2. For this period $= [\bar{r}_M - r_f] = 0.0617$, $\sigma_M = 0.1666$, and $r_f = 0.06$.

$$\sigma^{\text{micro}} = 0.4644$$

$$(\sigma^{\text{micro}})^2 = (0.4644)^2 = (\hat{\beta}^{\text{micro}})^2 \sigma_M^2 + (\sigma_{\varepsilon}^{\text{micro}})^2$$

$$(0.4644)^2 = (1.71)^2 (0.1666)^2 + (\sigma_{\varepsilon}^{\text{micro}})^2$$

$$\sigma_{\varepsilon}^{\text{micro}} = 0.3668 \text{ or } 36.68\%$$

$$\begin{aligned} E\tilde{r}^{\text{micro}} &= r_f + \hat{\beta}^{\text{micro}} [\bar{r}_M - r_f] \\ &= 0.06 + (1.71)[0.0617] \\ &= 0.166 \text{ or } 16.6\% \end{aligned}$$

This is the estimate of Microsoft's cost of capital using the CAPM.

B. Fama–French results

	$\tilde{r}_t^{\text{micro}} - r_f$	$= \hat{\alpha}^{\text{micro}}$	$+ \hat{\beta}_M^{\text{micro}} [\tilde{r}_t^M - r_f]$	$+ \hat{\beta}_{\text{SMB}}^{\text{micro}} [\widetilde{\text{SMB}}_t - r_f]$	$+ \hat{\beta}_{\text{HML}}^{\text{micro}} [\widetilde{\text{HML}}_t - r_f]$	$+ \tilde{\varepsilon}_t^{\text{micro}}$
Estimate	0.6335	1.096		-1.374		-1.389
Standard error	(1.25	(0.303)		(0.381)		(0.333)
t-statistic	0.505	3.62		-3.599		-4.165

$$(\bar{r}_M - r_f) = .0617$$

$$\widetilde{\text{SMB}}^A - r_f = .0212$$

$$\widetilde{\text{HML}}^A - r_f = .0379$$

$$\begin{aligned} E\tilde{r}^{\text{micro}} &= r_f + \hat{\beta}_M^{\text{micro}} [\bar{r}_M^A - r_f] + \hat{\beta}_{\text{SMB}}^{\text{micro}} (\widetilde{\text{SMB}}^A - r_f) + \hat{\beta}_{\text{HML}}^{\text{micro}} (\widetilde{\text{HML}}^A - r_f) \\ &= 0.06 + (1.1)[0.0617] + (-1.374)[0.0212] + (-1.39)(0.0379) \\ &= 0.0461, \text{ or } 4.61\% \end{aligned}$$

$E\tilde{r}^{\text{micro}}$ is the estimate of Microsoft's cost of capital using the Fama-French three factor model. There is a substantial difference between the CAPM and the Fama-French estimates.

An Intuitive Overview of Continuous Time Finance

Chapter Outline

15.1 Introduction	443
15.2 Random Walks and Brownian Motion	444
15.3 More General Continuous Time Processes	448
15.4 A Continuous Time Model of Stock Price Behavior	449
15.5 Simulation and Call Pricing	451
15.5.1 Ito processes	451
15.5.2 Binomial Model	453
15.6 Solving Stochastic Differential Equations: A First Approach	454
15.6.1 The Behavior of Stochastic Differentials	454
15.6.2 Ito's Lemma	456
15.6.3 The Black–Scholes Formula	457
15.7 A Second Approach: Martingale Methods	459
15.8 Applications	460
15.8.1 The Consumption–Savings Problem	460
15.8.2 An Application to Portfolio Analysis	461
15.8.2.1 <i>Digression to Discrete Time</i>	462
15.8.2.2 <i>Return to Continuous Time</i>	464
15.8.3 The Consumption CAPM in Continuous Time	466
15.9 Final Comments	467
References	467

15.1 Introduction

If we think of stock prices as arising from the equilibration of traders' demands and supplies, then the binomial model is implicitly one in which security trading occurs at discrete time intervals, however short, and this is, in fact, factually what actually happens. It will be mathematically convenient, however, to abstract from this intuitive setting and hypothesize that trading takes place "continuously." This is consistent with the notion of continuous compounding. But it is not fully realistic: It implies that an uncountable number

of individual transactions may transpire in any interval of time, however small, which is physically impossible.

Continuous time finance is principally concerned with techniques for the pricing of derivative securities under the fiction of continuous trading. These techniques frequently allow closed-form pricing solutions to be obtained—at the cost of working in a context that is less intuitive than discrete time. In the present chapter, we hope to convey some idea as to how this is done.

We will need first to develop a continuous time model of a stock's price evolution through time. Such a model must respect the statistical regularities that are known to characterize, empirically, equity returns as first formalized in Section 7.5.2:

1. Stock prices are lognormally distributed, which means that returns (continuously compounded) are normally distributed.
2. For short time horizons, stock returns are independently and identically distributed (iid) over nonoverlapping time intervals.

After we have faithfully represented these equity regularities in a continuous time setting, we will move on to a consideration of derivatives pricing. In doing so, we aim to give some idea how the principles of risk-neutral valuation carry over to this specialized setting. The discussion aims at intuition; no attempt is made to be mathematically complete.

In all cases, this intuition has its origins in the discrete time context. This leads us initially to review the notion of a random walk.

15.2 Random Walks and Brownian Motion

Consider a time horizon composed of N adjacent time intervals, each of duration Δt , and indexed successively by $t_0, t_1, t_2, \dots, t_N$, i.e.,

$$t_i - t_{i-1} = \Delta t, \quad i = 1, 2, \dots, N$$

We define a discrete time stochastic process, \tilde{x} , on this succession of time indices by

$$\begin{aligned} x(t_0) &= 0 \\ \tilde{x}(t_{j+1}) &= \tilde{x}(t_j) + \tilde{\varepsilon}(t_j)\sqrt{\Delta t}, \\ j &= 0, 1, 2, \dots, N-1 \end{aligned}$$

where, for all j , $\tilde{\varepsilon}(t_j) \sim N(0, 1)$. It is further assumed that the random factors $\tilde{\varepsilon}(t_j)$ are independent of one another which implies

$$E(\tilde{\varepsilon}(t_j)\tilde{\varepsilon}(t_i)) = 0, \quad i \neq j$$

This is a specific example of a random walk, particular in the sense that the uncertain disturbance term follows a specific distribution.¹

We are interested in understanding the behavior of a random walk over extended time periods. More precisely, we want to characterize the statistical properties of the incremental difference

$$x(t_k) - x(t_j) \text{ for any } j < k.$$

Clearly,

$$\tilde{x}(t_k) - x(t_j) = \sum_{i=j}^{k-1} \tilde{\varepsilon}(t_i) \sqrt{\Delta t}$$

Since the random disturbances $\tilde{\varepsilon}(t_i)$ all have mean zero,

$$E(\tilde{x}(t_k) - x(t_j)) = 0$$

Furthermore,

$$\begin{aligned} \text{var}(\tilde{x}(t_k) - x(t_j)) &= E\left(\sum_{i=j}^{k-1} \tilde{\varepsilon}(t_i) \sqrt{\Delta t}\right)^2 \\ &= E\left(\sum_{i=j}^{k-1} [\tilde{\varepsilon}(t_i)]^2 \sqrt{\Delta t}\right) \\ &\quad (\text{by independence}) \\ &= \sum_{i=j}^{k-1} (1) \Delta t = (k-j) \Delta t, \\ &\quad \text{since } E[\tilde{\varepsilon}(t_i)]^2 = 1. \end{aligned}$$

If we identify

$$x_{t_j} = \ln q_{t_j}^e$$

¹ A very simple random walk is of the form $\tilde{x}(t_{j+1}) = x(t_j) + \tilde{n}(t_j)$, where for all $j = 0, 1, 2 \dots$

$$\tilde{n}(t_j) = \begin{cases} +1, & \text{if a coin is flipped and heads appears} \\ -1, & \text{if a coin is flipped and tails appears.} \end{cases}$$

At each time interval $x(t_j)$, either increases or diminishes by one depending on the outcome of the coin toss. Suppose we think of $x(t_0) \equiv 0$ as representing the center of the sidewalk where an intoxicated person staggers one step to the right or to the left of the center in a manner that is consistent with independent coin flips (heads implies to the right). This example is the source of the term *random walk*.

where $q_{t_j}^e$ is the price of the stock at time t_j , then this simple random walk model becomes a candidate for our model of stock price evolution beginning from $t = 0$: At each node t_j , the logarithm of the stock's price is distributed normally, with mean $\ln q_{t_0}^e$ and variance $j \Delta t$.²

Since the discrete time random walk is so respectful of the empirical realities of stock returns, it is natural to seek its counterpart for “continuous time.” This is referred to as a *Brownian motion* (or a Wiener process), and it represents the limit of the discrete time random walk as we pass to continuous time, i.e., as $\Delta t \rightarrow 0$. It is represented symbolically by

$$dz = \tilde{\varepsilon}(t)\sqrt{dt}$$

where $\tilde{\varepsilon}(t) \sim N(0, 1)$, and for any times t, t' where $t \neq t'$, and $\tilde{\varepsilon}(t), \tilde{\varepsilon}(t')$ are independent. We used the word *symbolically* not only because the term dz does not represent a differential in the terminology of ordinary calculus, but because we make no attempt here to describe how such a limit is taken. Following what is commonplace notation in the literature, we don't write a \sim over z even though it represents a random quantity.

More formally, a stochastic process $z(t)$ defined on $[0, T]$ is a Brownian motion provided the following three properties are satisfied:

1. for any $t_1 < t_2$, $z(t_2) - z(t_1)$ is normally distributed with mean zero and variance $t_2 - t_1$;
2. for any $0 \leq t_1 < t_2 \leq t_3 < t_4$, $z(t_4) - z(t_3)$ is statistically independent of $z(t_2) - z(t_1)$; and
3. $z(t_0) \equiv 0$ with probability one.

A Brownian motion is a very unusual stochastic process, and we can only give a hint about what is actually transpiring as it evolves. Three of its properties are presented below:

1. First, a Brownian motion is a continuous process. If we were able to trace out a sample path $z(t)$ of a Brownian motion, we would not see any jumps.³
2. However, the sample path is not at all *smooth* and is, in fact, as “jagged as can be,” which we formalize by saying that it is nowhere differentiable. A function must be essentially smooth if it is to be differentiable. That is, if we magnify a segment of its time path sufficiently, it will appear approximately linear. This *smoothness* is totally absent in a Brownian motion.
3. Lastly, a Brownian motion is of *unbounded variation*. This is perhaps the least intuitive of its properties. This conveys the idea that if we could take one of those mileage wheels that are drawn along a route on a map to assess the overall distance (each revolution of the wheel corresponding to a fixed number of kilometers) and apply it to

² To be absolutely clear in our use of terminology, the random walk is a model of a stock's rate of return evolution, and thus, indirectly, its price evolution. Differences in logs of stock prices represent a rate of return.

³ At times such as the announcement of a takeover bid, stock prices exhibit jumps. We will not consider such *jump processes*, although considerable research effort has been devoted to studying them, and to the pricing of derivatives written on them.

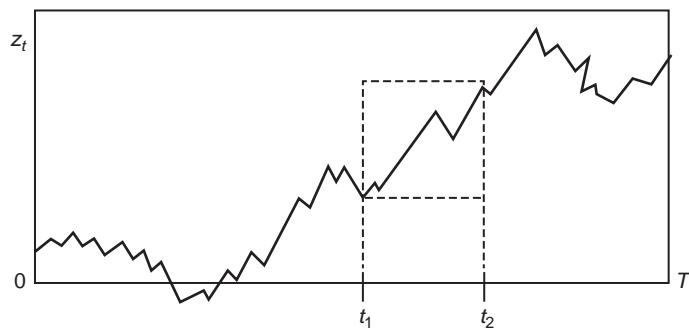


Figure 15.1
Approximate Brownian Motion Sample Path.

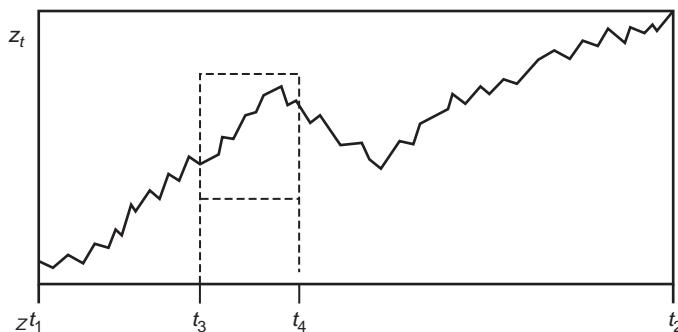


Figure 15.2
Approximate Brownian Motion Sample Path: Expanded Sub-Path of [Figure 15.1](#).

the sample path of a Brownian motion, no matter how small the time interval, the mileage wheel would record an *infinite distance* (if it ever got to the end of the path!).⁴

One way of visualizing such a process is to imagine a rough sketch of a particular sample path where we connect its position at a sequence of discrete time intervals by straight lines. [Figure 15.1](#) proposes one such path.

Suppose that we were next to enlarge the segment between time intervals t_1 and t_2 . We would find something on the order of [Figure 15.2](#).

Continue this process of taking a segment, enlarging it, taking another subsegment of that segment, enlarging it, etc. (in [Figure 15.2](#) we could next enlarge the segment from t_3 to t_4). Under a typical differentiable function of bounded variation, we would eventually be enlarging such a small segment that it would appear as a straight line. With a Brownian motion, however, this will never happen. No matter how much we enlarge even a segment that corresponds to an arbitrarily short time interval, the same “sawtooth” pattern will appear, and there will be many, many “teeth.”

⁴ This reference to a ‘mileage wheel’ certainly dates your authors (now see Google Maps).

A Brownian motion represents a special case of a continuous process with independent increments. For such processes, the standard deviation per unit of time becomes unbounded as the time interval becomes small:

$$\lim_{\Delta t \rightarrow 0} \frac{\sigma \sqrt{\Delta t}}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{\sigma}{\sqrt{\Delta t}} = \infty$$

No matter how small the time period, proportionately, a **lot** of variation remains. This constitutes our abstraction of a random walk to a context of continuous trading.⁵

15.3 More General Continuous Time Processes

A Brownian motion will be the principal building block of our description of the continuous time evolution of a stock's price—it will be the *engine* or *source* of the uncertainty. To it is often added a deterministic component intended to capture the “average” behavior through time of the process. Together we have something of the form

$$\begin{aligned} dx(t) &= a dt + b \tilde{\varepsilon}(t) \sqrt{dt} \\ &= a dt + b dz \end{aligned} \tag{15.1}$$

where the first component is the deterministic one and a is referred to as the drift term.

This is an example of a generalized Brownian motion or, to use more common terminology, a generalized Wiener process. If there were no uncertainty, $x(t)$ would evolve deterministically: if we integrate

$$\begin{aligned} dx(t) &= a dt, \text{ we obtain} \\ x(t) &= x(0) + at \end{aligned}$$

The solution to Eq. (15.1) is thus of the form

$$x(t) = x(0) + at + bz(t) \tag{15.2}$$

where the properties of $z(t)$ were articulated earlier (recall conditions 1, 2, and 3 of the definition). These imply that

$$\begin{aligned} E(x(t)) &= x(0) + at, \\ \text{var}(x(t)) &= b^2 t, \text{ and} \\ \text{SD}(x(t)) &= b\sqrt{t} \end{aligned}$$

⁵ The name Brownian motion comes from a nineteenth century physicist named Brown, who studied the behavior of dust particles floating on the surface of water. Under a microscope, dust particles are seen to move randomly about in a manner similar to the sawtooth pattern shown except that the motion can be in any 360° direction. The interpretation of the phenomena is that the dust particles experience the effect of random collisions by water molecules.

Equation (15.2) may be further generalized to allow the coefficients to depend upon the time and the current level of the process:

$$dx(t) = a(x(t), t)dt + b(x(t), t)dz \quad (15.3)$$

In this latter form, it is referred to as an Ito process after one of the earliest and most important developers of this field. An important issue in the literature—but one we will eschew—is to determine the conditions on $a(x(t), t)$ and $b(x(t), t)$ in order for Eq. (15.3) to have a solution. Equations (15.1) and (15.3) are generically referred to as *stochastic differential equations*.

Given this background, we now return to the original objective of modeling the behavior of a stock's price process.

15.4 A Continuous Time Model of Stock Price Behavior

Let us now restrict our attention only to those stocks that do not pay dividends, so that stock returns are exclusively identified with price changes (we will maintain this assumption throughout the chapter). Our basic discrete time model formulation is

$$\ln q^e(t + \Delta t) - \ln q^e(t) = \mu \Delta t + \sigma \tilde{\varepsilon} \sqrt{\Delta t} \quad (15.4)$$

Note that the stochastic process is imposed on differences in the logarithm of the stock's price. Equation (15.4) thus asserts that the continuously compounded return to the stock's ownership over the time period t to $t + \Delta t$ is distributed normally with mean $\mu \Delta t$ and variance $\sigma^2 \Delta t$.

This is clearly a lognormal model:

$$\ln(q^e(t + \Delta t)) \sim N(\ln q^e(t) + \mu \Delta t, \sigma \sqrt{\Delta t})$$

It is a more general formulation than a pure random walk as it admits the possibility that the mean increase in the logarithm of the price is positive. The continuous time analogue of Eq. (15.4) is

$$d \ln q^e(t) = \mu dt + \sigma dz \quad (15.5)$$

Following Eq. (15.2), it has the solution

$$\ln q^e(t) = \ln q^e(0) + \mu t + \sigma z(t) \quad (15.6)$$

where

$$\begin{aligned} E \ln q^e(t) &= \ln q^e(0) + \mu t, \text{ and} \\ \text{var } q^e(t) &= \sigma^2 t \end{aligned}$$

Since the $\ln q^e(t)$ on average grows linearly with t (so that, on average, $q^e(t)$ will grow exponentially), Eqs. (15.5) and (15.6) are, together, referred to as a geometric Brownian motion (GBM). It is clearly a lognormal process: $\ln q^e(t) \sim N(\ln q^e(0) + \mu t, \sigma \sqrt{t})$, and the parameters μ and σ can be estimated exactly as was discussed in Boxes 3.1 and 7.2 with the maintained assumption that time is measured in years.

While Eq. (15.6) is a complete description of the evolution of the logarithm of a stock's price, we are rather interested in the evolution of the price itself. Passing from a continuous time process on $\ln q^e(t)$ to one on $q^e(t)$ is not a trivial matter, however, and we need some additional background to make the conversion correctly. This is considered in the next few paragraphs.

The essence of lognormality is the idea that if a random variable \tilde{y} is distributed normally, then the random variable $\tilde{w} = e^{\tilde{y}}$ is distributed lognormally. Suppose, in particular, that $\tilde{y} \sim N(\mu_y, \sigma_y)$. A natural first question is: How are μ_w and σ_w related to μ_y and σ_y when $\tilde{w} = e^{\tilde{y}}$? We first note that

$$\mu_w \neq e^{\mu_y}, \text{ and } \sigma_w \neq e^{\sigma_y}$$

As noted back in Section 6.6, it is actually the case that:

$$\mu_w = e^{\mu_y + 1/2\sigma_y^2} \quad (15.7)$$

and

$$\sigma_w = e^{\mu_y + 1/2\sigma_y^2} (e^{\sigma_y^2} - 1)^{1/2}. \quad (15.8)$$

These formulae are not obvious, but we can at least shed some light on Eq. (15.8): Why should the variance of y have an impact on the mean of w ? To see why this is so, let us remind ourselves of the shape of the lognormal distribution, as found in Figure 15.3.

Suppose there is an increase in variance. Since this distribution is pinched off to the left at zero, a higher variance of y can only imply (within the same class of distributions) that probability is principally shifted to higher values of w . But this will have the simultaneous effect of increasing the mean of w . The variance of y and the mean of w cannot be specified independently. The mean and standard deviation of the lognormal variable \tilde{w} are thus each related to both the mean and variance of \tilde{y} as per the relationships in Eqs. (15.7) and (15.8).

These results allow us to express the mean and standard deviation of $q^e(t)$ (by analogy, \tilde{w}) in relation to $\ln q^e(t) + \mu t$ and $\sigma^2 t$ (by analogy, the mean and variance of \tilde{y}) via Eqs. (15.5) and (15.6):

$$\begin{aligned} E q^e(t) &= e^{\ln q^e(0) + (\mu + 1/2\sigma^2)t} \\ &= q^e(0) e^{(\mu + 1/2\sigma^2)t} \end{aligned} \quad (15.9)$$

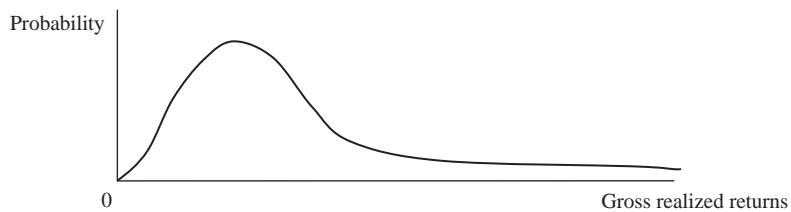


Figure 15.3
The lognormal density.

$$\begin{aligned} \text{SD}q^e(t) &= e^{\ln q^e(0) + (\mu + 1/2\sigma^2)t} (e^{\sigma^2 t} - 1)^{1/2} \\ &= q^e(0)e^{(\mu + 1/2\sigma^2)t} (e^{\sigma^2 t} - 1)^{1/2} \end{aligned} \quad (15.10)$$

We are now in a position, at least at an intuitive level, to pass from a stochastic differential equation describing the behavior of $\ln q^e(t)$ to one that governs the behavior of $q^e(t)$. If $\ln q^e(t)$ is governed by Eq. (15.5), then

$$\frac{dq^e(t)}{q^e(t)} = (\mu + 1/2\sigma^2)dt + \sigma dz(t) \quad (15.11)$$

where $dq^e(t)/q^e(t)$ can be interpreted as the instantaneous (stochastic) rate of price change. Rewriting Eq. (15.11) slightly differently yields

$$dq^e(t) = (\mu + 1/2\sigma^2)q^e(t)dt + \sigma q^e(t)dz(t) \quad (15.12)$$

which informs us that the stochastic differential equation governing the stock's price represents an Ito process since the coefficients of dt and $dz(t)$ are both time dependent (and stochastic).

We would also expect that if $q^e(t)$ were governed by

$$dq^e(t) = \mu q^e(t)dt + \sigma q^e(t)dz(t), \text{ then} \quad (15.13)$$

$$d \ln q^e(t) = (\mu + 1/2\sigma^2)dt + \sigma dz(t) \quad (15.14)$$

Equations (15.13) and (15.14) are fundamental to what follows.

15.5 Simulation and European Call Pricing

15.5.1 Ito processes

Ito processes and their constituents, most especially the Brownian motion, are difficult to grasp at this abstract level and it will assist our intuition to describe how we might simulate a discrete time approximation to them.

Suppose we have estimated $\hat{\mu}$ and $\hat{\sigma}$ for a stock's return process as per the web notes to Chapter 12. Recall that these estimates are derived from daily price data properly scaled up to reflect the fact that in this literature it is customary to measure time in years. We have two potential stochastic differential equations to guide us—[Eqs. \(15.13\) and \(15.14\)](#)—and each has a discrete time approximate counterpart.

i. Discrete time counterpart to [Eq. \(15.13\)](#)

If we approximate the stochastic differential $dq^e(t)$ by the change in the stock's price over a short interval of time Δt , we have

$$\begin{aligned} q^e(t + \Delta t) - q^e(t) &= \hat{\mu}q^e(t)\Delta t + \hat{\sigma}q^e(t)\tilde{\varepsilon}(t)\sqrt{\Delta t}, \text{ or} \\ q^e(t + \Delta t) &= q^e(t)[1 + \hat{\mu}\Delta t + \hat{\sigma}\tilde{\varepsilon}(t)\sqrt{\Delta t}] \end{aligned} \quad (15.15)$$

There is a problem with this representation, however, because for any $q^e(t)$, the price in the next period, $q^e(t + \Delta t)$, is normally distributed (recall that $\tilde{\varepsilon}(t) \sim N(0, 1)$) rather than lognormal as a correct match to the data requires. In particular, there is the unfortunate possibility that the price could go negative, although for small time intervals Δt , this is exceedingly unlikely.

ii. Discrete time counterpart to [Eq. \(15.14\)](#)

Approximating $d \ln q^e(t)$ by successive log values of the price over small time intervals Δt yields

$$\begin{aligned} \ln q^e(t + \Delta t) - \ln q^e(t) &= (\hat{\mu} - 1/2\hat{\sigma}^2)\Delta t + \hat{\sigma}\tilde{\varepsilon}(t)\sqrt{\Delta t}, \text{ or} \\ \ln q^e(t + \Delta t) &= \ln q^e(t) + (\hat{\mu} - 1/2\hat{\sigma}^2)\Delta t + \hat{\sigma}\tilde{\varepsilon}\sqrt{\Delta t} \end{aligned} \quad (15.16)$$

Here it is the logarithm of the price in period $t + \Delta t$ that is normally distributed, as required, and for this reason we'll limit ourselves to [Eq. \(15.16\)](#) and its successors. For simulation purposes, it is convenient to express [Eq. \(15.16\)](#) as

$$q^e(t + \Delta t) = q^e(t)e^{(\hat{\mu}-1/2\hat{\sigma}^2)\Delta t + \hat{\sigma}\tilde{\varepsilon}(t)\sqrt{\Delta t}} \quad (15.17)$$

It is easy to generate a possible sample path of price realizations for [Eq. \(15.17\)](#). First select an interval of time Δt , and the number of successive time periods of interest (this will be the length of the sample path), say N . Using a random number generator, next generate N successive draws from the standard normal distribution. By construction, these draws are independent and thus successive rates of return ($q^e(t + \Delta t)/q^e(t) - 1$) will be statistically independent of one another. Let this series of N draws be represented by $\{\varepsilon_j\}_{j=1}^N$. The corresponding sample path (or “time series”) of prices is thus created as per [Eq. \(15.18\)](#).

$$q^e(t_{j+1}) = q^e(t_j)e^{(\hat{\mu}-1/2\hat{\sigma}^2)\Delta t + \hat{\sigma}\tilde{\varepsilon}_j\sqrt{\Delta t}} \quad (15.18)$$

where $t_{j+1} = t_j + \Delta t$. This is not the price path we would use for derivatives pricing, however.

15.5.2 Binomial Model

Under the binomial model, European call valuation is undertaken in a context where the probabilities have been changed in such a way that all assets, including the underlying stock, earn the risk-free rate. The simulation-based counterpart to this transformation is to replace $\bar{\mu}$ by $\ln(1 + r_f)$ in Eqs. (15.17) and (15.18):

$$q^e(t + \Delta t) = q^e(t)e^{(\ln(1+r_f)-1/2\sigma^2)\Delta t + \sigma\tilde{\varepsilon}(t)\sqrt{\Delta t}} \quad (15.19)$$

where r_f is the 1-year risk-free rate (not continuously compounded) and $\ln(1 + r_f)$ is its continuously compounded counterpart.

How would we proceed to price a call in this simulation context? Since the value of the call at expiration is exclusively determined by the value of the underlying asset at that time, we first need a representative number of possible risk-neutral prices for the underlying asset at expiration. The entire risk-neutral sample path—as per Eq. (15.18)—is not required. By representative we mean enough prices so that their collective distribution is approximately lognormal. Suppose it was resolved to create J sample prices (to be even reasonably accurate, $J > 1000$) at expiration, T years from now. Given random draws $\{\varepsilon_k\}_{k=1}^J$ from $N(0, 1)$, the corresponding underlying stock price realizations are $\{q_k^e(T)\}_{k=1}^J$ as given by

$$q_k^e(T) = q^e(0)e^{(\ln(1+r_f)-1/2\sigma^2)T + \sigma\tilde{\varepsilon}_k\sqrt{T}} \quad (15.20)$$

For each of these prices, the corresponding call value at expiration is

$$C_j^T = \max\{0, q_j^e(T) - E\}, \quad j = 1, 2, \dots, J$$

The average expected payoff across all these possibilities is

$$C_{\text{Avg}}^T = \frac{1}{J} \sum_{j=1}^J C_j^T$$

Since under risk-neutral valuation, the expected payoff of any derivative asset in the span of the underlying stock and a risk-free bond is discounted back at the risk-free rate, our estimate of the call's value today (when the stock's price is $q^e(0)$) is

$$C^0 = e^{-\ln(1+r_f)T} C_{\text{Avg}}^T \quad (15.21)$$

In the case of the Asian option considered earlier (Chapter 11) or some other path-dependent option, a large number of sample paths would need to be generated since the exercise price of the option (and thus its value at expiration) is dependent upon the entire sample path of underlying asset prices leading to it.

Monte Carlo simulation, as the previous method is called, is not the only pricing technique where the underlying idea is related to the notion of risk-neutral valuation. There are ways that stochastic differential equations can be solved directly.

15.6 Solving Stochastic Differential Equations: A First Approach

Monte Carlo simulation employs the notion of risk-neutral valuation but it does not, of course, provide closed-form solutions for derivatives prices, such as the Black–Scholes formula for a European call.⁶ How are such closed-form expressions obtained? In what follows we provide a nontechnical outline of the first of two available methods. The context is unchanged: European call valuation on a non-dividend paying stock.

The idea is to obtain a partial differential equation whose solution, given the appropriate boundary condition, is the price of the call. This approach is due to Black and Scholes (1973) and, in a more general context, Merton (1973). The latter author's arguments will guide our discussion here.

In the same spirit as the replicating portfolio approach introduced in Section 13.4, Merton (1973) noted that the payoff to a call can be represented in continuous time by a portfolio of the underlying stock and a risk-free bond whose quantities are continuously adjusted. Given the stochastic differential equation that governs the stock's price (Eq. (15.13)) and another nonstochastic differential equation governing the bond's price evolution, it becomes possible to construct the stochastic differential equation governing the value of the replicating portfolio. This latter transformation is accomplished via an important theorem referred to in the literature as Ito's lemma. Using results from the stochastic calculus, this expression can be shown to imply that the value of the replicating portfolio must satisfy a particular partial differential equation. Together with the appropriate boundary condition (e.g., that $C(T) = \max\{q^e(T) - E, 0\}$), this partial differential equation (PDE) has a known solution—the Black–Scholes formula.

In what follows we begin with a brief overview of Merton's approach. This is illustrated in three steps.

15.6.1 The Behavior of Stochastic Differentials

In order to motivate what follows, we need to get a better idea of what the object $dz(t)$ means. It is clearly a random variable of some sort. We first explore its moments. Formally, $dz(t)$ is

⁶ The estimate obtained using a Monte Carlo simulation will very likely closely approximate the Black–Scholes value to a high degree of precision, however, if the number of simulated underlying stock prices is large ($\geq 10,000$) and the parameters r_f , E , σ , and T used in each method are identical.

$$\lim_{\Delta t \rightarrow 0} z(t + \Delta t) - z(t) \quad (15.22)$$

where we will not attempt to be precise as to how the limit is taken. We are reminded, however, that

$$E[z(t + \Delta t) - z(t)] = 0, \text{ and}$$

$$\text{var}[z(t + \Delta t) - z(t)] = (\sqrt{\Delta t})^2 = \Delta t, \text{ for all } \Delta t$$

It is not entirely surprising, therefore, that

$$E(dz(t)) \equiv \lim_{\Delta t \rightarrow 0} E[z(t + \Delta t) - z(t)] = 0 \text{ and} \quad (15.23)$$

$$\text{var}(dz(t)) \equiv \lim_{\Delta t \rightarrow 0} E[(z(t + \Delta t) - z(t))^2] = dt \quad (15.24)$$

The object $dz(t)$ may thus be viewed as denoting an infinitesimal random variable with zero mean and variance dt .

There are several other useful relationships:

$$E(dz(t)dz(t)) \equiv \text{var}(dz(t)) = dt \quad (15.25)$$

$$\text{var}(dz(t)dz(t)) \equiv \lim_{\Delta t \rightarrow 0} E[(z(t + \Delta t) - z(t))^4 - (\Delta t)^2] \approx 0 \quad (15.26)$$

$$E(dz(t)dt) = \lim_{\Delta t \rightarrow 0} E[(z(t + \Delta t) - z(t))\Delta t] = 0 \quad (15.27)$$

$$\text{var}(dz(t)dt) \equiv \lim_{\Delta t \rightarrow \infty} E[(z(t + \Delta t) - z(t))^2(\Delta t)^2] \approx 0 \quad (15.28)$$

Equations (15.28) and (15.26) imply, respectively, that Eqs. (15.25) and (15.27) are not only satisfied in expectation but with equality. Equation (15.25) is, in particular, quite surprising, as it argues that the square of a Brownian motion random process is effectively deterministic.

These results are frequently summarized by Table 15.1.

The expression $(dt)^2$ is negligible in the table in the sense that it is very much smaller than dt and we may treat it as zero.

Table 15.1: Products of Stochastic Differentials

	dz	dt
dz	dt	0
dt	0	0

The power of these results is apparent if we explore their implications for the computation of a quantity such as $(dq^e(t))^2$:

$$\begin{aligned}(dq^e(t))^2 &= (\mu dt + \sigma dz(t))^2 \\ &= \mu^2(dt)^2 + 2\mu\sigma dt dz(t) + \sigma^2(dz(t))^2 \\ &= \sigma^2 dt\end{aligned}$$

since, by the results in [Table 15.1](#), $(dt)(dt) = 0$ and $dt dz(t) = 0$.

