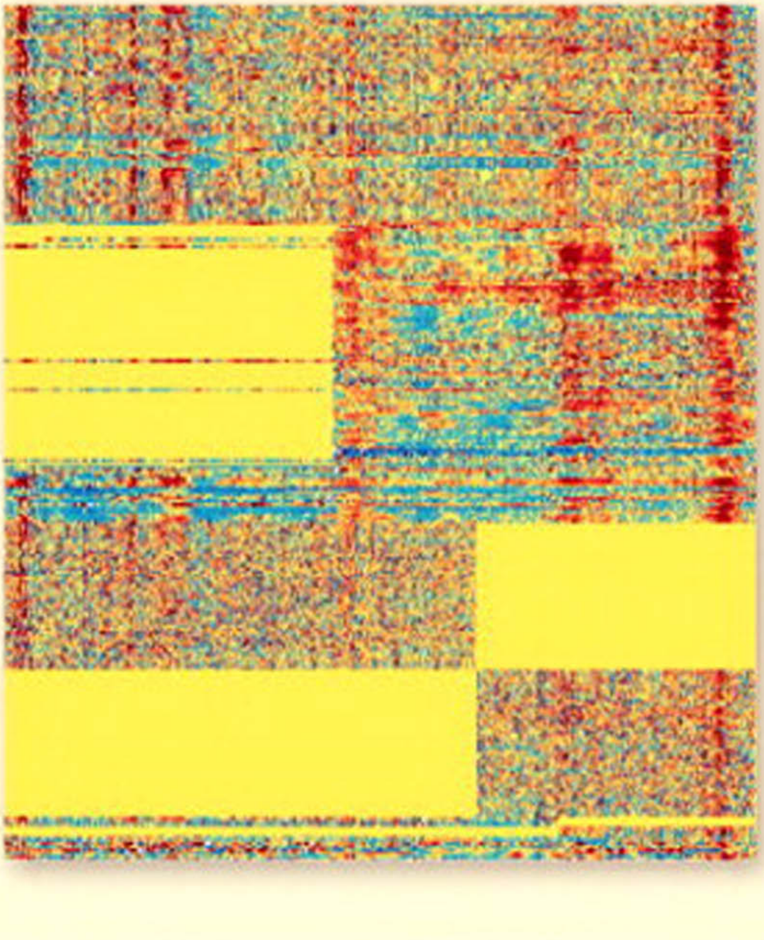
UPDATED EDITION

INTRODUCTION TO ECONOMETRICS

THIRD EDITION



James H. Stock
Mark W. Watson

Introduction to Econometrics

The Pearson Series in Economics

Abel/Bernanke/Croushore
*Macroeconomics**

Bade/Parkin
*Foundations of Economics**

Berck/Helfand
The Economics of the Environment

Bierman/Fernandez
Game Theory with Economic Applications

Blanchard
*Macroeconomics**

Blau/Ferber/Winkler
The Economics of Women, Men, and Work

Boardman/Greenberg/Vining/Weimer
Cost-Benefit Analysis

Boyer
Principles of Transportation Economics

Branson
Macroeconomic Theory and Policy

Bruce
Public Finance and the American Economy

Carlton/Perloff
Modern Industrial Organization

Case/Fair/Oster
*Principles of Economics**

Chapman
Environmental Economics: Theory, Application, and Policy

Cooter/Ulen
Law & Economics

Daniels/VanHoose
International Monetary & Financial Economics

Downs
An Economic Theory of Democracy

Ehrenberg/Smith
Modern Labor Economics

Farnham
Economics for Managers

Folland/Goodman/Stano
The Economics of Health and Health Care

Fort
Sports Economics

Froyen
Macroeconomics

Fusfeld
The Age of the Economist

Gerber
*International Economics**

González-Rivera
Forecasting for Economics and Business

Gordon
*Macroeconomics**

Greene
Econometric Analysis

Gregory
Essentials of Economics

Gregory/Stuart
Russian and Soviet Economic Performance and Structure

Hartwick/Olewiler
The Economics of Natural Resource Use

Heilbroner/Milberg
The Making of the Economic Society

Heyne/Boettke/Prychitko
The Economic Way of Thinking

Holt
Markets, Games, and Strategic Behavior

Hubbard/O'Brien
*Economics**
*Money, Banking, and the Financial System**

Hubbard/O'Brien/Rafferty
*Macroeconomics**

Hughes/Cain
American Economic History

Husted/Melvin
International Economics

Jehle/Reny
Advanced Microeconomic Theory

Johnson-Lans
A Health Economics Primer

Keat/Young/Erfle
Managerial Economics

Klein
Mathematical Methods for Economics

Krugman/Obstfeld/Melitz
*International Economics: Theory & Policy**

Laidler
The Demand for Money

Leeds/von Allmen
The Economics of Sports

Leeds/von Allmen/Schiming
*Economics**

Lynn
Economic Development: Theory and Practice for a Divided World

Miller
*Economics Today**
Understanding Modern Economics

Miller/Benjamin
The Economics of Macro Issues

Miller/Benjamin/North
The Economics of Public Issues

Mills/Hamilton
Urban Economics

Mishkin
*The Economics of Money, Banking, and Financial Markets**
*The Economics of Money, Banking, and Financial Markets, Business School Edition**
*Macroeconomics: Policy and Practice**

Murray
Econometrics: A Modern Introduction

O'Sullivan/Sheffrin/Perez
*Economics: Principles, Applications, and Tools**

Parkin
*Economics**

Perloff
*Microeconomics**
*Microeconomics: Theory and Applications with Calculus**

Perloff/Brander
*Managerial Economics and Strategy**

Phelps
Health Economics

Pindyck/Rubinfeld
*Microeconomics**

Riddell/Shackelford/Stamos/Schneider
Economics: A Tool for Critically Understanding Society

Roberts
The Choice: A Fable of Free Trade and Protection

Rohlf
Introduction to Economic Reasoning

Roland
Development Economics

Scherer
Industry Structure, Strategy, and Public Policy

Schiller
The Economics of Poverty and Discrimination

Sherman
Market Regulation

Stock/Watson
Introduction to Econometrics

Studenmund
Using Econometrics: A Practical Guide

Tietenberg/Lewis
Environmental and Natural Resource Economics
Environmental Economics and Policy

Todaro/Smith
Economic Development

Waldman/Jensen
Industrial Organization: Theory and Practice

Walters/Walters/Appel/Callahan/Centanni/Maex/O'Neill
Econversations: Today's Students Discuss Today's Issues

Weil
Economic Growth

Williamson
Macroeconomics

*denotes MyEconLab titles. Visit www.myeconlab.com to learn more.

Introduction to Econometrics

THIRD EDITION UPDATE

James H. Stock
Harvard University

Mark W. Watson
Princeton University

PEARSON

Boston Columbus Indianapolis New York San Francisco Hoboken
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Vice President, Product Management: Donna Battista
Acquisitions Editor: Christina Masturzo
Editorial Assistant: Christine Mallon
Vice President, Marketing: Maggie Moylan
Director, Strategy and Marketing: Scott Dustan
Manager, Field Marketing: Leigh Ann Sims
Product Marketing Manager: Alison Haskins
Executive Field Marketing Manager: Lori DeShazo
Senior Strategic Marketing Manager: Erin Gardner
Team Lead, Program Management: Ashley Santora
Program Manager: Carolyn Philips
Team Lead, Project Management: Jeff Holcomb
Project Manager: Liz Napolitano
Operations Specialist: Carol Melville
Cover Designer: Jon Boylan
Cover Art: Courtesy of Carolin Pflueger and the authors.
Full-Service Project Management, Design, and Electronic Composition: Cenveo® Publisher Services
Printer/Binder: Edwards Brothers Malloy
Cover Printer: Lehigh-Phoenix Color/Hagerstown
Text Font: 10/14 Times Ten Roman

About the cover: The cover shows a heat chart of 270 monthly variables measuring different aspects of employment, production, income, and sales for the United States, 1974–2010. Each horizontal line depicts a different variable, and the horizontal axis is the date. Strong monthly increases in a variable are blue and sharp monthly declines are red. The simultaneous declines in many of these measures during recessions appear in the figure as vertical red bands.

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on appropriate page within text.

Photo Credits: page 410 left: Henrik Montgomery/Pressens Bild/AP Photo; page 410 right: Paul Sakuma/AP Photo; page 428 left: Courtesy of Allison Harris; page 428 right: Courtesy of Allison Harris; page 669 top left: John McCombe/AP Photo; bottom left: New York University/AFP/Newscom; top right: Denise Applewhite/Princeton University/AP Photo; bottom right: Courtesy of the University of Chicago/AP Photo.

Library of Congress Cataloging-in-Publication Data

Stock, James H.
Introduction to econometrics/James H. Stock, Harvard University, Mark W. Watson, Princeton University.— Third edition update.
pages cm.—(The Pearson series in economics)
Includes bibliographical references and index.
ISBN 978-0-13-348687-2—ISBN 0-13-348687-7
1. Econometrics. I. Watson, Mark W. II. Title.
HB139.S765 2015
330.01'5195—dc23
2014018465

PEARSON

www.pearsonhighered.com

ISBN-10: 0-13-348687-7
ISBN-13: 978-0-13-348687-2

Brief Contents

This page intentionally left blank

Contents

PART THREE Further Topics in Regression Analysis

PART FIVE The Econometric Theory of Regression Analysis

Key Concepts

PART ONE Introduction and Review

PART TWO Fundamentals of Regression Analysis

This page intentionally left blank

General Interest Boxes

This page intentionally left blank

Preface

Econometrics can be a fun course for both teacher and student. The real world of economics, business, and government is a complicated and messy place, full of competing ideas and questions that demand answers. Is it more effective to tackle drunk driving by passing tough laws or by increasing the tax on alcohol? Can you make money in the stock market by buying when prices are historically low, relative to earnings, or should you just sit tight, as the random walk theory of stock prices suggests? Can we improve elementary education by reducing class sizes, or should we simply have our children listen to Mozart for 10 minutes a day? Econometrics helps us sort out sound ideas from crazy ones and find quantitative answers to important quantitative questions. Econometrics opens a window on our complicated world that lets us see the relationships on which people, businesses, and governments base their decisions.

Introduction to Econometrics is designed for a first course in undergraduate econometrics. It is our experience that to make econometrics relevant in an introductory course, interesting applications must motivate the theory and the theory must match the applications. This simple principle represents a significant departure from the older generation of econometrics books, in which theoretical models and assumptions do not match the applications. It is no wonder that some students question the relevance of econometrics after they spend much of their time learning assumptions that they subsequently realize are unrealistic so that they must then learn "solutions" to "problems" that arise when the applications do not match the assumptions. We believe that it is far better to motivate the need for tools with a concrete application and then to provide a few simple assumptions that match the application. Because the theory is immediately relevant to the applications, this approach can make econometrics come alive.

New to the Third Edition

- Updated treatment of standard errors for panel data regression
- Discussion of when and why missing data can present a problem for regression analysis
- The use of regression discontinuity design as a method for analyzing quasi-experiments

- Updated discussion of weak instruments
- Discussion of the use and interpretation of control variables integrated into the core development of regression analysis
- Introduction of the "potential outcomes" framework for experimental data
- Additional general interest boxes
- Additional exercises, both pencil-and-paper and empirical

This third edition builds on the philosophy of the first and second editions that applications should drive the theory, not the other way around.

One substantial change in this edition concerns inference in regression with panel data (Chapter 10). In panel data, the data within an entity typically are correlated over time. For inference to be valid, standard errors must be computed using a method that is robust to this correlation. The chapter on panel data now uses one such method, clustered standard errors, from the outset. Clustered standard errors are the natural extension to panel data of the heteroskedasticity-robust standard errors introduced in the initial treatment of regression analysis in Part II. Recent research has shown that clustered standard errors have a number of desirable properties, which are now discussed in Chapter 10 and in a revised appendix to Chapter 10.

Another substantial set of changes concerns the treatment of experiments and quasi-experiments in Chapter 13. The discussion of differences-in-differences regression has been streamlined and draws directly on the multiple regression principles introduced in Part II. Chapter 13 now discusses regression discontinuity design, which is an intuitive and important framework for the analysis of quasi-experimental data. In addition, Chapter 13 now introduces the potential outcomes framework and relates this increasingly commonplace terminology to concepts that were introduced in Parts I and II.

This edition has a number of other significant changes. One is that it incorporates a precise but accessible treatment of control variables into the initial discussion of multiple regression. Chapter 7 now discusses conditions for control variables being successful in the sense that the coefficient on the variable of interest is unbiased even though the coefficients on the control variables generally are not. Other changes include a new discussion of missing data in Chapter 9, a new optional calculus-based appendix to Chapter 8 on slopes and elasticities of nonlinear regression functions, and an updated discussion in Chapter 12 of what to do if you have weak instruments. This edition also includes new general interest boxes, updated empirical examples, and additional exercises.

The Updated Third Edition

- The time series data used in Chapters 14–16 have been extended through the beginning of 2013 and now include the Great Recession.
- The empirical analysis in Chapter 14 now focuses on forecasting the growth rate of real GDP using the term spread, replacing the Phillips curve forecasts from earlier editions.
- Several new empirical exercises have been added to each chapter. Rather than include all of the empirical exercises in the text, we have moved many of them to the Companion Website, **www.pearsonhighered.com/stock_watson**. This has two main advantages: first, we can offer more and more in-depth exercises, and second, we can add and update exercises between editions. We encourage you to browse the empirical exercises available on the Companion Website.

Features of This Book

Introduction to Econometrics differs from other textbooks in three main ways. First, we integrate real-world questions and data into the development of the theory, and we take seriously the substantive findings of the resulting empirical analysis. Second, our choice of topics reflects modern theory and practice. Third, we provide theory and assumptions that match the applications. Our aim is to teach students to become sophisticated consumers of econometrics and to do so at a level of mathematics appropriate for an introductory course.

Real-World Questions and Data

We organize each methodological topic around an important real-world question that demands a specific numerical answer. For example, we teach single-variable regression, multiple regression, and functional form analysis in the context of estimating the effect of school inputs on school outputs. (Do smaller elementary school class sizes produce higher test scores?) We teach panel data methods in the context of analyzing the effect of drunk driving laws on traffic fatalities. We use possible racial discrimination in the market for home loans as the empirical application for teaching regression with a binary dependent variable (logit and probit). We teach instrumental variable estimation in the context of estimating the demand elasticity for cigarettes. Although these examples involve economic reasoning, all can be understood with only a single introductory course in economics, and many can be understood without any previous economics coursework. Thus the instructor can focus on teaching econometrics, not microeconomics or macroeconomics.

We treat all our empirical applications seriously and in a way that shows students how they can learn from data but at the same time be self-critical and aware of the limitations of empirical analyses. Through each application, we teach students to explore alternative specifications and thereby to assess whether their substantive findings are robust. The questions asked in the empirical applications are important, and we provide serious and, we think, credible answers. We encourage students and instructors to disagree, however, and invite them to reanalyze the data, which are provided on the textbook's Companion Website (**www.pearsonhighered.com/stock_watson**).

Contemporary Choice of Topics

Econometrics has come a long way since the 1980s. The topics we cover reflect the best of contemporary applied econometrics. One can only do so much in an introductory course, so we focus on procedures and tests that are commonly used in practice. For example:

- ***Instrumental variables regression.*** We present instrumental variables regression as a general method for handling correlation between the error term and a regressor, which can arise for many reasons, including omitted variables and simultaneous causality. The two assumptions for a valid instrument—exogeneity and relevance—are given equal billing. We follow that presentation with an extended discussion of where instruments come from and with tests of overidentifying restrictions and diagnostics for weak instruments, and we explain what to do if these diagnostics suggest problems.

