



QUANTITATIVE METHODS, ECONOMICS

CFA[®] Program Curriculum
2024 • LEVEL 1 • VOLUME 1

©2023 by CFA Institute. All rights reserved. This copyright covers material written expressly for this volume by the editor/s as well as the compilation itself. It does not cover the individual selections herein that first appeared elsewhere. Permission to reprint these has been obtained by CFA Institute for this edition only. Further reproductions by any means, electronic or mechanical, including photocopying and recording, or by any information storage or retrieval systems, must be arranged with the individual copyright holders noted.

CFA®, Chartered Financial Analyst®, AIMR-PPS®, and GIPS® are just a few of the trademarks owned by CFA Institute. To view a list of CFA Institute trademarks and the Guide for Use of CFA Institute Marks, please visit our website at www.cfainstitute.org.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional service. If legal advice or other expert assistance is required, the services of a competent professional should be sought.

All trademarks, service marks, registered trademarks, and registered service marks are the property of their respective owners and are used herein for identification purposes only.

ISBN 978-1-953337-49-8 (paper)

ISBN 978-1-953337-23-8 (ebook)

May 2023

CONTENTS

How to Use the CFA Program Curriculum	xi
Errata	xi
Designing Your Personal Study Program	xi
CFA Institute Learning Ecosystem (LES)	xii
Feedback	xii
Quantitative Methods	
Learning Module 1	3
Rates and Returns	3
Introduction	3
Interest Rates and Time Value of Money	5
Determinants of Interest Rates	6
Rates of Return	8
Holding Period Return	8
Arithmetic or Mean Return	9
Geometric Mean Return	10
The Harmonic Mean	14
Money-Weighted and Time-Weighted Return	19
Calculating the Money Weighted Return	19
Annualized Return	28
Non-annual Compounding	29
Annualizing Returns	30
Continuously Compounded Returns	32
Other Major Return Measures and Their Applications	33
Gross and Net Return	33
Pre-Tax and After-Tax Nominal Return	34
Real Returns	34
Leveraged Return	36
<i>Practice Problems</i>	38
<i>Solutions</i>	42
Learning Module 2	45
The Time Value of Money in Finance	45
Introduction	45
Time Value of Money in Fixed Income and Equity	46
Fixed-Income Instruments and the Time Value of Money	47
Equity Instruments and the Time Value of Money	55
Implied Return and Growth	60
Implied Return for Fixed-Income Instruments	60
Equity Instruments, Implied Return, and Implied Growth	65
Cash Flow Additivity	69
Implied Forward Rates Using Cash Flow Additivity	71
Forward Exchange Rates Using No Arbitrage	74
Option Pricing Using Cash Flow Additivity	76
<i>Practice Problems</i>	81

<i>Solutions</i>	84
Learning Module 3	87
Statistical Measures of Asset Returns	87
Introduction	87
Measures of Central Tendency and Location	89
Measures of Central Tendency	90
Dealing with Outliers	92
Measures of Location	93
Measures of Dispersion	100
The Range	101
Mean Absolute Deviations	101
Sample Variance and Sample Standard Deviation	101
Downside Deviation and Coefficient of Variation	102
Measures of Shape of a Distribution	110
Correlation between Two Variables	117
Scatter Plot	117
Covariance and Correlation	118
Properties of Correlation	119
Limitations of Correlation Analysis	120
<i>Practice Problems</i>	126
<i>Solutions</i>	129
Learning Module 4	131
Probability Trees and Conditional Expectations	131
Introduction	131
Expected Value and Variance	132
Probability Trees and Conditional Expectations	134
Total Probability Rule for Expected Value	135
Bayes' Formula and Updating Probability Estimates	139
Bayes' Formula	140
<i>Practice Problems</i>	149
<i>Solutions</i>	150
Learning Module 5	151
Portfolio Mathematics	151
Introduction	151
Portfolio Expected Return and Variance of Return	153
Covariance	153
Correlation	156
Forecasting Correlation of Returns: Covariance Given a Joint Probability Function	161
Portfolio Risk Measures: Applications of the Normal Distribution	164
<i>Practice Problems</i>	171
<i>Solutions</i>	172
Learning Module 6	173
Simulation Methods	173
Introduction	173
Lognormal Distribution and Continuous Compounding	174
The Lognormal Distribution	174
Continuously Compounded Rates of Return	177