The object $dq^e(t)$ thus behaves in the manner of a random walk in that its variance is proportional to the length of the time interval.

We will use these results in the context of Ito's lemma.

15.6.2 Ito's Lemma

A statement of this fundamental result is presented in [Theorem 15.1](#).

Theorem All.2.1 (Ito's Lemma) Consider an Ito process $dx(t)$ of form $dx(t) = a(x(t), t)dt + b(x(t), t)dz(t)$, where $dz(t)$ is a Brownian motion and consider a process $y(t) = F(x(t), t)$.

Under quite general conditions, $y(t)$ satisfies the stochastic differential equation

$$dy(t) = \frac{\partial F}{\partial x} dx(t) + \frac{\partial F}{\partial t} dt + \frac{1}{2} \frac{\partial^2 F}{\partial x^2} (dx(t))^2 \quad (15.29)$$

The presence of the rightmost term (which would be absent in a standard differential equation) is due to the unique properties of a stochastic differential equation. Taking advantage of the results in [Table 15.1](#), let us specialize [Eq. \(15.29\)](#) to the standard Ito process where, for notational simplicity, we suppress the dependence of coefficients $a(\cdot)$ and $b(\cdot)$ on $x(t)$ and t :

$$\begin{aligned}dy(t) &= \frac{\partial F}{\partial x} (a dt + b dz(t)) + \frac{\partial F}{\partial t} dt \\ &\quad + \frac{1}{2} \frac{\partial^2 F}{\partial x^2} (a dt + b dz(t))^2 \\ &= \frac{\partial F}{\partial x} a dt + \frac{\partial F}{\partial x} b dz(t) + \frac{\partial F}{\partial x} dt \\ &\quad + \frac{1}{2} \frac{\partial^2 F}{\partial x^2} (a^2(dt)^2 + ab dt dz(t) + b^2(dz(t))^2)\end{aligned}$$

Note that $(dt)^2 = 0$, $dt dz(t) = 0$, and $(dz(t))^2 = dt$

Making these substitutions and collecting terms gives

$$dy(t) = \left(\frac{\partial F}{\partial x} a + \frac{\partial F}{\partial t} + \frac{1}{2} \frac{\partial^2 F}{\partial x^2} b^2 \right) dt + \frac{\partial F}{\partial x} bdz(t) \quad (15.30)$$

As a simple application, let us take as given

$$dq^e(t) = \mu q^e(t)dt + \sigma q^e(t)dz(t)$$

and try to derive the relationship for $d \ln q^e(t)$.

Here we have

$$a(q^e(t), t) \equiv \mu q^e(t),$$

$$b(q^e(t), t) \equiv \sigma q^e(t) \text{ and}$$

$$\frac{\partial F}{\partial q^e(t)} = \frac{1}{q^e(t)} \text{ and } \frac{\partial^2 F}{\partial q^e(t)^2} = -\frac{1}{q^e(t)^2}$$

Lastly $(\partial F(\))/(\partial t) = 0$.

Substituting these results into Eq. (15.30) yields

$$\begin{aligned} d \ln q^e(t) &= \left[\frac{1}{q^e(t)} \mu q^e(t) + 0 + \frac{1}{2} (-1) \left(\frac{1}{q^e(t)} \right)^2 (\sigma q^e(t))^2 \right] dt \\ &\quad + \frac{1}{q^e(t)} \sigma q^e(t) dz(t) \\ &= (\mu - 1/2\sigma^2)dt + \sigma dz(t) \end{aligned}$$

as was observed earlier (Eqn. 15.14).

This is the background.

15.6.3 The Black–Scholes Formula

In his derivation of the Black-Scholes formula, Merton (1973) requires four assumptions:

1. There are no market imperfections (perfect competition), transactions costs, taxes, short sales constraints, or any other impediment to the continuous trading of securities.
2. There is unlimited riskless borrowing and lending at the constant risk-free rate. If q^b is the period t price of a discount bond, then q^b is governed by the differential equation

$$\begin{aligned} dq^b(t) &= r_f q^b(t)dt \quad \text{or} \\ q^b(t) &= q^b(0) e^{r_f t} \end{aligned}$$

3. The underlying stock's price dynamics is given by a GBM of the form

$$\begin{aligned} dq^e(t) &= \mu q^e(t)dt + \sigma q^e(t)dz(t), \\ q^b(0) &> 0 \end{aligned}$$

4. There are no arbitrage opportunities across the financial markets in which the call, the underlying stock, or the discount bond are traded.

Attention is restricted to call pricing formulae, which are functions only of the stock's price currently and the time (so, for example, the possibility of past stock price dependence is ignored), i.e.,

$$C = C(q^e(t), t)$$

By a straightforward application of Ito's lemma, the call's price dynamics must be given by

$$\begin{aligned} dC &= \left[\mu q^e(t) \frac{\partial C}{\partial q^e(t)} + \frac{\partial C}{\partial t} + \frac{\sigma^2}{2} \frac{\partial^2 C}{\partial q^e(t)^2} \right] dt \\ &\quad + \sigma q^e(t) \frac{\partial C}{\partial q^e(t)} dz(t) \end{aligned}$$

which is of limited help since the form of $C(q^e(t), t)$ is precisely what is not known. The partials with respect to $q^e(t)$ and t of $C(q^e(t), t)$ must be somehow circumvented.

Following the replicating portfolio approach, Merton (1973) defines the value of the call in terms of the self-financing, continuously adjustable portfolio P composed of $\Delta(q^e(t), t)$ shares and $N(q^e(t), t)$ risk-free discount bonds:

$$V(q^e(t), t) = \Delta(q^e(t), t)q^e(t) + N(q^e(t), t)q^b(t) \quad (15.31)$$

By a straightforward application of Ito's lemma once again, the value of the portfolio must evolve according to (suppressing functional dependence in order to reduce the burdensome notation):

$$dV = \Delta dq^e + N dq^b + d\Delta q^e + dN q^b + (d\Delta)dq^e \quad (15.32)$$

Since $V()$ is assumed to be self-financing, any change in its value can only be due to changes in the values of the constituent assets and not in the numbers of them. Thus it must be that

$$dV = \Delta dq^e + N dq^b \quad (15.33)$$

which implies that the remaining terms in Eq. (15.32) are identically zero:

$$d\Delta q^e + dN q^b + (d\Delta)dq^e \equiv 0 \quad (15.34)$$

But both $\Delta(\cdot)$ and $N(\cdot)$ are functions of $q^e(t)$ and t and thus Ito's lemma can be applied to represent their evolution in terms of $dz(t)$ and dt . Using the relationships of Table 15.1 and collecting terms, both those preceding $dz(t)$ and those preceding dt must individually be zero.

Together these relationships imply that the value of the portfolio must satisfy partial differential:

$$1/2\sigma^2(q^e)^2Vq^eq^e + r_fq^eVq^e + V_t = r_fV \quad (15.35)$$

which can be shown to have as its solution the Black–Scholes formula when coupled with the terminal condition $V(q^e(T), T) = \max[0, q^e(T) - E]$.

15.7 A Second Approach: Martingale Methods

This method originated in the work of Harrison and Kreps (1979). It is popular as a methodology because it frequently allows for simpler computations than in the PDE approach. The underlying mathematics, however, are very complex and beyond the scope of this book. In order to convey a sense of what is going on, we present a brief heuristic argument that relies on the binomial abstraction.

Recall that in the binomial model, we undertook our pricing in a tree context where the underlying asset's price process had been modified. In particular, the true probabilities of the up and down states were replaced by the corresponding risk-neutral probabilities. All assets (including the underlying stock) displayed an expected return equal to the risk-free rate in the transformed setting.

Under GBM, the underlying price process is represented by an Ito stochastic differential equation of the form

$$dq^e(t) = \mu q^e(t)dt + \sigma q^e(t)dz(t) \quad (15.36)$$

In order to transform this price process into a risk-neutral setting, two changes must be made.

1. The expression μ defines the mean return and it must be replaced by r_f . Only with this substitution will the mean return on the underlying stock become r_f . Note that r_f denotes the corresponding continuously compounded risk-free rate.
2. The standard Brownian motion process must be modified. In particular, we replace dz by dz^* , where the two processes are related via the transformation:

$$dz^* = dz + (\mu - r_f)/\sigma$$

The transformed price process is thus

$$dq^e(t) = r_f q^e(t)dt + \sigma q^e(t)dz^*(t) \quad (15.37)$$

By Eq. (15.14), the corresponding process on $\ln q^e(t)$ is

$$d\ln q^e(t) = (r_f - 1/2\sigma^2)dt + \sigma dz^*(t) \quad (15.38)$$

Let T denote the expiration date of a simple European call option. In the same spirit as the binomial model, the price of a call must be the present value of its expected payoff at expiration under the transformed process.

Equation (15.38) informs us that in the transformed economy,

$$\ln\left(\frac{q^e(T)}{q^e(0)}\right) \sim N((r_f - (1/2)\sigma^2)T, \sigma^2 T) \quad (15.39)$$

Since

$$\text{Prob}_{\substack{\text{transformed} \\ \text{economy}}} (q^e(t) \geq E) = \text{Prob}_{\substack{\text{transformed} \\ \text{economy}}} (\ln q^e(t) \geq \ln E)$$

we can compute the call's value using the probability density implied by Eq. (15.39):

$$C = e^{-r_f T} \int_{\ln E}^{\infty} (e^s - E) f(s) ds$$

where $f(s)$ is the probability density of the log of the stock's price. This becomes

$$\begin{aligned} C &= e^{-r_f T} \left(\frac{1}{\sqrt{2\pi\sigma^2 T}} \right) \int_{\ln E}^{\infty} (e^s - E) \\ &\quad \times e^{-(s - \ln q^e(0) - r_f T + (\sigma^2 T/2))^2 / 2\sigma^2 T} ds \end{aligned} \quad (15.40)$$

which, when the integration is performed, yields the Black–Scholes formula.

15.8 Applications

We make reference to a number of applications that have been considered earlier in the text.

15.8.1 The Consumption–Savings Problem

This is a classic economic problem and we considered it fairly thoroughly in Chapter 4. Without the requisite math background, there is not a lot we can say about the continuous time analogue other than to set up the problem, but even that first step will be illuminating.

Suppose the risky portfolio (M) is governed by the following price process:

$$dq^M(t) = q^M(t)[\mu_M dt + \sigma_M dz(t)]$$

$q^M(0)$ given, and the risk-free asset by

$$dq^B(t) = r_f q^B(t)dt, \quad q^B(0) \text{ given}$$

If an investor has initial wealth $Y(0)$ and chooses to invest the proportion $w(t)$ (possibly continuously varying) in the risky portfolio, then his wealth $Y(t)$ will evolve according to

$$\begin{aligned} dY(t) &= Y(t)[w(t)(\mu_M - r_f) + r_f]dt \\ &\quad + Y(t)[w(t)\sigma dz(t)] - c(t)dt \end{aligned} \tag{15.41}$$

where $c(t)$ is his consumption path. With objective function

$$\max_{c(t), w(t)} E \int_0^T e^{-\gamma t} U(c(t))dt \tag{15.42a}$$

the investor's problem is one of maximizing Eq. (15.42a) subject to Eq. (15.41) and initial conditions on wealth and the constraint that $Y(t) \geq 0$ for all t .

A classic result allows us to transform this problem into one that can be solved much more easily:

$$\begin{aligned} \max_{c(t), w(t)} & E \int_0^T e^{-\gamma t} U(c(t))dt \\ \text{s.t. } & PV_0(c(t)) = E^* \int_0^T e^{-r_f t} c(t)dt \leq Y(0) \end{aligned} \tag{15.42b}$$

where E^* is the transformed risk-neutral measure under which the growth rate of the risky portfolio is r_f .

In what we have presented so far, all the notation is directly analogous to that of Chapter 4: $U(\cdot)$ is the investor's utility of (instantaneous) consumption, γ his (instantaneous) discount rate, and T his time horizon.

15.8.2 An Application to Portfolio Analysis

Here we hope to give a hint of how to extend the portfolio analysis of Chapters 5 and 6 to a setting where trading is (hypothetically) continuous and individual security returns follow GBMs.

Let there be $i = 1, 2, \dots, N$ equity securities, each of whose return is governed by the process in Eq. (15.43).

$$\frac{dq_i^e(t)}{q_i^e(t)} = \mu_i dt + \sigma_i dz_i(t) \tag{15.43}$$

where $\sigma > 0$. These processes may also be correlated with one another in a manner that we can represent precisely. Conducting a portfolio analysis in this setting has been found to have two principal advantages. First, it provides new insights concerning the implications of diversification for long-run portfolio returns and, second, it allows for an easier solution to certain classes of problems. We will note these advantages with the implicit understanding that the derived portfolio rules must be viewed as guides for practical applications. Literally interpreted they will imply, for example, continuous portfolio rebalancing—at an unbounded total expense, if the cost of doing each rebalancing is positive—which is absurd. In practice, one would rather employ them weekly or perhaps daily.

The stated objective is to maximize the expected rate of appreciation of a portfolio's value, or equivalently, to maximize its expected terminal value, which is the terminal wealth of the investor who owns it. Most portfolio managers would be familiar with this goal.

To get an idea of what this simplest criterion implies, and to make it more plausible in our setting, we first consider the discrete time equivalent (and, by implication) the discrete time approximation to GBM.

15.8.2.1 *Digression to Discrete Time*

Suppose a CRRA investor has initial wealth $Y(0)$ at time $t = 0$ and is considering investing in any or all of a set of stocks whose returns are iid. Since the rate of expected appreciation of the portfolio is its expected rate of return, and since the return distributions of the available assets are iid, the investor's optional portfolio proportions will be invariant to the level of his wealth, and the distribution of his portfolio's returns will itself be iid. At the conclusion of his planning horizon, T periods from the present, the investor's wealth will be

$$Y_T = Y_0 \prod_{s=1}^T \tilde{R}_s^P \quad (15.44)$$

where \tilde{R}_s^P denotes the (iid) gross portfolio return in period s . It follows that

$$\begin{aligned} \ln\left(\frac{Y_T}{Y_0}\right) &= \sum_{s=1}^T \ln \tilde{R}_s^P \text{ and} \\ \ln\left(\frac{Y_T}{Y_0}\right)^{1/T} &= \left(\frac{1}{T}\right) \sum_{s=1}^T \ln \tilde{R}_s^P \end{aligned} \quad (15.45)$$

Note that whenever we introduce the log we effectively convert to continuous compounding within the time period. As the number of periods in the time horizon grows without bound, $T \mapsto \infty$, by the law of large numbers,

$$\left(\frac{Y_T}{Y_0}\right)^{1/T} \mapsto e^{E \ln \tilde{R}^P} \text{ or} \quad (15.46)$$

$$Y_T \mapsto Y_0 e^{T E \ln \tilde{R}^P} \quad (15.47)$$

Consider an investor with a many-period time horizon who wishes to maximize her expected terminal wealth under continuous compounding. The relationship in Eq. (15.47) informs her that:

1. it is sufficient, under the aforementioned assumptions, for her to choose portfolio proportions that maximize $E \ln \tilde{R}^P$, the expected logarithm of the one-period return, and
2. by doing so the average growth rate of her wealth will approach a deterministic limit.

Before returning to the continuous time setting, let us present a brief classic example, one in which an investor must decide what fractions of his wealth to assign to a highly risky stock and to a risk-free asset (actually, the risk-free asset is equivalent to keeping money in a shoebox under the bed). For an amount Y_0 invested in either asset, the respective returns are found in Figure 15.4.

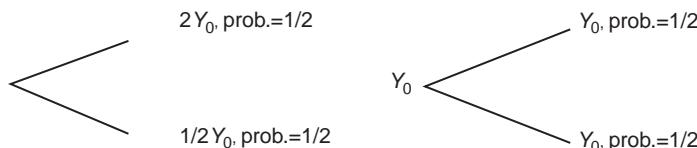


Figure 15.4
Two alternative asset returns.

Let w represent the proportion in the stock and note that the expected gross return to either asset under continuous compounding is zero:

Stock: $E \ln R^e = 1/2 \ln(2) + 1/2 \ln(1/2) = 0$

Shoebox: $E \ln R^{sb} = 1/2 \ln(1) + 1/2 \ln(1) = 0$.

With each asset paying the same expected return, and the stock being wildly risky, at first appearance the shoebox would seem the way to go. But according to Eq. (15.47), the investor ought to allocate his wealth between the two assets so as to maximize the expected log of the portfolio's one-period gross return:

$$\max_w E \ln \tilde{R}^P = \max_w \{1/2 \ln(2w + (1 - w)) + 1/2 \ln(1/2w + (1 - w))\}$$

A straightforward application of the calculus yields $w = 8/4$, with consequent portfolio returns in each state as shown in Figure 15.5.

$$\ln(1+w) = \ln(1.75) = 0.5596, \text{ prob. } = 1/2$$

$$\ln(1 - 1/2 \ln w) = \ln(0.625) = -0.47, \text{ prob. } = 1/2$$

Figure 15.5
Optimal portfolio returns in each state.

As a result, $E \ln \tilde{R}^P = 0.0448$ with an effective risk-free period return (for a very long time horizon) of 4.5% ($e^{0.448} = 1.045$).

This result is surprising and the intuition is not obvious. Briefly, the optimal proportions of $w = 3/4$ and $1-w = 1/4$ reflect the fact that by always keeping a fixed fraction of wealth in the risk-free asset, the worst wealth trajectories can be avoided. By frequent trading, although each asset has an expected return of zero, the indicated combination will yield an expected return that is strictly positive, and over a long time horizon, effectively riskless. As first noted in Chapter 13, frequent trading expands market opportunities.

15.8.2.2 Return to Continuous Time

The previous setup applies directly to a continuous time setting as all of the fundamental assumptions are satisfied. In particular, there are a very large number of periods (an uncountable number, in fact) and the returns to the various securities are iid through time. Let us make the added generalization that the individual asset returns are correlated through their Brownian motion components. By an application of Ito's lemma, we may write

$$\text{cov}(dz_i, dz_j) = E(dz_i(t)dz_j(t)) = \sigma_{ij}dt$$

where σ_{ij} denotes the (i,j) entry of the (instantaneous) variance–covariance matrix.

As has been our custom, denote the portfolio's proportions for the N assets by w_1, \dots, w_N and let the superscript P denotes the portfolio itself. As in earlier chapters, the process on the portfolio's instantaneous rate of return, $(dY^P(t))/(Y^P(t))$ will be the weighted average of the instantaneous constituent asset returns (as given in Eq. (15.43)):

$$\begin{aligned} \frac{dY^P(t)}{Y^P(t)} &= \sum_{i=1}^N w_i \frac{dq_i^e(t)}{q_i^e(t)} = \sum_{i=1}^N w_i(\mu_i dt + dz_i(t)) \\ &= \left(\sum_{i=1}^N w_i \mu_i \right) dt + \sum_{i=1}^N w_i dz_i(t) \end{aligned} \tag{15.48}$$

where the variance of the stochastic term is given by

$$\begin{aligned} E\left(\sum_{i=1}^N w_i dz_i(t)\right)^2 &= E\left\{\left(\sum_{i=1}^N w_i dz_i(t)\right)\left(\sum_{j=1}^N w_j dz_j(t)\right)\right\} \\ &= \left(\sum_{i=1}^N \sum_{j=1}^N w_i w_j \sigma_{ij}\right) dt \end{aligned}$$

Equation (15.48) describes the process on the portfolio's rate of return and we see that it implies that the portfolio's value, at any future time horizon T , will be lognormally distributed; furthermore, an uncountable infinity of periods will have passed. By analogy (and formally), our discrete time reflections suggest that an investor should, in this context, also choose portfolio proportion so as to maximize the mean growth rate, v_P , of the portfolio as given by

$$E\left\{\ln \frac{Y^P(T)}{Y(0)}\right\} = T v_P$$

Since the portfolio's value itself follows a Brownian motion (with drift $\sum_{i=1}^N w_i \mu_i$ and disturbance $\sum_{i=1}^N w_i dz_i$),

$$E\left[\ln \frac{Y^P(t)}{Y(0)}\right] = \left(\sum_{i=1}^N w_i \mu_i\right) t - \frac{1}{2} \left(\sum_{i=1}^N \sum_{j=1}^N w_i w_j \sigma_{ij}\right) t, \text{ and thus} \quad (15.49)$$

$$v_P = \frac{1}{T} E\left[\ln \frac{Y^P(t)}{Y(0)}\right] = \sum_{i=1}^N w_i \mu_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N w_i w_j \sigma_{ij} \quad (15.50)$$

The investor should choose portfolio proportions to maximize this latter quantity.

Without belaboring this development much further, it behooves us to recognize the message implicit in Eq. (15.50). This can be accomplished most straightforwardly in the context of an equally weighted portfolio, where each of the N assets is distributed independently of one another ($\sigma_{ij} = 0$ for $i \neq j$), and all have the same mean and variance ($(\mu_i, \sigma_i) = (\mu, \sigma)$ $i = 1, 2, \dots, N$).

In this case Eq. (15.50) reduces to

$$v_P = \mu - \left(\frac{1}{2N}\right) \sigma^2 \quad (15.51)$$

with the direct implication that the more identical stocks the investor adds to the portfolio the higher the mean instantaneous return. In this sense, it is useful to search for many similarly volatile stocks whose returns are independent of one another: by combining them in a portfolio where we continually (frequently) rebalance to maintain equal proportions,

not only will portfolio variance decline ($(1/2N) \sigma^2$), as in the discrete time case, but the mean return will also rise (which is *not* the case in discrete time!).

15.8.3 The Consumption CAPM in Continuous Time

Our final application concerns the consumption CAPM of Chapter 10, and the question we address is this: What is the equilibrium asset price behavior in a Mehra–Prescott asset pricing context when the growth rate in consumption follows a GBM? Specializing preferences to be of the customary form $U(c) = (c^{1-\gamma}/1 - \gamma)$, pricing relationship (10.4) reduces to

$$\begin{aligned} P_t &= E_t \left\{ Y_t \sum_{j=1}^{\infty} \delta^j x_{i+j}^{1-\gamma} \right\} \\ &= Y_t \sum_{j=1}^{\infty} \delta^j E_t \{x_{i+j}^{1-\gamma}\} \end{aligned}$$

where x_{i+j} is the growth rate in output (equivalently, consumption in the Mehra–Prescott economy) from period j to period $j + 1$.

We hypothesize that the growth rate x follows a GBM of the form

$$dx = \mu x dt + \sigma x dz$$

where we interpret x_{i+j} as the discrete time realization of $x(t)$ at time $i + j$.

One result from statistics is needed. Suppose \tilde{w} is lognormally distributed which we write $\tilde{w} \sim L(\xi, n)$ where $\xi = E \ln \tilde{w}$ and $\eta^2 = \text{var} \ln \tilde{w}$. Then for any real number q ,

$$E\{\tilde{w}^q\} = e^{q\xi + 1/2q^2\eta^2}$$

By the process on the growth rate just assumed, $x(t) \sim L((\mu - 1/2\sigma^2)t, \sigma\sqrt{t})$ so that at time $t + j$, $x_{t+j} \sim L((\mu - 1/2\sigma^2)j, \sigma\sqrt{j})$. By this result,

$$\begin{aligned} E_i\{x_{i+j}^{1-\gamma}\} &= e^{(1-\gamma)(\mu-1/2\sigma^2)j+1/2(1-\gamma)^2\sigma^2j} \\ &= e^{(1-\gamma)(\mu-1/2\gamma\sigma^2)j} \end{aligned}$$

and thus,

$$\begin{aligned} q_t &= Y_t \sum_{j=1}^{\infty} \delta^j e^{(1-\gamma)(\mu-1/2\gamma\sigma^2)j}, \\ &= Y_t \sum_{j=1}^{\infty} (\delta^j e^{(1-\gamma)(\mu-1/2\gamma\sigma^2)j}) \end{aligned}$$

which is well defined (the sum has a finite value) if $\beta e^{(1-\gamma)(\mu-1/2\gamma\sigma^2)} < 1$, which we will assume to be the case. Then

$$q_t = Y_t \frac{\beta e^{(1-\gamma)(\mu-1/2\gamma\sigma^2)}}{1 + \beta e^{(1-\gamma)(\mu-1/2\gamma\sigma^2)}} \quad (15.52)$$

This is an illustration of the fact that working in continuous time often allows convenient closed solutions.

These remarks are taken from Mehra and Sah (2002). There is much that may be said. In particular, there are many more extensions of CCAPM style models to a continuous time setting. Another issue we have not addressed is the sense in which a continuous time price process (e.g., Eq. (15.13)) can be viewed as an equilibrium price process in the sense of that concept as presented in this book.

15.9 Final Comments

Continuous time is clearly different from discrete time, but does its use (as a derivatives pricing tool) enrich our economic understanding of the larger financial and macroeconomic reality? That is not clear. It does, however, make available valuation formulae that are completely unaccessible using discrete time methodologies. Expression (15.52) is a case in point.

References

- Mehra, R., Sah, R., 2002. Mood fluctuations, projection bios, and volatility of equity prices. *J. Econ. Dynam. Control.* 26, 869–887.
- Harreson, J.M., Kreps, D., 1979. Martingales and arbitrage in multiperiod securities markets. *J. Econ Theory.* 20, 381–408.
- Merton, R., 1973. Theory of Rational Option Pricing. *Bell. J. Econ. Manage. Sci.* 4, 141–183.
- Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. *J. Polit Econ.* 81, 637–654.

Portfolio Management in the Long Run

Chapter Outline

16.1 Introduction	469
16.2 The Myopic Solution	472
16.3 Variations in the Risk-Free Rate	478
16.3.1 The Budget Constraint	479
16.3.2 The Optimality Equation	481
16.3.3 Optimal Portfolio Allocations	482
16.3.4 The Nature of the Risk-Free Asset	484
16.3.5 The Role of Bonds in Investor Portfolios	485
16.4 The Long-Run Behavior of Stock Returns	486
16.4.1 Solving for Optimal Portfolio Proportions in a Mean Reversion Environment	489
16.4.2 Strategic Asset Allocation	491
16.4.3 The Role of Stocks in Investor Portfolios	492
16.5 Background Risk: The Implications of Labor Income for Portfolio Choice	492
16.6 An Important Caveat	501
16.7 Another Background Risk: Real Estate	501
16.8 Conclusions	504
References	505

16.1 Introduction

The canonical portfolio problem (Section 5.1) and the modern portfolio theory (MPT) selection problem embedded in the CAPM are both one-period utility-of-terminal-wealth maximization problems.

As such, the advice to investors implicit in these theories is astonishingly straightforward:

1. *Be well diversified.* Conceptually, this recommendation implies that the risky portion of an investor's portfolio should resemble (be perfectly positively correlated with) the true market portfolio M . In practice, it usually means holding the risky component of invested wealth as a set of stock index funds, each one representing the stock market of a particular major market capitalization country with the relative proportions

dependent upon the relevant *ex ante* variance–covariance matrix estimated from recent historical data.

2. *Be on the capital market line.* That is, the investor should allocate his wealth between risk-free assets and the aforementioned major market portfolio in proportions that are consistent with his subjective risk tolerance. Implicit in this second recommendation is that the investor first estimates her coefficient of relative risk aversion as per Section 4.5 and then solves a joint savings–portfolio allocation problem of the form illustrated in Section 5.6.3. The risk-free rate used in these calculations is customarily a 1-year T-bill rate in the United States or its analogue elsewhere.

But what should the investor do next period after this period's risky portfolio return realization has been observed? Our one-period theory has nothing to say on this score except to invite the investor to repeat the above two-step process, possibly using an updated variance–covariance matrix and an updated risk-free rate. This is what is meant by the investor behaving myopically. Yet we are uneasy about leaving the discussion at this level. Indeed, a number of important considerations seem to be intentionally ignored by following such a set of recommendations.

1. There is some evidence (see Web Note 7.1) that equity return distributions have historically evolved in a pattern that is mean reverting. This is the property that a series of high-return realizations is on average followed by a series of low ones. This variation in conditional returns might reasonably be expected to influence intertemporal portfolio composition. The reader will recall that we first considered the implications of mean reversion in equity returns in Section 7.5.3 and noted there that it has the consequence of reducing long-run equity investment risk.

In Chapter 7 we also considered the implications of parameter uncertainty; in particular, possible uncertainty in the current mean stock return as well as uncertainty in the future mean returns. In the present chapter we will eschew this important consideration and assume all relevant means, variances and correlations are known and accurately estimated. The qualitative conclusions of this chapter are unlikely to be overturned by adding the sources of uncertainty we choose to ignore.

2. While known *ex ante* relative to the start of a period, the risk-free rate also varies through time (for the period 1928–1985 the standard deviation of the US T-bill rate is 5.67%). From the perspective of a long-term US investor, the one-period T-bill rate no longer represents a truly risk-free return. Can any asset be viewed as risk free from a multiperiod perspective?
3. Investors typically receive labor income, and this fact will likely affect both the quantity of investable savings and the risky/risk-free portfolio composition decision. The latter possibility follows from the observation that labor income may be viewed as the “dividend” on an implicit nontradable human capital asset, whose value may be differentially correlated with various risky assets in the investor's financial wealth

portfolio.¹ If labor income were risk free (tenured professors!), the presence of a high-value risk-free asset in the investor's overall wealth portfolio would likely tilt his security holdings in favor of a greater proportion in risky assets than would otherwise be the case.

4. There are other life cycle considerations: savings for the educational expenses of children, the gradual disappearance of the labor income asset as retirement approaches, and so on. How do these obligations and events impact portfolio choice?
5. There is also the issue of real estate. Not only does real estate (we are thinking of owner-occupied housing for the moment) provide a risk-free service flow, but it is also expensive for an investor to alter his stock of housing. How should real estate figure into an investor's multiperiod investment plan?
6. Other considerations abound. There are substantial taxes and transactions costs associated with rebalancing a portfolio of securities. Taking these costs into account, how frequently should a long-term investor optimally alter his portfolio's composition?

In this chapter, we propose to present some recent research regarding these issues. Our perspective is one in which investors live for many periods. (In the case of private universities, foundations or insurance companies, it is reasonable to postulate an infinite lifetime.) For the moment, we will set aside the issue of real estate and explicit transactions costs, and focus on the problem of a long-lived investor confronted with jointly deciding, on a period-by-period basis, not only how much he should save and consume out of current income, but also the mix of assets, risky and risk free, in which his wealth should be invested.

In its full generality, the problem confronting a multiperiod investor–saver with outside labor income is thus

$$\max_{\{a_t, S_t\}} E \left(\sum_{t=0}^T \delta^t U(\tilde{C}_t) \right) \quad (16.1)$$

$$\text{s.t. } C_T = S_{T-1}a_{T-1}(1 + \tilde{r}_T) + S_{T-1}(1 - a_{T-1})(1 + r_{f,T}) + \tilde{L}_T, \quad t = T$$

$$C_t + S_t \leq S_{t-1}a_{t-1}(1 + \tilde{r}_t) + S_{t-1}(1 - a_{t-1})(1 + r_{f,t}) + \tilde{L}_t, \quad 1 \leq t \leq T - 1$$

$$C_0 + S_0 \leq Y_0 + L_0, \quad t = 0$$

¹ In particular, the value of an investor's labor income asset is likely to be highly correlated with the return on the stock of the firm with which he is employed. From a wealth management perspective, basic intuition would suggest that the stock of one's employer should not be held in significant amounts.

where \tilde{L}_t denotes the investor's (possibly uncertain) period t labor income, $r_{f,t}$ the period risk-free rate, and \tilde{r}_t represents the period t return on the risky asset which we shall understand to mean a well-diversified stock portfolio.²

Equation (16.1) departs from our earlier notation in a number of ways that will be convenient for developments later in this chapter. In particular, C_t and S_t denote, respectively, period t consumption and savings rather than their lower case analogues (as in Chapter 5).

The fact that the risk-free rate is indexed by t admits the possibility that this quantity, though known at the start of a period, can vary from one period to the next. Lastly, a_t will denote the *proportion* of the investor's savings assigned to the risky asset (rather than the absolute amount as before).

All other notation is standard; Eq. (16.1) is simply the multiperiod version of the portfolio problem in Section 5.6.3 augmented by the introduction of labor income. In what follows we will also assume that all risky returns are lognormally distributed and that the investor's $U(C_t)$ is of the power utility constant relative risk aversion (CRRA) class. The latter is needed to make certain that risk aversion is independent of wealth. Although investors have become enormously wealthier over the past 200 years, risk-free rates and the return premium on stocks have not changed markedly, facts otherwise inconsistent with risk aversion dependent on wealth.

In its full generality, Eq. (16.1) is both very difficult to solve and begrudging of intuition. We thus restrict its scope and explore a number of special cases. The natural place to begin is to explore the circumstances under which the myopic solution of Section 5.3 carries over to the dynamic context of problem (16.1).