- ***Program evaluation.*** An increasing number of econometric studies analyze either randomized controlled experiments or quasi-experiments, also known as natural experiments. We address these topics, often collectively referred to as program evaluation, in Chapter 13. We present this research strategy as an alternative approach to the problems of omitted variables, simultaneous causality, and selection, and we assess both the strengths and the weaknesses of studies using experimental or quasi-experimental data.

- ***Forecasting.*** The chapter on forecasting (Chapter 14) considers univariate (autoregressive) and multivariate forecasts using time series regression, not large simultaneous equation structural models. We focus on simple and reliable tools, such as autoregressions and model selection via an information criterion, that work well in practice. This chapter also features a practically oriented treatment of stochastic trends (unit roots), unit root tests, tests for structural breaks (at known and unknown dates), and pseudo out-of-sample forecasting, all in the context of developing stable and reliable time series forecasting models.

- ***Time series regression.*** We make a clear distinction between two very different applications of time series regression: forecasting and estimation of dynamic causal effects. The chapter on causal inference using time series data (Chapter 15) pays careful attention to when different estimation methods, including generalized least squares, will or will not lead to valid causal inferences and when it is advisable to estimate dynamic regressions using OLS with heteroskedasticity- and autocorrelation-consistent standard errors.

Theory That Matches Applications

Although econometric tools are best motivated by empirical applications, students need to learn enough econometric theory to understand the strengths and limitations of those tools. We provide a modern treatment in which the fit between theory and applications is as tight as possible, while keeping the mathematics at a level that requires only algebra.

Modern empirical applications share some common characteristics: The data sets typically are large (hundreds of observations, often more); regressors are not fixed over repeated samples but rather are collected by random sampling (or some other mechanism that makes them random); the data are not normally distributed; and there is no *a priori* reason to think that the errors are homoskedastic (although often there are reasons to think that they are heteroskedastic).

These observations lead to important differences between the theoretical development in this textbook and other textbooks:

- ***Large-sample approach.*** Because data sets are large, from the outset we use large-sample normal approximations to sampling distributions for hypothesis testing and confidence intervals. In our experience, it takes less time to teach the rudiments of large-sample approximations than to teach the Student t and exact F distributions, degrees-of-freedom corrections, and so forth. This large-sample approach also saves students the frustration of discovering that, because of nonnormal errors, the exact distribution theory they just mastered is irrelevant. Once taught in the context of the sample mean, the large-sample approach to hypothesis testing and confidence intervals carries directly through multiple regression analysis, logit and probit, instrumental variables estimation, and time series methods.

- ***Random sampling.*** Because regressors are rarely fixed in econometric applications, from the outset we treat data on all variables (dependent and independent) as the result of random sampling. This assumption matches our initial applications to cross-sectional data, it extends readily to panel and time series data, and because of our large-sample approach, it poses no additional conceptual or mathematical difficulties.
- ***Heteroskedasticity.*** Applied econometricians routinely use heteroskedasticity-robust standard errors to eliminate worries about whether heteroskedasticity is present or not. In this book, we move beyond treating heteroskedasticity as an exception or a "problem" to be "solved"; instead, we allow for heteroskedasticity from the outset and simply use heteroskedasticity-robust standard errors. We present homoskedasticity as a special case that provides a theoretical motivation for OLS.

Skilled Producers, Sophisticated Consumers

We hope that students using this book will become sophisticated consumers of empirical analysis. To do so, they must learn not only how to use the tools of regression analysis but also how to assess the validity of empirical analyses presented to them.

Our approach to teaching how to assess an empirical study is threefold. First, immediately after introducing the main tools of regression analysis, we devote Chapter 9 to the threats to internal and external validity of an empirical study. This chapter discusses data problems and issues of generalizing findings to other settings. It also examines the main threats to regression analysis, including omitted variables, functional form misspecification, errors-in-variables, selection, and simultaneity—and ways to recognize these threats in practice.

Second, we apply these methods for assessing empirical studies to the empirical analysis of the ongoing examples in the book. We do so by considering alternative specifications and by systematically addressing the various threats to validity of the analyses presented in the book.

Third, to become sophisticated consumers, students need firsthand experience as producers. Active learning beats passive learning, and econometrics is an ideal course for active learning. For this reason, the textbook website features data sets, software, and suggestions for empirical exercises of different scopes.

Approach to Mathematics and Level of Rigor

Our aim is for students to develop a sophisticated understanding of the tools of modern regression analysis, whether the course is taught at a "high" or a "low" level of mathematics. Parts I through IV of the text (which cover the substantive material) are accessible to students with only precalculus mathematics. Parts I through IV have fewer equations and more applications than many introductory econometrics books and far fewer equations than books aimed at mathematical sections of undergraduate courses. But more equations do not imply a more sophisticated treatment. In our experience, a more mathematical treatment does not lead to a deeper understanding for most students.

That said, different students learn differently, and for mathematically well-prepared students, learning can be enhanced by a more explicitly mathematical treatment. Part V therefore contains an introduction to econometric theory that is appropriate for students with a stronger mathematical background. When the mathematical chapters in Part V are used in conjunction with the material in Parts I through IV, this book is suitable for advanced undergraduate or master's level econometrics courses.

Contents and Organization

There are five parts to *Introduction to Econometrics*. This textbook assumes that the student has had a course in probability and statistics, although we review that material in Part I. We cover the core material of regression analysis in Part II. Parts III, IV, and V present additional topics that build on the core treatment in Part II.

Part I

Chapter 1 introduces econometrics and stresses the importance of providing quantitative answers to quantitative questions. It discusses the concept of causality in statistical studies and surveys the different types of data encountered in econometrics. Material from probability and statistics is reviewed in Chapters 2 and 3, respectively; whether these chapters are taught in a given course or are simply provided as a reference depends on the background of the students.

Part II

Chapter 4 introduces regression with a single regressor and ordinary least squares (OLS) estimation, and Chapter 5 discusses hypothesis tests and confidence intervals in the regression model with a single regressor. In Chapter 6, students learn how they can address omitted variable bias using multiple regression, thereby estimating the effect of one independent variable while holding other independent variables constant. Chapter 7 covers hypothesis tests, including *F*-tests, and confidence intervals in multiple regression. In Chapter 8, the linear regression model is extended to models with nonlinear population regression functions, with a focus on regression functions that are linear in the parameters (so that the parameters can be estimated by OLS). In Chapter 9, students step back and learn how to identify the strengths and limitations of regression studies, seeing in the process how to apply the concepts of internal and external validity.

Part III

Part III presents extensions of regression methods. In Chapter 10, students learn how to use panel data to control for unobserved variables that are constant over time. Chapter 11 covers regression with a binary dependent variable. Chapter 12 shows how instrumental variables regression can be used to address a variety of problems that produce correlation between the error term and the regressor, and examines how one might find and evaluate valid instruments. Chapter 13 introduces students to the analysis of data from experiments and quasi-, or natural, experiments, topics often referred to as "program evaluation."

Part IV

Part IV takes up regression with time series data. Chapter 14 focuses on forecasting and introduces various modern tools for analyzing time series regressions, such as unit root tests and tests for stability. Chapter 15 discusses the use of time series data to estimate causal relations. Chapter 16 presents some more advanced tools for time series analysis, including models of conditional heteroskedasticity.

Part V

Part V is an introduction to econometric theory. This part is more than an appendix that fills in mathematical details omitted from the text. Rather, it is a self-contained treatment of the econometric theory of estimation and inference in the linear regression model. Chapter 17 develops the theory of regression analysis for a single regressor; the exposition does not use matrix algebra, although it does demand a higher level of mathematical sophistication than the rest of the text. Chapter 18 presents and studies the multiple regression model, instrumental variables regression, and generalized method of moments estimation of the linear model, all in matrix form.

Prerequisites Within the Book

Because different instructors like to emphasize different material, we wrote this book with diverse teaching preferences in mind. To the maximum extent possible, the chapters in Parts III, IV, and V are "stand-alone" in the sense that they do not require first teaching all the preceding chapters. The specific prerequisites for each chapter are described in Table I. Although we have found that the sequence of topics adopted in the textbook works well in our own courses, the chapters are written in a way that allows instructors to present topics in a different order if they so desire.

Sample Courses

This book accommodates several different course structures.

TABLE I Guide to Prerequisites for Special-Topic Chapters in Parts III, IV, and V

	Prerequisite parts or chapters								
	Part I	Part II		Part III		Part IV			Part V
Chapter	1–3	4–7, 9	8	10.1, 10.2	12.1, 12.2	14.1–14.4	14.5–14.8	15	17
10	X[a]	X[a]	X						
11	X[a]	X[a]	X						
12.1, 12.2	X[a]	X[a]	X						
12.3–12.6	X[a]	X[a]	X	X	X				
13	X[a]	X[a]	X	X	X				
14	X[a]	X[a]	b						
15	X[a]	X[a]	b			X			
16	X[a]	X[a]	b			X	X	X	
17	X	X	X						
18	X	X	X		X				X

This table shows the minimum prerequisites needed to cover the material in a given chapter. For example, estimation of dynamic causal effects with time series data (Chapter 15) first requires Part I (as needed, depending on student preparation, and except as noted in footnote a), Part II (except for Chapter 8; see footnote b), and Sections 14.1 through 14.4.

[a]Chapters 10 through 16 use exclusively large-sample approximations to sampling distributions, so the optional Sections 3.6 (the Student *t* distribution for testing means) and 5.6 (the Student *t* distribution for testing regression coefficients) can be skipped.
[b]Chapters 14 through 16 (the time series chapters) can be taught without first teaching Chapter 8 (nonlinear regression functions) if the instructor pauses to explain the use of logarithmic transformations to approximate percentage changes.

Standard Introductory Econometrics

This course introduces econometrics (Chapter 1) and reviews probability and statistics as needed (Chapters 2 and 3). It then moves on to regression with a single regressor, multiple regression, the basics of functional form analysis, and the evaluation of regression studies (all Part II). The course proceeds to cover regression with panel data (Chapter 10), regression with a limited dependent variable (Chapter 11), and instrumental variables regression (Chapter 12), as time permits. The course concludes with experiments and quasi-experiments in Chapter 13, topics that provide an opportunity to return to the questions of estimating causal effects raised at the beginning of the semester and to recapitulate core regression methods. *Prerequisites: Algebra II and introductory statistics*.

Introductory Econometrics with Time Series and Forecasting Applications

Like a standard introductory course, this course covers all of Part I (as needed) and Part II. Optionally, the course next provides a brief introduction to panel data (Sections 10.1 and 10.2) and takes up instrumental variables regression (Chapter 12, or just Sections 12.1 and 12.2). The course then proceeds to Part IV, covering forecasting (Chapter 14) and estimation of dynamic causal effects (Chapter 15). If time permits, the course can include some advanced topics in time series analysis such as volatility clustering and conditional heteroskedasticity (Section 16.5). *Prerequisites: Algebra II and introductory statistics*.

Applied Time Series Analysis and Forecasting

This book also can be used for a short course on applied time series and forecasting, for which a course on regression analysis is a prerequisite. Some time is spent reviewing the tools of basic regression analysis in Part II, depending on student preparation. The course then moves directly to Part IV and works through forecasting (Chapter 14), estimation of dynamic causal effects (Chapter 15), and advanced topics in time series analysis (Chapter 16), including vector autoregressions and conditional heteroskedasticity. An important component of this course is hands-on forecasting exercises, available to instructors on the book's accompanying website. *Prerequisites: Algebra II and basic introductory econometrics or the equivalent*.

Introduction to Econometric Theory

This book is also suitable for an advanced undergraduate course in which the students have a strong mathematical preparation or for a master's level course in econometrics. The course briefly reviews the theory of statistics and probability as necessary (Part I). The course introduces regression analysis using the nonmathematical, applications-based treatment of Part II. This introduction is followed by the theoretical development in Chapters 17 and 18 (through Section 18.5). The course then takes up regression with a limited dependent variable (Chapter 11) and maximum likelihood estimation (Appendix 11.2). Next, the course optionally turns to instrumental variables regression and generalized method of moments (Chapter 12 and Section 18.7), time series methods (Chapter 14), and the estimation of causal effects using time series data and generalized least squares (Chapter 15 and Section 18.6). *Prerequisites: Calculus and introductory statistics. Chapter 18 assumes previous exposure to matrix algebra.*

Pedagogical Features

This textbook has a variety of pedagogical features aimed at helping students understand, retain, and apply the essential ideas. *Chapter introductions* provide real-world grounding and motivation, as well as brief road maps highlighting the sequence of the discussion. *Key terms* are boldfaced and defined in context throughout each chapter, and *Key Concept boxes* at regular intervals recap the central ideas. *General interest boxes* provide interesting excursions into related topics and highlight real-world studies that use the methods or concepts being discussed in the text. A *Summary* concluding each chapter serves as a helpful framework for reviewing the main points of coverage. The questions in the *Review the Concepts* section check students' understanding of the core content, *Exercises* give more intensive practice working with the concepts and techniques introduced in the chapter, and *Empirical Exercises* allow students to apply what they have learned to answer real-world empirical questions. At the end of the textbook, the *Appendix* provides statistical tables, the *References* section lists sources for further reading, and a *Glossary* conveniently defines many key terms in the book.

Supplements to Accompany the Textbook

The online supplements accompanying the third edition update of *Introduction to Econometrics* include the Instructor's Resource Manual, Test Bank, and PowerPoint® slides with text figures, tables, and Key Concepts. The Instructor's Resource Manual includes solutions to all the end-of-chapter exercises, while the Test Bank, offered in Testgen, provides a rich supply of easily edited test problems and questions of various types to meet specific course needs. These resources are available for download from the Instructor's Resource Center at **www.pearsonhighered.com/stock_watson**.

Companion Website

The Companion Website, found at **www.pearsonhighered.com/stock_watson**, provides a wide range of additional resources for students and faculty. These resources include more and more in depth empirical exercises, data sets for the empirical exercises, replication files for empirical results reported in the text, practice quizzes, answers to end-of-chapter Review the Concepts questions and Exercises, and EViews tutorials.

MyEconLab

The third edition update is accompanied by a robust MyEconLab course. The MyEconLab course includes all the Review the Concepts questions as well as some Exercises and Empirical Exercises. In addition, the enhanced eText available in MyEconLab for the third edition update includes URL links from the Exercises and Empirical Exercises to questions in the MyEconLab course and to the data that accompanies them. To register for MyEconLab and to learn more, log on to **www.myeconlab.com**.

Acknowledgments

A great many people contributed to the first edition of this book. Our biggest debts of gratitude are to our colleagues at Harvard and Princeton who used early drafts of this book in their classrooms. At Harvard's Kennedy School of Government, Suzanne Cooper provided invaluable suggestions and detailed comments on multiple drafts. As a coteacher with one of the authors (Stock), she also helped vet much of the material in this book while it was being developed for a required course for master's students at the Kennedy School. We are also indebted to two other Kennedy School colleagues, Alberto Abadie and Sue Dynarski, for their patient explanations of quasi-experiments and the field of program evaluation and for their detailed comments on early drafts of the text. At Princeton, Eli Tamer taught from an early draft and also provided helpful comments on the penultimate draft of the book.

We also owe much to many of our friends and colleagues in econometrics who spent time talking with us about the substance of this book and who collectively made so many helpful suggestions. Bruce Hansen (University of Wisconsin–Madison) and Bo Honore (Princeton) provided helpful feedback on very early outlines and preliminary versions of the core material in Part II. Joshua Angrist (MIT) and Guido Imbens (University of California, Berkeley) provided thoughtful suggestions about our treatment of materials on program evaluation. Our presentation of the material on time series has benefited from discussions with Yacine Ait-Sahalia (Princeton), Graham Elliott (University of California, San Diego), Andrew Harvey (Cambridge University), and Christopher Sims (Princeton). Finally, many people made helpful suggestions on parts of the manuscript close to their area of expertise: Don Andrews (Yale), John Bound (University of Michigan), Gregory Chow (Princeton), Thomas Downes (Tufts), David Drukker (StataCorp.), Jean Baldwin Grossman (Princeton), Eric Hanushek (Hoover Institution), James Heckman (University of Chicago), Han Hong (Princeton), Caroline Hoxby (Harvard), Alan Krueger (Princeton), Steven Levitt (University of Chicago), Richard Light (Harvard), David Neumark (Michigan State University), Joseph Newhouse (Harvard), Pierre Perron (Boston University), Kenneth Warner (University of Michigan), and Richard Zeckhauser (Harvard).