	Monte Carlo Simulation	180
	Bootstrapping	185
	<i>Practice Problems</i>	189
	<i>Solutions</i>	190
Learning Module 7	Estimation and Inference	191
	Introduction	191
	Sampling Methods	193
	Simple Random Sampling	193
	Stratified Random Sampling	194
	Cluster Sampling	195
	Non-Probability Sampling	196
	Sampling from Different Distributions	198
	Central Limit Theorem and Inference	201
	The Central Limit Theorem	201
	Standard Error of the Sample Mean	203
	Bootstrapping and Empirical Sampling Distributions	205
	<i>Practice Problems</i>	210
	<i>Solutions</i>	211
Learning Module 8	Hypothesis Testing	213
	Introduction	213
	Hypothesis Tests for Finance	215
	The Process of Hypothesis Testing	215
	Tests of Return and Risk in Finance	220
	Test Concerning Differences between Means with Dependent Samples	224
	Test Concerning the Equality of Two Variances	225
	Parametric versus Nonparametric Tests	232
	Uses of Nonparametric Tests	233
	Nonparametric Inference: Summary	233
	<i>Practice Problems</i>	234
	<i>Solutions</i>	238
Learning Module 9	Parametric and Non-Parametric Tests of Independence	241
	Introduction	241
	Tests Concerning Correlation	242
	Parametric Test of a Correlation	243
	Non-Parametric Test of Correlation: The Spearman Rank Correlation Coefficient	247
	Tests of Independence Using Contingency Table Data	251
	<i>Practice Problems</i>	259
	<i>Solutions</i>	260
Learning Module 10	Simple Linear Regression	261
	Introduction	261
	Estimation of the Simple Linear Regression Model	263
	Introduction to Linear Regression	263

Estimating the Parameters of a Simple Linear Regression	266
Assumptions of the Simple Linear Regression Model	273
Assumption 1: Linearity	273
Assumption 2: Homoskedasticity	275
Assumption 3: Independence	277
Assumption 4: Normality	278
Hypothesis Tests in the Simple Linear Regression Model	280
Analysis of Variance	280
Measures of Goodness of Fit	281
Hypothesis Testing of Individual Regression Coefficients	283
Prediction in the Simple Linear Regression Model	293
ANOVA and Standard Error of Estimate in Simple Linear Regression	293
Prediction Using Simple Linear Regression and Prediction Intervals	295
Functional Forms for Simple Linear Regression	300
The Log-Lin Model	301
The Lin-Log Model	302
The Log-Log Model	304
Selecting the Correct Functional Form	305
<i>Practice Problems</i>	308
<i>Solutions</i>	321
Learning Module 11	325
Introduction to Big Data Techniques	325
Introduction	325
How Is Fintech used in Quantitative Investment Analysis?	326
Big Data	327
Advanced Analytical Tools: Artificial Intelligence and Machine Learning	330
Tackling Big Data with Data Science	333
Data Processing Methods	333
Data Visualization	334
Text Analytics and Natural Language Processing	335
<i>Practice Problems</i>	337
<i>Solutions</i>	338
Learning Module 12	339
Appendices A–E	339
Appendices A–E	339
Economics	
Learning Module 1	351
Firms and Market Structures	351
Introduction	351
Profit Maximization: Production Breakeven, Shutdown and Economies of Scale	354
Profit-Maximization, Breakeven, and Shutdown Points of Production	355
Breakeven Analysis and Shutdown Decision	357
The Shutdown Decision	358
Economies and Diseconomies of Scale with Short-Run and Long-Run Cost Analysis	362
Introduction to Market Structures	368