16.2 The Myopic Solution

With power utility, an investor's optimal savings to wealth ratio will be constant so that the key to a fully myopic decision rule will lie in the constancy of the a ratio. Intuitively, if the same portfolio decisions are to be made, a natural sufficient condition would be to guarantee that the investor is confronted by the same opportunities on a period-by-period basis. Accordingly, we assume the return environment is not changing through time; in other words that $r_{f,t} \equiv r_f$ is constant and $\{\tilde{r}_t\}$ is independently and identically distributed (i.i.d.). These assumptions guarantee that future prospects look the same period after period. Further exploration mandates that $L_t \equiv 0$. (With constant r_f , the value of this asset will

² This portfolio might be the market portfolio M but not necessarily. Consider the case in which the investor's labor income is paid by one of the firms in M . It is likely that this particular firm's shares would be underweighted (relative to M) in the investor's portfolio.

otherwise be monotonically declining, which is an implicit change in future wealth.) We summarize these considerations in the following theorem.

Theorem 16.1 (Merton, 1971) Consider the canonical multiperiod consumption–savings–portfolio allocation problem (16.1); suppose $U(\cdot)$ displays CRRA, r_f is constant, and $\{\tilde{r}_t\}$ is i.i.d. Then the proportion a_t is time invariant.³

This is an important result in the following sense. It delineates the conditions under which a pure static portfolio choice analysis may be generalized to a multiperiod context. The optimal portfolio choice—in the sense of the allocation decision between the risk-free and the risky asset—defined in a static one-period context will continue to characterize the optimal portfolio decision in the more natural multiperiod environment. The conditions that are imposed are easy to understand: if the returns on the risky asset were not independently distributed, today’s realization of the risky return would provide information about the future return distribution, which would almost surely affect the allocation decision.

Suppose, for example, that returns are positively correlated. Then a good realization today would suggest that high returns are more likely again tomorrow. It would be natural to take this into account by, say, increasing the share of the risky asset in the portfolio. (Beware, however, that, as the first sections of Chapter 5 illustrate, without extra assumptions on the shape of the utility function—beyond risk aversion—the more intuitive result may not generally obtain. We will be reminded of this in Chapter 17 where, in particular, the log utility agent will stand out as a benchmark.) The same can be said if the risk-free rate is changing through time. In a period of high-risk-free rates, the riskless asset will be more attractive, all other things equal.

The need for the other assumption—the CRRA utility specification—is a direct consequence of Theorem 5.5. With a utility form other than CRRA, Theorem 5.5 tells us that the share of wealth invested in the risky asset varies with the “initial” wealth level, i.e., the wealth level carried over from the last period. But in a multiperiod context, the investable wealth, i.e., the savings level, is sure to be changing over time, increasing when realized returns are favorable and decreasing otherwise. With a non-CRRA utility function, optimal portfolio allocations would consistently be affected by these changes.

Now let us illustrate the power of these ideas to evaluate an important practical problem. Consider the problem of an individual investor saving for retirement: at each period he must decide what fraction of his already accumulated wealth should be invested in stocks (understood to mean a well-diversified portfolio of risky assets) and risk-free bonds for the

³ If the investor’s period utility is log, it is possible to relax the independence assumption. This important observation, first made by Samuelson (1969), will be confirmed later on in this chapter.

next investment period. We will maintain the $L_t \equiv 0$ assumption. Popular wisdom in this area can be summarized in the following three assertions:

1. Early in life, the investor should invest nearly all of her wealth in stocks (stocks have historically outperformed risk-free assets over long—20-year—periods), while gradually shifting almost entirely into risk-free instruments as retirement approaches in order to avoid the possibility of a catastrophic loss.
2. If an investor is saving for a target level of wealth (such as, in the United States, college tuition payments for children), he should gradually reduce his holdings in stocks as his wealth approaches the target level in order to minimize the risk of a shortfall due to an unexpected market downturn.
3. Investors who are working and saving from their labor income should rely more heavily on stocks early in their working lives, not only because of the historically higher average returns that stocks provide but also because bad stock market returns, early on, can be offset by increased saving out of labor income in later years.

Following [Jagannathan and Kocherlakota \(1996\)](#), we wish to subject these assertions to the discipline imposed by a rigorous modeling perspective. Let us maintain the assumptions of [Theorem 16.1](#) and hypothesize that the risk-free rate is constant, that stock returns $\{\tilde{r}_t\}$ are i.i.d. (recall the random walk model of Chapter 7), and that the investor's utility function assumes the standard CRRA form.

To evaluate assertion (1), let us further simplify [Eq. \(16.1\)](#) by abstracting away from the consumption—savings problem. This amounts to assuming that the investor seeks to maximize the utility of his terminal wealth, Y_T in period T , the planned conclusion of his working life. As a result, $S_t = Y_t$ for every period $t < T$ (no intermediate consumption). Under CRRA, we know that the investor would invest the same fraction of his wealth in risky assets every period (disproving the assertion), but it is worthwhile to see how this comes about in a simple multiperiod setting.

Let \tilde{r} denote the (invariant) risky return distribution; the investor solves

$$\begin{aligned} \max_{\{a_t\}} \quad & E \left\{ \frac{(\widetilde{Y}_T)^{1-\gamma}}{1-\gamma} \right\} \\ \text{s.t.} \quad & \widetilde{Y}_T = a_{T-1} Y_{T-1} (1 + \tilde{r}_T) + (1 - a_{T-1}) Y_{T-1} (1 + r_f), \quad t = T \\ & \widetilde{Y}_T = a_{t-1} Y_{t-1} (1 + \tilde{r}_t) + (1 - a_{t-1}) Y_{t-1} (1 + r_f), \quad 1 \leq t \leq T-1 \\ & Y_0 \text{ given} \end{aligned}$$

Problems of this type are most appropriately solved by working backward: first solving for the $T-1$ decision, then solving for the $T-2$ decision conditional on the $T-1$ decision, and so on. In period $T-1$ the investor solves

$$\max_{a_{T-1}} E((1-\gamma)^{-1} \{ [a_{T-1} Y_{T-1}(1+\tilde{r}) + (1-a_{T-1}) Y_{T-1}(1+r_f)]^{(1-\gamma)} \})$$

The solution to this problem, $a_{T-1} \equiv \hat{a}$, satisfies the first-order condition

$$E\{\hat{a}(1+\tilde{r}) + (1-\hat{a})(1+r_f)\}^{-\gamma}(\tilde{r} - r_f) = 0$$

As expected, because of the CRRA assumption, the optimal fraction invested in stocks is independent of the period $T - 1$ wealth level. Given this result, we can work backward. In period $T - 2$, the investor rebalances his portfolio, knowing that in $T - 1$ he will invest the fraction \hat{a} in stocks. As such, this problem becomes

$$\begin{aligned} \max_{a_{T-2}} & E((1-\gamma)^{-1} \{ [a_{T-2} Y_{T-2}(1+\tilde{r}) + (1-a_{T-2}) Y_{T-2}(1+r_f)] \\ & [\hat{a}(1+\tilde{r}) + (1-\hat{a})(1+r_f)] \}^{(1-\gamma)}) \end{aligned} \quad (16.2)$$

Because stock returns are i.i.d., this objective function may be written as the product of expectations as per

$$\begin{aligned} & E[\hat{a}(1+\tilde{r}) + (1-\hat{a})(1+r_f)]^{(1-\gamma)}. \\ \max_{\{a_{T-2}\}} & E\{(1-\gamma)^{-1} [a_{T-2} Y_{T-2}(1+\tilde{r}) + (1-a_{T-2}) Y_{T-2}(1+r_f)]^{1-\gamma}\} \end{aligned} \quad (16.3)$$

Written in this way, the structure of the problem is no different from the prior one, and the solution is again $a_{T-2} \equiv \hat{a}$. Repeating the same argument, it must be the case that $a_t = \hat{a}$ in every period, a result that depends critically not only on the CRRA assumption (wealth factors out of the first-order condition) but also on the independence. The risky return realized in any period does not alter our belief about the future return distributions. There is no meaningful difference between the long-run (many periods) and the short-run (one period): agents invest the same fraction in stocks regardless of their portfolio's performance history. Assertion (1) is clearly not generally valid.

To evaluate our second assertion, and following again [Jagannathan and Kocherlakota \(1996\)](#), let us modify the agent's utility function to be of the form

$$U(Y_T) = \begin{cases} \frac{(Y_T - \bar{Y})^{1-\gamma}}{1-\gamma} & \text{if } Y_T \geq \bar{Y} \\ -\infty & \text{if } Y_T < \bar{Y} \end{cases}$$

where \bar{Y} is the target level of wealth. Under this formulation, it is absolutely essential that the target be achieved: as long as there exists a positive probability of failing to achieve the

target, the investor's expected utility-of-terminal wealth is $-\infty$. Accordingly, we must also require that

$$Y_0(1+r_f)^T > \bar{Y}$$

in other words, that the target can be attained by investing everything in risk-free assets. If such an inequality were not satisfied, then every strategy would yield an expected utility of $-\infty$, with the optimal strategy thus being indeterminate.

A straightforward analysis of this problem yields the following two-step solution:

- Step 1. always invest sufficient funds in risk-free assets to achieve the target wealth level with certainty.
- Step 2. invest a constant share a^* of any additional wealth in stock, where a^* is time invariant.

By this solution, the investor invests less in stocks than he would in the absence of a target, but since he invests in both stocks and bonds, his wealth will accumulate, on average, more rapidly than it would if invested solely at the risk-free rate, and the stock portion of his wealth will, on average, grow faster. As a result, the investor will typically use proportionally less of his resources to guarantee achievement of the target. And, over time, targeting will tend to *increase* the share of wealth in stocks, again contrary to popular wisdom!

In order to evaluate assertion (3), we must admit savings from labor income into the analysis. Let $\{L_t\}$ denote the stream of savings out of labor income. For simplicity, we assume that the stream of future labor income is fully known at date 0. The investor's problem is now:

$$\begin{aligned} \max_{\{a_t\}} \quad & E\left(\frac{(\widetilde{Y}_T)^{1-\gamma}}{1-\gamma}\right) \text{ s.t.} \\ \widetilde{Y}_T = & L_T + a_{T-1}Y_{T-1}(1+\tilde{r}_T) + (1-a_{T-1})Y_{T-1}(1+r_f), t=T \\ \widetilde{Y}_t \leq & L_t + a_{t-1}Y_{t-1}(1+\tilde{r}_t) + (1-a_{t-1})Y_{t-1}(1+r_f), 1 \leq t \leq T-1 \\ Y_0; \{L_t\}_{t=0}^T \quad & \text{given} \end{aligned}$$

We again abstract away from the consumption–savings problem and focus on maximizing the expected utility of terminal wealth.

In any period, the investor now has two sources of wealth: financial wealth, Y_t^F , where

$$Y_t^F = L_t + a_{t-1}Y_{t-1}(1+r_t) + (1-a_{t-1})Y_{t-1}(1+r_f)$$

(r_t is the period t realized value of \tilde{r}); and labor income wealth, Y_t^L , is measured by the present value of the future stream of labor income. As mentioned, we assume this income stream is risk free with present value,

$$Y_t^L = \frac{L_{t+1}}{(1+r_f)} + \dots + \frac{L_T}{(1+r_f)^{T-1}}$$

Since the investor continues to have CRRA preferences, he will, in every period, invest a constant fraction of his total wealth \hat{a} in stocks, where \hat{a} depends only upon his CRRA and the characteristics of the return distributions \tilde{r} and r_f , i.e.,

$$A_t = \hat{a}(Y_t^F + Y_t^L)$$

where A_t denotes the *amount* invested in the risky financial asset.

As the investor approaches retirement, his Y_t^L declines. In order to maintain the same fraction of wealth invested in risk-free assets, the fraction of financial wealth invested in stocks,

$$\frac{A_t}{Y_t^F} = \hat{a} \left(1 + \frac{Y_t^L}{Y_t^F} \right)$$

must decline on average. Here at least the assertion has theoretical support, but for a reason different from what is commonly asserted.

In what follows we will consider the impact on portfolio choice of a variety of changes to the myopic context just considered. In particular, we explore the consequences of relaxing the constancy of the risk-free rate and return independence for the aforementioned recommendations. In most (but not all) of the discussion, we will assume an infinitely lived investor ($T = \infty$ in problem (16.1)). Recall that this amounts to postulating that a finitely lived investor is concerned for the welfare of his descendants. In nearly every case it enhances tractability. As a device for tying the discussion together, we will also explore how robust the three investor recommendations just considered are to a more general return environment.

Our first modification admits a variable risk-free rate; the second generalizes the return generating process on the risky asset (no longer i.i.d. but “mean reverting”). Our remarks are largely drawn from a prominent publication (Campbell and Viceira, 2002).

Following the precedents established by these authors, it will prove convenient to log-linearize the investor’s budget constraint and optimality conditions. Simple and intuitive expression for optimal portfolio proportions typically results. Some of the underlying deviations are provided in an appendix available on this text’s web site; others are simply omitted when they are lengthy and complex and where an attractive intuitive interpretation is available.

In the next section, the risk-free rate is allowed to vary, though in a particularly structured way.

16.3 Variations in the Risk-Free Rate

Following [Campbell and Viceira \(2002\)](#), we specialize [Eq. \(16.1\)](#) to admit a variable risk-free rate. Other assumptions are:

- i. $L_t \equiv 0$ for all t ; there is no labor income so that all consumption comes from financial wealth alone.
- ii. $T = \infty$, i.e., we explore the infinite horizon version of [Eq. \(16.1\)](#); this allows a simplified description of the optimality conditions on portfolio choice.
- iii. All relevant return random variables are lognormal with constant variances and covariances. This is an admittedly strong assumption as it mandates that the return on the investor's portfolio has a constant variance and that the constituent assets have constant variances and covariances with the portfolio itself. Thus, the composition of the risky part of the investor's portfolio must itself be invariant. But this will be optimal only if the expected excess returns above the risk-free rate on these same constituent assets are also constant. Expected returns can vary over time, but, in effect, they must move in tandem with the risk-free rate. This assumption is considerably specialized, but it does allow for unambiguous conclusions.
- iv. The investor's period utility function is of the Epstein–Zin variety (cf. [Section 5.7.3](#)). In this case, the intertemporal optimality condition for [Eq. \(16.1\)](#) when $T = \infty$ and there are multiple risky assets can be expressed as (recall [Eqn. 10.3.2](#))

$$1 = E_t \left\{ \left[\delta \left(\frac{\tilde{C}_{t+1}}{C_t} \right)^{-\frac{1}{\rho}} \right]^\theta \left[\frac{1}{\tilde{R}_{P,t+1}} \right]^{1-\theta} \tilde{R}_{i,t+1} \right\} \quad (16.4)$$

where $\tilde{R}_{i,t}$ is the period t gross return on any available asset (risk free or otherwise, including the portfolio itself) and $\tilde{R}_{P,t}$ is the period t overall risky portfolio's gross return. Note that consumption C_t , and the various returns $\tilde{R}_{P,t}$ and $\tilde{R}_{i,t}$ are capitalized. We will henceforth denote the logs of these quantities by their respective lower case counterparts.⁴ [Equation \(16.4\)](#) is simply a restatement of [Eq. \(9.28\)](#), where γ is the risk-aversion parameter, ρ is the elasticity of intertemporal substitution, and $\theta = (1 - \gamma)/(1 - 1/\rho)$.

⁴ Recall from [Chapter 3, Box 31](#), that $r_t^c - \log R_t$ is the continuously compounded net rate of return over period t . When net returns are not large, $r_t^c = \log R_t = \log(1 + r_t) \approx r_t$: the net return and its continuously compounded counterpart are essentially the same.

Bearing in mind assumptions (i)–(iv), we now proceed first to the investor's budget constraint and then to his optimality condition. The plan is to log-linearize each expression. This will simplify the necessary development leading to our ultimate goal, Eq. (16.20).

16.3.1 The Budget Constraint

In a model with period consumption exclusively out of financial wealth, the intertemporal budget constraint is of the form

$$Y_{t+1} = (R_{P,t+1})(Y_t - C_t) \quad (16.5)$$

where the risky portfolio P potentially contains many risky assets; equivalently,

$$\frac{Y_t + 1}{Y_t} = (R_{P,t+1}) \left(1 - \frac{C_t}{Y_t} \right) \quad (16.6)$$

or, taking the log on both sides of the equation,

$$\Delta y_{t+1} = \log Y_{t+1} - \log Y_t = \log(R_{P,t+1}) + \log(1 - \exp(\log C_t - \log Y_t))$$

Recalling our identification of a lowercase variable with the log of that variable, we have

$$\Delta y_{t+1} = r_{P,t+1} + \log(1 - \exp(\tilde{c}_t - \tilde{y}_t))$$

Assuming that the $\log(C_t/Y_t)$ is not too variable (essentially this places us in the $\rho = \gamma = 1$ – the log utility case), then the rightmost term can be approximated around its mean to yield (see Campbell and Viceira, 2001a):

$$\Delta y_{t+1} = k_1 + r_{P,t+1} + \left(1 - \frac{1}{k_2} \right) (\tilde{c}_t - \tilde{y}_t) \quad (16.7)$$

where k_1 and $k_2 < 1$ are constants related to $\exp(E(\tilde{c}_t - \tilde{y}_t))$.

Speaking somewhat informally in a fashion that would identify the log of a variable with the variable itself, Eq. (16.7) simply states that wealth will be higher next period ($t + 1$) in a manner that depends on both the portfolio's rate of return ($r_{P,t+1}$) over the next period and on this period's consumption relative to wealth. If c_t greatly exceeds y_t , wealth next period cannot be higher!

We next employ an identity to allow us to rewrite Eq. (16.7) in a more useful way; it is

$$\Delta y_{t+1} = \Delta c_{t+1} + (c_t - y_t) - (c_{t+1} - y_{t+1}) \quad (16.8)$$

where $\Delta c_{t+1} = c_{t+1} - c_t$. Substituting the RHS of Eq. (16.8) into Eq. (16.7) and rearranging terms yields

$$(c_t - y_t) = k_2 k_1 + k_2(r_{P,t+1} - \Delta c_{t+1}) + k_2(c_{t+1} - y_{t+1}) \quad (16.9)$$

Equation (16.9) provides the same information as Eq. (16.7), albeit expressed differently. It states that an investor could infer his (log) consumption–wealth ratio ($c_t - y_t$) in period t from a knowledge of its corresponding value in period $t + 1$, ($c_{t+1} - y_{t+1}$), and his portfolio’s return (the growth rate of his wealth) relative to the growth rate of his consumption ($r_{P,t+1} - \Delta c_{t+1}$). (Note that our use of language again informally identifies a variable with its log.)

Equation (16.9) is a simple difference equation that can be solved forward to yield

$$c_t - y_t = \sum_{j=1}^{\infty} (k_2)^j (\tilde{r}_{P,t+j} - \Delta \tilde{c}_{t+j}) + \frac{k_2 k_1}{1 - k_2} \quad (16.10)$$

Equation (16.10) also has an attractive intuitive interpretation; a high (above average) consumption–wealth ratio ($(c_t - y_t)$ large and positive), i.e., a burst of consumption must be followed either by high returns on invested wealth or lowered future consumption growth. Otherwise the investor’s intertemporal budget constraint cannot be satisfied. But Eq. (16.10) holds *ex ante* relative to time t as well as *ex post*, its current form. Equation (16.11) provides the *ex ante* version:

$$c_t - y_t = E_t \sum_{j=1}^{\infty} (k_2)^j (\tilde{r}_{P,t+j} - \Delta \tilde{c}_{t+j}) + \frac{k_2 k_1}{1 - k_2} \quad (16.11)$$

Substituting this expression twice into the RHS of Eq. (16.8), substituting the RHS of Eq. (16.7) for the LHS of Eq. (16.8), and collecting terms yield our final representation for the log-linearized budget constraint equation:

$$c_{t+1} - E_t \tilde{c}_{t+1} = (E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{P,t+1+j} - (E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \Delta \tilde{c}_{t+1+j} \quad (16.12)$$

This equation again has an intuitive interpretation: if consumption in period $t + 1$ exceeds its period t expectation ($c_{t+1} > E_t \tilde{c}_{t+1}$, a positive consumption “surprise”), then this consumption increment must be “financed” either by an upward revision in expected future portfolio returns (the first term on the LHS of Eq. (16.12)) or a downward revision in future consumption growth (as captured by the second term on the LHS of Eq. (16.12)). If it were otherwise, the investor would receive “something for nothing”—as though his budget constraint could be ignored.

Since our focus is on deriving portfolio proportions and returns, it will be useful to be able to eliminate future consumption growth (the Δc_{t+1+j} terms) from the above equation and to replace it with an expression related only to returns. The natural place to look for such an equivalence is the investor's optimality Eq. (16.4), which directly relates the returns on his choice of optimal portfolio to his consumption experience, log-linearized so as to be in harmony with Eq. (16.12).

16.3.2 The Optimality Equation

The log-linearized version of Eq. (16.4) is

$$E_t \Delta c_{t+1} = \rho \log \delta + \rho E_t \tilde{r}_{P,t+1} + \frac{\theta}{2\rho} \text{var}_t(\Delta \tilde{c}_{t+1} - \rho \tilde{r}_{P,t+1}) \quad (16.13)$$

where we have specialized Eq. (16.4) somewhat by choosing the i th asset to be the portfolio itself so that $R_{i,t+1} = R_{P,t+1}$. The web appendix provides a derivation of this expression, but it is more important to grasp what it is telling us about an Epstein–Zin investor's optimal behavior: in our partial equilibrium setting where investors take return processes as given, Eq. (16.13) states that an investor's optimal expected consumption growth ($E_t(\Delta c_{t+1})$) is linearly (by the log-linear approximation) related to the time preference parameter δ (an investor with a bigger δ will save more and thus his expected consumption growth will be higher), the portfolio returns he expects to earn ($E_t \tilde{r}_{P,t+1}$), and the miscellaneous effects of uncertainty as captured by the final term $\theta/2\rho \text{var}_t(\Delta \tilde{c}_{t+1} - \rho \tilde{r}_{P,t+1})$. A high intertemporal elasticity of substitution ρ means that the investor is willing to experience a steeper consumption growth profile if there are incentives to do so, and thus ρ premultiplies both $\log \delta$ and $E_t \tilde{r}_{P,t+1}$. Lastly, if $\theta > 0$, an increase in the variance of consumption growth relative to portfolio returns leads to a greater expected consumption growth profile. Under this condition, the variance increase elicits greater precautionary savings in period t and thus a greater expected consumption growth rate.

Under assumption (iii) of this section, however, the variance term in Eq. (16.13) is constant, which leads to a much-simplified representation

$$E_t \Delta \tilde{c}_{t+1} = k_3 + \rho E_t \tilde{r}_{P,t+1} \quad (16.14)$$

where the constant k_3 incorporates both the constant variance and the time preference term $\rho \log \delta$. Substituting Eq. (16.14) into Eq. (16.11) in the most straightforward way and rearranging terms yield

$$c_t - y_t = (1 - \rho) E_t \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{P,t+j} + \frac{k_2(k_1 - k_2)}{1 - k_2} \quad (16.15)$$

Not surprisingly, Eq. (16.15) suggests that the investor's (log) consumption to wealth ratio (itself a measure of how willing he is to consume out of current wealth) depends linearly on future discounted portfolio returns, negatively if $\rho > 1$ and positively if $\rho < 1$ where ρ is his intertemporal elasticity of substitution. The value of ρ reflects the implied dominance of the substitution over the income effect. If $\rho < 1$, the income effect dominates: if portfolio returns increase, the investor can increase his consumption permanently without diminishing his wealth. If the substitution effect dominates ($\rho > 1$), however, the investor will reduce his current consumption in order to take advantage of the impending higher expected returns. Substituting Eq. (16.15) into Eq. (16.12) yields

$$c_{t+1} - E_t \tilde{c}_{t+1} = \tilde{r}_{P,t+1} - E_t \tilde{r}_{P,t+1} + (1 - \rho)(E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{P,t+1+j} \quad (16.16)$$

an equation that attributes period $t + 1$'s consumption surprise to (1) the unexpected contemporaneous component to the overall portfolio's return $\tilde{r}_{P,t+1} - E_t \tilde{r}_{P,t+1}$, plus (2) the revision in expectation of future portfolio returns, $(E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{P,t+1+j}$. This revision either encourages or reduces consumption, depending on whether, once again, the income or substitution effect dominates. This concludes the background on which the investor's optimal portfolio characterization rests. Note that Eq. (16.16) defines a relationship by which consumption may be replaced—in some other expression of interest—by a set of terms involving portfolio returns alone.

16.3.3 Optimal Portfolio Allocations

So far, we have not employed the assumption that the expected returns on all assets move in tandem with the risk-free rate; indeed the risk-free rate is not explicit in any of expressions (16.2)–(16.14). We address these issues presently.

In an Epstein–Zin context, recall that the risk premium on any risky asset over the safe asset, $E_t \tilde{r}_{P,t+1} - r_{f,t+1}$, is given by Eq. (9.32), which is recopied below:

$$E_t \tilde{r}_{t+1} - r_{f,t+1} + \frac{\sigma_t^2}{2} = \frac{\theta \text{cov}_t(\tilde{r}_{t+1}, \Delta \tilde{c}_{t+1})}{\rho} + (1 - \theta) \text{cov}_t(\tilde{r}_{t+1}, \tilde{r}_{P,t+1}) \quad (16.17)$$

where r_{t+1} denotes the return on the stock portfolio, and $r_{P,t+1}$ is the return on the portfolio of all the investor's assets, i.e., including the “risk-free” one. Note that implicit in assumption (iii) is the recognition that all variances and covariances are constant despite the time dependency in notation.

From expression (16.16), we see that the covariance of (log) consumption with any variable (and we have in mind its covariance with the risky return variable of Eq. (16.17)) may be replaced by the covariance of that variable with the portfolio's contemporaneous return plus $(1 - \rho)$ times the expectations revisions concerning future portfolio returns. Eliminating consumption from Eq. (16.17) in this way via a judicious insertion of Eq. (16.16) yields

$$E_t \tilde{r}_{t+1} - \tilde{r}_{f,t+1} + \frac{\sigma_t^2}{2} = \gamma \operatorname{cov}_t(\tilde{r}_{t+1}, \tilde{r}_{P,t+1}) + \\ (\gamma - 1) \operatorname{cov}_t \left(\tilde{r}_{t+1}, (E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{P,t+1+j} \right) \quad (16.18)$$

As noted in Campbell and Viceira (2002), Eqs. (16.16) and (16.18) delineate in an elegant way the consequences of the Epstein and Zin separation of time and risk preferences.

In particular, in Eq. (16.16), it is only the elasticity of intertemporal substitution parameter ρ that relates current consumption to future returns (and thus income)—a time preference effect—while, in Eq. (16.18), it is only γ , the risk-aversion coefficient, that appears to influence the risk premium on the risky asset.

If we further recall (assumption (iii)) that variation in portfolio expected returns must be exclusively attributable to variation in the risk-free rate, it follows logically that revisions in expectations of the portfolio expected returns must uniquely follow from revisions of expectations of the risk-free rate:

$$(E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{P,t+1+j} = (E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{f,t+1+j} \quad (16.19)$$

In a model with one risky asset (in effect, the risky portfolio whose composition we are a priori holding constant),

$$\operatorname{cov}_t(\tilde{r}_{t+1}, \tilde{r}_{P,t+1}) = a_t \sigma_t^2 = a_t \sigma_P^2, t$$

where a_t is, as before, the risky asset proportion in the portfolio.

Substituting both this latter expression and identification (16.19) into Eq. (16.18) and solving for a_t gives the optimal, time-invariant portfolio weight on the risky asset.

$$a_t \equiv a = \frac{1}{\gamma} \left[\frac{E_t \tilde{r}_{t+1} - r_{f,t+1} + \frac{\sigma_t^2}{2}}{\sigma_t^2} \right] \\ + \left(1 - \frac{1}{\gamma} \right) \frac{1}{\sigma_t^2} \operatorname{cov}_t \left(\tilde{r}_{t+1}, -(E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{f,t+1+j} \right), \quad (16.20)$$

our first portfolio result. In what follows we offer a set of interpretative comments related to it.

1. The first term in Eq. (16.20) represents the myopic portfolio demand for the risky asset—myopic in the sense that it describes the fraction of wealth invested in the risky portfolio when the investor ignores the possibility of future risk-free rate changes. In

particular, the risky portfolio proportion is inversely related to the investor's CRRA (γ) and positively related to the risk premium. Note, however, that these rate changes are the fundamental feature of this economy in the sense that variances are fixed and all expected risky returns move in tandem with the risk-free rate.

2. The second term in Eq. (16.20) captures the risky asset demand related to its usefulness for hedging intertemporal interest rate risk. The idea is as follows. We may view the risk-free rate as the "base line" return on the investor's wealth with the risky asset providing a premium on some fraction thereof. If expected future risk-free returns are revised downward, $[-(E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{f,t+1+j}$ increases], the investor's future income (consumption) stream will be reduced unless the risky asset's return increases to compensate. This will be so on average if the covariance term in Eq. (16.20) is positive. It is in this sense that risky asset returns (r_{t+1}) can hedge risk-free interest rate risk. If the covariance term is negative, however, risky asset returns tend only to magnify the consequences of a downward revision in expected future risk-free rates. As such, a long-term investor's holding of risky assets would be correspondingly reduced.

These remarks have their counterpart in asset price changes: if risk-free rates rise (bond prices fall), the investor would wish for changes in the risky portion of his portfolio to compensate via increased valuations.

3. As the investor becomes progressively more risk averse ($\gamma \mapsto \infty$), she will continue to hold stocks in her portfolio, but only because of their hedging qualities, and not because of any return premium they provide. An analogous myopic investor would hold no risky assets.
4. Note also that the covariance term in Eq. (16.20) depends on changes in expectations concerning the entire course of future interest rates. It thus follows that the investor's portfolio allocations will be much more sensitive to persistent changes in the expected risk-free rate than to transitory ones.

Considering all the complicated formulae that have been developed, the conclusions thus far are relatively modest. An infinitely lived investor principally consumes out of his portfolio's income, and he wishes to maintain a stable consumption series. To the extent that risky equity returns can offset (hedge) variations in the risk-free rate, investors are provided justification for increasing the share of their wealth invested in the high-return risky asset. This leads us to wonder if any asset can serve as a truly risk-free one for the long-term investor.

16.3.4 The Nature of the Risk-Free Asset

Implicit in the above discussion is the question of what asset, if any, best serves as the risk-free one. From the long-term investor's point of view, it clearly cannot be a short-term money market instrument (e.g., a T-bill) because its well-documented rate variation makes uncertain the future reinvestment rates that the investor will receive.

We are reminded at this juncture, however, that it is not the return risk *per se*, but the derived consumption risk that is of concern to investors. Viewed from the consumption perspective, a natural candidate for the risk-free asset is an indexed consol bond that pays (the real monetary equivalent of) one unit of consumption every period. [Campbell et al. \(1997\)](#) show that the (log) return on such a consol is given by

$$r_{c,t+1} = r_{f,t+1} + k_4 - (E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_5)^j \tilde{r}_{f,t+1+j} \quad (16.21)$$

where k_4 is a constant measuring the (constant) risk premium on the consol, and k_5 is another positive constant less than one. Suppose as well that we have an infinitely risk-averse investor ($\gamma = \infty$) so that [Eq. \(16.20\)](#) reduces to

$$a = \frac{1}{\sigma_t^2} \text{cov}_t \left(\tilde{r}_{t+1}, -(E_{t+1} - E_t) \sum_{j=1}^{\infty} (k_2)^j \tilde{r}_{f,t+1+j} \right) \quad (16.22)$$

and that the single risky asset is the consol bond ($r_{t+1} = r_{c,t+1}$). In this case (substituting [Eq. \(16.21\)](#) into [Eq. \(16.22\)](#) and observing that constants do not matter for the computing of covariances), $a \equiv 1$: the highly risk-averse investor will eschew short-term risk-free assets and invest entirely in indexed bonds. This alone will provide him with a risk-free consumption stream, although the value of the asset may change from period to period.

16.3.5 The Role of Bonds in Investor Portfolios

Now that we allow the risk-free rate to vary, let us return to the three life cycle portfolio recommendations mentioned in Section 16.2. Of course, the model—with an infinitely lived investor—is, by construction, not the appropriate one for the life cycle issues of recommendation 2, and, being without labor income, nothing can be said regarding recommendation 3 either. This leaves the first recommendation, which really concerns the portfolio of choice for long-term investors. The single message of this chapter subsection must be that conservative long-term investors should invest the bulk of their wealth in long-term index bonds. If such bonds are not available, then in an environment of low inflation risk, long-term government securities are a reasonable, second best substitute. For persons entering retirement—and likely to be very concerned about significant consumption risk—long-term real bonds should be the investment vehicle of choice.

This is actually a very different recommendation from the static one-period portfolio analysis that would argue for a large fraction of a conservative investor's wealth being assigned to risk-free assets (T-bills). Yet we know that short rate uncertainty, which the long-term investor would experience every time she rolled over her short-term instruments, makes such an investment strategy inadvisable for the long term.

16.4 The Long-Run Behavior of Stock Returns

Should the proportion of an investor's wealth invested in stocks differ systematically for long-term versus short-term investors? In either case, most of the attractiveness of stocks (by stocks we will continue to mean a well-diversified stock portfolio) to investors lies in their high excess returns (recall the equity premium puzzle of Chapter 10). But what about long- versus short-term equity risk, i.e., how does the *ex ante* return variance of an equity portfolio held for many periods compare with its variance in the short run?

The *ex post* historical return experience of equities versus other investments turns out to be quite unexpected in this regard.