Many people were very generous in providing us with data. The California test score data were constructed with the assistance of Les Axelrod of the Standards and Assessments Division, California Department of Education. We are grateful to Charlie DePascale, Student Assessment Services, Massachusetts Department of Education, for his help with aspects of the Massachusetts test score data set. Christopher Ruhm (University of North Carolina, Greensboro) graciously provided us with his data set on drunk driving laws and traffic fatalities. The research department at the Federal Reserve Bank of Boston deserves thanks for putting together its data on racial discrimination in mortgage lending; we particularly thank Geoffrey Tootell for providing us with the updated version of the data set we use in Chapter 9 and Lynn Browne for explaining its policy context. We thank Jonathan Gruber (MIT) for sharing his data on cigarette sales, which we analyze in Chapter 12, and Alan Krueger (Princeton) for his help with the Tennessee STAR data that we analyze in Chapter 13.

We thank several people for carefully checking the page proof for errors. Kerry Griffin and Yair Listokin read the entire manuscript, and Andrew Fraker, Ori Heffetz, Amber Henry, Hong Li, Alessandro Tarozzi, and Matt Watson worked through several chapters.

In the first edition, we benefited from the help of an exceptional development editor, Jane Tufts, whose creativity, hard work, and attention to detail improved

the book in many ways, large and small. Pearson provided us with first-rate support, starting with our excellent editor, Sylvia Mallory, and extending through the entire publishing team. Jane and Sylvia patiently taught us a lot about writing, organization, and presentation, and their efforts are evident on every page of this book. We extend our thanks to the superb Pearson team, who worked with us on the second edition: Adrienne D'Ambrosio (senior acquisitions editor), Bridget Page (associate media producer), Charles Spaulding (senior designer), Nancy Fenton (managing editor) and her selection of Nancy Freihofer and Thompson Steele Inc. who handled the entire production process, Heather McNally (supplements coordinator), and Denise Clinton (editor-in-chief). Finally, we had the benefit of Kay Ueno's skilled editing in the second edition. We are also grateful to the excellent third edition Pearson team of Adrienne D'Ambrosio, Nancy Fenton, and Jill Kolongowski, as well as Mary Sanger, the project manager with Nesbitt Graphics. We also wish to thank the Pearson team who worked on the third edition update: Christina Masturzo, Carolyn Philips, Liz Napolitano, and Heidi Allgair, project manager with Cenveo® Publisher Services.

We also received a great deal of help and suggestions from faculty, students, and researchers as we prepared the third edition and its update. The changes made in the third edition incorporate or reflect suggestions, corrections, comments, data, and help provided by a number of researchers and instructors: Donald Andrews (Yale University), Jushan Bai (Columbia), James Cobbe (Florida State University), Susan Dynarski (University of Michigan), Nicole Eichelberger (Texas Tech University), Boyd Fjeldsted (University of Utah), Martina Grunow, Daniel Hamermesh (University of Texas–Austin), Keisuke Hirano (University of Arizona), Bo Honore (Princeton University), Guido Imbens (Harvard University), Manfred Keil (Claremont McKenna College), David Laibson (Harvard University), David Lee (Princeton University), Brigitte Madrian (Harvard University), Jorge Marquez (University of Maryland), Karen Bennett Mathis (Florida Department of Citrus), Alan Mehlenbacher (University of Victoria), Ulrich Müller (Princeton University), Serena Ng (Columbia University), Harry Patrinos (World Bank), Zhuan Pei (Brandeis University), Peter Summers (Texas Tech University), Andrey Vasnov (University of Sydney), and Douglas Young (Montana State University). We also benefited from student input from F. Hoces dela Guardia and Carrie Wilson.

Thoughtful reviews for the third edition were prepared for Addison-Wesley by Steve DeLoach (Elon University), Jeffrey DeSimone (University of Texas at Arlington), Gary V. Engelhardt (Syracuse University), Luca Flabbi (Georgetown University), Steffen Habermalz (Northwestern University), Carolyn J. Heinrich (University of Wisconsin–Madison), Emma M. Iglesias-Vazquez (Michigan State

University), Carlos Lamarche (University of Oklahoma), Vicki A. McCracken (Washington State University), Claudiney M. Pereira (Tulane University), and John T. Warner (Clemson University). We also received very helpful input on draft revisions of Chapters 7 and 10 from John Berdell (DePaul University), Janet Kohlhase (University of Houston), Aprajit Mahajan (Stanford University), Xia Meng (Brandeis University), and Chan Shen (Georgetown University).

Above all, we are indebted to our families for their endurance throughout this project. Writing this book took a long time, and for them, the project must have seemed endless. They, more than anyone else, bore the burden of this commitment, and for their help and support we are deeply grateful.

Introduction to Econometrics

CHAPTER

1 Economic Questions and Data

Ask a half dozen econometricians what econometrics is, and you could get a half dozen different answers. One might tell you that econometrics is the science of testing economic theories. A second might tell you that econometrics is the set of tools used for forecasting future values of economic variables, such as a firm's sales, the overall growth of the economy, or stock prices. Another might say that econometrics is the process of fitting mathematical economic models to real-world data. A fourth might tell you that it is the science and art of using historical data to make numerical, or quantitative, policy recommendations in government and business.

In fact, all these answers are right. At a broad level, econometrics is the science and art of using economic theory and statistical techniques to analyze economic data. Econometric methods are used in many branches of economics, including finance, labor economics, macroeconomics, microeconomics, marketing, and economic policy. Econometric methods are also commonly used in other social sciences, including political science and sociology.

This book introduces you to the core set of methods used by econometricians. We will use these methods to answer a variety of specific, quantitative questions from the worlds of business and government policy. This chapter poses four of those questions and discusses, in general terms, the econometric approach to answering them. The chapter concludes with a survey of the main types of data available to econometricians for answering these and other quantitative economic questions.

1.1 Economic Questions We Examine

Many decisions in economics, business, and government hinge on understanding relationships among variables in the world around us. These decisions require quantitative answers to quantitative questions.

This book examines several quantitative questions taken from current issues in economics. Four of these questions concern education policy, racial bias in mortgage lending, cigarette consumption, and macroeconomic forecasting.

Question #1: Does Reducing Class Size Improve Elementary School Education?

Proposals for reform of the U.S. public education system generate heated debate. Many of the proposals concern the youngest students, those in elementary schools. Elementary school education has various objectives, such as developing social skills, but for many parents and educators, the most important objective is basic academic learning: reading, writing, and basic mathematics. One prominent proposal for improving basic learning is to reduce class sizes at elementary schools. With fewer students in the classroom, the argument goes, each student gets more of the teacher's attention, there are fewer class disruptions, learning is enhanced, and grades improve.

But what, precisely, is the effect on elementary school education of reducing class size? Reducing class size costs money: It requires hiring more teachers and, if the school is already at capacity, building more classrooms. A decision maker contemplating hiring more teachers must weigh these costs against the benefits. To weigh costs and benefits, however, the decision maker must have a precise quantitative understanding of the likely benefits. Is the beneficial effect on basic learning of smaller classes large or small? Is it possible that smaller class size actually has no effect on basic learning?

Although common sense and everyday experience may suggest that more learning occurs when there are fewer students, common sense cannot provide a quantitative answer to the question of what exactly is the effect on basic learning of reducing class size. To provide such an answer, we must examine empirical evidence—that is, evidence based on data—relating class size to basic learning in elementary schools.

In this book, we examine the relationship between class size and basic learning, using data gathered from 420 California school districts in 1999. In the California data, students in districts with small class sizes tend to perform better on standardized tests than students in districts with larger classes. While this fact is consistent with the idea that smaller classes produce better test scores, it might simply reflect many other advantages that students in districts with small classes have over their counterparts in districts with large classes. For example, districts with small class sizes tend to have wealthier residents than districts with large classes, so students in small-class districts could have more opportunities for learning outside the classroom. It could be these extra learning opportunities that lead to higher test scores, not smaller class sizes. In Part II, we use multiple regression analysis to isolate the effect of changes in class size from changes in other factors, such as the economic background of the students.

Question #2: Is There Racial Discrimination in the Market for Home Loans?

Most people buy their homes with the help of a mortgage, a large loan secured by the value of the home. By law, U.S. lending institutions cannot take race into account when deciding to grant or deny a request for a mortgage: Applicants who are identical in all ways except their race should be equally likely to have their mortgage applications approved. In theory, then, there should be no racial bias in mortgage lending.

In contrast to this theoretical conclusion, researchers at the Federal Reserve Bank of Boston found (using data from the early 1990s) that 28% of black applicants are denied mortgages, while only 9% of white applicants are denied. Do these data indicate that, in practice, there is racial bias in mortgage lending? If so, how large is it?

The fact that more black than white applicants are denied in the Boston Fed data does not by itself provide evidence of discrimination by mortgage lenders because the black and white applicants differ in many ways other than their race. Before concluding that there is bias in the mortgage market, these data must be examined more closely to see if there is a difference in the probability of being denied for *otherwise identical* applicants and, if so, whether this difference is large or small. To do so, in Chapter 11 we introduce econometric methods that make it possible to quantify the effect of race on the chance of obtaining a mortgage, *holding constant* other applicant characteristics, notably their ability to repay the loan.

Question #3: How Much Do Cigarette Taxes Reduce Smoking?

Cigarette smoking is a major public health concern worldwide. Many of the costs of smoking, such as the medical expenses of caring for those made sick by smoking and the less quantifiable costs to nonsmokers who prefer not to breathe secondhand cigarette smoke, are borne by other members of society. Because these costs are borne by people other than the smoker, there is a role for government intervention in reducing cigarette consumption. One of the most flexible tools for cutting consumption is to increase taxes on cigarettes.

Basic economics says that if cigarette prices go up, consumption will go down. But by how much? If the sales price goes up by 1%, by what percentage will the quantity of cigarettes sold decrease? The percentage change in the quantity demanded resulting from a 1% increase in price is the *price elasticity of demand*.

If we want to reduce smoking by a certain amount, say 20%, by raising taxes, then we need to know the price elasticity of demand to calculate the price increase necessary to achieve this reduction in consumption. But what is the price elasticity of demand for cigarettes?

Although economic theory provides us with the concepts that help us answer this question, it does not tell us the numerical value of the price elasticity of demand. To learn the elasticity, we must examine empirical evidence about the behavior of smokers and potential smokers; in other words, we need to analyze data on cigarette consumption and prices.

The data we examine are cigarette sales, prices, taxes, and personal income for U.S. states in the 1980s and 1990s. In these data, states with low taxes, and thus low cigarette prices, have high smoking rates, and states with high prices have low smoking rates. However, the analysis of these data is complicated because causality runs both ways: Low taxes lead to high demand, but if there are many smokers in the state, then local politicians might try to keep cigarette taxes low to satisfy their smoking constituents. In Chapter 12, we study methods for handling this "simultaneous causality" and use those methods to estimate the price elasticity of cigarette demand.

Question #4: By How Much Will U.S. GDP Grow Next Year?

It seems that people always want a sneak preview of the future. What will sales be next year at a firm that is considering investing in new equipment? Will the stock market go up next month, and, if it does, by how much? Will city tax receipts next year cover planned expenditures on city services? Will your microeconomics exam next week focus on externalities or monopolies? Will Saturday be a nice day to go to the beach?

One aspect of the future in which macroeconomists are particularly interested is the growth of real economic activity, as measured by real gross domestic product (GDP), during the next year. A management consulting firm might advise a manufacturing client to expand its capacity based on an upbeat forecast of economic growth. Economists at the Federal Reserve Board in Washington, D.C., are mandated to set policy to keep real GDP near its potential in order to maximize employment. If they forecast anemic GDP growth over the next year, they might expand liquidity in the economy by reducing interest rates or other measures, in an attempt to boost economic activity.

Professional economists who rely on precise numerical forecasts use econometric models to make those forecasts. A forecaster's job is to predict the future by using the past, and econometricians do this by using economic theory and statistical techniques to quantify relationships in historical data.

The data we use to forecast the growth rate of GDP are past values of GDP and the "term spread" in the United States. The *term spread* is the difference between long-term and short-term interest rates. It measures, among other things, whether investors expect short-term interest rates to rise or fall in the future. The term spread is usually positive, but it tends to fall sharply before the onset of a recession. One of the GDP growth rate forecasts we develop and evaluate in Chapter 14 is based on the term spread.

Quantitative Questions, Quantitative Answers

Each of these four questions requires a numerical answer. Economic theory provides clues about that answer—for example, cigarette consumption ought to go down when the price goes up—but the actual value of the number must be learned empirically, that is, by analyzing data. Because we use data to answer quantitative questions, our answers always have some uncertainty: A different set of data would produce a different numerical answer. Therefore, the conceptual framework for the analysis needs to provide both a numerical answer to the question and a measure of how precise the answer is.

The conceptual framework used in this book is the multiple regression model, the mainstay of econometrics. This model, introduced in Part II, provides a mathematical way to quantify how a change in one variable affects another variable, holding other things constant. For example, what effect does a change in class size have on test scores, *holding constant* or *controlling for* student characteristics (such as family income) that a school district administrator cannot control? What effect does your race have on your chances of having a mortgage application granted, *holding constant* other factors such as your ability to repay the loan? What effect does a 1% increase in the price of cigarettes have on cigarette consumption, *holding constant* the income of smokers and potential smokers? The multiple regression model and its extensions provide a framework for answering these questions using data and for quantifying the uncertainty associated with those answers.

1.2 Causal Effects and Idealized Experiments

Like many other questions encountered in econometrics, the first three questions in Section 1.1 concern causal relationships among variables. In common usage, an action is said to cause an outcome if the outcome is the direct result, or consequence, of that action. Touching a hot stove causes you to get burned; drinking water causes you to be less thirsty; putting air in your tires causes them to inflate; putting fertilizer on your tomato plants causes them to produce more tomatoes. Causality means that a specific action (applying fertilizer) leads to a specific, measurable consequence (more tomatoes).

Estimation of Causal Effects

How best might we measure the causal effect on tomato yield (measured in kilograms) of applying a certain amount of fertilizer, say 100 grams of fertilizer per square meter?

One way to measure this causal effect is to conduct an experiment. In that experiment, a horticultural researcher plants many plots of tomatoes. Each plot is tended identically, with one exception: Some plots get 100 grams of fertilizer per square meter, while the rest get none. Moreover, whether a plot is fertilized or not is determined randomly by a computer, ensuring that any other differences between the plots are unrelated to whether they receive fertilizer. At the end of the growing season, the horticulturalist weighs the harvest from each plot. The difference between the average yield per square meter of the treated and untreated plots is the effect on tomato production of the fertilizer treatment.

This is an example of a **randomized controlled experiment**. It is controlled in the sense that there are both a **control group** that receives no treatment (no fertilizer) and a **treatment group** that receives the treatment (100 g/m^2 of fertilizer). It is randomized in the sense that the treatment is assigned randomly. This random assignment eliminates the possibility of a systematic relationship between, for example, how sunny the plot is and whether it receives fertilizer so that the only systematic difference between the treatment and control groups is the treatment. If this experiment is properly implemented on a large enough scale, then it will yield an estimate of the causal effect on the outcome of interest (tomato production) of the treatment (applying 100 g/m^2 of fertilizer).