Learning Module 2 Understanding Business Cycles Introduction Overview of the Business Cycle Phases of the Business Cycle Leads and Lags in Business and Consumer Decision Making Market Conditions and Investor Behavior Credit Cycles Applications of Credit Cycles Consequences for Policy Economic Indicators over the Business Cycle The Workforce and Company Costs Fluctuations in Capital Spending Fluctuations in Inventory Levels Economic Indicators Types of Indicators Composite Indicators Leading Indicators Using Economic Indicators Other Composite Leading Indicators Surveys The Use of Big Data in Economic Indicators Nowcasting GDPNow <i>Practice Problems</i> <i>Solutions</i>	368 373 374 375 375 376 377 377 378 380 382 384 388 389 390 393 396
Learning Module 3 Fiscal Policy Introduction Introduction to Monetary and Fiscal Policy Roles and Objectives of Fiscal Policy	427 427 428 431

Learning Module 4	Monetary Policy	453
	Introduction	453
	Role of Central Banks	454
	Roles of Central Banks and Objectives of Monetary Policy	455
	The Objectives of Monetary Policy	457
	Monetary Policy Tools and Monetary Transmission	459
	Open Market Operations	460
	The Central Bank's Policy Rate	460
	Reserve Requirements	461
	The Transmission Mechanism	461
	Monetary Policy Objectives	464
	Inflation Targeting	464
	Central Bank Independence	465
	Credibility	465
	Transparency	466
	The Bank of Japan	469
	The US Federal Reserve System	469
	Exchange Rate Targeting	471
	Contractionary and Expansionary Monetary Policies and Their Limitations	473
	What's the Source of the Shock to the Inflation Rate?	474
	Limitations of Monetary Policy	474
	Interaction of Monetary and Fiscal Policy	480
	The Relationship Between Monetary and Fiscal Policy	480
	Practice Problems	485
	Solutions	487
Learning Module 5	Introduction to Geopolitics	489
	Introduction	489
	National Governments and Political Cooperation	492
	State and Non-State Actors	492
	Features of Political Cooperation	493
	Resource Endowment, Standardization, and Soft Power	495
	The Role of Institutions	496
	Hierarchy of Interests and Costs of Cooperation	497

Power of the Decision Maker	497
Political Non-Cooperation	498
Forces of Globalization	500
Features of Globalization	501
Motivations for Globalization	503
Costs of Globalization and Threats of Rollback	504
Threats of Rollback of Globalization	506
International Trade Organizations	507
Role of the International Monetary Fund	508
World Bank Group and Developing Countries	510
World Trade Organization and Global Trade	511
Assessing Geopolitical Actors and Risk	514
Archetypes of Country Behavior	514
The Tools of Geopolitics	520
The Tools of Geopolitics	520
Multifaceted Approaches	524
Geopolitical Risk and Comparative Advantage	525
Geopolitical Risk and the Investment Process	526
Types of Geopolitical Risk	526
Assessing Geopolitical Threats	529
Impact of Geopolitical Risk	531
Tracking Risks According to Signposts	532
Manifestations of Geopolitical Risk	533
Acting on Geopolitical Risk	535
<i>Practice Problems</i>	537
<i>Solutions</i>	539
Learning Module 6	541
International Trade	
Introduction	541
Benefits and Costs of Trade	542
Benefits and Costs of International Trade	543
Trade Restrictions and Agreements—Tariffs, Quotas, and Export Subsidies	545
Tariffs	546
Quotas	548
Export Subsidies	549
Trading Blocs and Regional Integration	551
Types Of Trading Blocs	552
Regional Integration	553
<i>Practice Problems</i>	557
<i>Solutions</i>	559
Learning Module 7	561
Capital Flows and the FX Market	
Introduction	561
The Foreign Exchange Market and Exchange Rates	562
Introduction and the Foreign Exchange Market	562
Market Participants	569
Market Composition	572
Exchange Rate Quotations	575

Exchange Rate Regimes: Ideals and Historical Perspective	579
The Ideal Currency Regime	579
Historical Perspective on Currency Regimes	580
A Taxonomy of Currency Regimes	582
Exchange Rates and the Trade Balance: Introduction	590
Capital Restrictions	592
<i>Practice Problems</i>	595
<i>Solutions</i>	596
Learning Module 8	
Exchange Rate Calculations	597
Introduction	597
Cross-Rate Calculations	598
Forward Rate Calculations	602
Arbitrage Relationships	603
Forward Discounts and Premiums	606
<i>Practice Problems</i>	610
<i>Solutions</i>	612
Glossary	G-1