From [Table 16.1](#), it is readily apparent that, historically, more than 20-year time horizons, stocks have never yielded investors a negative real annualized return, while for all other

Table 16.1: Minimum and maximum actual annualized real holding period returns for the period, 1802–1997; US securities markets and a variety of investment options^a

	Maximum Observed Return		Minimum Observed Return
Stocks	66.6%	One-year holding period	−38.6%
Bonds	35.1%		−21.9%
T-bills	23.7%		−15.6%
Stocks	41.0%	Two-year holding period	−31.6%
Bonds	24.7%		−15.9%
T-bills	21.6%		−15.1%
Stocks	26.7%	Five-year holding period	−11.0%
Bonds	17.7%		−10.1%
T-bills	14.9%		−8.2%
Stocks	16.9%	Ten-year holding period	−4.1% ^b
Bonds	12.4%		−5.4%
T-bills	11.6%		−5.1%
Stocks	12.6%	Twenty-year holding period	1.0%
Bonds	8.8%		−3.1%
T-bills	8.3%		−3.0%
Stocks	10.6%	Thirty-year holding period	2.6%
Bonds	7.4%		−2.0%
T-bills	7.6%		−1.8%

^aSource: [Siegel \(1998\)](#), Figure 2–1.

^bNote that beginning with a 10-year horizon, the minimum observed stock return exceeded the corresponding minimum bill and bond returns.

This table replicates the information in Fig. 7.6.

investment types this has been the case for some sample period. Are stocks in fact less risky than bonds for an appropriately “long run”?

In this section, we propose to explore this issue via an analysis of the following questions:

1. What are the intertemporal equity return patterns in order that the outcomes portrayed in [Table 16.1](#) are pervasive and not just represent the realizations of extremely low-probability events?
2. Given a resolution of (1), what are the implications for the portfolio composition of long-term versus short-term investors?
3. How does a resolution of questions (1) and (2) modify the myopic response to the long-run investment advice of [Section 16.2](#)?

It is again impossible to answer these questions in full generality. Following [Campbell and Viceira \(1999\)](#), we elect to examine investor portfolios composed of one risk-free and one risky asset (a diversified portfolio). Otherwise, the context is as follows:

- i. The investor is infinitely lived with Epstein–Zin preferences so that [Eq. \(16.4\)](#) remains as the investor’s intertemporal optimality condition; furthermore, the investor has no labor income.
- ii. The log real risk-free rate is constant from period to period. Under this assumption, all risk-free assets—long or short term—pay the same annualized return. The issues of [Section 16.3](#) thus cannot be addressed.
- iii. The equity return generating process builds on the following observations.

First, note that the cumulative log return over T periods under the benchmark i.i.d. assumption is given by

$$r_{t+1} + r_{t+2} + \dots + r_{t+T}$$

so that

$$\text{var}(\tilde{r}_{t+1} + \tilde{r}_{t+2} + \dots + \tilde{r}_{t+T}) = T \text{ var}(\tilde{r}_{t+1}) > T \text{ var}(\tilde{r}_{f,t+1})$$

For US data, $\text{var}(r_{t+1}) \approx (0.167)$ (taking the risky asset as the S&P 500 market index) and $\text{var}(\tilde{r}_{f,t+1}) = (0.057)^2$ (measuring the risk-free rate as the 1-year T-bill return). With a $T = 20$ -year time horizon, the observed range of annualized relative returns given by [Table 16.1](#) is thus extremely unlikely to have arisen from an i.i.d. process. What could be going on?

For an i.i.d. process, the large relative 20-year variance arises from the possibility of long sequences, respectively, of high and low returns. But if cumulative stock returns are to be less variable than bond returns at long horizons, some aspect of the return generating process must be discouraging these possibilities. That aspect is referred to as “mean reversion”: the tendency of high returns today to be followed by low returns tomorrow on

an expected basis and vice versa. It is one aspect of the “predictability” of stock returns and is well documented beyond the evidence in Table 16.1.⁵

Campbell and Viceira (1999) statistically model the mean reversion in stock returns in a particular way that facilitates the solution to the associated portfolio allocation problem. In particular, they assume the time variation in log return on the risky asset is captured by

$$r_{t+1} - E_t \tilde{r}_{t+1} = \tilde{u}_{t+1}, \tilde{u}_{t+1} \sim N(0, \sigma_u^2) \quad (16.23)$$

where u_{t+1} captures the unexpected risky return component or “innovation.” In addition, the expected premium on this risky asset is modeled as evolving according to

$$E_t \tilde{r}_{t+1} - r_f + \frac{\sigma_u^2}{2} = \tilde{x}_t \quad (16.24)$$

where x_t itself is a random variable following an AR(1) process with mean \bar{x} , persistence parameter ϕ , and random innovation $\tilde{\eta}_{t+1} \sim N(0, \sigma_\eta^2)$:⁶

$$\tilde{x}_{t+1} = \bar{x} + \phi(x_t - \bar{x}) + \tilde{\eta}_{t+1} \quad (16.25)$$

⁵ We introduced the notion of predictability back in footnote 14 of Chapter 10. Here we review the notion in more detail. Perhaps the most frequently cited predictive variable is $\log(D/P_t) = d_t - p_t$, the log of the dividend/price ratio at long horizons. In particular, regressions of the form

$$\tilde{r}_{t+k} \equiv \tilde{r}_{t+1} + \cdots + \tilde{r}_{t+k} = \beta_k(d_t - p_t) + \tilde{\varepsilon}_{t+k}$$

obtain an R^2 of an order of magnitude of 0.3. In the above expression r_{t+j} denotes the log return on the value weighted index portfolio comprising all NYSE, AMEX, and NASDAQ stocks in month $t+j$, d_t is the log of the sum of all dividends paid on the index over the entire year preceding period t , and P_t denotes the period t value of the index portfolio. See Campbell et al. (1997) for a detailed discussion. More recently, Santos and Veronesi (2006) have studied regressions whereby long horizon excess returns (above the risk-free rate) are predicted by lagged values of the (US data) aggregate labor income/consumption ratio:

$$r_{t+k} = \alpha_1 + \beta_k s_t^W + \varepsilon_{t+k}$$

where $s_t^W = w_t/c_t$; w_t is measured as period t total compensation to employees and c_t denotes consumption of nondurables plus services (quarterly data). For the period 1948–2001, for example, they obtain an adjusted R^2 of 0.42 for $k = 16$ quarters. Returns are computed in a manner identical to Campbell et al. (1997) just mentioned. The basic logic is as follows: when the labor income/consumption ratio is high, investors are less exposed to stock market fluctuations (equity income represents a small fraction of total consumption) and hence demand a lower premium. Stock prices are thus high. Since the s_t^W ratio is stationary (and highly persistent in the data), it will eventually return to its mean value, suggesting a lower future tolerance for risk, a higher risk premium, lower equity prices, and low future returns. Their statistical analysis concludes that the labor income/consumption ratio does indeed move in a direction opposite to long horizon returns. Campbell and Cochrane (1999) are also able to replicate many of the predictability results of the empirical literature using a slow-moving external habit formation model. When it comes down to it predictability would appear to be a direct consequence of mean reversion.

⁶ We are reminded that the presence of the $\sigma_u^2/2$ term in Eq. (16.24) follows from the fact that for any lognormal random variable \tilde{z} , $\log E_t \tilde{z}_{t+1} = E_t \log \tilde{z}_{t+1} + \frac{1}{2} \text{var}_t \log \tilde{z}_{t+1}$.

The x_t random variable thus moves slowly (depending on ϕ), with a tendency to return to its mean value. Lastly, mean reversion is captured by assuming $\text{cov}(\eta_{t+1}, u_{t+1}) = \sigma_{\eta u} < 0$, which translates, as per below, into a statement about risky return autocorrelations:

$$\begin{aligned} 0 > \sigma_{\eta u} &= \text{cov}(\tilde{u}_{t+1}, \tilde{\eta}_{t+1}) = \text{cov}_t[(\tilde{r}_{t+1} - E_t \tilde{r}_{t+1}), (\tilde{x}_{t+1} - \bar{x} - \phi(x_t - \bar{x}))] \\ &= \text{cov}_t(\tilde{r}_{t+1}, \tilde{x}_{t+1}) \\ &= \text{cov}_t\left(\tilde{r}_{t+1}, E_t \tilde{r}_{t+2} - r_f + \frac{\sigma_u^2}{2}\right) \\ &= \text{cov}_t\left(\tilde{r}_{t+1}, \tilde{r}_{t+2} - \tilde{u}_{t+2} - r_f + \frac{\sigma_u^2}{2}\right) \\ &= \text{cov}_t(\tilde{r}_{t+1}, \tilde{r}_{t+2}) \end{aligned}$$

a high return today reduces expected returns next period. Thus,

$$\text{var}_t(\tilde{r}_{t+1} + \tilde{r}_{t+2}) = 2 \text{ var}_t(\tilde{r}_{t+1}) + 2 \text{ cov}_t(\tilde{r}_{t+1}, \tilde{r}_{t+2}) < 2 \text{ var}_t(r_{t+1})$$

in contrast to the independence case. More generally, for all horizons k ,

$$\frac{\text{var}_t(\tilde{r}_{t+1} + \tilde{r}_{t+2} + \dots + \tilde{r}_{t+k})}{k \text{ var}_t(\tilde{r}_{t+1})} < 1$$

stocks will appear to be relatively safer to long-term investors.⁷ This concludes assumption (iii) and its interpretation. We next explore what this environment implies for optimal investment proportions.

16.4.1 Solving for Optimal Portfolio Proportions in a Mean Reversion Environment

Campbell and Viceira (1999) manipulate the log-linearized version of the optimality condition for Epstein–Zin utility (16.4) in conjunction with Eqs. (16.23)–(16.25) to obtain

$$a_t = a_0 + a_1 x_t \tag{16.26}$$

$$c_t - y_t = b_0 + b_1 x_t + b_2 x_t^2 \tag{16.27}$$

where a_t is the (time-varying) wealth proportion invested in the risky asset and Eq. (16.27) describes the behavior of the (log) consumption–wealth ratio. The terms a_0, a_1, b_0, b_1 ,

⁷ In the finance literature, the preceding inequality is often cited as the definition of mean reversion (for an example, see Mukherji (2011)).

b_2 , are constants; of special interest are those defining the time-varying risky asset proportion:

$$a_0 = \left(1 - \frac{1}{\gamma}\right) \left[\left(\frac{b_1}{1-\rho}\right) + 2\bar{x}(1-\phi)\left(\frac{b_2}{1-\rho}\right) \right] \left(\frac{-\sigma_{\eta u}}{\sigma_u^2}\right) \text{ and} \quad (16.28)$$

$$a_1 = \frac{1}{\gamma\sigma_u^2} + \left(1 - \frac{1}{\gamma}\right) \left[2\phi \frac{b_2}{(1-\rho)} \right] \left(\frac{-\sigma_{\eta u}}{\sigma_u^2}\right) \quad (16.29)$$

Note that in the case of a myopic investor ($\gamma = 1$) who holds stocks only because of the (conditional) excess returns they offer, $a_0 = 0$ and $a_1 = 1/\sigma_u^2$. This suggests that all the other terms present in the expressions defining a_0 and a_1 must be present to capture some aspect of a nonmyopic investor's *intertemporal hedging demand*.

What is meant by this latter expression in this particular context? With $\bar{x} > 0$, investors are typically long in stocks (the risky asset) in order to capture the excess returns they provide on average. Suppose in some period $t+1$, stock returns are high, meaning that stock prices rose a lot from t to $t+1$ (u_{t+1} is large). To keep the discussion less hypothetical, let's identify this event with the big run-up in stock prices in the late 1990s. Under the $\sigma_{\eta u} < 0$ assumption, expected future returns are likely to decline and perhaps even become negative (η_{t+1} is small, possibly negative so that x_{t+1} is small and thus, via Eq. (16.24), so is $E\tilde{r}_{t+1}$). Roughly speaking, this means stock prices are likely to decline—as they did in the 2000–2004 period.⁸ In anticipation of future price declines, long-term investors would rationally wish to assemble a short position in the risky portfolio, since this is the only way to enhance their wealth in the face of falling prices (r_f is constant by assumption). Most obviously, this is a short position in the risky portfolio itself, since negative returns must be associated with falling prices.

These thoughts are fully captured by Eqs. (16.26) and (16.27). Campbell and Viceira (2002) argue that the empirically relevant case is the one for which $\bar{x} > 0$, $b_1/(1-\rho) > 0$, $b_2/(1-\rho) > 0$, and $\sigma_{\eta u} < 0$. Under these circumstances, $a_0 > 0$, and $a_1 > 0$, for a sufficiently risk-averse investor ($\gamma > 1$). If u_{t+1} is large, then η_{t+1} is likely to be small—let's assume negative—and “large” in absolute value if $|\sigma_{\eta u}|$ is itself large. Via portfolio allocation Eq. (16.26), the optimal $a_t < 0$ —a short position in the risky asset.

This distinguishing feature of long-term risk-averse investors is made more striking if we observe that with $\sigma_{\eta u} < 0$, such an investor will maintain a position in the risky asset if average excess returns, $\bar{x} = 0$: even in this case $a_0 > 0$ (provided $\gamma > 1$). Thus, if $x_t = 0$ (no excess returns to the risky asset), the proportion of the investor's wealth in stocks is still positive. In a one-period CAPM investment universe, a mean–variance myopic investor

⁸ We have observed a similar pattern thus far in the 21st century. Stocks rose dramatically in value from 2004 until October 2008, and then declined precipitously with the onset of the financial crisis. In 2012 they started again rising in value and, as of this writing, are at a new peak with the S&P 500 index close to a level of 2000.

would invest nothing in stocks under these circumstances. Neither would the myopic expected utility maximizer of Theorem 4.1.

All this is to observe that a risk-averse rational long-term investor will use whatever means are open to him, including shorting stocks, when he (rationally) expects excess future stock returns to be sufficiently negative to warrant it. A major caveat to this line of reasoning, however, is that it cannot illustrate an equilibrium phenomenon: if all investors are rational and equally well informed about the process generating equity returns (16.23)–(16.25), then all will want simultaneously to go long or short. The latter, in particular, is not feasible from an equilibrium perspective.

16.4.2 Strategic Asset Allocation

The expression *strategic asset allocation* is suggestive not only of long-term investing (for which intertemporal hedging is a concern), but also of portfolio weights assigned to broad classes of assets (e.g., “stocks,” “long-term bonds”), each well diversified from the perspective of its own kind. This is exactly the setting of this chapter.

Can the considerations of this section, in particular, be conveniently contrasted with those of the preceding chapters? This is captured in [Figure 16.1](#) under the maintained

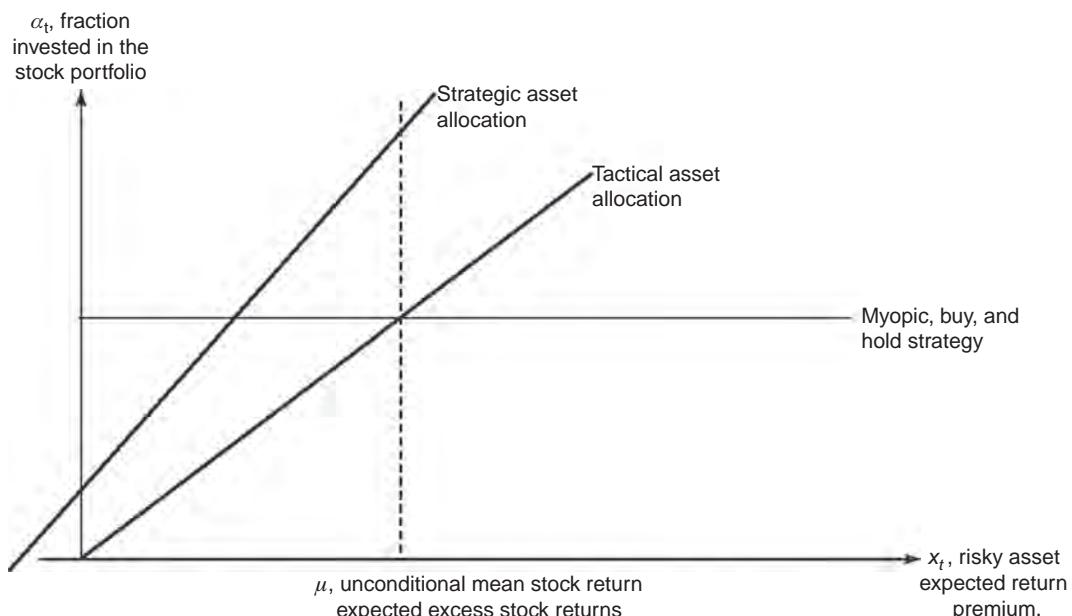


Figure 16.1

Alternative portfolio allocations.⁽ⁱ⁾

(i) Based on maintained assumptions (i)–(iii), and $\gamma > 1$.

assumptions of this subsection (itself a replica of Figure 4.1 in [Campbell and Viceira, 2002](#)). The myopic buy-and-hold strategy assumes a constant excess stock return equal to the true unconditional mean ($E_t \tilde{r}_{t+1} - r_f + \sigma_u^2/2 \equiv \bar{x}$) with the investor solving a portfolio allocation problem as per Theorem 4.1. The line marked “tactical asset allocation” describes the portfolio allocations for an investor who behaves as a one-period investor, conditional on his observation of x_t . Such an investor, by definition, will not take account of long-term hedging opportunities. Consistent with the CAPM recommendation, such an investor will elect $a_t = 0$ when $x_t = 0$ (no premium on the risky asset) but if the future looks good—even for just one period—he will increase his wealth proportions in the risky assets. Again, by construction, if $x_t = \bar{x}$, such an investor will adopt portfolio proportions consistent with the perpetually myopic investor.

Long-term “strategic” investors, with rational expectations vis-à-vis the return generating process (i.e., they know and fully take account of [Eqs. \(16.23\)–\(16.25\)](#)) will always elect to hold a greater proportion of their wealth in the risky portfolio than will be the case for the “tactical” asset allocator. In itself this is not entirely surprising, for only he is able to illustrate the “hedging demand.” But this demand is present in a very strong way; in particular, even if excess returns are zero, the strategic investor holds a positive wealth fraction in risky assets ($a_0 > 0$). Note also that the slope of the strategic asset allocation line exceeds that of the tactical asset allocation line. In the context of [Eqs. \(16.23\)–\(16.25\)](#), this is a reflection of the fact that $\phi = 0$ for the tactical asset allocator.

16.4.3 The Role of Stocks in Investor Portfolios

Suppose stocks are less risky in the long run because of mean reversion in stock returns. But does this necessarily imply a 100% stock allocation in perpetuity for long-term investors? Under the assumptions of [Campbell and Viceira \(2002\)](#), this is clearly not the case: long-term investors should be prepared to take advantage of mean reversion by *timing the market* in a manner illustrated in [Figure 16.1](#). But this in turn presumes the ability of investors to short stocks when their realized returns have recently been very high. Especially for small investors, shorting securities may entail prohibitive transactions costs. Even more significantly, this cannot represent an equilibrium outcome for all investors.

16.5 Background Risk: The Implications of Labor Income for Portfolio Choice

Background risks refer to uncertainties in the components of an investor’s income not directly related to his tradable financial wealth and, in particular, his stock–bond portfolio allocation. Labor income risk is a principal component of background risk; variations in

proprietary income (income from privately owned businesses) and in the value of owner-occupied real estate are the others.

In this section, we explore the significance of labor income risk for portfolio choice. It is a large topic and one that must be dealt with using models of varying complexity. The basic insight we seek to develop is as follows: an investor's labor income stream constitutes an element of his wealth portfolio. The desirability of the risky asset in the investor's portfolio will therefore depend not only upon its excess return (above the risk-free rate) relative to its variance (risk), but also the extent to which it can be used to hedge variations in the investor's labor income. Measuring how the proportion of an investor's financial wealth invested in the risky asset depends on its hedging attributes in the above sense is the principal focus of this section. Fortunately, it is possible to capture the basic insights in a very simple framework. As discussed in [Campbell and Viceira \(2002\)](#), that framework makes a number of assumptions:

- i. The investor has a one-period horizon, investing his wealth to enhance his consumption tomorrow (as such, the focus is on the portfolio allocation decision exclusively; there is no $t = 0$ simultaneous consumption—savings decision).
- ii. The investor receives labor income L_{t+1} , tomorrow, which for analytical simplicity is assumed to be lognormally distributed: $\log \tilde{L}_{t+1} \equiv \tilde{\ell}_{t+1} \sim N(l, \sigma_\ell^2)$.
- iii. There is one risk-free and one risky asset (a presumed-to-be well diversified portfolio). Following our customary notation, $(r_f = \log(R_f))$ and $\tilde{r}_{t+1} = \log(\tilde{R}_{t+1})$. Furthermore, $\tilde{r}_{t+1} - E_t \tilde{r}_{t+1} = \tilde{u}_{t+1}$ where $\tilde{u}_{t+1} \sim N(0, \sigma_u^2)$. The possibility is admitted that the risky asset return is correlated with labor income in the sense that $\text{cov}(\tilde{\ell}_{t+1}, \tilde{r}_{t+1}) = \sigma_{\ell u} \neq 0$.
- iv. The investor's period $t + 1$ utility function is of the CRRA-power utility type, with coefficient of relative risk aversion γ .

Since there is no labor—leisure choice, this model is implicitly one of fixed labor supply in conjunction with a random wage.

Accordingly, the investor solves the following problem:

$$\max_{\alpha_t} E_t \left[\delta \frac{\tilde{C}_{t+1}^{1-\gamma}}{1-\gamma} \right] \quad (16.30)$$

$$\text{s.t. } C_{t+1} = Y_t R_{P,t+1} + L_{t+1}$$

where

$$R_{P,t+1} = \alpha_t (R_{t+1} - R_f) + R_f \quad (16.31)$$

α_t represents the fraction of the investor's wealth assigned to the risky portfolio, and P denotes his overall wealth portfolio. As in nearly all of our problems to date, insights can be neatly obtained only if approximations are employed that take advantage of the

lognormal setup. In particular, we first need to modify the portfolio return expression (16.31).⁹ Since

$$\begin{aligned} R_{P,t+1} &= a_t R_{t+1} + (1 - a_t) R_f, \\ \frac{R_{P,t+1}}{R_f} &= 1 + a_t \left(\frac{R_{t+1}}{R_f} - 1 \right) \end{aligned}$$

Taking the log on both sides of this equation yields

$$r_{P,t+1} - r_f = \log [1 + a_t(\exp(r_{t+1} - r_f) - 1)] \quad (16.32)$$

The RHS of this equation can be approximated using a second-order Taylor expansion around $r_{P,t+1} - r_f = 0$, where the function to be approximated is

$$g_t(r_{P,t+1} - r_f) = \log [1 + a_t(\exp(r_{t+1} - r_f) - 1)]$$

By Taylor's theorem

$$g_t(r_{P,t+1} - r_f) \approx g_t(0) + g'_t(0)(r_{t+1} - r_f) + \frac{1}{2} g''_t(0)(r_{t+1} - r_f)^2$$

Clearly, $g_t(0) \equiv 0$; straightforward calculations (simple calculus) yield $g'_t(0) = a_t$, and $g''_t(0) = a_t(1 - a_t)$. Substituting into the Taylor expansion the indicated coefficient values, for the RHS of Eq. (16.32) yields

$$r_{P,t+1} - r_f = a_t(r_{t+1} - r_f) + \frac{1}{2} a_t(1 - a_t) \sigma_t^2$$

where $(r_{t+1} - r_f)^2$ is replaced by its conditional expectation. By the special form of the risky return generating process, $\sigma_t^2 = \sigma_u^2$, which yields

$$r_{P,t+1} = a_t(r_{t+1} - r_f) + r_f + \frac{1}{2} a_t(1 - a_t) \sigma_u^2 \quad (16.33)$$

We next modify the budget constraint to problem (16.30).

$$\frac{C_{t+1}}{L_{t+1}} = \frac{Y_t}{L_{t+1}} (R_{P,t+1}) + 1$$

or taking the log on both sides of the equation,

$$\begin{aligned} c_{t+1} - \ell_{t+1} &= \log[\exp(y_t + r_{P,t+1} - \ell_{t+1}) + 1] \\ &\approx k + \xi(y_t + r_{P,t+1} - \ell_{t+1}), \end{aligned} \quad (16.34)$$

⁹ The derivation to follow is presented in greater detail in Campbell and Viceira (2001b).

where k and ξ , $0 < \xi < 1$, are constants of approximation. Adding log labor income— ℓ_{t+1} —to both sides of the equation yields

$$c_{t+1} = k + \xi(y_t + r_{P,t+1}) + (1 - \xi)\ell_{t+1}, 1 > \xi > 0 \quad (16.35)$$

In other words, (log-) end of period consumption is a constant plus a weighted average of (log-) end-of-period financial wealth and (log-) labor income, with the weights ξ , $1 - \xi$ serving to describe the respective elasticities of consumption with respect to these individual wealth components.

So far, nothing has been said regarding optimality. [Equation \(16.30\)](#) is a one-period optimization problem. The first-order necessary and sufficient condition for this problem with respect to a_t , the proportion of financial wealth invested in the risky portfolio, is given by

$$E_t[\delta(\tilde{C}_{t+1})^{-\gamma}(\tilde{R}_{t+1})] = E_t[\delta(\tilde{C}_{t+1})^{-\gamma}(R_f)] \quad (16.36)$$

In log-linear form, [Eq. \(16.36\)](#) has the familiar form:

$$E_t(\tilde{r}_{t+1} - r_f) + \frac{1}{2}\sigma_t^2 = \gamma \text{cov}_t(\tilde{r}_{t+1}, \tilde{c}_{t+1})$$

Substituting the expression in [Eq. \(16.35\)](#) for c_{t+1} yields

$$E_t(\tilde{r}_{t+1} - r_f) + \frac{1}{2}\sigma_t^2 = \gamma \text{cov}_t(\tilde{r}_{t+1}, k + \xi(y_t + \tilde{r}_{P,t+1}) + (1 - \xi)\tilde{\ell}_{t+1})$$

After substituting [Eq. \(16.33\)](#) for $r_{P,t+1}$, we are left with

$$E_t(\tilde{r}_{t+1} - r_f) + \frac{1}{2}\sigma_t^2 = \gamma [\xi a_t \sigma_t^2 + (1 - \xi) \text{cov}_t(\tilde{\ell}_{t+1}, \tilde{r}_{P,t+1})]$$

from which we can solve directly for a_t .

Recall that our objective was to explore how the hedging (with respect to labor income) features of risky securities influence the proportion of financial wealth invested in the risky asset. Accordingly, it is convenient first to simplify the expression via the following identifications: let

- i. $\mu = E_t(\tilde{r}_{t+1} - r_f);$
- ii. $\sigma_t^2 = \sigma_u^2$, since $r_{t+1} - E_t \tilde{r}_{t+1} = u_{t+1};$
- iii. $\text{cov}(\tilde{\ell}_{t+1}, \tilde{r}_{t+1}) = \text{cov}(\tilde{\ell}_{t+1}, \tilde{r}_{t+1} - E_t \tilde{r}_{t+1}) = \text{cov}(\tilde{\ell}_{t+1}, \tilde{u}_{t+1}) = \sigma_{\ell u}.$

With these substitutions, the above expression reduces to

$$\mu + \frac{1}{2}\sigma_u^2 = \gamma[\xi a_t \sigma_u^2 + (1 - \xi)\sigma_{\ell u}].$$

Straightforwardly solving for a_t yields

$$a_t = \frac{1}{\xi} \left(\frac{\mu + \frac{\sigma_u^2}{2}}{\gamma \sigma_u^2} \right) + \left(1 - \frac{1}{\xi} \right) \frac{\sigma_{\ell u}}{\sigma_u^2}, \quad (16.37)$$

an expression with an attractive interpretation. The first term on the RHS of Eq. (16.37) represents the fraction in the risky asset if labor income is uncorrelated with the risky asset return ($\sigma_{\ell u} = 0$). It is positively related to the adjusted return premium ($\mu + \sigma_u^2/2$) and inversely related to the investor's risk-aversion coefficient γ . The second term represents the hedging component: if $\sigma_{\ell u} < 0$, then since $\xi < 1$, demand for the risky asset is enhanced since it can be employed to diversify away some of the investor's labor income risk. Or, to express the same idea from a slightly different perspective, if the investor's labor income has a "suitable" statistical pattern vis-à-vis the stock market, he can reasonably take on greater financial risk.

It is perhaps even more striking to explore further the case where $\sigma_{\ell u} = 0$: since $\xi < 1$, even in this case, the optimal fraction invested in the risky portfolio is

$$a_t = \frac{1}{\xi} \left(\frac{\mu + \frac{\sigma_u^2}{2}}{\gamma \sigma_u^2} \right) > \left(\frac{\mu + \frac{\sigma_u^2}{2}}{\gamma \sigma_u^2} \right),$$

where the rightmost ratio represents the fraction the investor places in the risky portfolio were there to be no labor income at all. If $\sigma_{\ell u} = 0$, then at least one of the following is true: $\text{corr}(u, \ell) = 0$ or $\sigma_\ell = 0$, and each leads to a slightly different interpretation of the optimal a_t . First, if $\sigma_\ell > 0$ (there is variation in labor income), then the independence of labor and equity income allows for a good deal of overall risk reduction, thereby implying a higher optimal risky asset portfolio weight. If $\sigma_\ell = 0$ —labor income is constant—then human capital wealth is a nontradable risk-free asset in the investor's overall wealth portfolio. Ceteris paribus, this also allows the investor to rebalance his portfolio in favor of a greater fraction held in risky assets. If, alternatively $\sigma_{\ell u} > 0$ —a situation in which the investor's income is closely tied to the behavior of the stock market—then the investor should correspondingly reduce his position in risky equities. In fact, if the investor's coefficient of relative risk aversion is sufficiently high and $\sigma_{\ell u}$ large and positive (say, if the investor's portfolio contained a large position in his own firm's stock) then $a_t < 0$, i.e., the investor should hold a short position in the overall equity market.

These remarks formalize, though in a very simple context, the idea that an investor's wage income stream represents an asset and that its statistical covariance with the equity portion of his portfolio should matter for his overall asset allocation. To the extent that variations in stock returns are offset by variations in the investor's wage income, stocks are effectively

less risky (so also is wage income less risky) and he can comfortably hold more of them. The reader may be suspicious, however, of the one-period setting. We remedy this next.

Viceira (2001) extends these observations to a multiperiod infinite horizon setting by adopting a number of special features. There is a representative investor–worker who saves for retirement and who must take account in his portfolio allocation decisions of the expected length of his retirement period. In any period, there is a probability π^r that the investor will retire; his probability of remaining employed and continuing to receive labor income is $\pi^e = 1 - \pi^r$, with constant probability period by period. With this structure of uncertainty, the expected number of periods until an investor's retirement period is $1/\pi^r$. Once retired (zero labor income), the period constant probability of death is π^d ; in a like manner the expected length of his retirement is $1/\pi^d$. Viceira (2001) also assumes that labor income is growing in the manner of

$$\Delta \ell_{t+1} = \log L_{t+1} - \log L_t = g + \tilde{\varepsilon}_{t+1} \quad (16.38)$$

where $g > 0$ and $\tilde{\varepsilon}_{t+1} \sim N(0, \sigma_\varepsilon^2)$. In expression (16.38), g represents the mean growth in labor income (for the United States this figure is approximately 2%) while $\tilde{\varepsilon}_t$ denotes random variations about the mean. The return on the risky asset is assumed to follow the same hypothetical process as in the prior example. In this case,

$$\sigma_{u\ell} = \text{cov}_t(\tilde{r}_{t+1}, \Delta \tilde{\ell}_{t+1}) = \text{cov}_t(\tilde{u}_{t+1}, \tilde{\varepsilon}_{t+1}) = \sigma_{ue}$$

With an identical asset structure as in the previous model, the investor's problem appears deceptively similar to Eq. (16.30):

$$\begin{aligned} & \max E_t \left(\sum_{i=0}^{\infty} \delta^i \frac{\tilde{C}_{t+i}^{1-\gamma}}{1-\gamma} \right) \\ & \text{s.t. } Y_{t+1} = (Y_t + L_t - C_t) R_{P,t+1} \end{aligned} \quad (16.39)$$

The notation in Eq. (16.39) is identical to that of the previous model. Depending on whether an agent is employed or retired, however, the first-order optimality condition will be different, reflecting the investor's differing probability structure looking forward. In Eqs. (16.40) and (16.41) to follow, the (not log) consumption is superscripted by e or r, depending on its enjoyment in the investor's period of employment or retirement, respectively. If the investor is retired, for any asset i (the portfolio P , or the risk-free asset):

$$1 = E_t \left[(1 - \pi^d) \delta \left(\frac{\tilde{C}_{t+1}^r}{C_t^r} \right)^{-\gamma} \tilde{R}_{i,t+1} \right] \quad (16.40)$$

The interpretation of Eq. (16.40) is more or less customary: the investor trades off the marginal utility lost in period t by investing one more consumption unit against

the expected utility gain in period $t + 1$ for having done so. The expectation is adjusted by the probability $(1 - \pi^d)$ that the investor is, in fact, still living in the next period. Analytically, its influence on the optimality condition is the same as a reduction in his subjective discount factor δ . If the investor is employed, but with positive probability of retirement and subsequent death, then each asset i satisfies:

$$1 = E_t \left\{ \left[\pi^e \delta \left(\frac{\tilde{C}_{t+1}^e}{C_t^e} \right)^{-\gamma} + (1 - \pi^e)(1 - \pi^d) \delta \left(\frac{\tilde{C}_{t+1}^r}{C_t^r} \right)^{-\gamma} \right] (\tilde{R}_{i,t+1}) \right\} \quad (16.41)$$

Equation (16.41)'s interpretation is analogous to Eq. (16.40) except that the investor must consider the likelihood of his two possible states next period: either he is employed (probability π^e) or retired and still living (probability $(1 - \pi^e)(1 - \pi^d)$). Whether employed or retired, these equations implicitly characterize the investor's optimal risk-free–risky portfolio proportions as those for which his expected utility gain to a marginal dollar invested in either one is the same.