In this book, the **causal effect** is defined to be the effect on an outcome of a given action or treatment, as measured in an ideal randomized controlled experiment. In such an experiment, the only systematic reason for differences in outcomes between the treatment and control groups is the treatment itself.

It is possible to imagine an ideal randomized controlled experiment to answer each of the first three questions in Section 1.1. For example, to study class size, one can imagine randomly assigning "treatments" of different class sizes to different groups of students. If the experiment is designed and executed so that the only systematic difference between the groups of students is their class size, then in theory this experiment would estimate the effect on test scores of reducing class size, holding all else constant.

The concept of an ideal randomized controlled experiment is useful because it gives a definition of a causal effect. In practice, however, it is not possible to perform ideal experiments. In fact, experiments are relatively rare in econometrics because often they are unethical, impossible to execute satisfactorily, or prohibitively expensive. The concept of the ideal randomized controlled experiment does, however, provide a theoretical benchmark for an econometric analysis of causal effects using actual data.

Forecasting and Causality

Although the first three questions in Section 1.1 concern causal effects, the fourth—forecasting the growth rate of GDP—does not. You do not need to know a causal relationship to make a good forecast. A good way to "forecast" whether it is raining is to observe whether pedestrians are using umbrellas, but the act of using an umbrella does not cause it to rain.

Even though forecasting need not involve causal relationships, economic theory suggests patterns and relationships that might be useful for forecasting. As we see in Chapter 14, multiple regression analysis allows us to quantify historical relationships suggested by economic theory, to check whether those relationships have been stable over time, to make quantitative forecasts about the future, and to assess the accuracy of those forecasts.

1.3 Data: Sources and Types

In econometrics, data come from one of two sources: experiments or nonexperimental observations of the world. This book examines both experimental and nonexperimental data sets.

Experimental Versus Observational Data

Experimental data come from experiments designed to evaluate a treatment or policy or to investigate a causal effect. For example, the state of Tennessee financed a large randomized controlled experiment examining class size in the 1980s. In that experiment, which we examine in Chapter 13, thousands of students were randomly assigned to classes of different sizes for several years and were given standardized tests annually.

The Tennessee class size experiment cost millions of dollars and required the ongoing cooperation of many administrators, parents, and teachers over several years. Because real-world experiments with human subjects are difficult to administer and to control, they have flaws relative to ideal randomized controlled experiments. Moreover, in some circumstances, experiments are not only expensive and difficult to administer but also unethical. (Would it be ethical to offer randomly selected teenagers inexpensive cigarettes to see how many they buy?) Because of these financial, practical, and ethical problems, experiments in economics are relatively rare. Instead, most economic data are obtained by observing real-world behavior.

Data obtained by observing actual behavior outside an experimental setting are called **observational data**. Observational data are collected using surveys, such as telephone surveys of consumers, and administrative records, such as historical records on mortgage applications maintained by lending institutions.

Observational data pose major challenges to econometric attempts to estimate causal effects, and the tools of econometrics are designed to tackle these challenges. In the real world, levels of "treatment" (the amount of fertilizer in the tomato example, the student–teacher ratio in the class size example) are not assigned at random, so it is difficult to sort out the effect of the "treatment" from other relevant factors. Much of econometrics, and much of this book, is devoted to methods for meeting the challenges encountered when real-world data are used to estimate causal effects.

Whether the data are experimental or observational, data sets come in three main types: cross-sectional data, time series data, and panel data. In this book, you will encounter all three types.

Cross-Sectional Data

Data on different entities—workers, consumers, firms, governmental units, and so forth—for a single time period are called **cross-sectional data**. For example, the data on test scores in California school districts are cross sectional. Those data are for 420 entities (school districts) for a single time period (1999). In general, the number of entities on which we have observations is denoted n; so, for example, in the California data set, $n = 420$.

The California test score data set contains measurements of several different variables for each district. Some of these data are tabulated in Table 1.1. Each row lists data for a different district. For example, the average test score for the first district ("district #1") is 690.8; this is the average of the math and science test scores for all fifth graders in that district in 1999 on a standardized test (the Stanford

TABLE 1.1 Selected Observations on Test Scores and Other Variables for California School Districts in 1999

Observation (District) Number	District Average Test Score (fifth grade)	Student–Teacher Ratio	Expenditure per Pupil ($)	Percentage of Students Learning English
1	690.8	17.89	$6385	0.0%
2	661.2	21.52	5099	4.6
3	643.6	18.70	5502	30.0
4	647.7	17.36	7102	0.0
5	640.8	18.67	5236	13.9
. . .	. . .	. . .	. . .	. . .
418	645.0	21.89	4403	24.3
419	672.2	20.20	4776	3.0
420	655.8	19.04	5993	5.0

Note: The California test score data set is described in Appendix 4.1.

Achievement Test). The average student–teacher ratio in that district is 17.89; that is, the number of students in district #1 divided by the number of classroom teachers in district #1 is 17.89. Average expenditure per pupil in district #1 is $6385. The percentage of students in that district still learning English—that is, the percentage of students for whom English is a second language and who are not yet proficient in English—is 0%.

The remaining rows present data for other districts. The order of the rows is arbitrary, and the number of the district, which is called the **observation number**, is an arbitrarily assigned number that organizes the data. As you can see in the table, all the variables listed vary considerably.

With cross-sectional data, we can learn about relationships among variables by studying differences across people, firms, or other economic entities during a single time period.

Time Series Data

Time series data are data for a single entity (person, firm, country) collected at multiple time periods. Our data set on the growth rate of GDP and the term spread in the United States is an example of a time series data set. The data set

TABLE 1.2 Selected Observations on the Growth Rate of GDP and the Term Spread in the United States: Quarterly Data, 1960:Q1–2013:Q1

Observation Number	Date (year:quarter)	GDP Growth Rate (% at an annual rate)	Term Spread (% per year)
1	1960:Q1	8.8%	0.6%
2	1960:Q2	–1.5	1.3
3	1960:Q3	1.0	1.5
4	1960:Q4	–4.9	1.6
5	1961:Q1	2.7	1.4
. . .	. . .	. . .	. . .
211	2012:Q3	2.7	1.5
212	2012:Q4	0.1	1.6
213	2013:Q1	1.1	1.9

Note: The United States GDP and term spread data set is described in Appendix 14.1.

contains observations on two variables (the growth rate of GDP and the term spread) for a single entity (the United States) for 213 time periods. Each time period in this data set is a quarter of a year (the first quarter is January, February, and March; the second quarter is April, May, and June; and so forth). The observations in this data set begin in the first quarter of 1960, which is denoted 1960:Q1, and end in the first quarter of 2013 (2013:Q1). The number of observations (that is, time periods) in a time series data set is denoted T. Because there are 213 quarters from 1960:Q1 to 2013:Q1, this data set contains $T = 213$ observations.

Some observations in this data set are listed in Table 1.2. The data in each row correspond to a different time period (year and quarter). In the first quarter of 1960, for example, GDP grew 8.8% at an annual rate. In other words, if GDP had continued growing for four quarters at its rate during the first quarter of 1960, the level of GDP would have increased by 8.8%. In the first quarter of 1960, the long-term interest rate was 4.5%, the short-term interest rate was 3.9%, so their difference, the term spread, was 0.6%.

By tracking a single entity over time, time series data can be used to study the evolution of variables over time and to forecast future values of those variables.

TABLE 1.3 Selected Observations on Cigarette Sales, Prices, and Taxes, by State and Year for U.S. States, 1985–1995

Observation Number	State	Year	Cigarette Sales (packs per capita)	Average Price per Pack (including taxes)	Total Taxes (cigarette excise tax + sales tax)
1	Alabama	1985	116.5	$1.022	$0.333
2	Arkansas	1985	128.5	1.015	0.370
3	Arizona	1985	104.5	1.086	0.362
⋮	⋮	⋮	⋮	⋮	⋮
47	West Virginia	1985	112.8	1.089	0.382
48	Wyoming	1985	129.4	0.935	0.240
49	Alabama	1986	117.2	1.080	0.334
⋮	⋮	⋮	⋮	⋮	⋮
96	Wyoming	1986	127.8	1.007	0.240
97	Alabama	1987	115.8	1.135	0.335
⋮	⋮	⋮	⋮	⋮	⋮
528	Wyoming	1995	112.2	1.585	0.360

Note: The cigarette consumption data set is described in Appendix 12.1.

Panel Data

Panel data, also called **longitudinal data**, are data for multiple entities in which each entity is observed at two or more time periods. Our data on cigarette consumption and prices are an example of a panel data set, and selected variables and observations in that data set are listed in Table 1.3. The number of entities in a panel data set is denoted n, and the number of time periods is denoted T. In the cigarette data set, we have observations on $n = 48$ continental U.S. states (entities) for $T = 11$ years (time periods) from 1985 to 1995. Thus there is a total of $n \times T = 48 \times 11 = 528$ observations.

KEY CONCEPT 1.1

Cross-Sectional, Time Series, and Panel Data

- Cross-sectional data consist of multiple entities observed at a single time period.
- Time series data consist of a single entity observed at multiple time periods.
- Panel data (also known as longitudinal data) consist of multiple entities, where each entity is observed at two or more time periods.

Some data from the cigarette consumption data set are listed in Table 1.3. The first block of 48 observations lists the data for each state in 1985, organized alphabetically from Alabama to Wyoming. The next block of 48 observations lists the data for 1986, and so forth, through 1995. For example, in 1985, cigarette sales in Arkansas were 128.5 packs per capita (the total number of packs of cigarettes sold in Arkansas in 1985 divided by the total population of Arkansas in 1985 equals 128.5). The average price of a pack of cigarettes in Arkansas in 1985, including tax, was \$1.015, of which 37¢ went to federal, state, and local taxes.

Panel data can be used to learn about economic relationships from the experiences of the many different entities in the data set and from the evolution over time of the variables for each entity.

The definitions of cross-sectional data, time series data, and panel data are summarized in Key Concept 1.1.

Summary

1. Many decisions in business and economics require quantitative estimates of how a change in one variable affects another variable.
2. Conceptually, the way to estimate a causal effect is in an ideal randomized controlled experiment, but performing such experiments in economic applications is usually unethical, impractical, or too expensive.
3. Econometrics provides tools for estimating causal effects using either observational (nonexperimental) data or data from real-world, imperfect experiments.
4. Cross-sectional data are gathered by observing multiple entities at a single point in time; time series data are gathered by observing a single entity at multiple points in time; and panel data are gathered by observing multiple entities, each of which is observed at multiple points in time.

Key Terms

randomized controlled experiment (6)
control group (6)
treatment group (6)
causal effect (6)
experimental data (7)
observational data (8)
cross-sectional data (8)
observation number (9)
time series data (9)
panel data (11)
longitudinal data (11)

Review the Concepts

1.1 Design a hypothetical ideal randomized controlled experiment to study the effect of hours spent studying on performance on microeconomics exams. Suggest some impediments to implementing this experiment in practice.

1.2 Design a hypothetical ideal randomized controlled experiment to study the effect on highway traffic deaths of wearing seat belts. Suggest some impediments to implementing this experiment in practice.

1.3 You are asked to study the casual effect of hours spent on employee training (measured in hours per worker per week) in a manufacturing plant on the productivity of its workers (output per worker per hour). Describe:

a. an ideal randomized controlled experiment to measure this causal effect;

b. an observational cross-sectional data set with which you could study this effect;

c. an observational time series data set for studying this effect; and

d. an observational panel data set for studying this effect.

CHAPTER

2 Review of Probability

This chapter reviews the core ideas of the theory of probability that are needed to understand regression analysis and econometrics. We assume that you have taken an introductory course in probability and statistics. If your knowledge of probability is stale, you should refresh it by reading this chapter. If you feel confident with the material, you still should skim the chapter and the terms and concepts at the end to make sure you are familiar with the ideas and notation.

Most aspects of the world around us have an element of randomness. The theory of probability provides mathematical tools for quantifying and describing this randomness. Section 2.1 reviews probability distributions for a single random variable, and Section 2.2 covers the mathematical expectation, mean, and variance of a single random variable. Most of the interesting problems in economics involve more than one variable, and Section 2.3 introduces the basic elements of probability theory for two random variables. Section 2.4 discusses three special probability distributions that play a central role in statistics and econometrics: the normal, chi-squared, and *F* distributions.

The final two sections of this chapter focus on a specific source of randomness of central importance in econometrics: the randomness that arises by randomly drawing a sample of data from a larger population. For example, suppose you survey ten recent college graduates selected at random, record (or "observe") their earnings, and compute the average earnings using these ten data points (or "observations"). Because you chose the sample at random, you could have chosen ten different graduates by pure random chance; had you done so, you would have observed ten different earnings and you would have computed a different sample average. Because the average earnings vary from one randomly chosen sample to the next, the sample average is itself a random variable. Therefore, the sample average has a probability distribution, which is referred to as its sampling distribution because this distribution describes the different possible values of the sample average that might have occurred had a different sample been drawn.

Section 2.5 discusses random sampling and the sampling distribution of the sample average. This sampling distribution is, in general, complicated. When the sample size is sufficiently large, however, the sampling distribution of the sample average is approximately normal, a result known as the central limit theorem, which is discussed in Section 2.6.

2.1 Random Variables and Probability Distributions

Probabilities, the Sample Space, and Random Variables

Probabilities and outcomes. The gender of the next new person you meet, your grade on an exam, and the number of times your computer will crash while you are writing a term paper all have an element of chance or randomness. In each of these examples, there is something not yet known that is eventually revealed.

The mutually exclusive potential results of a random process are called the **outcomes**. For example, your computer might never crash, it might crash once, it might crash twice, and so on. Only one of these outcomes will actually occur (the outcomes are mutually exclusive), and the outcomes need not be equally likely.

The **probability** of an outcome is the proportion of the time that the outcome occurs in the long run. If the probability of your computer not crashing while you are writing a term paper is 80%, then over the course of writing many term papers you will complete 80% without a crash.

The sample space and events. The set of all possible outcomes is called the **sample space**. An **event** is a subset of the sample space, that is, an event is a set of one or more outcomes. The event "my computer will crash no more than once" is the set consisting of two outcomes: "no crashes" and "one crash."

Random variables. A random variable is a numerical summary of a random outcome. The number of times your computer crashes while you are writing a term paper is random and takes on a numerical value, so it is a random variable.

Some random variables are discrete and some are continuous. As their names suggest, a **discrete random variable** takes on only a discrete set of values, like 0, 1, 2, . . . , whereas a **continuous random variable** takes on a continuum of possible values.

Probability Distribution of a Discrete Random Variable

Probability distribution. The **probability distribution** of a discrete random variable is the list of all possible values of the variable and the probability that each value will occur. These probabilities sum to 1.

For example, let M be the number of times your computer crashes while you are writing a term paper. The probability distribution of the random variable M is the list of probabilities of each possible outcome: The probability that $M = 0$, denoted $\Pr(M = 0)$, is the probability of no computer crashes; $\Pr(M = 1)$ is the probability of a single computer crash; and so forth. An example of a probability distribution for M is given in the second row of Table 2.1; in this distribution, if your computer crashes four times, you will quit and write the paper by hand. According to this distribution, the probability of no crashes is 80%; the probability of one crash is 10%; and the probability of two, three, or four crashes is, respectively, 6%, 3%, and 1%. These probabilities sum to 100%. This probability distribution is plotted in Figure 2.1.

Probabilities of events. The probability of an event can be computed from the probability distribution. For example, the probability of the event of one or two crashes is the sum of the probabilities of the constituent outcomes. That is, $\Pr(M = 1 \text{ or } M = 2) = \Pr(M = 1) + \Pr(M = 2) = 0.10 + 0.06 = 0.16$, or 16%.