How to Use the CFA Program Curriculum

The CFA® Program exams measure your mastery of the core knowledge, skills, and abilities required to succeed as an investment professional. These core competencies are the basis for the Candidate Body of Knowledge (CBOK™). The CBOK consists of four components:

- A broad outline that lists the major CFA Program topic areas (www.cfainstitute.org/programs/cfa/curriculum/cbok)
- Topic area weights that indicate the relative exam weightings of the top-level topic areas (www.cfainstitute.org/programs/cfa/curriculum)
- Learning outcome statements (LOS) that advise candidates about the specific knowledge, skills, and abilities they should acquire from curriculum content covering a topic area: LOS are provided in candidate study sessions and at the beginning of each block of related content and the specific lesson that covers them. We encourage you to review the information about the LOS on our website (www.cfainstitute.org/programs/cfa/curriculum/study-sessions), including the descriptions of LOS “command words” on the candidate resources page at www.cfainstitute.org.
- The CFA Program curriculum that candidates receive upon exam registration

Therefore, the key to your success on the CFA exams is studying and understanding the CBOK. You can learn more about the CBOK on our website: www.cfainstitute.org/programs/cfa/curriculum/cbok.

The entire curriculum, including the practice questions, is the basis for all exam questions and is selected or developed specifically to teach the knowledge, skills, and abilities reflected in the CBOK.

ERRATA

The curriculum development process is rigorous and includes multiple rounds of reviews by content experts. Despite our efforts to produce a curriculum that is free of errors, there are instances where we must make corrections. Curriculum errata are periodically updated and posted by exam level and test date online on the Curriculum Errata webpage (www.cfainstitute.org/en/programs/submit-errata). If you believe you have found an error in the curriculum, you can submit your concerns through our curriculum errata reporting process found at the bottom of the Curriculum Errata webpage.

DESIGNING YOUR PERSONAL STUDY PROGRAM

An orderly, systematic approach to exam preparation is critical. You should dedicate a consistent block of time every week to reading and studying. Review the LOS both before and after you study curriculum content to ensure that you have mastered the

applicable content and can demonstrate the knowledge, skills, and abilities described by the LOS and the assigned reading. Use the LOS self-check to track your progress and highlight areas of weakness for later review.

Successful candidates report an average of more than 300 hours preparing for each exam. Your preparation time will vary based on your prior education and experience, and you will likely spend more time on some study sessions than on others.

CFA INSTITUTE LEARNING ECOSYSTEM (LES)

Your exam registration fee includes access to the CFA Program Learning Ecosystem (LES). This digital learning platform provides access, even offline, to all of the curriculum content and practice questions and is organized as a series of short online lessons with associated practice questions. This tool is your one-stop location for all study materials, including practice questions and mock exams, and the primary method by which CFA Institute delivers your curriculum experience. The LES offers candidates additional practice questions to test their knowledge, and some questions in the LES provide a unique interactive experience.

PREREQUISITE KNOWLEDGE

The CFA® Program assumes basic knowledge of Economics, Quantitative Methods, and Financial Statements as presented in introductory university-level courses in Statistics, Economics, and Accounting. CFA Level I candidates who do not have a basic understanding of these concepts or would like to review these concepts can study from any of the three pre-read volumes.

FEEDBACK

Please send any comments or feedback to info@cfainstitute.org, and we will review your suggestions carefully.

Quantitative Methods

LEARNING MODULE

1

Rates and Returns

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, David E. Runkle, PhD, CFA, and Vijay Singal, PhD, CFA.

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Jacobs Levy Equity Management (USA). Vijay Singal, PhD, CFA, is at Virginia Tech (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	interpret interest rates as required rates of return, discount rates, or opportunity costs and explain an interest rate as the sum of a real risk-free rate and premiums that compensate investors for bearing distinct types of risk
<input type="checkbox"/>	calculate and interpret different approaches to return measurement over time and describe their appropriate uses
<input type="checkbox"/>	compare the money-weighted and time-weighted rates of return and evaluate the performance of portfolios based on these measures
<input type="checkbox"/>	calculate and interpret annualized return measures and continuously compounded returns, and describe their appropriate uses
<input type="checkbox"/>	calculate and interpret major return measures and describe their appropriate uses

INTRODUCTION

1

Interest rates are a critical concept in finance. In some cases, we assume a particular interest rate and in others, the interest rate remains the unknown quantity to determine. Although the pre-reads have covered the mechanics of time value of money problems, here we first illustrate the underlying economic concepts by explaining the meaning and interpretation of interest rates and then calculate, interpret, and compare different return measures.