Viceira (2001) log-linearizes these equations and their associated budget constraints to obtain the following expressions for log consumption and the optimal risky portfolio weight in both retirement and employment; for a retired investor:

$$c_t^r = b_0^r + b_1^r y_t \text{ and} \quad (16.42)$$

$$a^r = \frac{\mu + \frac{\sigma_u^2}{2}}{\gamma b_1^r \sigma_u^2} \quad (16.43)$$

where $b_1^r = 1$ and b_0^r is a complicated (in terms of the model's parameters) constant of no immediate concern; for an employed investor, the corresponding expressions are

$$c_t^e = b_0^e + b_1^e y_t + (1 - b_1^e) \ell_t \quad (16.44)$$

$$a^e = \frac{\mu + \frac{\sigma_u^2}{2}}{\gamma \bar{b}_1 \sigma_u^2} - \left(\frac{\pi^e (1 - b_1^e)}{\bar{b}_1} \right) \frac{\sigma_{eu}}{\sigma_u^2} \quad (16.45)$$

with $0 < b_1^e < 1$, $\bar{b}_1 = \pi^e b_1^e + (1 - \pi^e) b_1^r$, and b_0^e , again, a complex constant whose precise form is not relevant for the discussion.

These formulas allow a number of observations:

1. Since $b_1^r > b_1^e$, (log) consumption is more sensitive to (log) wealth changes for the retired (Eq. (16.42)) as compared with the employed (Eq. (16.44)). This is not surprising as the employed can hedge this risk via his labor income. The retired cannot.

2. As in the prior model with labor income, there are two terms that together comprise the optimal risky asset proportions for the employed, a^e . The first $(\mu + (\sigma_u^2/2))/\gamma \bar{b}_1 \sigma_u^2$ reflects the proportion when labor income is independent of risky returns ($\sigma_{eu} = 0$). The second, $-(\pi^e(1 - b_1^e)/\bar{b}_1)\sigma_{eu}/\sigma_u^2$ accounts for the hedging component. If $\sigma_{eu} < 0$, then the hedge that labor income provides to the risky component of the investor's portfolio is more powerful: the optimal a^e is thus higher, while the opposite is true if $\sigma_{eu} > 0$. With a longer expected working life (greater π^e), the optimal hedging component is also higher: the present value of the gains to diversification provided by labor income variation correspondingly increases. Note also that the hedging feature is very strong in the sense that even if the mean equity premium, $\mu = 0$, the investor will retain a positive allocation in risky assets purely for their diversification effect vis-à-vis labor income.
3. Let us next separate the hedging effect by assuming $\sigma_\varepsilon = 0$ (and thus $\sigma_{eu} = 0$). Since $b_1^e < b_1^r$, $\bar{b}_1 < b_1^r$,

$$a^e = \frac{\mu + \frac{\sigma_u^2}{2}}{\gamma \bar{b}_1 \sigma_u^2} > \frac{\mu + \frac{\sigma_u^2}{2}}{\gamma \bar{b}_1^r \sigma_u^2} = a^r$$

for any level of risk aversion γ : even if labor income provides no hedging services, the employed investor will hold a greater fraction of his wealth in the risky portfolio than will the retired investor. This is the labor income wealth effect. *Ceteris paribus*, a riskless labor income stream contributes a valuable riskless asset, and its presence allows the investor to tilt the financial component of his wealth in favor of a greater proportion in stocks. It also enhances his average consumption suggesting less aversion to risk. If $\sigma_{eu} = 0$ because $\rho_{eu} = 0$, $a^e > a^r$ can be justified on the basis of diversification alone. This latter comment is strengthened (weakened) when greater (lesser) diversification is possible: $a_{eu} < 0$ ($\sigma_{eu} > 0$).

Before summarizing these thoughts, we return to a consideration of the initial three life cycle portfolio recommendations. Strictly speaking, Eq. (16.39) is not a life cycle model. Life cycle considerations can be dealt with to a good approximation, however, if we progressively resolve Eq. (16.39) for a variety of choices of π^e , π^d . If π^d is increased, the expected length of the period of retirement falls. If π^r is increased (π^e decreased), it is as if the investor's expected time to retirement was declining as he "aged." Campbell and Viceira (2002) present the results of such an exercise, which we report in Table 16.2 for a selection of realistic risk-aversion parameters.

Here we find more general theoretical support for at least the first and third of our original empirical observations. As investors approach retirement, the fraction of their wealth invested in the risky portfolio does decline strongly. Note also that it is optimal for mildly risk-averse young investors ($\gamma = 2$) to short dramatically the risk-free asset to buy more of the risky one in order to "capture" the return supplement inherent in the equity premium. In

Table 16.2: Optimal percentage allocation to stocks^{a,b}

		Employed			Retired
		Expected time to retirement (years)			
		35	25	10	5
Panel A: $\text{corr}(\tilde{r}_{P,t+1}, \Delta\tilde{l}_{t+1}) = 0$					
$\gamma = 2$	184	156	114	97	80
$\gamma = 5$	62	55	42	37	32
Panel B: $\text{corr}(\tilde{r}_{P,t+1}, \Delta\tilde{l}_{t+1}) = 0.35$					
$\gamma = 2$	155	136	116	93	80
$\gamma = 5$	42	39	35	33	32

^a $\tilde{r}_f = 0.02$, $E\tilde{r}_{P,t+1} - \tilde{r}_{f,t+1} = \mu = 0.04$, $\sigma_u = 0.157$, $g = 0.03$, $\sigma_\varepsilon = 0.10$.

^bTable 16.2 is a subset of Table 6.1 in [Campbell and Viceira \(2002\)](#).

actual practice, however, such a leverage level is unlikely to be feasible for young investors without a high level of collateral assets. However, the “pull of the premium” is so strong that even retired persons with $\gamma = 5$ (the upper bound for which there is empirical support) will retain roughly one-third of their wealth in the risky equity index. In this sense, the latter aspect of the first empirical assertion is not borne out, at least for this basic preference specification.

We conclude this section by summarizing the basic points.

1. Riskless labor income creates a tilt in investor portfolios toward risky equities. This is not surprising, for the labor income stream in this case contributes a risk-free asset in the investor’s portfolio. There are two effects going on. One is a wealth effect: *ceteris paribus*, an investor with a labor income stream is wealthier than an investor without one, and with CRRA utility some of that additional wealth will be assigned to equities. This is complemented by a pure portfolio effect: the risk-free asset alters overall portfolio proportions in a way that is manifest as an increased share of financial wealth in risky assets.
2. These same effects are strengthened by the hedging effect if $\sigma_{\varepsilon u} \leq 0$ (effectively, this means $\sigma_{rl} \leq 0$). Stocks and risky labor income covary in a way that each reduces the effective risk of the other. Only if $\sigma_{\varepsilon u}$ is large and positive will the presence of labor income risk reduce the fraction of financial wealth invested in risky assets.
3. Although not discussed explicitly, the ability of an investor to adjust her labor supply—and thus her labor income—only enhances these effects. In this case, the investor can elect not only to save more but also to work more if she experiences an unfavorably risky return realization. Her ability to hedge adverse risky return realizations is thus enriched, and stocks appear effectively less risky.

16.6 An Important Caveat

We again acknowledge the concerns of Chapter 7. The accuracy and usefulness of the notions developed in the preceding sections, especially as regards applications of the formulas for practical portfolio allocations, should not be overemphasized. Their usefulness depends in every case on the accuracy of the forecast means, variances, and covariances, which represent the inputs to them: garbage in; garbage out still applies! Unfortunately, these quantities—especially expected risky returns—have been notoriously difficult to forecast accurately, even 1 year in advance. Errors in these estimates can have substantial significance for risky portfolio proportions, as these are generally computed using a formula of the generic form

$$\mathbf{a}_t = \frac{1}{\gamma} \sum^{-1} (\mathbf{E}_t \mathbf{r}_{t+1} - \mathbf{r}_{f,t+1} \mathbf{1})$$

where boldface letters represent vectors and Σ^{-1} is a matrix of “large” numbers. Errors in $\mathbf{E}_t \mathbf{r}_{t+1}$, the return vector forecasts, are magnified accordingly in the portfolio proportion choice.

In a recent paper, [Garlappi et al. \(2009\)](#) evaluate a number of complex portfolio strategies against a simple equal-portfolio-weights-buy-and-hold strategy. Using the same data set as [Campbell and Viceira \(2002\)](#) use, the equal weighting strategy tends to dominate all the others, simply because, under this strategy, the forecast return errors (which tend to be large) do not affect the portfolio’s makeup.

16.7 Another Background Risk: Real Estate

In this final section, we explore the impact of real estate holdings on an investor’s optimal stock–bond allocations. As before, our analysis will be guided by two main principles: (1) all assets—including human capital wealth—should be explicitly considered as components of the investor’s overall wealth portfolio and (2) it is the correlation structure of cash flows from these various income sources that will be paramount for the stock–bond portfolio proportions. Residential real estate is important because it represents roughly half of the US aggregate wealth, and it is not typically included in empirical stand-ins for the US market portfolio M .

Residential real estate also has features that make it distinct from pure financial assets. In particular, it provides a stream of housing services that are inseparable from the house itself. Houses are indivisible assets: one may acquire a small house but not one-half of a house. Such indivisibilities effectively place minimum bounds on the amount of real estate that can be acquired. Furthermore, houses cannot be sold without paying a substantial transactions fee, variously estimated to be between 8% and 10% of the value of the unit being exchanged. As the purchase of a home is typically a leveraged transaction, most lenders require minimum “down payments” or equity investments by the purchaser in the

house. Lastly, investors may be forced to sell their houses for totally exogenous reasons, such as a job transfer to a new location.

[Cocco \(2005\)](#) studies the stock–bond portfolio allocation problem in the context of a model with the above features, which is otherwise very similar to the ones considered thus far in this chapter. Recall that our perspective is one of partial equilibrium where, in this section, we seek to understand how the ownership of real estate influences an investor’s other asset holdings, given assumed return processes on the various assets. In the following, we highlight certain aspects of [Cocco’s \(2005\)](#) modeling of the investor’s problem.

The investor’s objective function, in particular, is

$$\max_{\{S_t, B_t, D_t, FC_t\}} E \left\{ \sum_{t=0}^T \beta^t \frac{(\tilde{C}_t^\theta \tilde{H}_t^{1-\theta})^{1-\gamma}}{1-\gamma} + \beta^T \frac{(\tilde{Y}_{T+1})^{1-\gamma}}{1-\gamma} \right\}$$

where, as before, C_t is his period t (nondurable) consumption (not logged; in fact, no variables will be logged in the problem description), H_t denotes period t housing services (presumed proportional to housing stock with a proportionality constant of one), and Y_t the investor’s period t wealth. Under this formulation, nondurable consumption and housing services complement one another with the parameter θ describing the relative preference of one to the other.¹⁰ Investor risk sensitivity to variations in the nondurable consumption–housing joint consumption decision is captured by γ (the investor displays CRRA with respect to the composite consumption product). The rightmost term, $((Y_{T+1})^{1-\gamma})/1 - \gamma$, is to be interpreted as a bequest function: the representative investor receives utility from nonconsumed terminal wealth, which is presumed to be bequeathed to the next generation, with the same risk preference structure applying to this quantity as well. In order to capture the idea that houses are indivisible assets, [Cocco \(2005\)](#) imposes a minimum size constraint

$$H_t \geq H_{\min}$$

to capture the fact that transactions costs are involved in changing one’s stock of housing, the agent is assumed to receive only

$$(1 - \lambda)P_t H_{t-1}$$

if he sells his housing stock H_{t-1} , in period t for a price P_t . In his calibration, λ —the magnitude of the transaction cost—is fixed at 0.08, a level for which there is substantial empirical support in US data. Note the apparent motivation for a bequest motive: given the minimum housing stock constraint, an investor in the final period of his life would otherwise own a significant level of housing stock for which the disposition at his death would be ambiguous.¹¹

¹⁰ The idea is simply that an investor will “enjoy his dinner more if he eats it in a warm and spacious house.”

¹¹ An alternative device for dealing with this modeling feature would be to allow explicitly for reverse mortgages.

Let \tilde{R}_t , R_f , and R_D denote the gross random exogenous return on equity, (constant) risk-free rate, and the (constant) mortgage interest rate. If the investor elects not to alter his stock of housing in period t relative to $t - 1$, his budget constraint for that period is

$$S_t + B_t = \tilde{R}_t S_{t-1} + R_f B_t - R_D D_{t-1}^M + L_t - C_t - \chi_t^{FC} F - \Omega P_t H_{t-1} + D_t^M = Y_t \quad (16.46)$$

where the notation is suggestive: S_t and B_t denote his period t stock and bond holdings, respectively, D_t^M the level of period t mortgage debt, Ω is a parameter measuring the maintenance cost of home ownership, and F is a fixed cost of participating in the financial markets. The indicator function χ_t^{FC} assumes values $\chi_t^{FC} = 1$, if the investor alters his stock or bondholdings relative to period $t - 1$ and 0 otherwise. This device is meant to capture the cost of participating in the securities markets. In the event the investor wishes to alter his stock of housing in period t , his budget constraint is modified in the to-be-expected way (most of it is unchanged except for the addition of the costs of trading houses):

$$S_t + B_t = Y_t + (1 - \lambda)P_t H_{t-1} - P_t H_t \text{ and} \quad (16.47)$$

$$D_t \leq (1 - d)P_t H_t \quad (16.48)$$

The additional terms in Eq. (16.47) relative to Eq. (16.46) are simply the net proceeds from the sale of the “old” house, $(1 - \lambda)P_t H_{t-1}$, less the costs of the “new” one $P_t H_t$. Constraint (16.48) reflects the down payment equity requirement and the consequent limits to mortgage debt. (In his simulation Cocco, 2005, chooses $d = 0.15$.)

Cocco (2005) numerically solves the above problem given various assumptions on the return and house price processes that are calibrated to historical data. In particular, he allows for house prices and aggregate labor income shocks to be perfectly positively correlated and for labor income to have both random and determinate components.¹²

¹² In particular, Cocco (2005) assumes

$$\tilde{L}_t = \begin{cases} f(t) + \tilde{u}_t, & t \leq T \\ f(t) & t > T \end{cases}$$

where T is the retirement date and the deterministic component $f(t)$ is chosen to replicate the hump shape earnings pattern typically observed. The random component \tilde{u}_t has aggregate ($\tilde{\eta}_t$) and idiosyncratic components ($\tilde{\omega}_t$) where

$$\tilde{u}_t = \tilde{\eta}_t + \tilde{\omega}_t \text{ and} \quad (16.49)$$

$$\tilde{\eta}_t = \kappa_\eta \tilde{P}_t \quad (16.50)$$

where P_t is the log of the average house price. In addition, he assumes $R_f = 1.02$ is fixed for the period $[0, T]$, as is the mortgage rate $R_D = 1.04$. The return on equity follows

$$r_t = \log(\tilde{R}_t) = E \log \tilde{R} + \tilde{l}_t$$

with $\tilde{l}_t \sim N(0; \sigma_l^2)$, $\sigma_l = 0.1674$, and $E \log \tilde{R} = 0.10$.

Cocco (2005) uses his model to comment upon a number of outstanding financial puzzles, of which we will review three: (1) considering the magnitude of the equity premium and the mean reversion in equity returns, why do all investors not hold at least some of their wealth as a well-diversified equity portfolio? Simulations of the model reveal that the minimum housing level H_{\min} (which is calculated at US\$20,000US\$) in conjunction with the down payment requirement make it nonoptimal for lower labor income investors to pay the fixed costs of entering the equity markets. This is particularly the case for younger investors who remain liquidity constrained. (2) While the material in Section 16.5 suggests that the investors' portfolio share invested in stocks should decrease in later life (as the value of labor income wealth declines), the empirical literature finds that for most investors the portfolio share invested in stocks is increasing over their life cycle. Cocco's (2005) model implies that the share in equity investments increases over the life cycle. As noted above, early in life, housing investments keep investors' liquid assets low, and they choose not to participate in the markets. More surprisingly, he notes that the presence of housing can prevent a decline in the share invested in stocks as investors age: as housing wealth increases, investors are more willing to accept equity risk as that risk is not highly correlated with this component. Lastly, (3) Cocco deals with the cross-sectional observation that the extent of leveraged mortgage debt is highly positively correlated with equity asset holdings. His model is able to replicate this phenomenon as well because of the consumption dimension of housing: investors with more human capital acquire more expensive houses and thus borrow more. Simultaneously, the relatively less risky human capital component induces a further tilt toward stock in high labor income investor portfolios.

16.8 Conclusions

The analysis in this chapter has brought us conceptually to the state of the art in MPT. It is distinguished by (1) the comprehensive array of asset classes that must be explicitly considered in order properly to understand an investor's *financial* asset allocations. Labor income (human capital wealth) and residential real estate are two principal cases in point. To some extent, these two asset classes provide conflicting influences on an investor's stock–bond allocations. On the one hand, as relatively riskless human capital diminishes as an investor ages, then *ceteris paribus*, his financial wealth allocation to stocks should fall. On the other hand, if his personal residence has dramatically increased in value over the investor's working years, this fact argues for increased equity holdings given the low correlation between equity and real estate returns. Which effectively dominates is unclear. (2) Long-run portfolio analysis is distinguished by its consideration of security return paths beyond the standard one-period-ahead mean, variance, and covariance characterization. Mean reversion in stock returns suggests intertemporal hedging opportunities, as does the long-run variation in the risk-free rate.

References

- Campbell, J., Cochrane, J., 1999. By force of habit: a consumption based explanation of aggregate stock market behavior. *J. Pol. Econ.* 107, 205–251.
- Campbell, J., Lo, A., MacKinlay, C., 1997. *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NY.
- Campbell, J., Viceira, L., 1999. Consumption and portfolio decisions when expected returns are time varying. *Q. J. Econ.* 114, 433–495.
- Campbell, J., Viceira, L. 2001a. Appendix to Strategic Asset Allocation, <<http://kuznets.fas.harvard.edu/~Campbell/papers.html>>.
- Campbell, J., Viceira, L., 2001b. Who should buy long-term bonds?. *Am. Econ. Rev.* 91, 99–127.
- Campbell, J., Viceira, L., 2002. *Strategic Asset Allocation*. Oxford University Press, New York, NY.
- Cocco, D., 2005. Portfolio choice in the presence of housing. *Rev. Finan. Stud.* 18, 491–533.
- Garlappi, L., De Miguel, V., Uppal, R., 2009. Optimal versus naive diversification: how inefficient is the $1/N$ portfolio strategy? *The Review of Financial Studies*. 22, 1915–1953.
- Jagannathan, R., Kocherlakota, N.R., 1996. Why should older people invest less in stocks than younger people? *Federal Reserve Bank of Minneapolis Quarterly Review*. Summer, 11–23.
- Merton, R.C., 1971. Optimum consumption and portfolio rules in a continuous time model. *J. Econ. Theory*. 3, 373–413.
- Mukherji, S., 2011. Are stock returns still mean-reverting? *Review of Financial Economics*. 20, 22–27.
- Samuelson, P.A., 1969. Lifetime portfolio selection by dynamic stochastic programming. *Rev. Econ. Stat.* 51, 239–246.
- Santos, T., Veronesi, P., 2006. Labor Income and Predictable Stock Returns. *Review of Financial Studies* 19, 1–44.
- Siegel, J., 1998. *Stocks for the Long Run*. McGraw-Hill, New York, NY.
- Viceira, L., 2001. Optimal portfolio choice for long-term investors with nontradable labor income. *J. Finan.* 56, 433–470.

Financial Structure and Firm Valuation in Incomplete Markets

Chapter Outline

17.1 Introduction 507

- 17.1.1 What Securities Should a Firm Issue if the Value of the Firm is to be Maximized? 508
- 17.1.2 What Securities Should a Firm Issue if it is to Grow as Rapidly as Possible? 508

17.2 Financial Structure and Firm Valuation 508

- 17.2.1 Financial Structure F_1 510
- 17.2.2 Financial Structure F_2 512

17.3 Arrow–Debreu and Modigliani–Miller 514

17.4 On the Role of Short Selling 516

17.5 Financing and Growth 518

- 17.5.1 No Contingent Claims Markets 519
- 17.5.2 Contingent Claims Trading 519
- 17.5.3 Incomplete Markets 521
- 17.5.4 Complete Contingent Claims 522

17.6 Conclusions 524

References 524

Appendix: Details of the Solution of the Contingent Claims Trade Case of Section 17.5 525

17.1 Introduction

We have thus far motivated the creation of financial markets by the fundamental need of individuals to transfer income across states of nature and across time periods. In Chapter 9 (Section 9.5), we initiated a discussion of the possibility of market failure in financial innovation. There we raised the possibility that coordination problems in the sharing of the benefits and the costs of setting up a new market could result in the failure of a Pareto-improving market to materialize. In reality, however, the bulk of traded securities are issued by firms with the view of raising capital for investment purposes, rather than by private individuals. It is thus legitimate to explore the incentives for security issuance, taking the viewpoint of the corporate sector. This is what we do in this chapter. Doing so involves touching upon a set of fairly wide and not fully understood topics. One of them is the issue

of *security design*. This term refers to the various forms financial contracts can take (and to their properties), in particular, in the context of managing the relationship between a firm and its managers on the one hand, and financiers and owners on the other. We will not touch on these incentive issues here but will first focus on the following two questions.

17.1.1 What Securities Should a Firm Issue if the Value of the Firm is to be Maximized?

This question is, of course, central to standard financial theory and is usually resolved under the heading [Modigliani–Miller \(MM\) Theorem \(1958\)](#). The *MM* Theorem tells us that under a set of appropriate conditions, if markets are complete, the financial decisions of the firm are irrelevant (recall our discussion in Chapter 2). Absent any tax considerations in particular, whether the firm is financed by debt or equity has no impact on its valuation. Here we go one step further and rephrase the question in a context where markets are incomplete and a firm's financing decision modifies the set of available securities. In such a world, the firm's financing decisions are important for individuals as they may affect the possibilities offered to them for transferring income across states. In this context, is it still the case that the firm's financing decisions are irrelevant for its valuation? If not, can we be sure that the interests of the firm's owners as regards the firm's financing decisions coincide with the interests of society at large?

In a second step, we cast the same *security design* issue in the context of intertemporal investment that can be loosely connected with the finance and growth issues touched upon in Chapter 1. Specifically, we raise the following complementary question.

17.1.2 What Securities Should a Firm Issue if it is to Grow as Rapidly as Possible?

We first discuss the connection between the supply of savings and the financial market structure and then consider the problem of a firm wishing to raise capital from the market. The questions raised are important: Is the financial structure relevant for a firm's ability to obtain funds to finance its investments? If so, are the interests of the firm aligned with those of society?

17.2 Financial Structure and Firm Valuation

Our discussion will be phrased in the context of the following simple example. We assume the existence of a unique firm owned by an entrepreneur who wishes only to consume at date $t = 0$; for this entrepreneur, $U'(c_0) > 0$. The assumption of a single entrepreneur circumvents the problem of *shareholder unanimity*: If markets are incomplete, the firm's objective does not need to be the maximization of market value: shareholders cannot reallocate income across all dates and states as they may wish. By definition, there are

missing markets. But then shareholders may well have differing preferred payment patterns by the firm—over time and across states—depending on the specificities of their own endowments. One shareholder, for example, may prefer investment project A because it implies the firm will flourish and pay high dividends in future circumstances when he himself will otherwise have a low income. Another shareholder may prefer the firm to undertake some other investment project or to pay higher current dividends because her personal circumstances are different. Furthermore, there may be no markets where the two shareholders could insure one another.

The firm's financial structure consists of a finite set of claims against the firm's period 1 output. These securities are assumed to exhaust the returns to the firm in each state of nature. Since the entrepreneur wishes to consume only in period 0, and yet his firm creates consumption goods only in period 1, he will want to sell claims against period 1 output in exchange for consumption in period 0.

The other agents in our economy are agents 1 and 2 of the standard Arrow–Debreu setting of Chapter 9, and we retain the same general assumptions:

1. There are two dates: 0, 1.
2. At date 1, N possible states of nature, indexed $\theta = 1, 2, \dots, N$, with probabilities π_θ , may be realized. In fact, for nearly all that we wish to illustrate $N = 2$ is sufficient.
3. There is one consumption good.
4. Besides the entrepreneur, there are two consumer-investors, indexed $k = 1, 2$, with preferences given by

$$U_0^k(c_0^k) + \delta^k \sum_{\theta=1}^N \pi_\theta U^k(c_\theta^k) = \alpha c_\theta^k + E \ln c_\theta^k$$

and endowments $e_0^k, (e_\theta^k)_{\theta=1,2,\dots,N}$. We interpret c_θ^k to be the consumption of agent k if state θ should occur, and c_0^k his period zero consumption. Agents' period utility functions are all assumed to be concave, α is the constant date 0 marginal utility, which, for the moment, we will specify to be 0.1, and the discount factor is unity (there is no time discounting). The endowment matrix for the two agents is assumed to be as given in Table 17.1.

Table 17.1: Endowment matrix

	Date $t = 0$	Date $t = 1$	
		State $\theta = 1$	State $\theta = 2$
Agent $k = 1$	4	1	5
Agent $k = 2$	4	5	1

Table 17.2: Cash flows at date $t = 1$

	$\theta = 1$	$\theta = 2$
<i>Firm</i>	2	2

Each state has probability $\frac{1}{2}$ (equally likely), and consumption in period 0 cannot be stored and carried over into period 1. Keeping matters as simple as possible, let us further assume that the cash flows to the firm are the same in each state of nature, as seen in [Table 17.2](#).

At least two different financial structures could be written against this output vector:

$$F_1 = \{(2,2)\} \text{—pure equity,}^1$$

$$F_2 = \{(2,0), (0,2)\} \text{—Arrow–Debreu securities.}^2$$

From our discussion in Chapter 9, we expect financial structure F_2 to be more desirable to agents 1 and 2, because it better allows them to effect income (consumption) stabilization: F_2 amounts to a complete market structure with the two required Arrow–Debreu securities. Let us compute the value of the firm (what the claims to its output could be sold for) under both financial structures. Note that the existence of either set of securities affords an opportunity to shift consumption between periods. This situation is fundamentally different, in this way, from the pure reallocation examples in the pure exchange economies of Chapter 9.

17.2.1 Financial Structure F_1

Let q^e denote the price (in terms of date 0 consumption) of equity—security $\{(2, 2)\}$ —and let z_1, z_2 , respectively, be the quantities demanded by agents 1 and 2. In equilibrium, $z_1 + z_2 = 1$ since there is one unit of equity issued; holding z units of equity entitles the owner to a dividend of $2z$ in both states 1 and 2.

Agent 1 solves:

$$\max_{q^e z_1 \leq 4} (0.1)(4 - q^e z_1) + \frac{1}{2}[\ln(1 + 2z_1) + \ln(5 + 2z_1)]$$

Agent 2 solves:

$$\max_{q^e z_2 \leq 4} (0.1)(4 - q^e z_2) + \frac{1}{2}[\ln(5 + 2z_2) + \ln(1 + 2z_2)]$$

¹ Equity is risk free here. This is the somewhat unfortunate consequence of our symmetry assumption (same output in the two date $t = 1$ states). The reader may want to check that our message carries over with a state $\theta = 2$ output of 3.

² We could assume, equivalently, that the firm issues two units of the two conceivable *pure* Arrow–Debreu securities, $\{(1, 0), (0, 1)\}$.

Table 17.3: Equilibrium allocation

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	$4 - 3\frac{1}{3}$	1 + 1	5 + 1
Agent 2	$4 - 3\frac{1}{3}$	5 + 1	1 + 1

Assuming an interior solution, the first-order conditions (FOCs) for agents 1 and 2 are, respectively,

$$z_1: \left(\frac{1}{10} \right) q^e = \frac{1}{2} \left[\frac{2}{1+2z_1} \right] + \frac{1}{2} \left[\frac{2}{5+2z_1} \right]$$

$$\frac{q^e}{10} = \left[\frac{1}{1+2z_1} + \frac{1}{5+2z_1} \right]$$

$$z_2: \frac{q^e}{10} = \left[\frac{1}{5+2z_2} + \frac{1}{5+2z_2} \right]$$

Clearly, $z_1 = z_2 = \frac{1}{2}$, and $\frac{q^e}{10} = [1/(1+1) + 1/(5+1)] = [1/2 + 1/6] = \frac{2}{3}$ or $q^e = \frac{20}{3}$. Thus, $V_{F_1} = q^e = \frac{20}{3} = 6\frac{2}{3}$, and the resulting equilibrium allocation is displayed in [Table 17.3](#).

Agents are thus willing to pay a large proportion of their period 1 consumption in order to increase period 2 consumption. On balance, agents (except the entrepreneur) wish to shift income from the present (where $MU = \alpha = 0.1$) to the future and now there is a device by which they may do so.

Since markets are incomplete in this example, the competitive equilibrium need not be Pareto optimal. That is the case here. There is no way to equate the ratios of the two agents' marginal utilities across the two states: In state 1, the MU ratio is $\frac{1/2}{1/6} = 3$ while it is $\frac{1/2}{1/6} = \frac{1}{3}$ in state 2. A transfer of one unit of consumption from agent 2 to agent 1 in state 1 in exchange for one unit of consumption in the other direction in state 2 would obviously be Pareto improving. Such a transfer cannot, however, be effected with the limited set of financial instruments available. This is the reality of incomplete markets.

Note that our economy is one of three agents: agents 1 and 2, and the original firm owner. From another perspective, the equilibrium allocation under F_1 is not a Pareto optimum because a redistribution of wealth between agents 1 and 2 could be effected, making them both better off in *ex ante* expected utility terms while not reducing the utility of the firm owner (which is, presumably, directly proportional to the price he receives for the firm). In particular, the allocation that dominates the one achieved under F_1 is given in [Table 17.4](#).

Table 17.4: A Pareto-superior allocation

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	$\frac{2}{3}$	4	4
Agent 2	$\frac{2}{3}$	4	4
Owner	$6\frac{2}{3}$	0	0

17.2.2 Financial Structure F_2

This is a complete Arrow–Debreu financial structure. It will be notationally clearer here if we deviate from our usual notation and denote the securities as $X = (2, 0)$, $W = (0, 2)$ with prices q_X , q_W , respectively (q_X thus corresponds to the price of two units of the state-1 Arrow–Debreu security, while q_W is the price of two units of the state-2 Arrow–Debreu security), and quantities $z_X^1, z_X^2, z_W^1, z_W^2$. The problems confronting the agents are as follows.

Agent 1 solves:

$$\max_{(z_X^1, z_W^1)} \left(\frac{1}{10} (4 - q_X z_X^1 - q_W z_W^1) + \left[\frac{1}{2} \ln(1 + 2z_X^1) + \frac{1}{2} \ln(5 + 2z_W^1) \right] \right)$$

$$q_X z_X^1 + q_W z_W^1 \leq 4$$

Agent 2 solves:

$$\max_{(z_X^2, z_W^2)} \left(\frac{1}{10} (4 - q_X z_X^2 - q_W z_W^2) + \left[\frac{1}{2} \ln(5 + 2z_X^2) + \frac{1}{2} \ln(5 + 2z_W^2) \right] \right)$$

$$q_X z_X^2 + q_W z_W^2 \leq 4$$

The FOCs are:

$$\text{Agent 1: } \begin{cases} (i) \frac{1}{10} q_X = \frac{1}{2} \left(\frac{1}{1 + 2z_X^1} \right) 2 \\ (ii) \frac{1}{10} q_W = \frac{1}{2} \left(\frac{1}{5 + 2z_W^1} \right) 2 \end{cases}$$

$$\text{Agent 2: } \begin{cases} (iii) \frac{1}{10} q_X = \frac{1}{2} \left(\frac{1}{1 + 2z_X^2} \right) 2 \\ (iv) \frac{1}{10} q_W = \frac{1}{2} \left(\frac{1}{5 + 2z_W^2} \right) 2 \end{cases}$$

$$\text{By equation (i): } \frac{1}{10}q_X = \frac{1}{1+2z_X^1} \Rightarrow 1+2z_X^1 = \frac{10}{q_X} \Rightarrow z_X^1 = \frac{5}{q_X} - \frac{1}{2}$$

$$\text{By equation (iii): } \frac{1}{10}q_X = \frac{1}{5+2z_X^2} \Rightarrow 5+2z_X^2 = \frac{10}{q_X} \Rightarrow z_X^2 = \frac{5}{q_X} - \frac{5}{2}$$

With one security of each type issued:

$$\begin{aligned} z_X^1 + z_X^2 &= 1(z_X^1 \geq 0, z_X^2 \geq 0) \\ \frac{5}{q_X} - \frac{1}{2} + \frac{5}{q_X} - \frac{5}{2} &= 1 \Rightarrow \frac{10}{q_X} = 4 \Rightarrow q_X = \frac{10}{4} \end{aligned}$$

Similarly, $q_W = \frac{10}{4}$ (by symmetry) and $V_F = q_X + q_W = \frac{10}{4} + \frac{10}{4} = \frac{20}{4} = 5$.