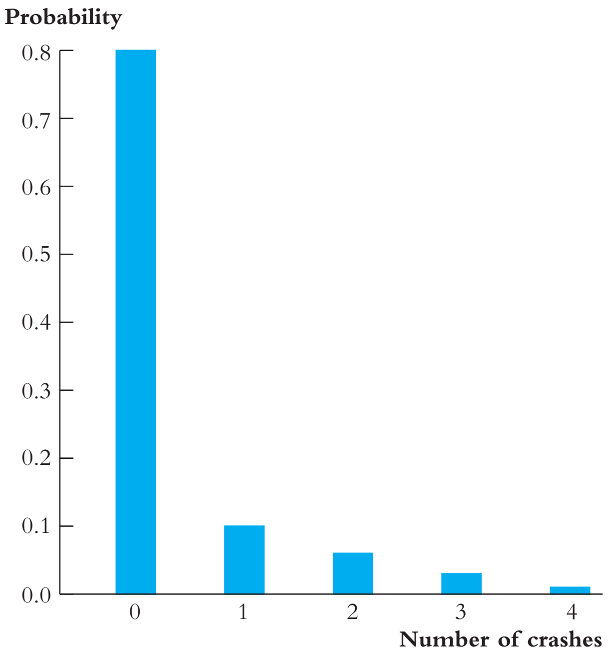
Cumulative probability distribution. The **cumulative probability distribution** is the probability that the random variable is less than or equal to a particular value. The last row of Table 2.1 gives the cumulative probability distribution of the random variable M. For example, the probability of at most one crash, $\Pr(M \leq 1)$, is 90%, which is the sum of the probabilities of no crashes (80%) and of one crash (10%).

TABLE 2.1 Probability of Your Computer Crashing *M* Times

	Outcome (number of crashes)				
	0	**1**	**2**	**3**	**4**
Probability distribution	0.80	0.10	0.06	0.03	0.01
Cumulative probability distribution	0.80	0.90	0.96	0.99	1.00

FIGURE 2.1 Probability Distribution of the Number of Computer Crashes

The height of each bar is the probability that the computer crashes the indicated number of times. The height of the first bar is 0.8, so the probability of 0 computer crashes is 80%. The height of the second bar is 0.1, so the probability of 1 computer crash is 10%, and so forth for the other bars.



A cumulative probability distribution is also referred to as a **cumulative distribution function**, a **c.d.f.**, or a **cumulative distribution**.

The Bernoulli distribution. An important special case of a discrete random variable is when the random variable is binary, that is, the outcomes are 0 or 1. A binary random variable is called a **Bernoulli random variable** (in honor of the seventeenth-century Swiss mathematician and scientist Jacob Bernoulli), and its probability distribution is called the **Bernoulli distribution**.

For example, let G be the gender of the next new person you meet, where $G = 0$ indicates that the person is male and $G = 1$ indicates that she is female. The outcomes of G and their probabilities thus are

$$G = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p, \end{cases} \tag{2.1}$$

where p is the probability of the next new person you meet being a woman. The probability distribution in Equation (2.1) is the Bernoulli distribution.

FIGURE 2.2 Cumulative Distribution and Probability Density Functions of Commuting Time

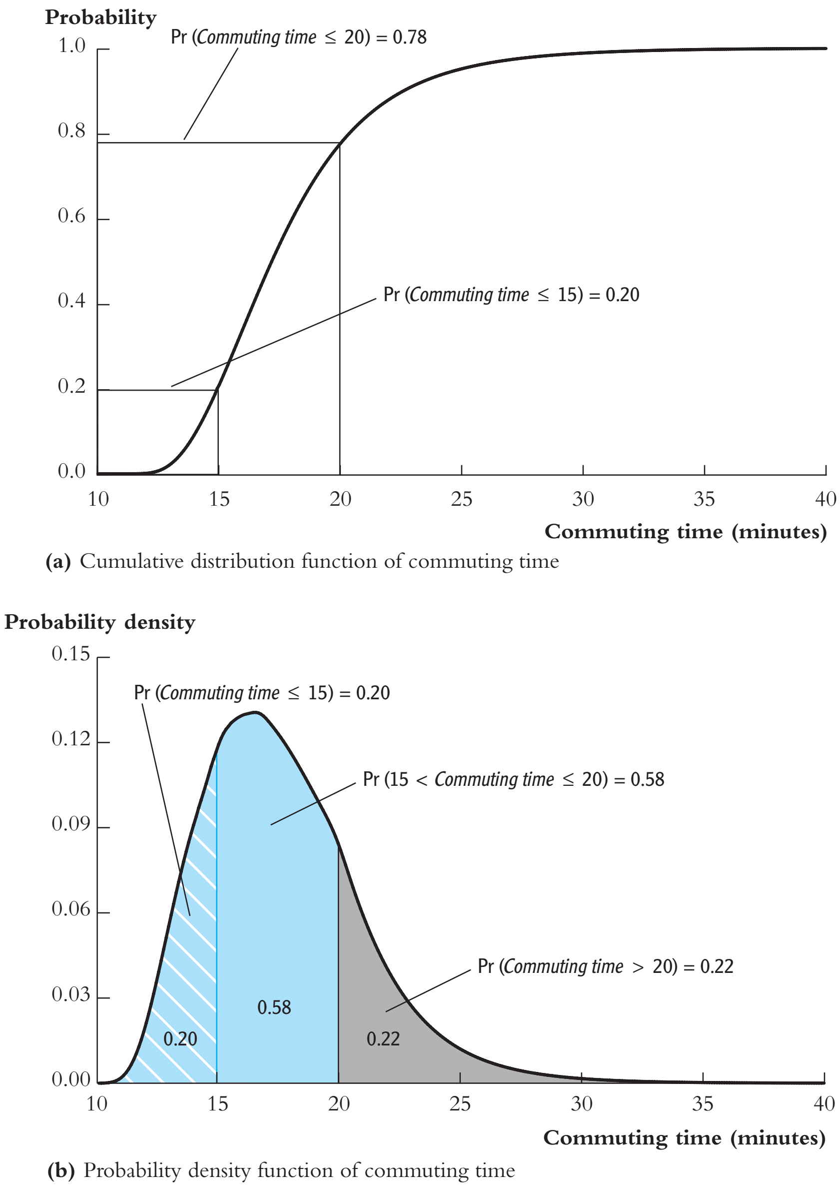


Figure 2.2a shows the cumulative probability distribution (or c.d.f.) of commuting times. The probability that a commuting time is less than 15 minutes is 0.20 (or 20%), and the probability that it is less than 20 minutes is 0.78 (78%). Figure 2.2b shows the probability density function (or p.d.f.) of commuting times. Probabilities are given by areas under the p.d.f. The probability that a commuting time is between 15 and 20 minutes is 0.58 (58%) and is given by the area under the curve between 15 and 20 minutes.

Probability Distribution of a Continuous Random Variable

Cumulative probability distribution. The cumulative probability distribution for a continuous variable is defined just as it is for a discrete random variable. That is, the cumulative probability distribution of a continuous random variable is the probability that the random variable is less than or equal to a particular value.

For example, consider a student who drives from home to school. This student's commuting time can take on a continuum of values and, because it depends on random factors such as the weather and traffic conditions, it is natural to treat it as a continuous random variable. Figure 2.2a plots a hypothetical cumulative distribution of commuting times. For example, the probability that the commute takes less than 15 minutes is 20% and the probability that it takes less than 20 minutes is 78%.

Probability density function. Because a continuous random variable can take on a continuum of possible values, the probability distribution used for discrete variables, which lists the probability of each possible value of the random variable, is not suitable for continuous variables. Instead, the probability is summarized by the **probability density function**. The area under the probability density function between any two points is the probability that the random variable falls between those two points. A probability density function is also called a **p.d.f.**, a **density function**, or simply a **density**.

Figure 2.2b plots the probability density function of commuting times corresponding to the cumulative distribution in Figure 2.2a. The probability that the commute takes between 15 and 20 minutes is given by the area under the p.d.f. between 15 minutes and 20 minutes, which is 0.58, or 58%. Equivalently, this probability can be seen on the cumulative distribution in Figure 2.2a as the difference between the probability that the commute is less than 20 minutes (78%) and the probability that it is less than 15 minutes (20%). Thus the probability density function and the cumulative probability distribution show the same information in different formats.

2.2 Expected Values, Mean, and Variance

The Expected Value of a Random Variable

Expected value. The **expected value** of a random variable Y, denoted $E(Y)$, is the long-run average value of the random variable over many repeated trials or occurrences. The expected value of a discrete random variable is computed as a weighted average of the possible outcomes of that random variable, where the weights are the probabilities of that outcome. The expected value of Y is also called the **expectation** of Y or the **mean** of Y and is denoted μ_Y.

For example, suppose you loan a friend \$100 at 10% interest. If the loan is repaid, you get \$110 (the principal of \$100 plus interest of \$10), but there is a risk of 1% that your friend will default and you will get nothing at all. Thus the amount you are repaid is a random variable that equals \$110 with probability 0.99 and equals \$0 with probability 0.01. Over many such loans, 99% of the time you would be paid back \$110, but 1% of the time you would get nothing, so on average you would be repaid $\$110 \times 0.99 + \$0 \times 0.01 = \$108.90$. Thus the expected value of your repayment (or the "mean repayment") is \$108.90.

As a second example, consider the number of computer crashes M with the probability distribution given in Table 2.1. The expected value of M is the average number of crashes over many term papers, weighted by the frequency with which a crash of a given size occurs. Accordingly,

$$E(M) = 0 \times 0.80 + 1 \times 0.10 + 2 \times 0.06 + 3 \times 0.03 + 4 \times 0.01 = 0.35. \quad (2.2)$$

That is, the expected number of computer crashes while writing a term paper is 0.35. Of course, the actual number of crashes must always be an integer; it makes no sense to say that the computer crashed 0.35 times while writing a particular term paper! Rather, the calculation in Equation (2.2) means that the average number of crashes over many such term papers is 0.35.

The formula for the expected value of a discrete random variable Y that can take on k different values is given as Key Concept 2.1. (Key Concept 2.1 uses "summation notation," which is reviewed in Exercise 2.25.)

KEY CONCEPT 2.1

Expected Value and the Mean

Suppose the random variable Y takes on k possible values, $y_1, \ldots, y_k$, where y_1 denotes the first value, y_2 denotes the second value, and so forth, and that the probability that Y takes on y_1 is p_1, the probability that Y takes on y_2 is p_2, and so forth. The expected value of Y, denoted $E(Y)$, is

$$E(Y) = y_1 p_1 + y_2 p_2 + \cdots + y_k p_k = \sum_{i=1}^{k} y_i p_i, \quad (2.3)$$

where the notation $\sum_{i=1}^{k} y_i p_i$ means "the sum of $y_i p_i$ for i running from 1 to k." The expected value of Y is also called the mean of Y or the expectation of Y and is denoted μ_Y.

Expected value of a Bernoulli random variable. An important special case of the general formula in Key Concept 2.1 is the mean of a Bernoulli random variable. Let G be the Bernoulli random variable with the probability distribution in Equation (2.1). The expected value of G is

$$E(G) = 1 \times p + 0 \times (1 - p) = p. \tag{2.4}$$

Thus the expected value of a Bernoulli random variable is p, the probability that it takes on the value "1."

Expected value of a continuous random variable. The expected value of a continuous random variable is also the probability-weighted average of the possible outcomes of the random variable. Because a continuous random variable can take on a continuum of possible values, the formal mathematical definition of its expectation involves calculus and its definition is given in Appendix 17.1.

The Standard Deviation and Variance

The variance and standard deviation measure the dispersion or the "spread" of a probability distribution. The **variance** of a random variable Y, denoted $\text{var}(Y)$, is the expected value of the square of the deviation of Y from its mean: $\text{var}(Y) = E[(Y - \mu_Y)^2]$.

Because the variance involves the square of Y, the units of the variance are the units of the square of Y, which makes the variance awkward to interpret. It is therefore common to measure the spread by the **standard deviation**, which is the square root of the variance and is denoted σ_Y. The standard deviation has the same units as Y. These definitions are summarized in Key Concept 2.2.

Variance and Standard Deviation

KEY CONCEPT 2.2

The variance of the discrete random variable Y, denoted σ_Y^2, is

$$\sigma_Y^2 = \text{var}(Y) = E[(Y - \mu_Y)^2] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 p_i. \tag{2.5}$$

The standard deviation of Y is σ_Y, the square root of the variance. The units of the standard deviation are the same as the units of Y.

For example, the variance of the number of computer crashes M is the probability-weighted average of the squared difference between M and its mean, 0.35:

$$\begin{aligned} \mathrm{var}(M) = {} & (0-0.35)^2 \times 0.80 + (1-0.35)^2 \times 0.10 + (2-0.35)^2 \times 0.06 \\ & + (3-0.35)^2 \times 0.03 + (4-0.35)^2 \times 0.01 = 0.6475. \end{aligned} \tag{2.6}$$

The standard deviation of M is the square root of the variance, so $\sigma_M = \sqrt{0.64750} \cong 0.80$.

Variance of a Bernoulli random variable. The mean of the Bernoulli random variable G with probability distribution in Equation (2.1) is $\mu_G = p$ [Equation (2.4)], so its variance is

$$\mathrm{var}(G) = \sigma_G^2 = (0-p)^2 \times (1-p) + (1-p)^2 \times p = p(1-p). \tag{2.7}$$

Thus the standard deviation of a Bernoulli random variable is $\sigma_G = \sqrt{p(1-p)}$.

Mean and Variance of a Linear Function of a Random Variable

This section discusses random variables (say, X and Y) that are related by a linear function. For example, consider an income tax scheme under which a worker is taxed at a rate of 20% on his or her earnings and then given a (tax-free) grant of \$2000. Under this tax scheme, after-tax earnings Y are related to pre-tax earnings X by the equation

$$Y = 2000 + 0.8X. \tag{2.8}$$

That is, after-tax earnings Y is 80% of pre-tax earnings X, plus \$2000.

Suppose an individual's pre-tax earnings next year are a random variable with mean μ_X and variance σ_X^2. Because pre-tax earnings are random, so are after-tax earnings. What are the mean and standard deviations of her after-tax earnings under this tax? After taxes, her earnings are 80% of the original pre-tax earnings, plus \$2000. Thus the expected value of her after-tax earnings is

$$E(Y) = \mu_Y = 2000 + 0.8\mu_X. \tag{2.9}$$

The variance of after-tax earnings is the expected value of $(Y - \mu_Y)^2$. Because $Y = 2000 + 0.8X$, $Y - \mu_Y = 2000 + 0.8X - (2000 + 0.8\mu_X) = 0.8(X - \mu_X)$.

Thus $E[(Y - \mu_Y)^2] = E\{[0.8(X - \mu_X)]^2\} = 0.64E[(X - \mu_X)^2]$. It follows that $\text{var}(Y) = 0.64\text{var}(X)$, so, taking the square root of the variance, the standard deviation of Y is

$$\sigma_Y = 0.8\sigma_X. \tag{2.10}$$

That is, the standard deviation of the distribution of her after-tax earnings is 80% of the standard deviation of the distribution of pre-tax earnings.

This analysis can be generalized so that Y depends on X with an intercept a (instead of \$2000) and a slope b (instead of 0.8) so that

$$Y = a + bX. \tag{2.11}$$

Then the mean and variance of Y are

$$\mu_Y = a + b\mu_X \quad \text{and} \tag{2.12}$$

$$\sigma_Y^2 = b^2\sigma_X^2, \tag{2.13}$$

and the standard deviation of Y is $\sigma_Y = b\sigma_X$. The expressions in Equations (2.9) and (2.10) are applications of the more general formulas in Equations (2.12) and (2.13) with $a = 2000$ and $b = 0.8$.