LEARNING MODULE OVERVIEW

- An interest rate, r , can have three interpretations: (1) a required rate of return, (2) a discount rate, or (3) an opportunity cost. An interest rate reflects the relationship between differently dated cash flows.
- An interest rate can be viewed as the sum of the real risk-free interest rate and a set of premiums that compensate lenders for bearing distinct types of risk: an inflation premium, a default risk premium, a liquidity premium, and a maturity premium.
- The nominal risk-free interest rate is approximated as the sum of the real risk-free interest rate and the inflation premium.
- A financial asset's total return consists of two components: an income yield consisting of cash dividends or interest payments, and a return reflecting the capital gain or loss resulting from changes in the price of the financial asset.
- A holding period return, R , is the return that an investor earns for a single, specified period of time (e.g., one day, one month, five years).
- Multiperiod returns may be calculated across several holding periods using different return measures (e.g., arithmetic mean, geometric mean, harmonic mean, trimmed mean, winsorized mean). Each return computation has special applications for evaluating investments.
- The choice of which of the various alternative measurements of mean to use for a given dataset depends on considerations such as the presence of extreme outliers, outliers that we want to include, whether there is a symmetric distribution, and compounding.
- A money-weighted return reflects the actual return earned on an investment after accounting for the value and timing of cash flows relating to the investment.
- A time-weighted return measures the compound rate of growth of one unit of currency invested in a portfolio during a stated measurement period. Unlike a money-weighted return, a time-weighted return is not sensitive to the timing and amount of cashflows and is the preferred performance measure for evaluating portfolio managers because cash withdrawals or additions to the portfolio are generally outside of the control of the portfolio manager.
- Interest may be paid or received more frequently than annually. The periodic interest rate and the corresponding number of compounding periods (e.g., quarterly, monthly, daily) should be adjusted to compute present and future values.
- Annualizing periodic returns allows investors to compare different investments across different holding periods to better evaluate and compare their relative performance. With the number of compounding periods per year approaching infinity, the interest is compounded continuously.
- Gross return, return prior to deduction of managerial and administrative expenses (those expenses not directly related to return generation), is an appropriate measure to evaluate the comparative performance of an asset manager.

- Net return, which is equal to the gross return less managerial and administrative expenses, is a better return measure of what an investor actually earned.
- The after-tax nominal return is computed as the total return minus any allowance for taxes on dividends, interest, and realized gains.
- Real returns are particularly useful in comparing returns across time periods because inflation rates may vary over time and are particularly useful for comparing investments across time periods and performance between different asset classes with different taxation.
- Leveraging a portfolio, via borrowing or futures, can amplify the portfolio's gains or losses.

INTEREST RATES AND TIME VALUE OF MONEY

2



interpret interest rates as required rates of return, discount rates, or opportunity costs and explain an interest rate as the sum of a real risk-free rate and premiums that compensate investors for bearing distinct types of risk

The time value of money establishes the equivalence between cash flows occurring on different dates. As cash received today is preferred to cash promised in the future, we must establish a consistent basis for this trade-off to compare financial instruments in cases in which cash is paid or received at different times. An **interest rate (or yield)**, denoted r , is a rate of return that reflects the relationship between differently dated – timed – cash flows. If USD 9,500 today and USD 10,000 in one year are equivalent in value, then $\text{USD } 10,000 - \text{USD } 9,500 = \text{USD } 500$ is the required compensation for receiving USD 10,000 in one year rather than now. The interest rate (i.e., the required compensation stated as a rate of return) is $\text{USD } 500/\text{USD } 9,500 = 0.0526$ or 5.26 percent.