So we see that V_F has declined from $6\frac{2}{3}$ in the F_1 case to 5. Let us further examine this result. Consider the allocations implied by the complete financial structure:

$$z_X^1 = \frac{5}{q_X} - \frac{1}{2} = \frac{5}{5/2} - \frac{1}{2} = 2 - \frac{1}{2} = 1\frac{1}{2}$$

$$z_X^2 = \frac{5}{q_X} - \frac{5}{2} = \frac{5}{5/2} - \frac{5}{2} = 2 - \frac{5}{2} = -\frac{1}{2}$$

$$z_W^1 = -\frac{1}{2}, z_W^2 = 1\frac{1}{2} \text{ by symmetry}$$

Thus, agent 1 wants to short sell security 2 while agent 2 wants to short sell security 1. Of course, in the case of financial structure $F_1(2, 2)$, there was no possibility of short selling since every agent in equilibrium must have the same security holdings. The post-trade allocation is found in [Table 17.5](#). This, unsurprisingly, constitutes a Pareto optimum.³

Table 17.5: Post-trade allocation

$t = 0$
$\text{Agent 1: } 4 - (1\frac{1}{2})q_X + \frac{1}{2}q_W = 4 - \frac{3}{2}(\frac{10}{4}) + \frac{1}{2}(\frac{10}{4}) = 4 - \frac{10}{4} = 1\frac{1}{2}$
$\text{Agent 2: } 4 + \frac{1}{2}q_X - \frac{3}{2}q_W = 4 + \frac{1}{2}(\frac{10}{4}) - \frac{3}{2}(\frac{10}{4}) = 4 - \frac{10}{4} = 1\frac{1}{2}$
$t = 1$
$\text{Agent 1: } (1, 5) + 1\frac{1}{2}(2, 0) - \frac{1}{2}(0, 2) = (4, 4)$
$\text{Agent 2: } (5, 1) + (-\frac{1}{2})(2, 0) + 1\frac{1}{2}(0, 2) = (4, 4)$

³ Note that our example also illustrates the fact that the addition of new securities in a financial market does not necessarily improve the welfare of *all* participants. Indeed, the firm owner is made worse off by the transition from F_1 to F_2 .

We have thus reached an important result that we summarize in [Propositions 17.1 and 17.2](#).

Proposition 17.1 When markets are incomplete, the MM Theorem fails to hold and the financial structure of the firm may affect its valuation by the market.

Proposition 17.2 When markets are incomplete, it may not be in the interest of a value-maximizing firm to issue the socially optimal set of securities.

In our example the issuing of the *right* set of securities by the firm leads to completing the market and making a Pareto optimal allocation attainable. The impact of the financial decision of the firm on the set of markets available to individuals in the economy places us outside the realm of the MM Theorem; indeed, the value of the firm is not left unaffected by the choice of financing. Moreover, it appears that it is not, in this situation, in the private interest of the firm's owner to issue the socially optimal set of securities. Our example thus suggests that there is no reason to necessarily expect that value-maximizing firms will issue the set of securities society would find preferable.⁴

17.3 Arrow–Debreu and Modigliani–Miller

In order to understand why V_F declines when the firm issues the richer set of securities, it is useful to draw on our work on Arrow–Debreu pricing (Chapter 9). Think of the economy under financial structure F_2 . This is a complete Arrow–Debreu structure in which we can use the information on equilibrium endowments to recompute the pure Arrow–Debreu prices as per [Eq. \(17.1\)](#),

$$q_\theta = \frac{\delta \pi_\theta \frac{\partial U^k}{\partial c_\theta^k}}{\frac{\partial U^k}{\partial c_0^k}}, \quad \theta = 1, 2 \quad (17.1)$$

which, in our example, given the equilibrium allocation (four units of commodity in each state for both agents) reduces to

$$q_\theta = \frac{1\left(\frac{1}{2}\right)\left(\frac{1}{4}\right)}{0.1} = \frac{5}{4}, \quad \theta = 1, 2$$

which corresponds, of course, to

$$q_X = q_W = \frac{10}{4}$$

and to $V_F = 5$.

⁴ The reader may object that our example is just that, an example. Because it helps us reach results of a negative nature, this example is, however, a fully general *counterexample*, ruling out the proposition that the MM Theorem continues to hold and that firms' financial structure decisions will always align with the social interest.

This Arrow–Debreu complete markets equilibrium is unique: this is generically the case in an economy such as ours, implying there are no other allocations satisfying the required conditions and no other possible prices for the Arrow–Debreu securities. This implies the MM proposition as the following reasoning illustrates. In our example, the firm is a mechanism to produce two units of output in date 1, both in states 1 and 2. Given that the date 0 price of one unit of the good in state 1 at date 1 is $\frac{5}{4}$ and the price of one unit of the good in state 2 at date 1 is $\frac{5}{4}$ as well, it must of necessity be that the price (value) of the firm is four times $\frac{5}{4}$, i.e., 5. In other words, absent any romantic love for this firm, no one will pay more than five units of the current consumption good (which is the numeraire) for the title of ownership to this production mechanism, knowing that the same bundle of goods can be obtained for five units of the numeraire by purchasing two units of each Arrow–Debreu security. The converse reasoning guarantees that the firm will also not sell for less. The value of the firm is thus given by its fundamentals and is independent of the specific set of securities the entrepreneur chooses to issue: This is the essence of the MM Theorem!

Now let us try to understand how this reasoning is affected when markets are incomplete and why, in particular, the value of the firm is higher in that context. The intuition is as follows. In the incomplete market environment of financial structure F_1 , security $\{(2, 2)\}$ is desirable for two reasons: to transfer income across time *and* to reduce date 1 consumption risk. In this terminology, the firm in the incomplete market environment is more than a mechanism to produce two units of output in either states of nature in date 1. The security issued by the entrepreneur is also the only available vehicle to reduce second-period consumption risk. Individual consumers are willing to pay something, i.e., to sacrifice current consumption, to achieve such risk reduction. To see that trading of security $\{(2, 2)\}$ provides some risk reduction in the former environment, we need only compare the range of date 1 utilities across states after trade and before trade for agent 1 (agent 2 is symmetric). See [Table 17.6](#).

The premium paid for the equity security, over and above the value of the firm in complete markets, thus originates in the dual role it plays as a mechanism for consumption risk smoothing and as a title to two units of output in each future state. A question remains: Given that the entrepreneur, by his activity and security issuance, plays this dual role, why can't he reap the corresponding rewards independently of the security structure he chooses to issue? In other words, why is it that his incentives are distorted away from the socially optimal financial structure? To understand this, note that if any amount of Arrow–Debreu-like securities, such as in $F_2 = \{(2, 0), (0, 2)\}$ is issued, no matter how

Table 17.6: Agent 1 state utilities under F_1

	Before Trade	$\{(2, 2)\}; z^1 = 0.5$ (Equilibrium Allocation)
<i>State 1</i>	$U^1(c_1^1) = \ln 1 = 0$	$U^1(c_1^1) = \ln 2 = 0.693$
<i>State 2</i>	$U^1(c_2^1) = \ln 5 = 1.609$ Difference = 1.609	$U^1(c_2^1) = \ln 6 = 1.792$ Difference = 1.099

Table 17.7: Allocation when the two agents trade Arrow–Debreu securities among themselves

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	4	3	3
Agent 2	4	3	3

small, the market for such securities has effectively been created. With no further trading restrictions, the agents can then supply additional amounts of these securities to one another. This has the effect of empowering them to trade, entirely independently of the magnitude of the firm's security issuance, to the following endowment allocation (Table 17.7).

In effect, investors can eliminate all *second-period* endowment uncertainty *themselves*. Once this has been accomplished and markets are effectively completed (because there is no further demand for across-state income redistribution, it is irrelevant to the investor whether the firm issues $\{(2, 2)\}$ or $\{(2, 0), (0, 2)\}$, since either package is equally appropriate for transferring income *across time periods*). Were $\{(2, 0), (0, 2)\}$ to be the package of securities issued, each agent would buy equal amounts of $(2, 0)$, and $(0, 2)$, and effectively repackage them as $(2, 2)$. To do otherwise would be to reintroduce date 1 endowment uncertainty. Thus, the relative value of the firm under either financial structure, $\{(2, 2)\}$ or $\{(2, 0), (0, 2)\}$, is determined solely by whether the security $(2, 2)$ is worth more to the investors in the environment of period 2 endowment uncertainty or when all risk has been eliminated as in the environment noted previously.

Said otherwise, once the markets have been completed, the value of the firm is fixed at 5 as we have seen before, and there is nothing the entrepreneur can do to appropriate the extra insurance premium. If investors can eliminate all the risk *themselves* (via short selling), there is no premium to be paid to the firm, in terms of value enhancement, for doing so. This is confirmed if we examine the value of the firm when security $\{(2, 2)\}$ is issued *after* the agents have traded among themselves to equal second-period allocation $(3, 3)$. In this case $V_F = 5$ also.

There is another lesson to be gleaned from this example and that leads us back to the CAPM. One implication of the CAPM was that securities could not be priced in isolation: their prices and rates of return depended on their interactions with other securities as measured by the covariance. This example follows in that tradition by confirming that the value of the securities issued by the firm is not independent of the other securities available on the market or which the investors can *themselves* create.

17.4 On the Role of Short Selling

From another perspective (as noted in [Allen and Gale, 1994](#)), short selling expands the supply of securities and provides additional opportunities for risk sharing, but in such a way

that the benefits are not internalized by the innovating firm. When deciding what securities to issue, however, the firm only takes into account the impact of the security issuance on its own value; in other words, it only considers those benefits it can internalize. Thus, in an incomplete market setting, the firm may not issue the socially optimal package of securities.

It is interesting to consider the consequence of forbidding or making it impossible for investors to increase the supply of securities (2, 0) and (0, 2) via short selling. Accordingly, let us impose a no-short-selling condition (by requiring that all holdings of all securities by all agents are positive). Agent 1 wants to short sell (0, 2); agent 2 wants to short sell (2, 0). We thus know that the constrained optimum will have (simply setting $z = 0$ wherever the unconstrained optimum had a negative z and anticipating the market clearing condition):

$$\begin{aligned} z_X^2 &= 0 \quad z_W^1 = 0 \\ z_X^1 &= 1 \quad z_W^2 = 1 \end{aligned}$$

$$\frac{1}{10}q_X = MU_1 = \frac{1}{2} \left(\frac{1}{1+2(1)} \right) 2 = \frac{1}{3}$$

$$\frac{1}{10}q_W = MU_2 = \frac{1}{2} \left(\frac{1}{1+2(1)} \right) 2 = \frac{1}{3}$$

$$q_X = \frac{10}{3}, q_W = \frac{10}{3}$$

$$V_F = \frac{20}{3} = 6\frac{2}{3}$$

which coincides with the valuations when the security (2, 2) was issued.

The fact that V_F increases when short sales are prohibited is not surprising since it reduces the supply of securities (2, 0) and (0, 2). With demand unchanged, both q_X and q_W increase, and with it, V_F . In some sense, now the firm has a monopoly in the issuance of (2, 0) and (0, 2), and that monopoly position has value. All this is in keeping with the general reasoning developed previously. While it is, therefore, not surprising that the value of the firm has risen with the imposition of the short sales constraint, the fact that its value has returned precisely to what it was when it issued $\{(2, 2)\}$ is striking and possibly somewhat of a coincidence.

Is the ruling out of short selling realistic? In practice, short selling on the US stock exchanges is costly, and only a very limited amount of it occurs. The reason for this is that the short seller must deposit as collateral with the lending institution, as much as 100% of the value of the securities he borrows to short sell. Under current practice in the United States, the interest on this deposit is less than the T-bill rate even for the largest

participants, and for small investors it is near zero. There are other exchange-imposed restrictions on short selling. On the NYSE, for example, investors are forbidden to short sell on a down-tick in the stock's price.⁵

17.5 Financing and Growth

Now we must consider our second set of issues, which we may somewhat more generally characterize as follows: How does the degree of completeness in the securities markets affect the level of capital accumulation? This is a large topic, touched upon in our introductory chapter, for which there is little existing theory. Once again we pursue our discussion in the context of examples.

Example 17.1 Our first example serves to illustrate the fact that, although a more complete set of markets is unambiguously good for welfare, it is not necessarily so for growth. Consider the following setup. Agents own firms (have access to a productive technology) while also being able to trade state-contingent claims with one another (net supply is zero). We retain the two-agent, two-period setting. Agents have state-contingent consumption endowments in the second period. They also have access to a productive technology which, for every k units of period one consumption foregone, produces \sqrt{k} in date 1 in either state of nature⁶ (Table 17.8).

The agent endowments are given in Table 17.9, and the agent preference orderings are now (identically) given by

$$EU(c_0, c_\theta) = \ln(c_0) + \frac{1}{2} \ln(c_1) + \frac{1}{2} \ln(c_2)$$

Table 17.8: The return from investing k units

$t = 1$	
θ_1 \sqrt{k}	θ_2 \sqrt{k}

⁵ Brokers must obtain permission from clients to borrow their shares and relend them to a short seller. In the early part of 2000, a number of high-technology firms in the United States asked their shareholders to deny such permission as it was argued short sellers were depressing prices! Of course, if a stock's price begins rising, short sellers may have to enter the market to buy shares to cover their short position. This boosts the share price even further.

⁶ Such a technology may not look very interesting at first sight! But, at the margin, agents may be very grateful for the opportunity it provides to smooth consumption across time periods.

Table 17.9: Agent endowments

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	3	5	1
Agent 2	3	1	5
Prob(θ_1) = Prob(θ_2) = $\frac{1}{2}$			

In this context, we compute the agents' optimal savings levels under two alternative financial structures. In one case, there is a complete set of contingent claims, and in the other, the productive technology is the only possibility for redistributing purchasing power across states (as well as across time) among the two agents.

17.5.1 No Contingent Claims Markets

Each agent acts autonomously and solves

$$\max_k \ln(3 - k) + \frac{1}{2} \ln(5 + \sqrt{k}) + \frac{1}{2} (1 + \sqrt{k})$$

Assuming an interior solution, the optimal level of savings k^* solves

$$-\frac{1}{3 - k^*} + \left\{ \frac{1}{2} \left(\frac{1}{5 + \sqrt{k^*}} \right) \frac{1}{2} (k^*)^{-\frac{1}{2}} + \frac{1}{2} \left(\frac{1}{1 + \sqrt{k^*}} \right) \frac{1}{2} (k^*)^{-\frac{1}{2}} \right\} = 0$$

which, after several simplifications, yields

$$3(k^*)^{\frac{3}{2}} + 15k^* + 7\sqrt{k^*} - 9 = 0$$

The solution to this equation is $k^* = 0.31$. With two agents in the economy, *economy-wide* savings are 0.62. Let us now compare this result with the case in which the agents also have access to contingent claims markets.

17.5.2 Contingent Claims Trading

Let q_1 be the price of a security that pays one unit of consumption if state 1 occurs, and let q_2 be the price of a security that pays one unit of consumption if state 2 occurs. Similarly, let $z_1^1, z_2^1, z_1^2, z_2^2$ denote, respectively, the quantities of these securities demanded by agents 1 and 2. These agents continue to have simultaneous access to the technology.

Agent 1 solves

$$\max_{k_1, z_1^1, z_2^1} \ln(3 - k_1 - q_1 z_1^1 - q_2 z_2^1) + \frac{1}{2} \ln(5 + \sqrt{k_1} + z_1^1) + \frac{1}{2} \ln(1 + \sqrt{k_1} + z_2^1)$$

Agent 2's problem is essentially the same:

$$\max_{k_2, z_1^2, z_2^2} \ln(3 - k_2 - q_1 z_1^2 - q_2 z_2^2) + \frac{1}{2} \ln(1 + \sqrt{k_2} + z_1^2) + \frac{1}{2} \ln(5 + \sqrt{k_2} + z_2^2)$$

By symmetry, in equilibrium

$$\begin{aligned} k_1 &= k_2; q_1 = q_2; \\ z_1^1 &= z_2^2 = -z_1^2, z_2^1 = z_1^2 = -z_2^2 \end{aligned}$$

Using these facts and the FOCs (see the Appendix), it can be directly shown that

$$-2 = z_1^1$$

and, it then follows that $k_1 = 0.16$. Thus, *total savings* = $k_1 + k_2 = 2k_1 = 0.32$.

Savings have thus been substantially reduced. This result also generalizes to situations of more general preference orderings, and to the case where the uncertainty in the states is in the form of uncertainty in the production technology rather than in the investor endowments. The explanation for this phenomenon is relatively straightforward, and it parallels the mechanism at work in the previous sections. With the opening of contingent claims markets, the agents can eliminate all second-period risk. In the absence of such markets, it is real investment that alone must provide for any risk reduction as well as for income transference across time periods—a dual role. In a situation of greater uncertainty, resulting from the absence of contingent claims markets, more is saved and the extra savings take, necessarily, the form of productive capital: there is a precautionary demand for capital. [Jappelli and Pagano \(1994\)](#) found traces of a similar behavior in Italy prior to recent measures of financial deregulation.

Example 17.2 This result also suggests that if firms want to raise capital in order to invest for date 1 output, it may not be value maximizing to issue a more complete set of securities, an intuition we confirm in our second example.

Consider a firm with access to a technology with the output pattern found in [Table 17.10](#).

Investor endowments are given in [Table 17.11](#). Their preference orderings are both of the form

$$EU(c_1, c_\theta) = \frac{1}{12}c_0 + \frac{1}{2} \ln(c_1) + \frac{1}{2} \ln(c_2)$$

Table 17.10: The firm's technology

$t = 0$	$t = 1$	
	θ_1	θ_2
$-k$	\sqrt{k}	\sqrt{k}

Table 17.11: Investor endowments

	$t = 0$	$t = 1$	
		θ_1	θ_2
Agent 1	12	$\frac{1}{2}$	10
Agent 2	12	10	$\frac{1}{2}$

17.5.3 Incomplete Markets

Suppose a security of the form $(1, 1)$ is traded, at a price q ; agents 1 and 2 demand, respectively, z_1 and z_2 . The agent maximization problems that define their demand are as follows:

Agent 1:

$$\max_{qz_1 \leq 12} \left(\frac{1}{12} \right) (12 - qz_1) + \frac{1}{2} \ln \left(\frac{1}{2} + z_1 \right) + \frac{1}{2} \ln (10 + z_1)$$

Agent 2:

$$\max_{qz_2 \leq 12} \left(\frac{1}{12} \right) (12 - qz_2) + \frac{1}{2} \ln (10 + z_2) + \frac{1}{2} \ln \left(\frac{1}{2} + z_2 \right)$$

It is obvious that $z_1 = z_2$ at equilibrium. The FOCs are (again assuming an interior solution):

$$\text{Agent 1: } \frac{q}{12} = \frac{1}{2} \frac{1}{(\frac{1}{2} + z_1)} + \frac{1}{2} \frac{1}{(10 + z_1)}$$

$$\text{Agent 2: } \frac{q}{12} = \frac{1}{2} \frac{1}{(10 + z_2)} + \frac{1}{2} \frac{1}{(\frac{1}{2} + z_2)}$$

In order for the technological constraint to be satisfied, it must also be that

$$\begin{aligned}[q(z_1 + z_2)]^{\frac{1}{2}} &= z_1 + z_2, \text{ or} \\ q &= z_1 + z_2 = 2z_1 \text{ as noted earlier}\end{aligned}$$

Substituting for q in the first agent's FOC gives

$$\begin{aligned}\frac{2z_1}{12} &= \frac{1}{2} \frac{1}{\left(\frac{1}{2} + z_1\right)} + \frac{1}{2(10 + z_1)} \text{ or} \\ 0 &= z_1^3 + 10.5z_1^2 - z_1 - 31.5\end{aligned}$$

Trial and error gives $z_1 = 1.65$. Thus, $q = 3.3$ and total investment is $q = z_1 + z_2 = (3.3)(3.3) = 10.89 = V_F$; date 1 output in each state is thus $\sqrt{10.89} = 3.3$.

17.5.4 Complete Contingent Claims

Now suppose securities $R = (1, 0)$ and $S = (0, 1)$ are traded at prices q_R and q_S and denote quantities demanded, respectively, as $z_R^1, z_R^2, z_S^1, z_S^2$. The no-short sales assumption is retained. With this assumption, agent 1 buys only R , while agent 2 buys only security S . Each agent thus prepares himself for his worst possibility.

Agent 1:

$$\begin{aligned}\max \left(\frac{1}{12} \right) (12 - q_R z_R^1) + \frac{1}{2} \ln \left(\frac{1}{2} + z_R^1 \right) + \frac{1}{2} \ln(10) \\ 0 \leq q_R z_R^1\end{aligned}$$

Agent 2:

$$\begin{aligned}\max \left(\frac{1}{12} \right) (12 - q_S z_S^2) + \frac{1}{2} \ln(10) + \frac{1}{2} \ln \left(\frac{1}{2} + z_S^2 \right) \\ 0 \leq q_S z_S^2\end{aligned}$$

The FOCs are thus

$$\text{Agent 1: } \frac{q_R}{12} = \frac{1}{2} \frac{1}{\left(\frac{1}{2} + z_R^1\right)}$$

$$\text{Agent 2: } \frac{q_S}{12} = \frac{1}{2} \frac{1}{\left(\frac{1}{2} + z_S^2\right)}$$

Clearly, $q_R = q_S$ by symmetry, and $z_R^1 = z_S^2$; by the technological constraints:

$$(q_R z_R^1 + q_S z_S^2)^{\frac{1}{2}} = \left(\frac{z_R^1 + z_S^2}{2} \right) \text{ or}$$

$$q_R = \frac{z_R^1}{2}$$

Solving for z_R^1 :

$$\frac{q_R}{12} = \frac{z_R^1}{24} = \frac{1}{2} \frac{1}{\left(\frac{1}{2} + z_R^1 \right)} = \frac{1}{1 + 2z_R^1}$$

$$z_R^1(1 + 2z_R^1) = 24$$

$$z_R^1 = \frac{-1 \pm \sqrt{1 - 4(2)(-24)}}{4} = \frac{-1 \pm \sqrt{1 + 192}}{4} = \frac{-1 \pm 13.892}{4}$$

(taking positive root)

$$z_R^1 = 3.223$$

$$z_S^2 = 3.223$$

$$q_R = 1.61, \text{ and}$$

$$q_R(z_R^1 + z_S^2) = 1.61(6.446) = 10.378 = V_F$$

As suspected, this is less than what the firm could raise issuing only (1, 1).

Much in the spirit of our discussion of [Section 17.2](#), this example illustrates the fact that, for a firm wishing to maximize the amount of capital levied from the market, it may not be a good strategy to propose contracts leading to a (more) complete set of markets. This is another example of the failure of the MM Theorem in a situation of incomplete markets, and the reasoning is the same as before: in incomplete markets, the firm's value is not necessarily equal to the value, computed at Arrow–Debreu prices, of the portfolio of goods it delivers in future date states. This is because the security it issues may, in addition, be valued by market participants for its unintended role as an insurance mechanism, a role that disappears if markets are complete. In the growth context of our last examples, this may mean that more savings will be forthcoming when markets are incomplete, a fact that may lead a firm wishing to raise capital from the markets to refrain from issuing the optimal set of securities.

17.6 Conclusions

We have reached a number of conclusions in this chapter.

1. In an incomplete market context, it may not be value maximizing for firms to offer the socially optimal (complete) set of securities. This follows from the fact that, in a production setting, securities can be used not only to handle risk reduction but also to transfer income across dates. The value of a security will depend upon its usefulness in accomplishing these alternative tasks.
2. The value of securities issued by the firm is not independent of the supply of similar securities issued by other market participants. To the extent that others can increase the supply of a security initially issued by the firm (via short selling), its value will be reduced.
3. Finally, welfare is promoted by the issuance of a more complete set of markets, but growth may not be.⁷ As a result, it may not be in the best interest of a firm aiming at maximizing the amount of capital it wants to raise, to issue the most socially desirable set of securities.

All these results show that if markets are incomplete, the link between private interests and social optimality is considerably weakened. Here lies the intellectual foundation for financial market regulation and supervision.

References

- Allen, F., Gale, D., 1994. Financial Innovation and Risk Sharing. MIT Press, Cambridge, MA.
- Hart, O., 1975. On the optimality of equilibrium when market structure is incomplete. *J. Econ. Theory*. 11, 418–443.
- Jappelli, T., Pagano, M., 1994. Savings, growth and liquidity constraints. *Q. J. Econ.* 109, 83–109.
- Modigliani, F., Miller, M., 1958. The cost of capital, corporation finance, and the theory of investment. *Am. Econ. Rev.* 48, 261–297.

⁷ The statement regarding welfare is strictly true only when financial innovation achieves full market completeness. [Hart \(1975\)](#) shows that it is possible that everyone is made worse off when the markets become more complete but not fully complete (say, going from 9 to 10 linearly independent securities when 15 would be needed to make the markets complete).

Appendix: Details of the Solution of the Contingent Claims Trade Case of Section 17.5

Agent 1 solves:

$$\begin{aligned} \max_{k_1, z_1^1, z_2^1} & \ln(3 - k_1 - q_1 z_1^1 - q_2 z_2^1) + \frac{1}{2} \ln(5 + \sqrt{k_1} + z_1^1) + \frac{1}{2} \ln(1 + \sqrt{k_1} + z_2^1) \\ k_1: & \frac{-1}{3 - k_1 - q_1 z_1^1 - q_2 z_2^1} + \frac{1}{2} \left(\frac{1}{5 + \sqrt{k_1} + z_1^1} \right) \frac{1}{2} k_1^{-\frac{1}{2}} + \frac{1}{2} \left(\frac{1}{1 + \sqrt{k_1} + z_2^1} \right) \frac{1}{2} k_1^{-\frac{1}{2}} = 0 \quad (17.2) \\ z_1^1: & \frac{-q_1}{3 - k_1 - q_1 z_1^1 - q_2 z_2^1} + \frac{1}{2} \left(\frac{1}{5 + \sqrt{k_1} + z_1^1} \right) = 0 \quad (17.3) \\ z_2^1: & \frac{-q_2}{3 - k_1 - q_1 z_1^1 - q_2 z_2^1} + \frac{1}{2} \left(\frac{1}{1 + \sqrt{k_1} + z_2^1} \right) = 0 \quad (17.4) \end{aligned}$$

Agent 2's problem and FOC are essentially the same:

$$\begin{aligned} \max_{k_2, z_1^2, z_2^2} & \ln(3 - k_2 - q_1 z_1^2 - q_2 z_2^2) + \frac{1}{2} \ln(1 + \sqrt{k_2} + z_1^2) + \frac{1}{2} \ln(5 + \sqrt{k_2} + z_2^2) \\ k_2: & \frac{-1}{3 - k_2 - q_1 z_1^2 - q_2 z_2^2} + \frac{1}{2} \left(\frac{1}{1 + \sqrt{k_2} + z_1^2} \right) \frac{1}{2} k_2^{-\frac{1}{2}} + \frac{1}{2} \left(\frac{1}{5 + \sqrt{k_2} + z_2^2} \right) \frac{1}{2} k_2^{-\frac{1}{2}} = 0 \quad (17.5) \\ z_1^2: & \frac{-q_1}{3 - k_2 - q_1 z_1^2 - q_2 z_2^2} + \frac{1}{2} \left(\frac{1}{1 + \sqrt{k_2} + z_1^2} \right) = 0 \quad (17.6) \\ z_2^2: & \frac{-q_2}{3 - k_2 - q_1 z_1^2 - q_2 z_2^2} + \frac{1}{2} \left(\frac{1}{5 + \sqrt{k_2} + z_2^2} \right) = 0 \quad (17.7) \end{aligned}$$

By symmetry, in equilibrium

$$\begin{aligned} k_1 &= k_2; q_1 = q_2; \\ z_1^1 &= z_2^2 = -z_1^2, z_2^1 = z_1^2 - z_2^2 \end{aligned}$$

By Eqs. (17.3) and (17.6), using the fact that $z_1^1 + z_2^1 = z_2^2 + z_1^2$:

$$\frac{1}{5 + \sqrt{k_1} + z_1^1} = \frac{1}{1 + \sqrt{k_2} + z_1^2}$$

Eqs. (17.4) and (17.7):

$$\frac{1}{1 + \sqrt{k_1} + z_2^1} = \frac{1}{5 + \sqrt{k_2} + z_2^2}$$

The equations defining k_1 and z_1^1 are thus reduced to

$$k_1: \frac{1}{3 - k_1 - q_1 z_1^1 - q_2 z_2^2} + \frac{1}{4 \sqrt{k_1}} \left(\frac{1}{5 + \sqrt{k_1} + z_1^1} \right) + \frac{1}{4 \sqrt{k_1}} \left(\frac{1}{1 + \sqrt{k_1} - z_1^1} \right) = 0 \quad (17.8)$$

$$z_1^1: \frac{1}{5 + \sqrt{k_1} + z_1^1} = \frac{1}{1 + \sqrt{k_1} - z_1^1} \quad (17.9)$$

Solving for k_1 , z_1^1 , yields from Eq. (17.8)

$$\begin{aligned} 1 + \sqrt{k_1} - z_1^1 &= 5 + \sqrt{k_1} + z_1^1 \\ -4 &= 2z_1^1 \\ -2 &= z_1^1 \end{aligned}$$

Substituting this value into Eq. (17.8) gives

$$\begin{aligned} \frac{1}{3 - k_1} &= \frac{1}{4 \sqrt{k_1}} \left\{ \frac{1}{5 + \sqrt{k_1} - 2} + \frac{1}{1 + \sqrt{k_1} + 2} \right\} \\ \frac{1}{3 - k_1} &= \frac{1}{4 \sqrt{k_1}} \left\{ \frac{1}{3 + \sqrt{k_1}} \right\} \\ 4\sqrt{k_1} \left\{ 3 + \sqrt{k_1} \right\} &= 2(3 - k_1), \text{ or simplifying} \\ -6 + 12\sqrt{k_1} + 6k_1 &= 0 \\ -1 + 2\sqrt{k_1} + k_1 &= 0 \end{aligned}$$

Let $X = \sqrt{k_1}$

$$X = \frac{-2 \pm \sqrt{4 - 4(1)(-1)}}{2} = \frac{-2 \pm \sqrt{8}}{2}$$

$$X = -1 + \sqrt{2} = -1 + 1.4$$

$$X = 0.4$$

$k_1 = 0.16$ and total savings = $k_1 + k_2 = 2k_1 = 0.32$

Financial Equilibrium with Differential Information

Chapter Outline

18.1 Introduction	527
18.2 On the Possibility of an Upward-Sloping Demand Curve	529
18.3 An Illustration of the Concept of REE: Homogeneous Information	530
18.4 Fully Revealing REE: An Example	535
18.5 The Efficient Market Hypothesis	539
References	542
Appendix: Bayesian Updating with the Normal Distribution	542

18.1 Introduction

The fact that investors often disagree about expected future returns or the evaluation of the risks associated with specific investments is probably the source of the majority of financial trading volume. Yet we have said very little so far about the possibility of such disagreements and, more generally, about differences in investors' information. In fact, two of the equilibrium models we have reviewed have explicitly assumed that investors have identical information sets. In the case of the capital asset pricing model (CAPM), it is assumed that all investors' expectations are summarized by the same vector of expected returns and the same variance–covariance matrix. It is this assumption that gives relevance to the single efficient frontier. Similarly, the assumption of a single representative decision maker in the consumption CAPM (CCAPM) is akin to assuming the existence of a large number of investors endowed with identical preferences and information sets.¹ The rational expectations hypothesis, which is part of the CCAPM, necessarily implies that, at equilibrium, all investors share the same objective views about future returns.

Both the Arbitrage Pricing Theory (APT) and the Martingale pricing models are nonstructural models which, by construction, are agnostic about the background information

¹ Box 10.1 discussed the extent to which this interpretation can be relaxed as far as utility functions are concerned.

(or preferences) of the investors. In a sense these theories go beyond the homogeneous information assumption, but without being explicit as to the specific implications of such an extension. The Arrow–Debreu model is a structural model equipped to deal, at least implicitly, with heterogeneously informed agents. In particular, it can accommodate general utility representations defined on state-contingent commodities where, in effect, the assumed state probabilities are embedded in the specific form taken by the individual's utility function.² Thus, while agents must agree on the relevant states of the world, they could disagree on their probabilities. We did not exploit this degree of generality, however, and typically made our arguments on the basis of time-additive and state-additive utility functions with explicit, investor-homogeneous, state probabilities.

In this chapter, we relax the assumption that all agents in the economy have the same subjective probabilities about states of nature or the same expectations about returns, or that they know the objective probability distributions. In so doing we open a huge and fascinating, yet incomplete, chapter in financial economics, part of which was selectively reviewed in Chapter 2. We will again be very selective in the topics we choose to address under this heading and will concentrate on the issue of market equilibrium with differentially informed traders. This is in keeping with the spirit of this book and enables us to revisit the last important pillar of traditional financial theory left untouched thus far: the efficient market hypothesis.