Other Measures of the Shape of a Distribution

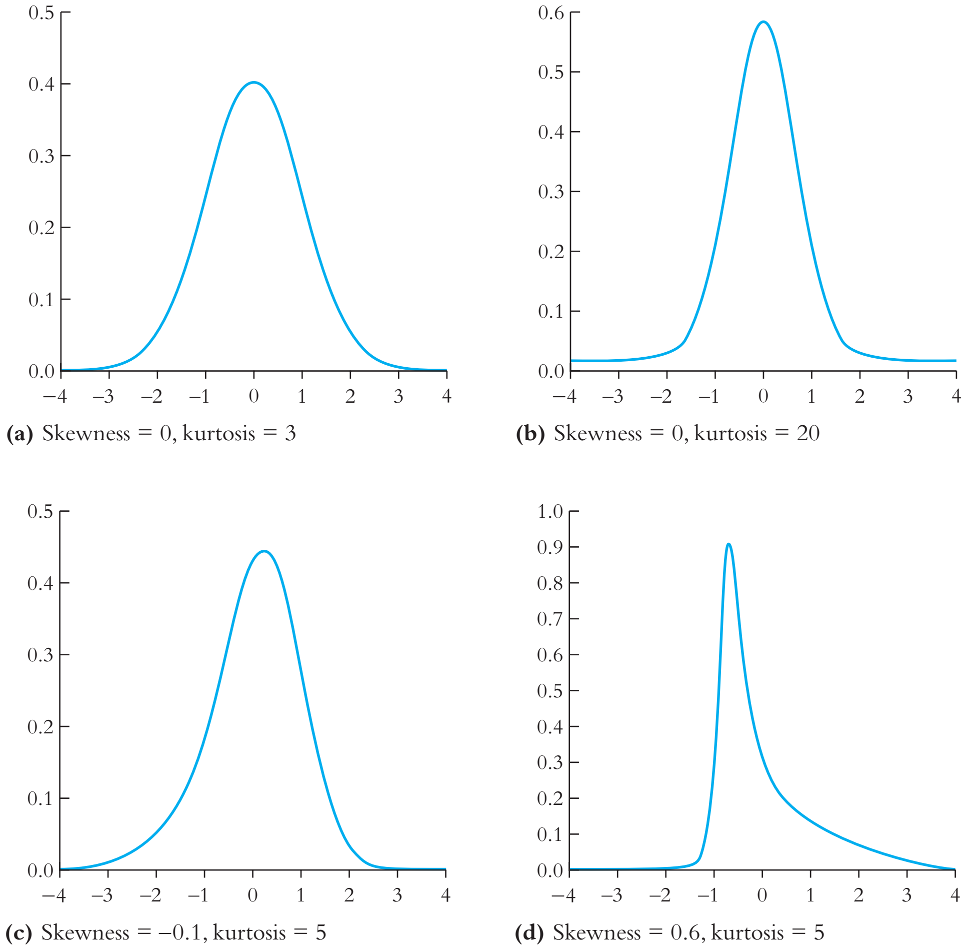
The mean and standard deviation measure two important features of a distribution: its center (the mean) and its spread (the standard deviation). This section discusses measures of two other features of a distribution: the skewness, which measures the lack of symmetry of a distribution, and the kurtosis, which measures how thick, or "heavy," are its tails. The mean, variance, skewness, and kurtosis are all based on what are called the **moments of a distribution**.

Skewness. Figure 2.3 plots four distributions, two which are symmetric (Figures 2.3a and 2.3b) and two which are not (Figures 2.3c and 2.3d). Visually, the distribution in Figure 2.3d appears to deviate more from symmetry than does the distribution in Figure 2.3c. The skewness of a distribution provides a mathematical way to describe how much a distribution deviates from symmetry.

The **skewness** of the distribution of a random variable Y is

$$\text{Skewness} = \frac{E[(Y - \mu_Y)^3]}{\sigma_Y^3}, \tag{2.14}$$

FIGURE 2.3 Four Distributions with Different Skewness and Kurtosis



All of these distributions have a mean of 0 and a variance of 1. The distributions with skewness of 0 (a and b) are symmetric; the distributions with nonzero skewness (c and d) are not symmetric. The distributions with kurtosis exceeding 3 (b–d) have heavy tails.

where σ_Y is the standard deviation of Y. For a symmetric distribution, a value of Y a given amount above its mean is just as likely as a value of Y the same amount below its mean. If so, then positive values of $(Y - \mu_Y)^3$ will be offset on average (in expectation) by equally likely negative values. Thus, for a symmetric distribution, $E[(Y - \mu_Y)^3] = 0$; the skewness of a symmetric distribution is zero. If a distribution is not symmetric, then a positive value of $(Y - \mu_Y)^3$ generally is not offset on average by an equally likely negative value, so the skewness is nonzero for a distribution that is not symmetric. Dividing by σ_Y^3 in the denominator of Equation (2.14) cancels the units of Y^3 in the numerator, so the skewness is unit free; in other words, changing the units of Y does not change its skewness.

Below each of the four distributions in Figure 2.3 is its skewness. If a distribution has a long right tail, positive values of $(Y - \mu_Y)^3$ are not fully offset by negative values, and the skewness is positive. If a distribution has a long left tail, its skewness is negative.

Kurtosis. The **kurtosis** of a distribution is a measure of how much mass is in its tails and, therefore, is a measure of how much of the variance of Y arises from extreme values. An extreme value of Y is called an **outlier**. The greater the kurtosis of a distribution, the more likely are outliers.

The kurtosis of the distribution of Y is

$$\text{Kurtosis} = \frac{E[(Y - \mu_Y)^4]}{\sigma_Y^4}. \tag{2.15}$$

If a distribution has a large amount of mass in its tails, then some extreme departures of Y from its mean are likely, and these departures will lead to large values, on average (in expectation), of $(Y - \mu_Y)^4$. Thus, for a distribution with a large amount of mass in its tails, the kurtosis will be large. Because $(Y - \mu_Y)^4$ cannot be negative, the kurtosis cannot be negative.

The kurtosis of a normally distributed random variable is 3, so a random variable with kurtosis exceeding 3 has more mass in its tails than a normal random variable. A distribution with kurtosis exceeding 3 is called **leptokurtic** or, more simply, heavy-tailed. Like skewness, the kurtosis is unit free, so changing the units of Y does not change its kurtosis.

Below each of the four distributions in Figure 2.3 is its kurtosis. The distributions in Figures 2.3b–d are heavy-tailed.

Moments. The mean of Y, $E(Y)$, is also called the first moment of Y, and the expected value of the square of Y, $E(Y^2)$, is called the second moment of Y. In general, the expected value of Y^r is called the **r^{th} moment** of the random variable Y. That is, the r^{th} moment of Y is $E(Y^r)$. The skewness is a function of the first, second, and third moments of Y, and the kurtosis is a function of the first through fourth moments of Y.

2.3 Two Random Variables

Most of the interesting questions in economics involve two or more variables. Are college graduates more likely to have a job than nongraduates? How does the distribution of income for women compare to that for men? These questions concern the distribution of two random variables, considered together (education and employment status in the first example, income and gender in the second). Answering such questions requires an understanding of the concepts of joint, marginal, and conditional probability distributions.

Joint and Marginal Distributions

Joint distribution. The **joint probability distribution** of two discrete random variables, say X and Y, is the probability that the random variables simultaneously take on certain values, say x and y. The probabilities of all possible (x, y) combinations sum to 1. The joint probability distribution can be written as the function $\Pr(X = x, Y = y)$.

For example, weather conditions—whether or not it is raining—affect the commuting time of the student commuter in Section 2.1. Let Y be a binary random variable that equals 1 if the commute is short (less than 20 minutes) and equals 0 otherwise and let X be a binary random variable that equals 0 if it is raining and 1 if not. Between these two random variables, there are four possible outcomes: it rains and the commute is long $(X = 0, Y = 0)$; rain and short commute $(X = 0, Y = 1)$; no rain and long commute $(X = 1, Y = 0)$; and no rain and short commute $(X = 1, Y = 1)$. The joint probability distribution is the frequency with which each of these four outcomes occurs over many repeated commutes.

An example of a joint distribution of these two variables is given in Table 2.2. According to this distribution, over many commutes, 15% of the days have rain and a long commute $(X = 0, Y = 0)$; that is, the probability of a long, rainy commute is 15%, or $\Pr(X = 0, Y = 0) = 0.15$. Also, $\Pr(X = 0, Y = 1) = 0.15$,

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

$\Pr(X = 1, Y = 0) = 0.07$, and $\Pr(X = 1, Y = 1) = 0.63$. These four possible outcomes are mutually exclusive and constitute the sample space so the four probabilities sum to 1.

Marginal probability distribution. The **marginal probability distribution** of a random variable Y is just another name for its probability distribution. This term is used to distinguish the distribution of Y alone (the marginal distribution) from the joint distribution of Y and another random variable.

The marginal distribution of Y can be computed from the joint distribution of X and Y by adding up the probabilities of all possible outcomes for which Y takes on a specified value. If X can take on l different values $x_1, \ldots, x_l$, then the marginal probability that Y takes on the value y is

$$\Pr(Y = y) = \sum_{i=1}^{l} \Pr(X = x_i, Y = y). \tag{2.16}$$

For example, in Table 2.2, the probability of a long rainy commute is 15% and the probability of a long commute with no rain is 7%, so the probability of a long commute (rainy or not) is 22%. The marginal distribution of commuting times is given in the final column of Table 2.2. Similarly, the marginal probability that it will rain is 30%, as shown in the final row of Table 2.2.

Conditional Distributions

Conditional distribution. The distribution of a random variable Y conditional on another random variable X taking on a specific value is called the **conditional distribution** of Y given X. The conditional probability that Y takes on the value y when X takes on the value x is written $\Pr(Y = y \mid X = x)$.

For example, what is the probability of a long commute ($Y = 0$) if you know it is raining ($X = 0$)? From Table 2.2, the joint probability of a rainy short commute is 15% and the joint probability of a rainy long commute is 15%, so if it is raining a long commute and a short commute are equally likely. Thus the probability of a long commute ($Y = 0$), conditional on it being rainy ($X = 0$), is 50%, or $\Pr(Y = 0 \mid X = 0) = 0.50$. Equivalently, the marginal probability of rain is 30%; that is, over many commutes it rains 30% of the time. Of this 30% of commutes, 50% of the time the commute is long (0.15/0.30).

In general, the conditional distribution of Y given $X = x$ is

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}. \tag{2.17}$$

TABLE 2.3 Joint and Conditional Distributions of Computer Crashes (M) and Computer Age (A)

A. Joint Distribution

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
Old computer ($A = 0$)	0.35	0.065	0.05	0.025	0.01	0.50
New computer ($A = 1$)	0.45	0.035	0.01	0.005	0.00	0.50
Total	0.80	0.10	0.06	0.03	0.01	1.00

B. Conditional Distributions of *M* given *A*

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
$\Pr(M \mid A = 0)$	0.70	0.13	0.10	0.05	0.02	1.00
$\Pr(M \mid A = 1)$	0.90	0.07	0.02	0.01	0.00	1.00

For example, the conditional probability of a long commute given that it is rainy is $\Pr(Y = 0 \mid X = 0) = \Pr(X = 0, Y = 0)/\Pr(X = 0) = 0.15/0.30 = 0.50$.

As a second example, consider a modification of the crashing computer example. Suppose you use a computer in the library to type your term paper and the librarian randomly assigns you a computer from those available, half of which are new and half of which are old. Because you are randomly assigned to a computer, the age of the computer you use, A ($= 1$ if the computer is new, $= 0$ if it is old), is a random variable. Suppose the joint distribution of the random variables M and A is given in Part A of Table 2.3. Then the conditional distribution of computer crashes, given the age of the computer, is given in Part B of the table. For example, the joint probability $M = 0$ and $A = 0$ is 0.35; because half the computers are old, the conditional probability of no crashes, given that you are using an old computer, is $\Pr(M = 0 \mid A = 0) = \Pr(M = 0, A = 0)/\Pr(A = 0) = 0.35/0.50 = 0.70$, or 70%. In contrast, the conditional probability of no crashes given that you are assigned a new computer is 90%. According to the conditional distributions in Part B of Table 2.3, the newer computers are less likely to crash than the old ones; for example, the probability of three crashes is 5% with an old computer but 1% with a new computer.

Conditional expectation. The **conditional expectation** of Y given X, also called the **conditional mean** of Y given X, is the mean of the conditional distribution of Y given X. That is, the conditional expectation is the expected value of Y, computed using the conditional distribution of Y given X. If Y takes on k values $y_1, \ldots, y_k$, then the conditional mean of Y given $X = x$ is

$$E(Y \mid X = x) = \sum_{i=1}^{k} y_i \Pr(Y = y_i \mid X = x). \tag{2.18}$$

For example, based on the conditional distributions in Table 2.3, the expected number of computer crashes, given that the computer is old, is $E(M \mid A = 0) = 0 \times 0.70 + 1 \times 0.13 + 2 \times 0.10 + 3 \times 0.05 + 4 \times 0.02 = 0.56$. The expected number of computer crashes, given that the computer is new, is $E(M \mid A = 1) = 0.14$, less than for the old computers.

The conditional expectation of Y given $X = x$ is just the mean value of Y when $X = x$. In the example of Table 2.3, the mean number of crashes is 0.56 for old computers, so the conditional expectation of Y given that the computer is old is 0.56. Similarly, among new computers, the mean number of crashes is 0.14, that is, the conditional expectation of Y given that the computer is new is 0.14.

The law of iterated expectations. The mean of Y is the weighted average of the conditional expectation of Y given X, weighted by the probability distribution of X. For example, the mean height of adults is the weighted average of the mean height of men and the mean height of women, weighted by the proportions of men and women. Stated mathematically, if X takes on the l values $x_1, \ldots, x_l$, then

$$E(Y) = \sum_{i=1}^{l} E(Y \mid X = x_i)\Pr(X = x_i). \tag{2.19}$$

Equation (2.19) follows from Equations (2.18) and (2.17) (see Exercise 2.19).

Stated differently, the expectation of Y is the expectation of the conditional expectation of Y given X,

$$E(Y) = E[E(Y \mid X)], \tag{2.20}$$

where the inner expectation on the right-hand side of Equation (2.20) is computed using the conditional distribution of Y given X and the outer expectation is computed using the marginal distribution of X. Equation (2.20) is known as the **law of iterated expectations**.

For example, the mean number of crashes M is the weighted average of the conditional expectation of M given that it is old and the conditional expectation of

M given that it is new, so $E(M) = E(M \mid A = 0) \times \Pr(A = 0) + E(M \mid A = 1) \times \Pr(A = 1) = 0.56 \times 0.50 + 0.14 \times 0.50 = 0.35$. This is the mean of the marginal distribution of M, as calculated in Equation (2.2).

The law of iterated expectations implies that if the conditional mean of Y given X is zero, then the mean of Y is zero. This is an immediate consequence of Equation (2.20): if $E(Y \mid X) = 0$, then $E(Y) = E[E(Y \mid X)] = E[0] = 0$. Said differently, if the mean of Y given X is zero, then it must be that the probability-weighted average of these conditional means is zero, that is, the mean of Y must be zero.

The law of iterated expectations also applies to expectations that are conditional on multiple random variables. For example, let X, Y, and Z be random variables that are jointly distributed. Then the law of iterated expectations says that $E(Y) = E[E(Y \mid X, Z)]$, where $E(Y \mid X, Z)$ is the conditional expectation of Y given both X and Z. For example, in the computer crash illustration of Table 2.3, let P denote the number of programs installed on the computer; then $E(M \mid A, P)$ is the expected number of crashes for a computer with age A that has P programs installed. The expected number of crashes overall, $E(M)$, is the weighted average of the expected number of crashes for a computer with age A and number of programs P, weighted by the proportion of computers with that value of both A and P.

Exercise 2.20 provides some additional properties of conditional expectations with multiple variables.