Interest rates can be thought of in three ways:

- First, they can be considered *required rates of return*—that is, the minimum rate of return an investor must receive to accept an investment.
- Second, interest rates can be considered *discount rates*. In the previous example, 5.26 percent is the discount rate at which USD 10,000 in one year is equivalent to USD 9,500 today. Thus, we use the terms “interest rate” and “discount rate” almost interchangeably.
- Third, interest rates can be considered *opportunity costs*. An **opportunity cost** is the value that investors forgo by choosing a course of action. In the example, if the party who supplied USD 9,500 had instead decided to spend it today, he would have forgone earning 5.26 percent by consuming rather than saving. So, we can view 5.26 percent as the opportunity cost of current consumption.

Determinants of Interest Rates

Economics tells us that interest rates are set by the forces of supply and demand, where investors supply funds and borrowers demand their use. Taking the perspective of investors in analyzing market-determined interest rates, we can view an interest rate r as being composed of a real risk-free interest rate plus a set of premiums that are required returns or compensation for bearing distinct types of risk:

$$r = \text{Real risk-free interest rate} + \text{Inflation premium} + \text{Default risk premium} + \\ \text{Liquidity premium} + \text{Maturity premium. (1)}$$

- The **real risk-free interest rate** is the single-period interest rate for a completely risk-free security if no inflation were expected. In economic theory, the real risk-free rate reflects the time preferences of individuals for current versus future real consumption.
- The **inflation premium** compensates investors for expected inflation and reflects the average inflation rate expected over the maturity of the debt. Inflation reduces the purchasing power of a unit of currency—the amount of goods and services one can buy with it.
- The **default risk premium** compensates investors for the possibility that the borrower will fail to make a promised payment at the contracted time and in the contracted amount.
- The **liquidity premium** compensates investors for the risk of loss relative to an investment's fair value if the investment needs to be converted to cash quickly. US Treasury bills (T-bills), for example, do not bear a liquidity premium because large amounts of them can be bought and sold without affecting their market price. Many bonds of small issuers, by contrast, trade infrequently after they are issued; the interest rate on such bonds includes a liquidity premium reflecting the relatively high costs (including the impact on price) of selling a position.
- The **maturity premium** compensates investors for the increased sensitivity of the market value of debt to a change in market interest rates as maturity is extended, in general (holding all else equal). The difference between the interest rate on longer-maturity, liquid Treasury debt and that on short-term Treasury debt typically reflects a positive maturity premium for the longer-term debt (and possibly different inflation premiums as well).

The sum of the real risk-free interest rate and the inflation premium is the nominal risk-free interest rate:

The nominal risk-free interest rate reflects the combination of a real risk-free rate plus an inflation premium:

$$(1 + \text{nominal risk-free rate}) = (1 + \text{real risk-free rate})(1 + \text{inflation premium}).$$

In practice, however, the nominal rate is often approximated as the sum of the real risk-free rate plus an inflation premium:

$$\text{Nominal risk-free rate} = \text{Real risk-free rate} + \text{inflation premium.}$$

Many countries have short-term government debt whose interest rate can be considered to represent the nominal risk-free interest rate over that time horizon in that country. The French government issues BTs, or negotiable fixed-rate discount Treasury bills (Bons du Trésor à taux fixe et à intérêts précomptés), with maturities of up to one year. The Japanese government issues a short-term Treasury bill with maturities of 6 and 12 months. The interest rate on a 90-day US T-bill, for example, represents the nominal risk-free interest rate for the United States over the next three

months. Typically, interest rates are quoted in annual terms, so the interest rate on a 90-day government debt security quoted at 3 percent is the annualized rate and not the actual interest rate earned over the 90-day period.

Whether the interest rate we use is a required rate of return, or a discount rate, or an opportunity cost, the rate encompasses the real risk-free rate and a set of risk premia that depend on the characteristics of the cash flows. The foundational set of premia consist of inflation, default risk, liquidity risk, and maturity risk. All these premia vary over time and continuously change, as does the real risk-free rate. Consequently, all interest rates fluctuate, but how much they change depends on various economic fundamentals—and on the expectation of how these various economic fundamentals can change in the future.

EXAMPLE 1

Determining Interest Rates

Exhibit 1 presents selected information for five debt securities. All five investments promise only a single payment at maturity. Assume that premiums relating to inflation, liquidity, and default risk are constant across all time horizons.