The import of differential information for understanding financial markets, institutions, and contracts, however, goes much beyond market efficiency. Since [Akerlof \(1970\)](#), asymmetric information—a situation where agents are differentially informed with, moreover, one subgroup having *superior* information—is known potentially to lead to the failure of a market to exist. This *lemons* problem is a relevant one in financial markets: one may be reluctant to purchase a stock from a better-informed intermediary, or, a fortiori, from the primary issuer of a security who may be presumed to have the best information about the exact value of the underlying assets. One may suspect that the issuer would be unwilling to sell at a price lower than the fundamental value of the asset. What is called the *winner's curse* is applicable here: if the transaction is concluded, i.e., if the better-informed owner has agreed to sell, is it not likely that the buyer will have paid too much for the asset? This reasoning might go some way toward explaining the fact that capital raised by firms in equity markets is such a small proportion of total firm financing (on this issue, see [Greenwald and Stiglitz, 1993](#)).

Asymmetric information may also explain the phenomenon of credit rationing. The idea here is that it may not be to the advantage of a lender, confronted with a demand for funds larger than he can accommodate, to increase the interest rate he charges as would be

² Such preference structures are, strictly speaking, not expected utility.

required to balance supply and demand: in doing so, the lender may alter the pool of applicants in an unfavorable way. Specifically, this possibility depends on the plausible hypothesis that the lender does not know the degree of riskiness of the projects for which borrowers need funds and that, in the context of a debt contract, a higher hurdle rate may eliminate the less profitable, but consequently, also the less risky, projects. It is easy to construct cases where the creditor is worse off lending his funds at a higher rate because at the high rate the pool of borrowers becomes riskier ([Stiglitz and Weiss, 1981](#)).

Asymmetric information has also been used to explain the prevalence of debt contracts relative to contingent claims. We have used the argument before (Chapter 9): states of nature are often costly to ascertain and verify for one of the parties in a contract. As a result, when two parties enter into a contract, it may be more efficient to stipulate noncontingent payments most of the time, thus economizing on verification costs. Only states leading to bankruptcy or default are recognized as resulting in different rights and obligations for the parties involved ([Townsend, 1979](#)).

These are only a few of the important issues that can be addressed with the asymmetric information assumption. A full review would deserve a whole book in itself. One reason for the need to be selective is that there is a lack of a unifying framework in this literature. It has often proceeded with a set of specific examples rather than more encompassing models. We refer interested readers to [Hirshleifer and Riley \(1992\)](#) for a broader review of this fascinating and important topic in financial economics.

18.2 On the Possibility of an Upward-Sloping Demand Curve

There are plenty of reasons to believe that differences in information and beliefs constitute an important motivation for trading in financial markets. It is extremely difficult to rationalize observed trading volumes in a world of homogeneously informed agents. The main reason for having neglected what is without doubt an obvious fact is that our equilibrium concept, borrowed from traditional supply and demand analysis (the standard notion of Walrasian equilibrium), must be thoroughly updated once we allow for heterogeneous information.

The intuition is as follows. The Walrasian equilibrium price is necessarily some function of the orders placed by traders. Suppose that traders are heterogeneously informed and that their private information set is a relevant determinant of their orders. The equilibrium price will, therefore, reflect and, in that sense, transmit at least a fraction of the privately held information. In this case, the equilibrium price is not only a signal of relative scarcity, as in a Walrasian world; it also reflects the agents' information. In this context, the price quoted for a commodity or a security may be high because the demand for it is objectively high and/or the supply is low. But it may also be high because a group of investors has private

information suggestive that the commodity or security in question will be expensive tomorrow. Of course, this information about the future value of the item is of interest to all. Presumably, except for liquidity reasons, no one will want to sell something today at a low price that will likely be of much higher value tomorrow. This means that when the price quoted on the market is high (in the fiction of standard microeconomics, when the *Walrasian auctioneer* announces a high price), a number of market participants will realize that they have sent in their orders on the basis of information that is probably not shared by the *rest of the market*. Depending on the confidence they place in their own information, they may then want to revise their orders, and to do so in a paradoxical way: because the announced price is higher than they thought it would be, they may want to buy more! Fundamentally, this means that what was thought to be the equilibrium price is not, in fact, an equilibrium.

This is a new situation, and it requires a departure from the Walrasian equilibrium concept. In this chapter, we will develop these ideas with the help of an example. We first illustrate the notion of a rational expectations equilibrium (REE), a concept we have used more informally in preceding chapters (e.g., Chapter 10), in a context where all participants share the same information. We then extend it to encompass situations where agents are heterogeneously informed. We provide an example of a fully revealing REE, which may be deemed to be the formal representation of the notion of an *informationally efficient market*. We conclude by discussing some weaknesses of this equilibrium concept and possible extensions.

18.3 An Illustration of the Concept of REE: Homogeneous Information

Let us consider the joint equilibrium of a spot market for a given commodity and its associated futures market. The context is the familiar now and then, two-date economy. The single commodity is traded at date 1.³ Viewed from date 0, the date at which producers must make their production decisions, the demand for this commodity, emanating from final users, is stochastic. It can be represented by a linear demand curve shocked by a random term as in

$$D(p, \tilde{\eta}) = a - cp + \tilde{\eta}$$

where $D(\cdot)$ represents the quantity demanded, p is the (spot) price for the commodity in question, a and c are positive constants, and $\tilde{\eta}$ is a stochastic demand-shifting element.⁴ This latter quantity is centered at (has mean value) zero, at which point the demand curve assumes its average position, and it is normally distributed with variance $\sigma_{\tilde{\eta}}^2$. In other words, $h(\tilde{\eta}) = N(0; \sigma_{\tilde{\eta}}^2)$ where $h(\cdot)$ is the probability density function on $\tilde{\eta}$. See Figure 18.1

³ The rest of this chapter closely follows [Danthine \(1978\)](#).

⁴ Looking forward, the demand for heating oil next winter is stochastic because the severity of the winter is impossible to predict in advance.

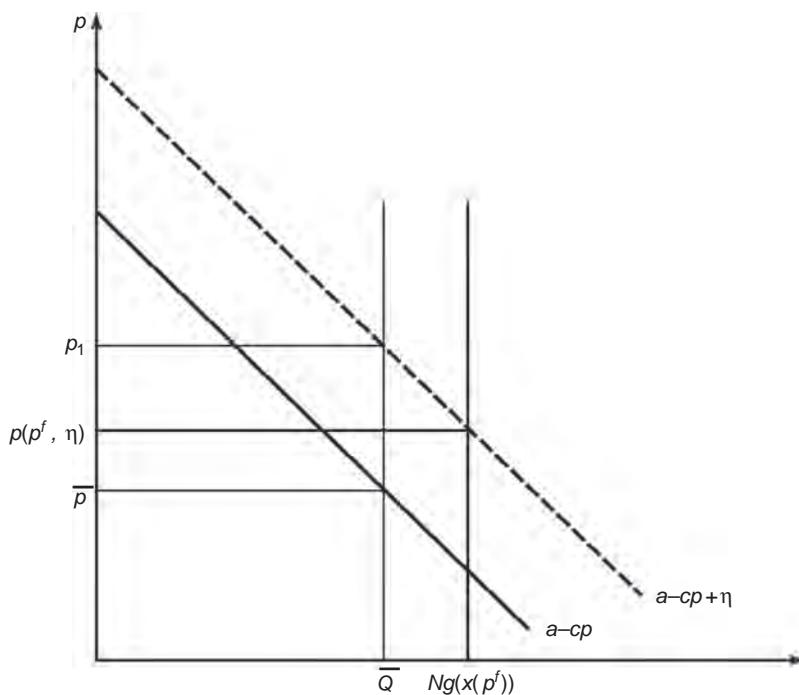


Figure 18.1
Equilibrium with a stochastic demand curve.

for an illustration. At date 0, the N producers decide on their input level x —the input price is normalized at 1—knowing that $g(x)$ units of output will then be available after a one-period production lag at date 1. The production process is thus nonstochastic, and the only uncertainty originates from the demand side. Because of the latter feature, the future sale price \tilde{p} is unknown at the time of the input decision.

We shall assume the existence of a futures or forward market⁵ that our producers may use for hedging or speculative purposes. Specifically, let $f > 0 (< 0)$ be the short (long) futures position taken by the representative producer, i.e., the quantity of output sold (bought) for future delivery at the future (or forward) price p^f .

⁵ The term *futures market* is normally reserved for a market for future delivery taking place in the context of an organized exchange. A *forward market* refers to private exchanges of similar contracts calling for the future delivery of a commodity or financial instrument. While knowledge of the creditworthiness and honesty of the counterparty is of essence in the case of forward contracts, a futures market is anonymous. The exchange is the relevant counterparty for the two sides in a contract. It protects itself and ensures that both parties' engagements will be fulfilled by demanding initial guarantee deposits as well as issuing daily *margin calls* to the party against whose position the price has moved. In a two-date setting, thus in the absence of interim price changes, the notion of margin calls is not relevant, and it is not possible to distinguish futures from forwards.

Here we shall assume that the good traded in the futures market (i.e., specified as acceptable for delivery in the futures contract) is the same as the commodity exchanged on the spot market. For this reason, arbitrageurs will ensure that, at date 1, the futures and the spot price will be exactly identical. In the language of futures markets, the *basis* is constantly equal to zero, and there is thus no *basis risk*.

Under these conditions, the typical producer's cash flow \tilde{y} is

$$\tilde{y} = \tilde{p}g(x) - x + (p^f - \tilde{p})f$$

which can also be written as

$$\tilde{y} = \tilde{p}(g(x) - f) - x + p^f f$$

It is seen that by setting $f = g(x)$, i.e., by selling forward the totality of his production, the producer can eliminate all his risks. Although this need not be his optimal futures position, the feasibility of shedding all risks explains the separation result that follows (much in the spirit of the CAPM: diversifiable risk is not priced).

Let us assume that producers maximize the expected utility of their future cash flow where $U'(\cdot) > 0$ and $U''(\cdot) < 0$:

$$\max_{\substack{x, f \\ x \geq 0}} EU(\tilde{y})$$

Differentiating with respect to x and f successively, and assuming an interior solution, we obtain the following two first-order conditions (FOCs):

$$x: E[U_1(\tilde{y})\tilde{p}] = \frac{1}{g_1(x)}EU_1(\tilde{y}) \quad (18.1)$$

$$f: E[U_1(\tilde{y})\tilde{p}] = p^f EU_1(\tilde{y}) \quad (18.2)$$

which together imply

$$p^f = \frac{1}{g_1(x)} \quad (18.3)$$

Equation (18.3) is remarkable because it says that the optimal input level should be such that the marginal cost of production is set equal to the (known) futures price p^f , the latter replacing the expected spot price as the appropriate production signal. The futures-price-equals-marginal-cost condition is also worth noticing because it implies that, despite the uncertain context in which they operate, producers should not factor in a risk premium when computing their optimal production decision.

For us, a key implication of this result is that, since the supply level will directly depend on the futures price quoted at date 0, the equilibrium spot price at date 1 will be a function of the futures price realized one period earlier. Indeed, writing $x = x(p^f)$ and $g(x) = g(x(p^f))$ to highlight the implications of Eq. (18.3) for the input and output levels, the supply–equals–demand condition for the date 1 spot market reads

$$Ng(x(p^f)) = a - cp + \tilde{\eta}$$

which implicitly defines the equilibrium (date 1) spot price as a function of the date 0 value taken by the futures price, or

$$\tilde{p} = p(p^f, \tilde{\eta}) \quad (18.4)$$

It is clear from Eq. (18.4) that the structure of our problem is such that the probability distribution on \tilde{p} cannot be spelled out independently of the value taken by p^f .

Consequently, it would not be meaningful to assume expectations for \tilde{p} , on the part of producers or futures market speculators, which would not take account of this fundamental link between the two prices. This observation, which is a first step toward the definition of a rational expectation equilibrium, can be further developed by focusing now on the futures market itself.

Let us assume that, in addition to the N producers, n speculators take positions on the futures market. We define speculators by their exclusive involvement in the futures markets; in particular they have no position in the underlying commodity. Accordingly, their cash flows are simply

$$\tilde{z}_i = (p^f - \tilde{p})b_i$$

where b_i is the futures position ($b_i > 0$ represents a short position; $b_i < 0$ represents a long position) taken by speculator i . Suppose for simplicity that their preferences are represented by a linear mean–variance utility function of their cash flows:

$$W(\tilde{z}_i) = E(\tilde{z}_i) - \frac{\chi}{2} \text{var}(\tilde{z}_i)$$

where χ represents the (Arrow–Pratt) absolute risk-aversion index of the representative speculator. We shall similarly specialize the utility function of producers. The assumption of a linear mean–variance utility representation is, in fact, equivalent to hypothesizing an exponential (constant absolute risk aversion, CARA) utility function such as

$$W(\tilde{z}) = -e^{-x/2\tilde{z}}$$

if the context is such that the argument of the function, z , is normally distributed. This hypothesis will be verified at the equilibrium of our model.

Under these hypotheses, it is easy to verify that the optimal futures position of speculator i is

$$b_i = \frac{p^f - E(\tilde{p}|p^f)}{\chi \operatorname{var}(\tilde{p}|p^f)} \quad (18.5)$$

where the conditioning in the expectation and variance operators is made necessary by Eq. (18.4). The form of Eq. (18.5) is not surprising. It implies that the optimal futures position selected by a speculator will have the same sign as the expected difference between the futures price and the expected spot price, i.e., a speculator will be short ($b > 0$) if and only if the futures price at which he sells is larger than the spot price at which he expects to be able to unload his position tomorrow. As to the size of his position, it will be proportional to the expected difference between the two prices, which is indicative of the size of the expected return, and inversely related to the perceived riskiness of the speculation, measured by the product of the variance of the spot price with the Arrow–Pratt coefficient of risk aversion. More risk-averse speculators will assume smaller positions, everything else being the same.

Under a linear mean–variance specification of preferences, the producer's objective function becomes

$$\max_{\substack{x, f \\ x \geq 0}} E(\tilde{p}|p^f)(g(x) - f - x) + p^f f - \frac{\xi}{2}(g(x) - f)^2 \operatorname{var}(\tilde{p}|p^f)$$

where ξ is the absolute risk-aversion measure for producers.

With this specification of the objective function, Eq. (18.2), the FOC with respect to f , becomes

$$f = g(x(p^f)) + \frac{p^f - E(\tilde{p}|p^f)}{\xi \operatorname{var}(\tilde{p}|p^f)} \equiv f(p^f) \quad (18.6)$$

which is the second part of the separation result alluded to previously. The optimal futures position of the representative producer consists in selling forward the totality of his production ($g(x)$) and then readjusting by a component that is simply the futures position taken by a speculator with the same degree of risk aversion. To see this, compare the last term in Eq. (18.6) with Eq. (18.5). A producer's actual futures position can be viewed as the sum of these two terms. He may under-hedge, i.e., sell less than his future output at the futures price. This is so if he anticipates paying an insurance premium in the form of a sale price (p^f) lower than the spot price he expects to prevail tomorrow. But he could as well over-hedge and sell forward more than his total future output. That is, if he considers the current futures price to be a high enough price, he may be willing to speculate on it, selling high at the futures price what he hopes to buy low tomorrow on the spot market.

Putting together speculators' and producers' positions, we find that the futures market-clearing condition becomes

$$\sum_{i=1}^n b_i + Nf = 0 \text{ or} \\ n \left\{ \frac{p^f - E(\tilde{p}|p^f)}{\chi \text{ var}(\tilde{p}|p^f)} \right\} + N \left\{ \frac{p^f - E(\tilde{p}|p^f)}{\xi \text{ var}(\tilde{p}|p^f)} \right\} + Ng(x(p^f)) = 0 \quad (18.7)$$

which must be solved for the equilibrium futures price p^f . Equation (18.7) makes clear that the equilibrium futures price p^f is dependent on the expectations held on the future spot price \tilde{p} ; we have previously emphasized the dependence on p^f of expectations about \tilde{p} . This apparently circular reasoning can be resolved under the rational expectations hypothesis, which consists of assuming that individuals have learned to understand the relationship summarized in Eq. (18.4), i.e.,

$$E(\tilde{p}|p^f) = E[p(p^f, \tilde{\eta})|p^f], \text{ var}(\tilde{p}|p^f) = \text{var}[p(p^f, \tilde{\eta})|p^f] \quad (18.8)$$

Definition 18.1 In the context of this section, an REE is defined as

1. a futures price p^f solving Eq. (18.7) given Eq. (18.8), and the distributional assumption made on η , and
2. a spot price p solving Eq. (18.4) given p^f and the realization of η .

The first part of the definition indicates that the futures price equilibrates the futures market at date 0 when agents rationally anticipate the effective condition under which the spot market will clear tomorrow and make use of the objective probability distribution on the stochastic parameter $\tilde{\eta}$. Given the supply of the commodity available tomorrow (itself a function of the equilibrium futures price quoted today), and given the particular value taken by $\tilde{\eta}$ (i.e., the final position of the demand curve), the second part specifies that the spot price clears the date 1 spot market.

18.4 Fully Revealing REE: An Example

Let us pursue this example one step further and assume that speculators have access to privileged information in the following sense: Before the futures exchange opens, speculator i , ($i = 1, \dots, n$), observes some unbiased approximation v_i to the future realization of the variable $\tilde{\eta}$. The signal v_i can be viewed as the future η itself plus an error of observation ω_i . The latter is specific to speculator i , but all speculators are similarly imprecise in the information they manage to gather. Thus,

$$v_i = \eta + \omega_i \text{ where the } \tilde{\omega}_i \text{'s are i.i.d. } N(0; \sigma_\omega^2)$$

across agents and across time periods.

This relationship can be interpreted as follows: η is a summary measure of the mood of consumers or of other conditions affecting demand. Speculators can obtain advanced information as to the particular value of this realization for the relevant period through, for instance, a survey of consumer's intentions or a detailed weather forecast (assuming the latter influences demand). These observations are not without errors, but (regarding these two periods as only one occasion of a multiperiod process where learning has been taking place), speculators are assumed to be sufficiently skilled to avoid systematic biases in their evaluations. In the present model, this advance information is freely available to them.

Under these conditions, Eq. (18.5) becomes

$$b_i = \frac{p^f - E(\tilde{p}|p^f; v_i)}{\chi \operatorname{var}(\tilde{p}|p^f; v_i)} \equiv b(p^f; v_i)$$

where we make explicit the fact that both the expected value and the variance of the spot price are affected by the advance piece of information obtained by speculator i . The appendix details how these expectations can actually be computed, but this needs not occupy us for the moment.

Formally, Eq. (18.6) is unchanged, so that the futures market-clearing condition can be written

$$Nf(p^f) + \sum_{i=1}^n b(p^f; v_i) = 0$$

It is clear from this equation that the equilibrium futures price will be affected by the “elements” of information gathered by speculators. In fact, under appropriate regularity conditions, the market-clearing equation implicitly defines a function

$$p^f = l(v_1, v_2, \dots, v_n) \quad (18.9)$$

that formalizes this link and thus the information content of the equilibrium futures price.

All this implies that there is more than meets the eye in the conditioning on p^f of $E(\tilde{p}|p^f)$ and $\operatorname{var}(\tilde{p}|p^f)$. So far the reasoning for this conditioning was given by Eq. (18.4): a higher p^f stimulates supply from $g(x(p^f))$ and thus affects the equilibrium spot price. Now a higher p^f also indicates high v_i s on average, thus transmitting information about the future realization of $\hat{\eta}$. The real implications of this link can be understood by reference to Figure 18.1. In the absence of advance information, supply will be geared to the average demand conditions. \bar{Q} represents this average supply level, leading to a spot price \bar{p} under conditions of average demand ($\eta = 0$). If suppliers receive no advance warning of an abnormally high demand level, an above-average realization $\hat{\eta}$ requires a high price p_1 to balance supply and demand. If, on the other hand, speculators' advance information is

transmitted to producers via the futures price, supply increases in anticipation of the high demand level and the price increase is mitigated.

We are now in a position to provide a precise answer to the question that has occupied us since [Section 18.2](#): How much information is transmitted by the equilibrium price p^f ? It will not be a fully general one. Our model has the nature of an example because it presumes specific functional forms. The result we will obtain certainly stands at one extreme on the spectrum of possible answers; it can be considered as a useful benchmark. In what follows, we will construct, under the additional simplification $g(x) = \alpha x^{\frac{1}{2}}$, a consistent equilibrium in which the futures price is itself a summary of *all the information* there is to obtain, a summary that, in an operational sense, is fully equivalent to the complete list of signals obtained by all speculators. More precisely, we will show that the equilibrium futures price is an invertible (linear) function of $\sum v_j$ and that, indeed, it clears the futures market given that everyone realizes this property and bases his orders on the information he can thus extract. This result is important because $\sum v_j$ is a sufficient statistic for the entire vector (v_1, v_2, \dots, v_n) . While we will precisely define the notion of a “sufficient statistic” shortly, here it simply suggests that the sum contains as much relevant information for the problem at hand as the entire vector, in the sense that knowing the sum leads to placing the same market orders as knowing the whole vector. Ours is thus a context where the answer to our question is: *All* the relevant information is aggregated in the equilibrium price and is revealed freely to market participants. The REE is thus *fully revealing*!

Let us proceed and make these assertions precise. Under the assumed technology, $g(x) = \alpha x^{\frac{1}{2}}$, [Eqs. \(18.3\), \(18.4\), and \(18.8\)](#) become, respectively,

$$g(x(p^f)) = \frac{\alpha^2}{2} p^f$$

$$p(p^f, \tilde{\eta}) = A - Bp^f + \frac{1}{c}\tilde{\eta}$$

$$\text{with } A = \frac{a}{c}, B = \frac{N\alpha^2}{c} \frac{2}{2}$$

$$E(\tilde{p}|p^f) = A - Bp^f + \frac{1}{c}E(\tilde{\eta}|p^f)$$

$$\text{var}(\tilde{p}|p^f) = \frac{1}{c^2} \text{var}(\tilde{\eta}|p^f)$$

The informational structure is as follows. Considering the market as a whole, an experiment has been performed consisting of observing the values taken by n independent drawings of some random variable \bar{v} , where $\bar{v} = \eta + \tilde{w}$ and \tilde{w} is $N(0, \sigma_w^2)$. The results are summarized in

the vector $v = (v_1, v_2, \dots, v_n)$ or, as we shall demonstrate, in the sum of the v_j 's, $\sum v_j$, which is a *sufficient statistic* for $v = (v_1, v_2, \dots, v_n)$. The latter expression means that conditioning expectations on $\sum v_j$ or on $\sum v_j$ and the whole vector of v yields the same posterior distribution for $\tilde{\eta}$. In other words, the entire vector does not contain any information that is not already present in the sum. Formally, we have [Definition 18.2](#).

Definition 18.2 $\sum v_j$ is a sufficient statistic for $v = (v_1, v_2, \dots, v_n)$ relative to the distribution $h(\eta)$ if and only if $h(\tilde{\eta} | \sum v_j, v) = h(\tilde{\eta} | \sum v_j)$.

Being a function of the observations (see [Eq. \(18.9\)](#)), p^f is itself a statistic used by traders in calibrating their probabilities. The question is: How good a statistic can be? How well can the futures price summarize the information available to the market? As promised, we now display an equilibrium where the price p^f is a sufficient statistic for the information available to the market, i.e., it is invertible for the sufficient statistic $\sum v_j$. In that case, knowledge of p^f is equivalent to the knowledge of $\sum v_j$, and farmers' and speculators' expectations coincide. If the futures price has this revealing property, expectations held at equilibrium by all agents must be (see the appendix to this chapter for details)

$$E(\tilde{\eta}|p^f) = E(\tilde{\eta}|v_j, p^f) = E(\tilde{\eta} | \sum v_j) = \frac{\sigma_\eta^2}{n\sigma_\eta^2 + \sigma_w^2} \sum v_j \quad (18.10)$$

$$\text{var}(\tilde{\eta}|p^f) = \text{var}(\tilde{\eta}|v_j, p^f) = \frac{\sigma_w^2 \sigma_\eta^2}{n\sigma_\eta^2 + \sigma_w^2} \quad (18.11)$$

[Equations \(18.10\) and \(18.11\)](#) make clear that conditioning on the futures price would, under our hypothesis, be equivalent to conditioning on $\sum v_j$, the latter being, of course, superior information relative to the single piece of individual information, v_i , initially obtained by speculator i . Using these expressions for the expectations in [Eq. \(18.7\)](#), one can show after a few tedious manipulations that, as announced, the market-clearing futures price has the form

$$p^f = F + L \sum v_j \quad (18.12)$$

where

$$F = \frac{(N_\chi + n\xi)A}{(N_\chi + n\xi)(B + 1) + N\alpha^2\xi\chi \frac{1}{c^2} \frac{\sigma_w^2 \sigma_\eta^2}{n\sigma_\eta^2 + \sigma_w^2}} \quad \text{and}$$

$$L = \frac{1}{c} \frac{\sigma_w^2 \sigma_\eta^2}{n\sigma_\eta^2 + \sigma_w^2} \frac{F}{A}$$

Equation (18.12) shows the equilibrium price p^f to be proportional to $\sum v_j$ and thus a sufficient statistic as postulated. It satisfies our definition of an equilibrium. It is a market-clearing price, the result of speculators' and farmers' maximizing behavior, and it corresponds to an equilibrium state of expectations. That is, when Eq. (18.12) is the hypothesized functional relationship between p^f and v , this relationship is indeed realized given that each agent then appropriately extracts the information $\sum v_j$ from the announcement of the equilibrium price.

18.5 The Efficient Market Hypothesis

The result obtained in Section 18.4 is without doubt extreme. It is interesting, however, as it stands as the paragon of the concept of market efficiency. Here is a formal and precise context in which the valuable pieces of information held by heterogeneously informed market participants are aggregated and freely transmitted to all via the trading process. This outcome is reminiscent of the statements made earlier in the century by the famous liberal economist F. von Hayek who celebrated the virtues of the market as an information aggregator (von Hayek, 1945). It must also correspond to what Fama (1970) intended when introducing the concept of strong form efficiency, defined as a situation where market prices fully reflect all publicly *and* privately held information.

The reader may recall that Fama (1970) also introduced the notions of *weak-form efficiency*, covering situations where market prices fully and instantaneously reflect the information included in historical prices, and of *semi-strong form efficiency* where prices, in addition, reflect all publicly available information (of whatever nature). A securities market equilibrium such as the one described in Chapter 10 under the heading of the CCAPM probably best captures what one can understand as semi-strong efficiency: agents are rational in the sense of being expected utility maximizers, they are homogeneously informed (so that all information is indeed publicly held), and they efficiently use all the relevant information when defining their asset holdings. In the CCAPM, no agent can systematically “beat the market”, a largely accepted hallmark of an efficient market equilibrium, provided “beating the market” is appropriately defined in terms of both risk and return.

The concept of a Martingale, first used in Chapters 12 and 13, has long constituted another hallmark of market efficiency. It is useful here to provide a formal definition.

Definition 18.3 A stochastic process \tilde{x}_t is a Martingale with respect to an information set Φ_t if

$$E(\tilde{x}_{t+1} | \Phi_t) = x_t \quad (18.13)$$

It is a short step from this notion of a Martingale to the assertion that one cannot beat the market, which is the case if the current price of a stock is the best predictor of its future price. The latter is likely to be the case if market participants indeed make full use of all available information: In that situation, future price changes can only be unpredictable. An equation like Eq. (18.13) cannot be true exactly for stock prices as stock returns would then be zero on average. It is clear that what could be a Martingale under the previous intuitive reasoning would be a price series normalized to take account of dividends and a normal expected return for holding stock. To get an idea of what this would mean, let us refer to the price equilibrium Eq. (10.2) of the CCAPM

$$U_1(Y_t)q_t^e = \delta E_t\{U_1(Y_{t+1})(q_{t+1}^e + Y_{t+1})\} \quad (18.14)$$

Making the assumption of risk neutrality, one obtains

$$q_t^e = \delta E_t(q_{t+1}^e + Y_{t+1}) \quad (18.15)$$

If we entertain, for a moment, the possibility of a nondividend-paying stock, $Y_t \equiv 0$, then Eq. (18.14) indeed implies that the normalized series $x_t = \delta' p_t$ satisfies Eq. (18.13) and is thus a Martingale. This normalization implies that the expected return on stockholding is constant and equal to the risk-free rate. In the case of a dividend-paying stock, a similar, but slightly more complicated, normalization yields the same result.

The main points of this discussion are (1) that a pure Martingale process requires adjusting the stock price series to take account of dividends and the existence of a positive normal return and (2) that the Martingale property is a mark of market efficiency only under a strong hypothesis of risk neutrality that includes, as a corollary, the property that expected return to stockholding is constant. The large empirical literature on market efficiency has not always been able to take account appropriately of these qualifications. See LeRoy (1989) for an in-depth survey of this issue.

Our model of the previous section is more ambitious, addressing as it does, the concept of strong form efficiency. Its merit is to underline what it takes for this extreme concept to be descriptive of reality, thus also helping to delineate its limits. Two of these limits deserve mentioning. The first one arises once one attempts, plausibly, to dispense with the hypothesis that speculators are able to obtain their elements of privileged information costlessly. If information is free, it is difficult to see why all speculators would not get all the relevant information, thus reverting to a model of homogeneous information. However, the spirit of our example is that resources are needed to collect information and that speculators are those market participants specializing in this costly search process. Yet why should speculator i expand resources to obtain private information v_i when the equilibrium price will freely reveal to him the sufficient statistic $\sum v_j$, which by itself is more

informative than the information he could gather at a cost? The very fact that the equilibrium REE price is fully revealing implies that individual speculators have no use for their own piece of information, with the obvious corollary that they will not be prepared to spend a penny to obtain it. On the other hand, if speculators are not endowed with privileged information, there is no way the equilibrium price will be the celebrated information aggregator and transmitter. In turn, if the equilibrium price is not informative, it may well pay for speculators to obtain valuable private information. We are thus trapped in a vicious circle that results in the nonexistence of equilibrium, an outcome [Grossman and Stiglitz \(1980\)](#) have logically dubbed “the impossibility of informationally efficient markets.”

Another limitation of the conceptual setup of [Section 18.4](#) resides in the fact that the hypotheses required for the equilibrium price to be fully revealing are numerous and particularly severe. The rational expectations hypothesis includes, as always, the assumption that market participants understand the environment in which they operate. This segment of the hypothesis is particularly demanding in the context of our model, and it is crucial for agents to be able to extract sufficient statistics from the equilibrium futures price. By that we mean that, for individual agents to be in position to read all the information concealed in the equilibrium price, they need to know exactly the number of uninformed and informed agents and their respective degrees of risk aversion, which must be identical within each agent class. The information held by the various speculators must have identical precision (i.e., an error term with the same variance), and none of the market participants can be motivated by liquidity considerations. All in all, these requirements are simply too strong to be plausibly met in real-life situations. Although the real-life complications may be partly compensated for by the fact that trading is done on a repeated, almost continuous basis, it is more reasonable to assume that the fully revealing equilibrium is the exception rather than the rule. See the recent paper by Vives (2014) on this score.

The more normal situation is certainly one where some, but not all, information is aggregated and transmitted by the equilibrium price. In such an equilibrium, the incentives to collect information remain, although if the price is too good a transmitter, they may be significantly reduced. The nonexistence-of-equilibrium problem uncovered by Grossman and Stiglitz is then more a curiosum than a real source of worry. Equilibria with partial transmission of information have been described in the literature under the heading *noisy rational expectation equilibrium*. The apparatus is quite a bit messier than the one in the reference case discussed in [Section 18.4](#), and we will not explore it further (see [Hellwig, 1980](#) for a first step in this direction). Suffice it to say that this class of models serves as the basis for the branch of financial economics known as *market microstructure* which strives to explain the specific forms and rules underlying asset trading in a competitive market environment. The reader is referred to [O’Hara \(1997\)](#) for a broad coverage of these topics.

References

- Akerlof, G., 1970. The market for lemons: qualitative uncertainty and the market mechanism. *Q. J. Econ.* 89, 488–500.
- Danthine, J.-P., 1978. Information, futures prices and stabilizing speculation. *J. Econ. Theory.* 17, 79–98.
- Fama, E., 1970. Efficient capital markets: a review of theory and empirical work. *J. Finan.* 25, 383–417.
- Greenwald, B., Stiglitz, J.E., 1993. Financial market imperfections and business cycles. *Q. J. Econ.* 108, 77–115.
- Grossman, S., Stiglitz, J.E., 1980. On the impossibility of informationally efficient markets. *Am. Econ. Rev.* 70 (3), 393–408.
- von Hayek, F.H., 1945. The use of knowledge in society. *Am. Econ. Rev.* 35, 519–530.
- Hellwig, M.F., 1980. On the aggregation of information in competitive markets. *J. Econ. Theory.* 26, 279–312.
- Hirshleifer, J., Riley, J.G., 1992. *The Analytics of Uncertainty and Information*. Cambridge University Press, Cambridge.
- LeRoy, S.F., 1989. Efficient capital markets and martingales. *J. Econ. Lit.* 27, 1583–1621.
- O'Hara, M., 1997. *Market Microstructure Theory*. Basil Blackwell, Malden, MA.
- Stiglitz, J.E., Weiss, A., 1981. Credit rationing in markets with imperfect information. *Am. Econ. Rev.* 71, 393–410.
- Townsend, R., 1979. Optimal contracts and competitive markets with costly state verification. *J. Econ. Theory.* 21, 417–425.
- Vives, X., 2014. On the possibility of informationally efficient markets. IESE Business School working paper.