Conditional variance. The variance of Y conditional on X is the variance of the conditional distribution of Y given X. Stated mathematically, the **conditional variance** of Y given X is

$$\operatorname{var}(Y \mid X = x) = \sum_{i=1}^{k} [y_i - E(Y \mid X = x)]^2 \Pr(Y = y_i \mid X = x). \quad (2.21)$$

For example, the conditional variance of the number of crashes given that the computer is old is $\operatorname{var}(M \mid A = 0) = (0 - 0.56)^2 \times 0.70 + (1 - 0.56)^2 \times 0.13 + (2 - 0.56)^2 \times 0.10 + (3 - 0.56)^2 \times 0.05 + (4 - 0.56)^2 \times 0.02 \cong 0.99$. The standard deviation of the conditional distribution of M given that $A = 0$ is thus $\sqrt{0.99} = 0.99$. The conditional variance of M given that $A = 1$ is the variance of the distribution in the second row of Panel B of Table 2.3, which is 0.22, so the standard deviation of M for new computers is $\sqrt{0.22} = 0.47$. For the conditional distributions in Table 2.3, the expected number of crashes for new computers (0.14) is less than that for old computers (0.56), and the spread of the distribution of the number of crashes, as measured by the conditional standard deviation, is smaller for new computers (0.47) than for old (0.99).

Independence

Two random variables X and Y are **independently distributed**, or **independent**, if knowing the value of one of the variables provides no information about the other. Specifically, X and Y are independent if the conditional distribution of Y given X equals the marginal distribution of Y. That is, X and Y are independently distributed if, for all values of x and y,

$$\Pr(Y = y \mid X = x) = \Pr(Y = y) \quad \text{(independence of } X \text{ and } Y\text{)}. \tag{2.22}$$

Substituting Equation (2.22) into Equation (2.17) gives an alternative expression for independent random variables in terms of their joint distribution. If X and Y are independent, then

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y). \tag{2.23}$$

That is, the joint distribution of two independent random variables is the product of their marginal distributions.

Covariance and Correlation

Covariance. One measure of the extent to which two random variables move together is their covariance. The **covariance** between X and Y is the expected value $E[(X - \mu_X)(Y - \mu_Y)]$, where μ_X, where μ_X is the mean of X and μ_Y is the mean of Y. The covariance is denoted $\text{cov}(X, Y)$ or σ_{XY}. If X can take on l values and Y can take on k values, then the covariance is given by the formula

$$\begin{aligned}\text{cov}(X, Y) &= \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_j - \mu_X)(y_i - \mu_Y)\Pr(X = x_j, Y = y_i).\end{aligned} \tag{2.24}$$

To interpret this formula, suppose that when X is greater than its mean (so that $X - \mu_X$ is positive), then Y tends be greater than its mean (so that $Y - \mu_Y$ is positive), and when X is less than its mean (so that $X - \mu_X < 0$), then Y tends to be less than its mean (so that $Y - \mu_Y < 0$). In both cases, the product $(X - \mu_X) \times (Y - \mu_Y)$ tends to be positive, so the covariance is positive. In contrast, if X and Y tend to move in opposite directions (so that X is large when Y is small, and vice versa), then the covariance is negative. Finally, if X and Y are independent, then the covariance is zero (see Exercise 2.19).

Correlation. Because the covariance is the product of X and Y, deviated from their means, its units are, awkwardly, the units of X multiplied by the units of Y. This "units" problem can make numerical values of the covariance difficult to interpret.

The correlation is an alternative measure of dependence between X and Y that solves the "units" problem of the covariance. Specifically, the **correlation** between X and Y is the covariance between X and Y divided by their standard deviations:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.25}$$

Because the units of the numerator in Equation (2.25) are the same as those of the denominator, the units cancel and the correlation is unitless. The random variables X and Y are said to be **uncorrelated** if $\text{corr}(X, Y) = 0$.

The correlation always is between –1 and 1; that is, as proven in Appendix 2.1,

$$-1 \leq \text{corr}(X, Y) \leq 1 \quad \text{(correlation inequality)}. \tag{2.26}$$

Correlation and conditional mean. If the conditional mean of Y does not depend on X, then Y and X are uncorrelated. That is,

$$\text{if } E(Y \mid X) = \mu_Y, \text{ then } \text{cov}(Y, X) = 0 \text{ and } \text{corr}(Y, X) = 0. \tag{2.27}$$

We now show this result. First suppose that Y and X have mean zero so that $\text{cov}(Y, X) = E[(Y - \mu_Y)(X - \mu_X)] = E(YX)$. By the law of iterated expectations [Equation (2.20)], $E(YX) = E[E(YX \mid X)] = E[E(Y \mid X)X] = 0$ because $E(Y \mid X) = 0$, so $\text{cov}(Y, X) = 0$. Equation (2.27) follows by substituting $\text{cov}(Y, X) = 0$ into the definition of correlation in Equation (2.25). If Y and X do not have mean zero, first subtract off their means, then the preceding proof applies.

It is *not* necessarily true, however, that if X and Y are uncorrelated, then the conditional mean of Y given X does not depend on X. Said differently, it is possible for the conditional mean of Y to be a function of X but for Y and X nonetheless to be uncorrelated. An example is given in Exercise 2.23.

The Mean and Variance of Sums of Random Variables

The mean of the sum of two random variables, X and Y, is the sum of their means:

$$E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y. \tag{2.28}$$

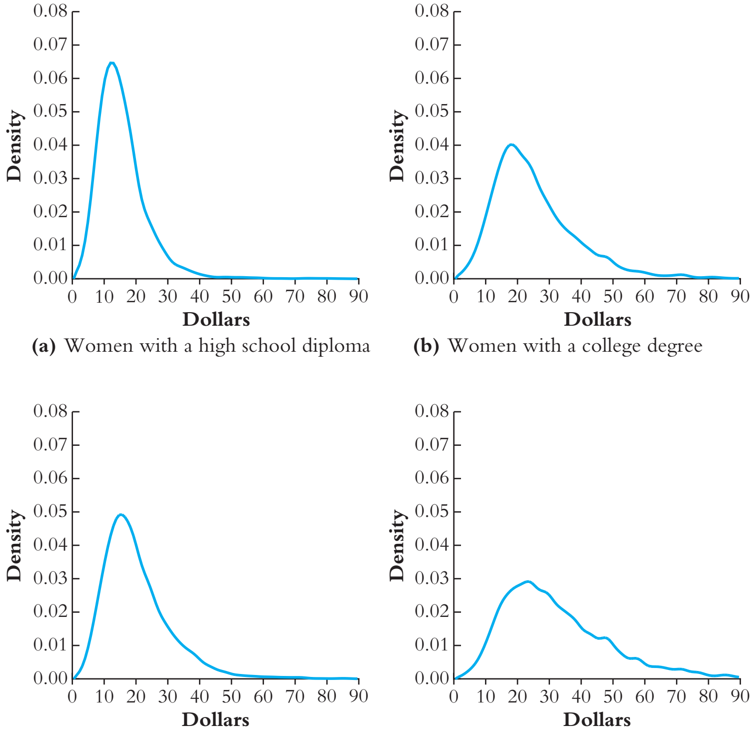
The Distribution of Earnings in the United States in 2012

Some parents tell their children that they will be able to get a better, higher-paying job if they get a college degree than if they skip higher education. Are these parents right? Does the distribution of earnings differ between workers who are college graduates and workers who have only a high school diploma, and, if so, how? Among workers with a similar education, does the distribution of earnings for men and women differ? For example, do the best-paid college-educated women earn as much as the best-paid college-educated men?

One way to answer these questions is to examine the distribution of earnings of full-time workers, conditional on the highest educational degree achieved (high school diploma or bachelor's degree) and on gender. These four conditional distributions are shown in Figure 2.4, and the mean, standard deviation, and

FIGURE 2.4 Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2012, Given Education Level and Gender

The four distributions of earnings are for women and men, for those with only a high school diploma (a and c) and those whose highest degree is from a four-year college (b and d).



(a) Women with a high school diploma

(b) Women with a college degree

(c) Men with a high school diploma

(d) Men with a college degree

continued on next page

TABLE 2.4 Summaries of the Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2012 Given Education Level and Gender

	Mean	Standard Deviation	Percentile 25%	50% (median)	75%	90%
(a) Women with high school diploma	\$15.49	\$8.42	\$10.10	\$14.03	\$18.75	\$24.52
(b) Women with four-year college degree	25.42	13.81	16.15	22.44	31.34	43.27
(c) Men with high school diploma	20.25	11.00	12.92	17.86	24.83	33.78
(d) Men with four-year college degree	32.73	18.11	19.61	28.85	41.68	57.30

Average hourly earnings are the sum of annual pretax wages, salaries, tips, and bonuses divided by the number of hours worked annually.

some percentiles of the conditional distributions are presented in Table 2.4.[1] For example, the conditional mean of earnings for women whose highest degree is a high school diploma—that is, $E(Earnings|Highest\ degree = high\ school\ diploma,\ Gender = female)$—is \$15.49 per hour.

The distribution of average hourly earnings for female college graduates (Figure 2.4b) is shifted to the right of the distribution for women with only a high school degree (Figure 2.4a); the same shift can be seen for the two groups of men (Figure 2.4d and Figure 2.4c). For both men and women, mean earnings are higher for those with a college degree (Table 2.4, first numeric column). Interestingly, the spread of the distribution of earnings, as measured by the standard deviation, is greater for those with a college degree than for those with a high school diploma. In addition, for both men and women, the 90th percentile of earnings is much higher for workers with a college degree than for workers with only a high school diploma. This final comparison is consistent with the parental admonition that a college degree opens doors that remain closed to individuals with only a high school diploma.

Another feature of these distributions is that the distribution of earnings for men is shifted to the right of the distribution of earnings for women. This "gender gap" in earnings is an important—and, to many, troubling—aspect of the distribution of earnings. We return to this topic in later chapters.

[1]The distributions were estimated using data from the March 2013 Current Population Survey, which is discussed in more detail in Appendix 3.1.

Means, Variances, and Covariances of Sums of Random Variables

KEY CONCEPT 2.3

Let X, Y, and V be random variables, let μ_X and σ_X^2 be the mean and variance of X, let σ_{XY} be the covariance between X and Y (and so forth for the other variables), and let a, b, and c be constants. Equations (2.29) through (2.35) follow from the definitions of the mean, variance, and covariance:

$$E(a + bX + cY) = a + b\mu_X + c\mu_Y, \tag{2.29}$$

$$\text{var}(a + bY) = b^2\sigma_Y^2, \tag{2.30}$$

$$\text{var}(aX + bY) = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \tag{2.31}$$

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2, \tag{2.32}$$

$$\text{cov}(a + bX + cV, Y) = b\sigma_{XY} + c\sigma_{VY}, \tag{2.33}$$

$$E(XY) = \sigma_{XY} + \mu_X\mu_Y, \tag{2.34}$$

$$|\text{corr}(X, Y)| \leq 1 \text{ and } |\sigma_{XY}| \leq \sqrt{\sigma_X^2\sigma_Y^2} \text{ (correlation inequality).} \tag{2.35}$$

The variance of the sum of X and Y is the sum of their variances plus two times their covariance:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}. \tag{2.36}$$

If X and Y are independent, then the covariance is zero and the variance of their sum is the sum of their variances:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) = \sigma_X^2 + \sigma_Y^2$$
$$\text{(if } X \text{ and } Y \text{ are independent).} \tag{2.37}$$

Useful expressions for means, variances, and covariances involving weighted sums of random variables are collected in Key Concept 2.3. The results in Key Concept 2.3 are derived in Appendix 2.1.

2.4 The Normal, Chi-Squared, Student *t*, and *F* Distributions

The probability distributions most often encountered in econometrics are the normal, chi-squared, Student *t*, and *F* distributions.

The Normal Distribution

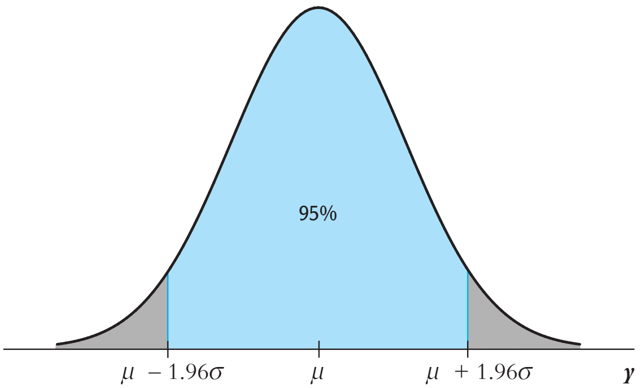
A continuous random variable with a **normal distribution** has the familiar bell-shaped probability density shown in Figure 2.5. The function defining the normal probability density is given in Appendix 17.1. As Figure 2.5 shows, the normal density with mean μ and variance σ^2 is symmetric around its mean and has 95% of its probability between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

Some special notation and terminology have been developed for the normal distribution. The normal distribution with mean μ and variance σ^2 is expressed concisely as "$N(\mu, \sigma^2)$." The **standard normal distribution** is the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ and is denoted $N(0, 1)$. Random variables that have a $N(0, 1)$ distribution are often denoted Z, and the standard normal cumulative distribution function is denoted by the Greek letter Φ; accordingly, $\Pr(Z \leq c) = \Phi(c)$, where c is a constant. Values of the standard normal cumulative distribution function are tabulated in Appendix Table 1.

To look up probabilities for a normal variable with a general mean and variance, we must **standardize the variable** by first subtracting the mean, then by dividing

FIGURE 2.5 The Normal Probability Density

The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.



Computing Probabilities Involving Normal Random Variables

KEY CONCEPT 2.4

Suppose Y is normally distributed with mean μ and variance σ^2; in other words, Y is distributed $N(\mu, \sigma^2)$. Then Y is standardized by subtracting its mean and dividing by its standard deviation, that is, by computing $Z = (Y - \mu)/\sigma$.

Let c_1 and c_2 denote two numbers with $c_1 < c_2$ and let $d_1 = (c_1 - \mu)/\sigma$ and $d_2 = (c_2 - \mu)/\sigma$. Then

$$\Pr(Y \leq c_2) = \Pr(Z \leq d_2) = \Phi(d_2), \tag{2.38}$$

$$\Pr(Y \geq c_1) = \Pr(Z \geq d_1) = 1 - \Phi(d_1), \tag{2.39}$$

$$\Pr(c_1 \leq Y \leq c_2) = \Pr(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1). \tag{2.40}$$

The normal cumulative distribution function Φ is tabulated in Appendix Table 1.

the result by the standard deviation. For example, suppose Y is distributed $N(1, 4)$—that is, Y is normally distributed with a mean of 1 and a variance of 4. What is the probability that $Y \leq 2$—that is, what is the shaded area in Figure 2.6a? The standardized version of Y is Y minus its mean, divided by its standard deviation, that is, $(Y - 1)/\sqrt{4} = \frac{1}{2}(Y - 1)$. Accordingly, the random variable $\frac{1}{2}(Y - 1)$ is normally distributed with mean zero and variance one (see Exercise 2.8); it has the standard normal distribution shown in Figure 2.6b. Now $Y \leq 2$ is equivalent to $\frac{1}{2}(Y - 1) \leq \frac{1}{2}(2 - 1)$—that is, $\frac{1}{2}(Y - 1) \leq \frac{1}{2}$. Thus,

$$\Pr(Y \leq 2) = \Pr[\tfrac{1}{2}(Y - 1) \leq \tfrac{1}{2}] = \Pr(Z \leq \tfrac{1}{2}) = \Phi(0.5) = 0.691, \tag{2.41}$$

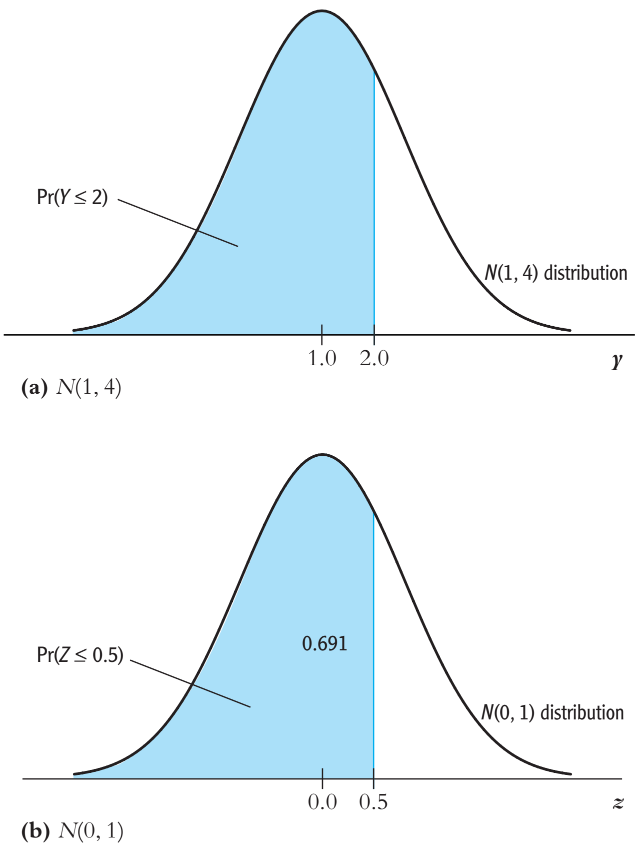
where the value 0.691 is taken from Appendix Table 1.