Exhibit 1: Investments Alternatives and Their Characteristics

Investment	Maturity (in years)	Liquidity	Default Risk	Interest Rate (%)
1	2	High	Low	2.0
2	2	Low	Low	2.5
3	7	Low	Low	r_3
4	8	High	Low	4.0
5	8	Low	High	6.5

Based on the information in Exhibit 1, address the following:

1. Explain the difference between the interest rates offered by Investment 1 and Investment 2.

Solution:

Investment 2 is identical to Investment 1 except that Investment 2 has low liquidity. The difference between the interest rate on Investment 2 and Investment 1 is 0.5 percent. This difference in the two interest rates represents a liquidity premium, which represents compensation for the lower liquidity of Investment 2 (the risk of loss relative to an investment's fair value if the investment needs to be converted to cash quickly).

2. Estimate the default risk premium affecting all securities.

Solution:

To estimate the default risk premium, identify two investments that have the same maturity but different levels of default risk. Investments 4 and 5 both have a maturity of eight years but different levels of default risk. Investment 5, however, has low liquidity and thus bears a liquidity premium relative to Investment 4. From Part A, we know the liquidity premium is 0.5 percent. The difference between the interest rates offered by Investments 5 and 4 is 2.5 percent ($6.5\% - 4.0\%$), of which 0.5 percent is a liquidity premium. This

implies that 2.0 percent ($2.5\% - 0.5\%$) must represent a default risk premium reflecting Investment 5's relatively higher default risk.

3. Calculate upper and lower limits for the unknown interest rate for Investment 3, r_3 .

Solution:

Investment 3 has liquidity risk and default risk comparable to Investment 2, but with its longer time to maturity, Investment 3 should have a higher maturity premium and offer a higher interest rate than Investment 2. Therefore, the interest rate on Investment 3, r_3 , should thus be above 2.5 percent (the interest rate on Investment 2).

If the liquidity of Investment 3 was high, Investment 3 would match Investment 4 except for Investment 3's shorter maturity. We would then conclude that Investment 3's interest rate should be less than the interest rate offered by Investment 4, which is 4 percent. In contrast to Investment 4, however, Investment 3 has low liquidity. It is possible that the interest rate on Investment 3 exceeds that of Investment 4 despite Investment 3's shorter maturity, depending on the relative size of the liquidity and maturity premiums. However, we would expect r_3 to be less than 4.5 percent, the expected interest rate on Investment 4 if it had low liquidity ($4\% + 0.5\%$, the liquidity premium). Thus, we should expect the interest rate offered by Investment 3 to be between 2.5 percent and 4.5 percent.

3

RATES OF RETURN



calculate and interpret different approaches to return measurement over time and describe their appropriate uses

Financial assets are frequently defined in terms of their return and risk characteristics. Comparison along these two dimensions simplifies the process of building a portfolio from among all available assets. In this lesson, we will compute, evaluate, and compare various measures of return.

Financial assets normally generate two types of return for investors. First, they may provide periodic income through cash dividends or interest payments. Second, the price of a financial asset can increase or decrease, leading to a capital gain or loss.

Some financial assets provide return through only one of these mechanisms. For example, investors in non-dividend-paying stocks obtain their return from price movement only. Other assets only generate periodic income. For example, defined benefit pension plans and retirement annuities make income payments over the life of a beneficiary.

Holding Period Return

Returns can be measured over a single period or over multiple periods. Single-period returns are straightforward because there is only one way to calculate them. Multiple-period returns, however, can be calculated in various ways and it is important to be aware of these differences to avoid confusion.

A **holding period return**, R , is the return earned from holding an asset for a single specified period of time. The period may be one day, one week, one month, five years, or any specified period. If the asset (e.g., bond, stock) is purchased today, time ($t = 0$), at a price of 100 and sold later, say at time ($t = 1$), at a price of 105 with no dividends or other income, then the holding period return is 5 percent $[(105 - 100)/100]$. If the asset also pays income of two units at time ($t = 1$), then the total return is 7 percent. This return can be generalized and shown as a mathematical expression in which P is the price and I is the income, as follows:

$$R = \frac{(P_1 - P_0) + I_1}{P_0}, \quad (1)$$

where the subscript indicates the time of the price or income; ($t = 0$) is the beginning of the period; and ($t = 1$) is the end of the period. The following two observations are important.