Appendix: Bayesian Updating with the Normal Distribution

Theorem A18.1 If we assume \tilde{x} and \tilde{y} are two normally distributed vectors with

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}, V\right)$$

with matrix of variances and covariances

$$V = \begin{pmatrix} V_{xx} & V_{xy} \\ V_{xy} & V_{yy} \end{pmatrix}$$

then the distribution of \tilde{x} conditional on the observation $\tilde{y} = y^0$ is normal with mean $\bar{x} + V_{xx}V_{yy}^{-1}(y^0 - \bar{y})$ and covariance matrix $V_{xx} - V_{xy}V_{yy}^{-1}V_{xy}$.

Applications

Let $\tilde{v}_i = \tilde{\eta} + \tilde{\omega}_i$.

If $\begin{pmatrix} \tilde{\eta} \\ \tilde{\omega}_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\eta^2 & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 + \sigma_\omega^2 \end{pmatrix}\right)$, then

$$E(\tilde{\eta}|v_i) = 0 + \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\omega^2} v_i$$

$$V(\tilde{\eta}|v_i) = \sigma_\eta^2 - \frac{\sigma_\eta^4}{\sigma_\eta^2 + \sigma_\omega^2} = \frac{\sigma_\eta^2 \sigma_\omega^2}{\sigma_\eta^2 + \sigma_\omega^2}$$

If $\begin{pmatrix} \tilde{\eta} \\ \sum \tilde{v}_i \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\eta^2 & \sigma_\eta^2 \\ n\sigma_\eta^2 & n^2\sigma_\eta^2 + n\sigma_\omega^2 \end{pmatrix} \right)$, then

$$E(\tilde{\eta} | \sum v_i) = 0 + \frac{n\sigma_\eta^2}{n^2\sigma_\eta^2 + n\sigma_\omega^2} (\sum v_i) = \frac{\sigma_\eta^2}{n\sigma_\eta^2 + \sigma_\omega^2} (\sum v_i)$$

$$\text{var}(\tilde{\eta} | \sum v_i) = \sigma_\eta^2 - n\sigma_\eta^2 \frac{1}{n^2\sigma_\eta^2 + n\sigma_\omega^2} n\sigma_\eta^2 = \frac{\sigma_\eta^2 \sigma_\omega^2}{n\sigma_\eta^2 + \sigma_\omega^2}$$

Index

Note: Page numbers followed by “*b*,” “*f*,” and “*t*” refer to boxes, figures, and tables, respectively.

A

- Absolute prudence, 127, 129
- Absolute risk aversion index, 533
- A–D securities.
 - See* Arrow–Debreu securities
- Agency costs and capital structure, 48–49
- Allais paradox, 72–74, 135
- APT. *See* Arbitrage pricing theory (APT)
- Arbitrage bounds on value of claim, 371
- Arbitrage perspective of Arrow–Debreu pricing, 323
 - abstract setting, 337–342
 - deriving term structure, 330–335
 - extracting from option prices example, 352*b*
 - first approximation, 343–344
 - forward prices and forward rates, 358–359
 - law of one price, 327
 - market completeness and complex securities, 326–330
 - in multiperiod setting, 352–357
 - recovering from option prices, 345–351
 - state claim, pricing, 352*t*
 - state-contingent claim prices in risk-free world, constructing, 330–335

- synthesizing state-contingent claims, 343–344
- using options to complete the market, 337–342
- Value Additivity Theorem, 335–337
 - valuing uncertain cash-flow stream, 330
- Arbitrage pricing theory (APT), 34–37, 247
 - advantage of, 436–437
 - Arrow–Debreu model, 418–419
 - capital budgeting, 441–442
 - and CAPM, 421–424
 - factor-mimicking portfolios, 428–436
 - factor models, 419–421
 - financial structure and firm valuation and, 510, 514–516, 523
 - formal statement, 424–426
 - graphical interpretation of, 438–441
 - investor preferences and, 527–528
 - macroeconomic factor models, 426–428
 - market model using, 420–421
 - momentum portfolios, 434–436
 - multifactor models, 421–424
 - second illustration, 421–424
 - size and value factors of Fama and French, 428–433
 - for stock or portfolio selection, 436–437

- Arbitrage *versus* equilibrium, 35–37
- Arrow–Debreu equilibrium pricing theory, 34–35, 247
 - arbitrage perspective.
 - See* Arbitrage perspective of Arrow–Debreu pricing
- CAPM *versus*, 247
- competitive equilibrium and Pareto optimality
 - illustrated, 250–257
- constructing representative agent and, 274
- economy, 248–250, 384–386
- implementing Pareto optimal allocations, 260–263
- interior *versus* corner solutions, 253*b*
- market failure possibility, 260–263
- Pareto optimality and risk sharing, 257–259
- risk-neutral measures and, 370, 372
- securities. *See* Arrow–Debreu securities
- Arrow–Debreu pricing approach, 422–423
- Arrow–Debreu securities, 16, 249, 256, 282–283, 325–326
 - allocations when two agents trade among themselves, 511*t*
- arbitrage approach and, 36–37
- completeness and, 327, 387–388

- Arrow–Debreu securities
(Continued)
 complex security and, 327
 definition, 249
 design of, 325–326
 financial structure and firm valuation, 509–510, 515
 price construction in risk-free world, 330–335
 prices, 252–254
 risk-free discount bonds as, 330
 in risk-free setting, 330
 synthesizing, 343–344
 trading, 519–520, 525–526
 using options to complete the market and, 337–342
- Value Additivity Theorem and, 336
- Arrow–Debreu state-contingent claims with CCAPM pricing, 281–286
- Arrow–Pratt measure, 119, 533–534
- Asset allocation, strategic, 491–492
- Asset correlations, 190–191
- Asset definition, 32
- Asset pricing challenges, 31
 arbitrage *versus* equilibrium, 35–37
 banks, 49–51
 capital structure, 44–45
 capital structure and agency costs, 48–49
 discounting risky cash flows, 33–35
 models and stylized facts, 39–44
 pecking order theory of investment financing, 49
 question of financial theory, main, 31–33
 risk premia, decomposing, 37–38
 taxes and capital structure, 46–48
 valuing risk-free cash flow, 32
- Asset returns, properties of U.S., 287^t
- Asymmetric information, 7, 529
- B**
- Background risk, 492–500
 “Bank runs” 50
 Banks, 49–51
 Bansal and Yaron economy, 307
 Bayesian updating with normal distribution, 542–543
 Behavioral finance, 75–85
 framing, 76–78
 overconfidence, 84–85
 Prospect Theory, 78–84
- Benchmark random walk model, 198
- Binomial model, 453–454
 of derivates valuation, 397–407
- Black–Scholes formula, 457–459
- Black–Scholes option pricing formula, 346–348
 obtaining call prices, 344, 352
 pricing securities with, 348^b
- Bond volatility puzzle, 44
- Bonds and their role in investor portfolios, 485
- (Book value of equity)/(market value of equity) (BE/ME) ratio, 42
- Brownian motion, 444–448
- Business cycle and financial intermediation, 18–19
- Butterfly spread, 345–346
- C**
- Call-related securities and speculation, 406
- Capital asset pricing model (CAPM), 34–35, 207, 420, 490–492
 and arbitrage pricing theory (APT), 421–424
 Arrow–Debreu model *versus*, 247
 background for deriving the zero-beta of, 224–227
 characterizing efficient portfolios, 222–224
 consumer-based, fundamental equations of, 277–278
 consumption testing, 293–295
- empirical assessment of, 231–239
 Banz (1981) and the “Size Effect” 234
 Fama and French, 234–235
 Fama–MacBeth two-step regression procedure, 232–234, 245
 volatility anomalies, 235–239
- investors’ expectations assumption, 527
- mathematics of portfolio frontier, 217–222, 242–244
- Modern Portfolio Theory and, 209–210
- portfolio management and, 469
 pricing securities in isolation and, 516
- risky asset as, 116–117
- Security Market Line, 213, 214^f, 227
- Sharpe–Lintner–Mossin, 228
- standard, 229–230
- systematic risk and, 345
- traditional approach, 210–213
- valuing risky cash flows with, 214–217
- zero-beta, 224–228
- Capital budgeting, 441–442
- Capital market line (CML), 211–212, 211^f
- Capital structure, 45–46
 agency costs and, 48–49
 taxes and, 46–48
- CAPM. *See* Capital asset pricing model (CAPM)
- CARA. *See* Constant absolute risk aversion (CARA)
- Cash flow(s)
 discounting risky, 33–35
 valuing risk-free, 32^t
 valuing risky cash flows with CAPM, 214–217
- Cass–Stiglitz Theorem, 144
- CCAPM. *See* Consumption capital asset pricing model
- CEQ return. *See* Certainty equivalent (CEQ) return

- Certainty equivalent (CEQ) return, 193–194
- Choice theory under certainty introduction, 61–66
- Closed-form pricing, 444
- CML. *See* Capital market line (CML)
- Commitment strategy, 136
- Compensating precautionary premium, definition, 129
- Competitive equilibrium, 27–29, 250–257
- Complete contingent claims, 521–522
- Completeness definition, 327
- Complex security definition, 327
- Constant absolute risk aversion (CARA), 120, 124, 533
- Constant relative risk aversion (CRRA), 146–147, 287–288
- Consumption capital asset pricing model, 269
- constructing a representative agent, 272^b
 - discrete time infinite horizon economies in, 388–390
 - efficient market hypothesis and, 539
 - empirical validity of, 286–293
 - with Epstein–Zin utility, 303–313
 - equity premium puzzle, 286–293, 303–304
 - exchange (endowment) economy. *See* Exchange (endowment) economy
 - habit formation, 302–303
 - Hansen–Jagannathan bounds, 293–295
 - infinitely lived representative agent, 270–271
 - Lucas fruit tree economy, 276, 304
 - no-trade equilibrium concept, 271–275
 - pricing Arrow–Debreu state-contingent claims with, 281–286
- properties of lognormal distribution, some, 320–322
- rational expectations hypothesis, 527
- representative agent hypothesis and its notice of equilibrium, 270–275
- risk-neutral pricing in, 390–397
- and risk-neutral valuation, 285–286
- semi-strong form efficiency and, 539
- solving with growth, 319–320 testing, 293–295
- Consumption CAPM in continuous time, 466–467
- Consumption preference for smooth, 4–5, 4^b
- Consumption–savings decision, 145
- Consumption–savings problem, 460–461
- Contingent claims.
- See* Arrow–Debreu securities
- Continuous time finance, intuitive overview of, 443
- applications, 460–467
 - behavior of stochastic differentials, 454–456
 - binomial model, 453–454
 - Black–Scholes formula, 457–459
 - Brownian motion, 444–448
 - consumption CAPM in Continuous Time, 466–467
 - consumption–savings problem, 460–461
 - general continuous time processes, 448–449
 - Ito processes, 451–452
 - Ito's lemma, 456–457
 - Martingale methods, 459–460
 - to portfolio analysis, 461–466
 - random walks, 444–448
 - simulation and call pricing, 451–454
 - stochastic differential equations, solving, 454–459
- stock price behavior, continuous time model of, 449–451
- Convertible bonds, 49
- Corporate finance and asset pricing, 44–45
- Credit rationing, 528–529
- Cross-border capital flows, 187
- CRRA. *See* Constant relative risk aversion (CRRA)
- ## D
- Date-contingent claim prices, 334, 334^t
- Date-state, 327
- Debt/equity (D/E) ratio, 47, 50
- Demand curve, possibility of upward-sloping, 529–530
- Derivatives valuation binomial model, 397–407
- Digression to discrete time, 462–464
- Disaster state, 299–301
- Discount bond(s) *versus* date claim prices, 334^t
- as Arrow–Debreu claims, 334^t
 - replicating cash flow, 334^t
- Diversification process, 6–7
- Dominance, 56
- mean–variance, 57
 - second-order stochastic (SSD), 101–102
 - state-by-state, 56–58, 98
 - stochastic, 98–102, 104
- Dybvig's evaluation of dynamic trading strategies, 410–414
- ## E
- Economic growth and financial system, 8–12
- Economic rationality, 61–62
- Economic relationships, changing, 187
- Edgeworth–Bowley box, 18–22, 24, 27–28, 28^f, 29^f
- Efficient frontier, 152–158, 165–166, 171–179, 189, 210
- of risky assets, 183

Epstein–Zin utility
 CCAPM with, 303–313
 Bansal and Yaron model,
 306–308
 beyond representative agent
 and rational expectations,
 313–317
 Dufresne et al model,
 308–313
 long-run behavior of stock
 returns and, 487
 variations in risk free rate and,
 478–485

Equally weighted portfolios,
 193–194

Equilibrium and no arbitrage
 opportunities, 375–377

Equilibrium pricing kernel,
 293

Equilibrium *versus* arbitrage,
 35–37

Equity premium, 40–42

Equity premium puzzle, 41–42,
 286–293, 303–304,
 316–317

Equity trading strategies,
 frequently cited, 413–414

Exchange (endowment) economy,
 275–281

- Arrow–Debreu exchange economies and, 275
- equilibrium price function, calculating, 290^b
- formal consumption CAPM, 281
- interpreting exchange equilibrium, 278–281
- Lucas fruit tree economy and, 276
- model, 275–278
- rational expectations economy and, 276
- Executive compensation, application to, 111–112
- Expected utility construct, 70
- theorem, 66–74
- unsettling observation about, 105–106

F

Factor models, of APT, 419–421

Factor-mimicking portfolios, of APT, 428–436

25 Fama–French value-weighted portfolios, 429–430

Financial “repression” 12

Financial accelerator theory,
 18–19

Financial crisis, 19–22, 191–192
 return standard deviations, 192

Financial equilibrium with differential information, 527

- Bayesian updating with normal distribution, 542–543
- homogeneous information, 530–535
- lemons problem, 528
- market hypothesis, efficient, 539–541
- REE concept illustration, 530–535
- REE example, fully revealing, 535–539
- upward-sloping demand curve possibility, 529–530

Financial intermediation and business cycle, 18–19

Financial markets and institutions, role of, 1

- desynchronization, 6–7
- economic growth and, 8–12
- financial crises, 19–22
- intermediation and business cycle, 18–19
- risk dimension, 6–7
- screening and monitoring functions, 7–8
- social welfare and, 12–18
- time dimension, 3–6

Financial structure and firm valuation in incomplete markets, 507

- Arrow–Debreu and Modigliani–Miller, 514–516
- complete contingent claims, 521–522

contingent claims trading,
 519–520, 525–526

example, 508–514

growth and financing, 518–523

incomplete markets, 521–522

no contingent claims market, 519

short selling role, 516–518

Financial system

- definition, 3–4
- function, primary, 3–4
- time dimension, 3–6

Financial theory, main question of, 31–33

Forward market definition, 531

Forward prices and forward rates, 358–359

Framing, 76–78

Futures market definition, 531

G

GBM. *See* Geometric Brownian motion (GBM)

GDP. *See* Gross domestic product (GDP)

General competitive equilibrium, 14

General continuous time processes, 448–449

General equilibrium theory

- competitive equilibrium, 27–29
- conditions, 13–14
- Pareto optimal allocations, 25–27
- results of, 14

Geometric Brownian motion (GBM), 450

GNP. *See* Gross national product (GNP)

Graphical interpretation of APT, 438–441

Great Depression, 19

“Great Moderation” 182–183

Great Recession, 19–20, 182–183

Gross domestic product (GDP), 8–10, 20–22

Gross national product (GNP), 8–10

H

- Habit formation, 302–303
 Hansen–Jagannathan bounds, 293–295
 HML factors, 428–430
 Home bias puzzle, 157–158

I

- IARA. *See* Increasing absolute risk aversion (IARA)
 “Immediate gratification” 135–136
 Implied volatility estimate, 409–410
 Incomplete markets. *See* Financial structure and firm valuation in incomplete markets
 Increasing absolute risk aversion (IARA), 148
 International stock market cross-correlations, 188–189
 Intertemporal hedging demand, 490
 Intertemporal stock return behavior through time, 197–200
 Investing close to home, 64b
 Ito processes, 451–452
 Ito’s lemma, 456–457

J

- Jarque–Bera statistic, 151
 Jensen’s Inequality, 94
 Joint equity premium and risk-free rate puzzles, resolving, 298
 Joint saving–portfolio problem, 129–130

K

- Kurtosis, 150–151

L

- Law of one price, 327
 Lehman Brothers bankruptcy filing, 191–192
 Leverage on an asset’s risk and return, 106–112
 Levered equity, 107
 Long investment horizons, riskiness of stocks for, 195–203

- long- and short-run equity riskiness, 195–197
 predictive perspective, 201–203
 random walk model, 197–200
 Long-run risks model, 307–308
 Lottery stocks, 237
 Lucas, R.E., Jr., 40
 Lucas fruit tree economy, 276, 304, 396

M

- Macroeconomic factor models, of APT, 426–428
 Marginal rate of substitution competitive equilibrium and Pareto optimality and, 252–253, 255
 consumption capital asset pricing model and, 297–298
 exchange (endowment) economy model and, 278
 intertemporal, 278, 325–326
 Pareto optimality and risk sharing and, 258–259
 pricing Arrow–Debreu state-contingent claims with CCAPM and, 288
 solving CCAPM with growth and, 319
 Market completeness and complex securities, 326–330
 Market failure possibility, 260–263
 Market microstructure, 541
 Martingale measure in discrete time, 387
 binomial model of derivatives valuation, 397–407
 Black–Scholes formula, 407–410
 CCAPM setting, 388–390
 continuous time, 407–410
 Dybvig’s evaluation of dynamic trading strategies, 410–414
 infinite horizon economies, 388–390
 risk-neutral pricing in CCAPM, 390–397

- risk-neutral valuation
 discounting at term
 structure of multiperiod discount bond, 414–415
 Martingale methods, 459–460
 Martingale pricing theory (measure), 36–37, 361
 application, 377–383
 definition, 539–540
 in discrete time. *See* Martingale measure in discrete time
 efficient market hypothesis and, 540

- equilibrium and no arbitrage opportunities, 375–377
 incompleteness, 372–375
 investor preferences and, 527–528
 maximizing expected utility of terminal wealth, 377–383
 notation, definitions, and basic results, 364–369
 numeral example, 381–383
 portfolio investment and risk-neutral probabilities, 377–379
 setting and intuition, 362–363
 solving portfolio problem, 380–381
 uniqueness, 369–372

- Mean preserving spread, 102–104
 Means and variances in practice, computing, 59b

- Mean–variance-based portfolio selection rules, 193

- Microsoft Excel to construct portfolio efficient frontier, using, 158, 173*t*, 174*f*, 175*f*, 176*f*, 178*f*

- Minimum variance frontier, 153–155

- MM theorem.
See Modigliani–Miller (MM) theorem

- Models and stylized facts, 39–44
 equity premium, 40–42
 term structure, 43–44
 value premium, 42–43

- Modern portfolio theory (MPT), 140, 143, 181, 187, 205
 capital asset pricing model and, 209–210
 constructing efficient frontier, 171–179
 consumption–savings decision, 145
 gains from diversification and efficient frontier, 152–158
 indifference curves under quadratic utility or normally distributed returns, 166–170
 individual decision problem sequentially, 145–146
 long run portfolio management and, 504
 mean–variance dominance and, 57
 normality-of-returns assumption, refining, 149–152
 selection problem, 469
 separation theorem, 158–159
 shape of efficient frontier, 171–172
 stochastic dominance and diversification, 159–165
 utility functions, 144–149
M
 Modigliani–Miller (MM) theorem, 37, 45
 essence of, 515
 failure in incomplete markets of, 523
 financial structure and firm valuation in incomplete markets and, 508, 514–516
Momentum portfolios, of APT, 434–436
 stock or portfolio selection, APT for, 436–437
Moral hazard, 7
MPT. *See* Modern portfolio theory (MPT)
Multifactor models, of APT, 421–424
Multiperiod setting.
 Arrow–Debreu pricing in, 352–357
- N**
 Naïve diversification, 194
 Net present value (NPV), 48
 Normality-of-returns assumption, refining, 149–152
 No-trade equilibrium concept, 271–275
 NPV. *See* Net present value (NPV)
- O**
 Operating leverage, 315
 Opportunity description set in mean-variance space, 152–158
 Out of sample portfolio returns, 193–194
 Overconfidence, 84–85
- P**
 Pareto optimum (Pareto-efficient), 14
 achieving, 14
 allocations, 25–27
 and competitive equilibrium illustrated, 250–257
 constructing representative agent and, 272–273
 definition, 14
 financial structure and firm valuation and, 511, 513
 implementing allocations, 260–263
 market failure possibility and, 260–263
 risk sharing and, 257–259
 Pecking order theory of investment financing, 49
 “Peso problem” 299–300
 Portfolio analysis, 461–466
 Portfolio management in long run, 469
 advice to investors, 469–470
 background risk, 492–500
 behavior of stock returns, 486–492
 budget constraint, 479–481
 considerations, 470–471
 diversification, 469–470
- holding period returns table, 486f
 implications of labor income for portfolio choice, 492–500
 individual investor saving for retirement problem, 473–474, 477
 multiperiod investor-saver problem, 471–472
 myopic solution, 472–478
 nature of risk free asset, 484–485
 optimal allocations, 482–484
 optimality equation, 481–482
 real estate as background risk, 501–504
 role of bonds in investor portfolios, 485
 role of stocks in investor portfolios, 492
 solving in mean reversion environment, 489–491
 strategic asset allocation, 491–492
 variations in risk free rate, 478–485
 Portfolio turnover, measure of, 193–194
Portfolio(s)
 allocation and risk aversion, 116–118
 of basic underlying assets, 341
 composition, risk aversion, and wealth, 118–121
 composition, risky, 122–124
 efficient, characterizing, 222–224
 equally weighted, 193–194
 frontier, 217–223, 242–244
 investment and risk-neutral possibilities, 377–379
 market, 345
 optimal, 158–159, 210
 payoff diagrams, 346f, 347f
 states of nature and, 345
 stocks role in, 492
 zero-covariance, 224–227
 Preference reversal phenomenon, 73
 Pricing kernel, 295

- Prospect Theory, 78–84
preference orderings with connections to, 83–84
- Prudence
absolute, 127, 129
illustrating, 128–129
relative, 127
- Q**
- Quadratic utility, 166–170
- Quasi-hyperbolic discounting, 135–137
- R**
- Random walk model, 197–200
- Random walks, 444–448
- Rational economic models, 41–42
- Rational expectations economy, 276
- Rational Expectations Equilibrium (REE), 530, 541
concept illustration, 530–535
definition, 276
example, fully revealing, 535–541
noisy, 541
- Rational Expectations hypothesis, 527
- Rationality of collective decision making, 74b
- Real estate risk, 501–504
- REE. *See* Rational Expectations Equilibrium
- Relative Prudence, 127
- Representative agent, constructing, 272b
- Return distributions, 149
- Return standard deviations, during financial crisis, 192
- Return in continuous time, 464–466
- Risk
aversion to, 6
market, 212–213
spreading, 11
systematic, 212–213
undiversifiable, 212–213
- Risk and time preferences, separating, 137–139
- Risk aversion, measuring risk and, 87, 93
absolute risk aversion, 90–92
assessing degree, 97–98
definition, 87–88
expected utility, unsettling observation about, 105–106
interpreting, 90–93
investment behavior affected by, 89
leverage and risk, applications, 106–112
mean preserving spreads, 102–104
premium and certainty equivalence, risk, 94–97
relative risk aversion, 90, 92–93, 97–98
representing, 13b
stochastic dominance concept, 98–102
- Risk aversion and investment decisions, 115, 181.
See also Modern Portfolio Theory (MPT)
- canonical portfolio problem, 116–117, 144
- consequences of parameter uncertainty, 183–187
- constant absolute risk aversion (CARA), 120, 124, 275, 533
- declining absolute risk aversion (DARA), 120
- equally weighted portfolios, 193–194
- “Great Moderation” 182–183
- “Great Recession” 182–183
- increasing absolute risk aversion (IARA), 121, 148
- joint saving–portfolio problem, 129–130
- long investment horizons, riskiness of stocks for long- and short-run equity riskiness, 195–197
- predictive perspective, 201–203
- random walk model, 197–200
- modern portfolio theory and. *See* Modern portfolio theory
- portfolio composition, risky, 122–124
- prudence, illustrating, 128–129
- risk-free *versus* risky assets, 116–118
- risk-neutral investors, 121–122
- savings and riskiness of returns, 124–128
- savings behavior, 124–130
- Sharpe ratio, 193–194
- stock market return data, trends and cycles in, 187–192
- asset correlations in cyclical periods of high volatility, 190–191
- financial crisis, 191–192
- international stock market cross-correlations, 188–189
- VNM-expected utility representation, generalizing, 130–139
- Risk free asset, nature of, 484–485
- Risk free rate variations, 478–485
budget constraint, 479–481
nature of, 484–485
optimality equation, 481–482
optimal portfolio allocations, 482–484
role of bonds in investor portfolios, 485
- Risk neutral investors, 93
- Risk Neutral Valuation Model, 247, 405, 409, 414–415
- Risk premia, decomposing, 37–38
- Risk premium and certainty equivalence, 94–97
- Risk sharing and Pareto optimality, 257–259
- Risk shifting, 6–7
- Risk tolerance, 90
- Risk-averse investor, 105
- Risk-free rate puzzle, 304
- Risk-neutral investors, 121–122

- Risk-neutral pricing in CCAPM, 390–397
- Risk-neutral probability, 370, 372
definition, 377
fundamental securities in economy relationship to, 377
portfolio investment and, 377–379
- Risk-neutral valuations, 263–266, 285–286
- Risky assets, efficient frontier of, 183
- Risky situations, making choices in, 53
Allais paradox, 72–74
behavioral finance, 75–85
choice theory under certainty, 61–66
expected utility theorem, 66–72
preliminaries, 56–60
prerequisite, 61–63
- S**
- Savings and growth in developing countries chart, 8/*f*
- Savings behavior and risk aversion, 124–130
illustrating prudence, 128–129
joint saving–portfolio problem, 129–130
riskiness of returns and, 124–128
- Savings–portfolio composition problem, 147
- Sawtooth pattern, 447
- SDF. *See* Stochastic discount factor (SDF)
- Security design, 507–508
- Security market line (SML), 213, 214/*f*, 227, 422
- Semi-strong form efficiency, 539
- Separation theorem, 158–159, 210
- Shareholder unanimity, 508–509
- Sharpe ratio, 58–59, 109, 193–194
- Sharpe–Lintner–Mossin CAPM.
See Zero-beta capital asset pricing model
- Short selling role, 516–518
- Simulation and call pricing, 451–454
binomial model, 453–454
Ito processes, 451–452
- Size and value factors of Fama and French, 428–433
- Skewness, 150–151
- SMB factor, 428–430
- SMI. *See* Swiss Market Index (SMI)
- SML. *See* Security market line (SML)
- Social welfare and financial markets, 12–18
- Speculation and call-related securities, 406
- State-contingent claim prices, constructing, 330–335.
See also Arrow–Debreu securities
- States of nature
definition, 6–7, 15
dominance and, 56
financial equilibrium with differential information and, 528
financial markets and social welfare and, 16–17
as future time periods, 330
market completeness and complex securities and, 326
market portfolio and, 357–358
multiperiod cash flow and, 353, 355–356
using options to complete the market and, 337–338
- Stochastic differential equations, 449
behavior of stochastic differentials, 454–456
- Black–Scholes formula, 457–459
- Ito's lemma, 456–457
solving, 454–459
- Stochastic discount factor (SDF), 293, 295–298
- Stochastic dominance and diversification, 159–165
- Stochastic dominance concept, 98–102
- Stochastic volatility, 306
- Stock market return data, trends and cycles in, 187–192
asset correlations in cyclical periods of high volatility, 190–191
financial crisis, 191–192
international stock market cross-correlations, 188–189
- Stock of habit, 302
- Stock price behavior, continuous time model of, 449–451
- Stock returns, 197–198
- Stock returns long-run behavior, 486–492
role of stocks in investor portfolios, 492
solving for optimal portfolio proportions in mean reversion environment, 489–491
strategic asset allocation, 491–492
- Stylized facts, 39–44
- Swiss Market Index (SMI), 357
- Systematic risk, 212–213
- T**
- Taxes and capital structure, 46–48
- Taylor series approximation, 147
- Taylor's theorem, 91–92, 494
- Term structure, 43–44
of interest rates, 333, 335
- Terminal wealth, maximizing expected utility of, 377–383
numerical example, 381–383
- portfolio investment and risk-neutral probabilities, 377–379
solving portfolio problem, 380–381
- Theory of risk-neutral valuation.
See Martingale pricing theory (measure)
- Time as key element in finance, 3–6
- Time-consistent planning, preferences that guarantee, 133–137

-
- quasi-hyperbolic discounting, 135–137
- Transferring funds from savers to investors, 9–10
- Two-fund theorem, 159
- U**
- Uncertainty resolution, preferences for the timing of, 131–133
- Utility-of-money functions, 93, 146
- V**
- Value Additivity Theorem, 335–337
- Value premium, 42–43
- Valuing risk-free cash flow, 32*t*
- VNM utility function. *See* von Neumann–Morgenstern (VNM) utility function
- von Neumann–Morgenstern (VNM) utility function, 70–71, 390
- preferences for the timing of uncertainty resolution, 131–133
- preferences guaranteeing time-consistent planning, 133–137
- quasi-hyperbolic discounting, 135–137
- risk aversion and, 87–88, 92–93, 112, 115–117, 139–140
- separating risk and time preferences, 137–139
- W**
- Walrasian equilibrium, 529–530
- Weak-form efficiency, 539
- Wiener process, 446
- Winner’s curse, 528
- Z**
- Zero net-investment portfolio, 418–419
- Zero-beta capital asset pricing model, 224–228

List of Frequently Used Symbols and Notation

A text such as *Intermediate Financial Theory* is, by nature, relatively notation intensive. We have adopted a strategy to minimize the notational burden within each individual chapter at the cost of being, at times, inconsistent in our use of symbols across chapters. We list here a set of symbols regularly used with their specific meaning. At times, however, we have found it more practical to use some of the listed symbols to represent a different concept. In other instances, clarity required making the symbolic representation more precise (e.g., by being more specific as to the time dimension of an interest rate).

Roman Alphabet

a	Amount invested in the risky asset; in Chapter 16, fraction of wealth invested in the risky asset or portfolio
A^T	Transpose of the matrix (or vector) A
c	Consumption; in Chapter 16 only, consumption is represented by C , while c represents $\ln C$
c_θ^k	Consumption of agent k in state of nature θ
CE	Certainty equivalent
C_A	Price of an American call option
C_e	Price of a European call option
d	Dividend rate or amount
Δ	Number of shares in the replicating portfolio (Chapter 13)
E	The expectations operator
e_θ^k	Endowment of agent k in state of nature θ
f	Futures position (Chapter 18);
p^f	Price of a futures contract (Chapter 18)
F, G	Cumulative distribution functions associated with densities:
f, g	Probability density functions

K	The strike or exercise price of an option
$K(\tilde{x})$	Kurtosis of the random variable \tilde{x}
L	A lottery
\tilde{m}	Pricing kernel
M	The market portfolio
MU_θ^k	Marginal utility of agent k in state θ
p	Price of an arbitrary asset
\mathbf{P}	Measure of Absolute Prudence
q	Arrow–Debreu price
q^b	Price of risk-free discount bond, occasionally denoted q^{rf}
q^e	Price of equity
r_f	Rate of return on a risk-free asset
R_f	Gross rate of return on a risk-free asset
\tilde{r}	Rate of return on a risky asset
\tilde{R}	Gross rate of return on a risky asset
R_A	Absolute risk aversion coefficient
R_R	Relative risk aversion coefficient
s	Usually denotes the amount saved
S	In the context of discussing options, used to denote the price of the underlying stock
$S(\tilde{x})$	Skewness of the random variable x
\mathbb{T}	Transition matrix
U	Utility of money function, or, simply, the utility function
\mathbb{U}, \mathbb{V}	von Neumann–Morgenstern utility function
V	Usually denotes variance–covariance matrix of asset returns; occasionally is used as another utility function symbol; may also signify value as in
V_P	The value of portfolio P or
V_F	The value of the firm
w_i	Portfolio weight of asset i in a given portfolio
Y_0	Initial wealth

Greek Alphabet

α	Intercept coefficient in the market model (alpha)
β	The slope coefficient in the market model (beta)
δ	Time discount factor
η	Elasticity
λ	Lagrange multiplier

μ	Mean
π_θ	State probability of state θ
π_θ^{RN}	Risk-neutral probability of state θ
Π	Risk premium
$\rho(\tilde{x}, \tilde{y})$	Correlation of random variables x and y
ρ	Elasticity of intertemporal substitution parameter (Chapter 15)
σ	Standard deviation
σ_{ij}	Covariance between random variables i and j
θ	Index for state of nature
Ω	Rate of depreciation of physical capital
ψ	Compensating precautionary premium

Numerals and Other Terms

1	Vector of ones
$>$	Is strictly preferred to
\geq	Is preferred to (non strictly, that is allowing for indifference)
<i>GBM</i>	Geometric Brownian Motion stochastic process
<i>FSD</i>	First-order stochastic dominance
<i>SSD</i>	Second-order stochastic dominance
<i>VNM</i>	von Neuman-Morgenstern (utility function)
<i>CAPM</i>	Capital Asset Pricing Model
<i>MPT</i>	Modern Portfolio Theory