The same approach can be applied to compute the probability that a normally distributed random variable exceeds some value or that it falls in a certain range. These steps are summarized in Key Concept 2.4. The box "A Bad Day on Wall Street" presents an unusual application of the cumulative normal distribution.

The normal distribution is symmetric, so its skewness is zero. The kurtosis of the normal distribution is 3.

FIGURE 2.6 Calculating the Probability That $Y \leq 2$ When Y Is Distributed $N(1, 4)$

To calculate $\Pr(Y \leq 2)$, standardize Y, then use the standard normal distribution table. Y is standardized by subtracting its mean ($\mu = 1$) and dividing by its standard deviation ($\sigma = 2$). The probability that $Y \leq 2$ is shown in Figure 2.6a, and the corresponding probability after standardizing Y is shown in Figure 2.6b. Because the standardized random variable, $(Y - 1)/2$, is a standard normal (Z) random variable, $\Pr(Y \leq 2) = \Pr\left(\frac{Y-1}{2} \leq \frac{2-1}{2}\right) = \Pr(Z \leq 0.5)$. From Appendix Table 1, $\Pr(Z \leq 0.5) = \Phi(0.5) = 0.691$.



The multivariate normal distribution. The normal distribution can be generalized to describe the joint distribution of a set of random variables. In this case, the distribution is called the **multivariate normal distribution**, or, if only two variables are being considered, the **bivariate normal distribution**. The formula for the bivariate normal p.d.f. is given in Appendix 17.1, and the formula for the general multivariate normal p.d.f. is given in Appendix 18.1.

The multivariate normal distribution has four important properties. If X and Y have a bivariate normal distribution with covariance σ_{XY} and if a and b are two constants, then $aX + bY$ has the normal distribution:

$$aX + bY \text{ is distributed } N(a\mu_X + b\mu_Y,\ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})$$
$$(X, Y \text{ bivariate normal}). \tag{2.42}$$

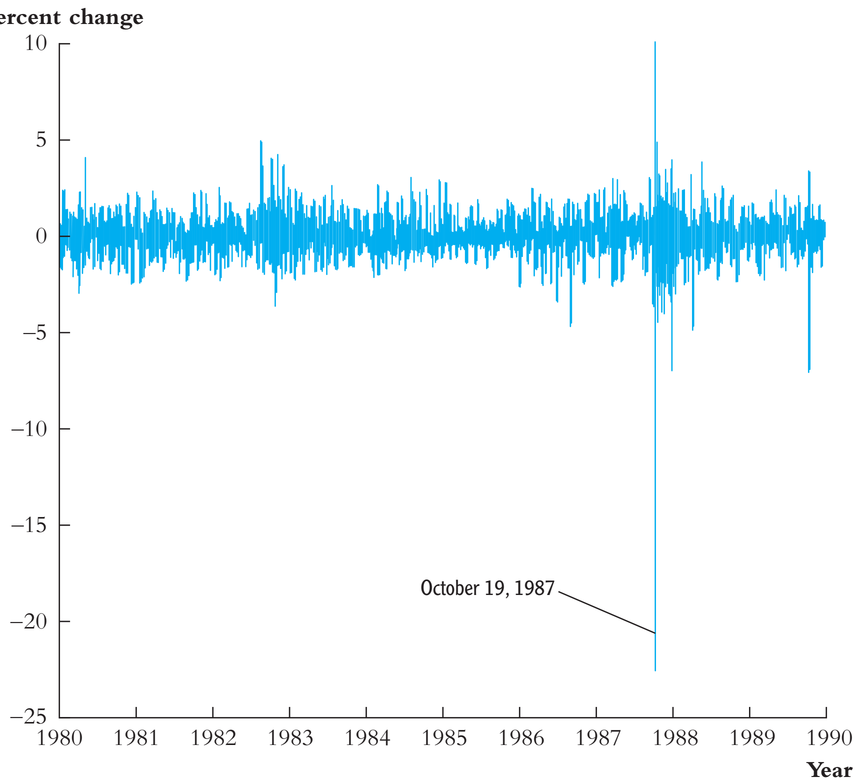
A Bad Day on Wall Street

On a typical day the overall value of stocks traded on the U.S. stock market can rise or fall by 1% or even more. This is a lot—but nothing compared to what happened on Monday, October 19, 1987. On "Black Monday," the Dow Jones Industrial Average (an average of 30 large industrial stocks) fell by 22.6%! From January 1, 1980, to December 31, 2012, the standard deviation of daily percentage price changes on the Dow was 1.12%, so the drop of 22.6% was a negative return of 20(= 22.6/1.12) standard deviations. The enormity of this drop can be seen in Figure 2.7, a plot of the daily returns on the Dow during the 1980s.

If daily percentage price changes are normally distributed, then the probability of a change of at least 20 standard deviations is $\Pr(|Z| \geq 20) = 2 \times \Phi(-20)$. You will not find this value in Appendix Table 1, but you can calculate it using a computer (try it!). This probability is 5.5×10^{-89}, that is, 0.000 . . . 00055, where there are a total of 88 zeros!

FIGURE 2.7 Daily Percentage Changes in the Dow Jones Industrial Average in the 1980s

From 1980 through 2012, the average percentage daily change of "the Dow" index was 0.04% and its standard deviation was 1.12%. On October 19, 1987—"Black Monday"—the Dow fell 22.6%, or more than 20 standard deviations.



continued on next page

How small is 5.5×10^{-89}? Consider the following:

- The world population is about 7 billion, so the probability of winning a random lottery among all living people is about one in 7 billion, or 1.4×10^{-10}.
- The universe is believed to have existed for 14 billion years, or about 5×10^{17} seconds, so the probability of choosing a particular second at random from all the seconds since the beginning of time is 2×10^{-18}.
- There are approximately 10^{43} molecules of gas in the first kilometer above the earth's surface. The probability of choosing one at random is 10^{-43}.

Although Wall Street *did* have a bad day, the fact that it happened at all suggests its probability was more than 5.5×10^{-89}. In fact, there have been many days—good and bad—with stock price changes too large to be consistent with a normal distribution with a constant variance. Table 2.5 lists the ten largest daily percentage price changes in the Dow Jones Industrial Average in the 8325 trading days between January 1, 1980, and December 31, 2012, along with the standardized change using the mean and variance over this period. All ten changes exceed 6.4 standard deviations, an extremely rare event if stock prices are normally distributed.

Clearly, stock price percentage changes have a distribution with heavier tails than the normal distribution. For this reason, finance professionals use other models of stock price changes. One such model treats stock price changes as normally distributed with a variance that evolves over time, so periods like October 1987 and the financial crisis in the fall of 2008 have higher volatility than others (models with time-varying variances are discussed in Chapter 16). Other models abandon the normal distribution in favor of distributions with heavier tails, an idea popularized in Nassim Taleb's 2007 book, *The Black Swan*. These models are more consistent with the very bad—and very good—days we actually see on Wall Street.

TABLE 2.5 The Ten Largest Daily Percentage Changes in the Dow Jones Industrial Index, 1980–2012, and the Normal Probability of a Change at Least as Large

Date	Percentage Change (x)	Standardized Change $Z = (x - \mu)/\sigma$	Normal Probability of a Change at Least This Large $\Pr(\lvert Z\rvert \geq z) = 2\Phi(-z)$
October 19, 1987	–22.6	–20.2	5.5×10^{-89}
October 13, 2008	11.1	9.9	6.4×10^{-23}
October 28, 2008	10.9	9.7	3.8×10^{-22}
October 21, 1987	10.1	9.0	1.8×10^{-19}
October 26, 1987	–8.0	–7.2	5.6×10^{-13}
October 15, 2008	–7.9	–7.1	1.6×10^{-12}
December 01, 2008	–7.7	–6.9	4.9×10^{-12}
October 09, 2008	–7.3	–6.6	4.7×10^{-11}
October 27, 1997	–7.2	–6.4	1.2×10^{-10}
September 17, 2001	–7.1	–6.4	1.6×10^{-10}

More generally, if n random variables have a multivariate normal distribution, then any linear combination of these variables (such as their sum) is normally distributed.

Second, if a set of variables has a multivariate normal distribution, then the marginal distribution of each of the variables is normal [this follows from Equation (2.42) by setting $a = 1$ and $b = 0$].

Third, if variables with a multivariate normal distribution have covariances that equal zero, then the variables are independent. Thus, if X and Y have a bivariate normal distribution and $\sigma_{XY} = 0$, then X and Y are independent. In Section 2.3 it was shown that if X and Y are independent, then, regardless of their joint distribution, $\sigma_{XY} = 0$. If X and Y are jointly normally distributed, then the converse is also true. This result—that zero covariance implies independence—is a special property of the multivariate normal distribution that is not true in general.

Fourth, if X and Y have a bivariate normal distribution, then the conditional expectation of Y given X is linear in X; that is, $E(Y|X = x) = a + bx$, where a and b are constants (Exercise 17.11). Joint normality implies linearity of conditional expectations, but linearity of conditional expectations does not imply joint normality.

The Chi-Squared Distribution

The chi-squared distribution is used when testing certain types of hypotheses in statistics and econometrics.

The **chi-squared distribution** is the distribution of the sum of m squared independent standard normal random variables. This distribution depends on m, which is called the degrees of freedom of the chi-squared distribution. For example, let Z_1, Z_2, and Z_3 be independent standard normal random variables. Then $Z_1^2 + Z_2^2 + Z_3^2$ has a chi-squared distribution with 3 degrees of freedom. The name for this distribution derives from the Greek letter used to denote it: A chi-squared distribution with m degrees of freedom is denoted χ_m^2.

Selected percentiles of the χ_m^2 distribution are given in Appendix Table 3. For example, Appendix Table 3 shows that the 95th percentile of the χ_m^2 distribution is 7.81, so $\Pr(Z_1^2 + Z_2^2 + Z_3^3 \leq 7.81) = 0.95$.

The Student *t* Distribution

The **Student *t* distribution** with m degrees of freedom is defined to be the distribution of the ratio of a standard normal random variable, divided by the square root of an independently distributed chi-squared random variable with m degrees of freedom divided by m. That is, let Z be a standard normal random variable, let W be a random variable with a chi-squared distribution with m degrees of freedom, and let Z and W be independently distributed. Then the random variable $Z/\sqrt{W/m}$ has a Student t distribution (also called the ***t* distribution**) with m degrees of freedom. This distribution is denoted t_m. Selected percentiles of the Student t distribution are given in Appendix Table 2.

The Student t distribution depends on the degrees of freedom m. Thus the 95th percentile of the t_m distribution depends on the degrees of freedom m. The Student t distribution has a bell shape similar to that of the normal distribution, but when m is small (20 or less), it has more mass in the tails—that is, it is a "fatter" bell shape than the normal. When m is 30 or more, the Student t distribution is well approximated by the standard normal distribution and the t_∞ distribution equals the standard normal distribution.

The *F* Distribution

The ***F* distribution** with m and n degrees of freedom, denoted $F_{m,n}$, is defined to be the distribution of the ratio of a chi-squared random variable with degrees of freedom m, divided by m, to an independently distributed chi-squared random variable with degrees of freedom n, divided by n. To state this mathematically, let W be a chi-squared random variable with m degrees of freedom and let V be a chi-squared random variable with n degrees of freedom, where W and V are independently distributed. Then $\frac{W/m}{V/n}$ has an $F_{m,n}$ distribution—that is, an F distribution with numerator degrees of freedom m and denominator degrees of freedom n.

In statistics and econometrics, an important special case of the F distribution arises when the denominator degrees of freedom is large enough that the $F_{m,n}$ distribution can be approximated by the $F_{m,\infty}$ distribution. In this limiting case, the denominator random variable V/n is the mean of infinitely many squared standard normal random variables, and that mean is 1 because the mean of a squared standard normal random variable is 1 (see Exercise 2.24). Thus the $F_{m,\infty}$ distribution is the distribution of a chi-squared random variable with m degrees of freedom, divided by m: W/m is distributed $F_{m,\infty}$. For example, from Appendix Table 4, the 95th percentile of the $F_{3,\infty}$ distribution is 2.60, which is the same as the 95th percentile of the χ^2_3 distribution, 7.81 (from Appendix Table 2), divided by the degrees of freedom, which is 3 ($7.81/3 = 2.60$).

The 90th, 95th, and 99th percentiles of the $F_{m,n}$ distribution are given in Appendix Table 5 for selected values of m and n. For example, the 95th percentile of the $F_{3,30}$ distribution is 2.92, and the 95th percentile of the $F_{3,90}$ distribution is 2.71. As the denominator degrees of freedom n increases, the 95th percentile of the $F_{3,n}$ distribution tends to the $F_{3,\infty}$ limit of 2.60.

2.5 Random Sampling and the Distribution of the Sample Average

Almost all the statistical and econometric procedures used in this book involve averages or weighted averages of a sample of data. Characterizing the distributions of sample averages therefore is an essential step toward understanding the performance of econometric procedures.

This section introduces some basic concepts about random sampling and the distributions of averages that are used throughout the book. We begin by discussing random sampling. The act of random sampling—that is, randomly drawing a sample from a larger population—has the effect of making the sample average itself a random variable. Because the sample average is a random variable, it has a probability distribution, which is called its sampling distribution. This section concludes with some properties of the sampling distribution of the sample average.

Random Sampling

Simple random sampling. Suppose our commuting student from Section 2.1 aspires to be a statistician and decides to record her commuting times on various days. She selects these days at random from the school year, and her daily commuting time has the cumulative distribution function in Figure 2.2a. Because these days were selected at random, knowing the value of the commuting time on one of these randomly selected days provides no information about the commuting time on another of the days; that is, because the days were selected at random, the values of the commuting time on each of the different days are independently distributed random variables.

The situation described in the previous paragraph is an example of the simplest sampling scheme used in statistics, called **simple random sampling**, in which n objects are selected at random from a **population** (the population of commuting days) and each member of the population (each day) is equally likely to be included in the sample.

The n observations in the sample are denoted $Y_1, \dots, Y_n$, where Y_1 is the first observation, Y_2 is the second observation, and so forth. In the commuting example, Y_1 is the commuting time on the first of her n randomly selected days and Y_i is the commuting time on the i^{th} of her randomly selected days.

Because the members of the population included in the sample are selected at random, the values of the observations $Y_1, \dots, Y_n$ are themselves random. If