- We computed a capital gain of 5 percent and an income yield of 2 percent in this example. For ease of illustration, we assumed that the income is paid at time $t = 1$. If the income was received before $t = 1$, our holding period return may have been higher if we had reinvested the income for the remainder of the period.
- Return can be expressed in decimals (0.07), fractions (7/100), or as a percent (7 percent). They are all equivalent.

A holding period return can be computed for a period longer than one year. For example, an analyst may need to compute a one-year holding period return from three annual returns. In that case, the one-year holding period return is computed by compounding the three annual returns:

$$R = [(1 + R_1) \times (1 + R_2) \times (1 + R_3)] - 1,$$

where R_1 , R_2 , and R_3 are the three annual returns.

Arithmetic or Mean Return

Most holding period returns are reported as daily, monthly, or annual returns. When assets have returns for multiple holding periods, it is necessary to normalize returns to a common period for ease of comparison and understanding. There are different methods for aggregating returns across several holding periods. The remainder of this section presents various ways of computing average returns and discusses their applicability.

The simplest way to compute a summary measure for returns across multiple periods is to take a simple arithmetic average of the holding period returns. Thus, three annual returns of -50 percent, 35 percent, and 27 percent will give us an average of 4 percent per year $= \left(\frac{-50\% + 35\% + 27\%}{3}\right)$. The arithmetic average return is easy to compute and has known statistical properties.

In general, the arithmetic or mean return is denoted by \bar{R}_i and given by the following equation for asset i , where R_{it} is the return in period t and T is the total number of periods:

$$\bar{R}_i = \frac{R_{i1} + R_{i2} + \dots + R_{iT-1} + R_{iT}}{T} = \frac{1}{T} \sum_{t=1}^T R_{it}. \quad (2)$$

Geometric Mean Return

The arithmetic mean return assumes that the amount invested at the beginning of each period is the same. In an investment portfolio, however, even if there are no cash flows into or out of the portfolio the base amount changes each year. The previous year's earnings must be added to the beginning value of the subsequent year's investment—these earnings will be “compounded” by the returns earned in that subsequent year. We can use the geometric mean return to account for the compounding of returns.

A geometric mean return provides a more accurate representation of the growth in portfolio value over a given time period than the arithmetic mean return. In general, the geometric mean return is denoted by \bar{R}_{Gi} and given by the following equation for asset i :

$$\begin{aligned}\bar{R}_{Gi} &= \sqrt[T]{(1 + R_{i1}) \times (1 + R_{i2}) \times \dots \times (1 + R_{iT-1}) \times (1 + R_{iT})} - 1 \\ &= \sqrt[T]{\prod_{t=1}^T (1 + R_t)} - 1 ,\end{aligned}\quad (3)$$

where R_{it} is the return in period t and T is the total number of periods.

In the example in the previous section, we calculated the arithmetic mean to be 4.00 percent. Using Equation 4, we can calculate the geometric mean return from the same three annual returns:

$$\bar{R}_{Gi} = \sqrt[3]{(1 - 0.50) \times (1 + 0.35) \times (1 + 0.27)} - 1 = -0.0500.$$

Exhibit 2 shows the actual return for each year and the balance at the end of each year using actual returns.

Exhibit 2: Portfolio Value and Performance

	Actual Return for the Year (%)	Year-End Amount	Year-End Amount Using Arithmetic Return of 4%	Year-End Amount Using Geometric Return of -5%
Year 0		EUR1.0000	EUR1.0000	EUR1.0000
Year 1	-50	0.5000	1.0400	0.9500
Year 2	35	0.6750	1.0816	0.9025
Year 3	27	0.8573	1.1249	0.8574

Beginning with an initial investment of EUR1.0000, we will have a balance of EUR0.8573 at the end of the three-year period as shown in the fourth column of Exhibit 2. Note that we compounded the returns because, unless otherwise stated, we earn a return on the balance as of the end of the prior year. That is, we will receive a return of 35 percent in the second year on the balance at the end of the first year, which is only EUR0.5000, not the initial balance of EUR1.0000. Let us compare the balance at the end of the three-year period computed using geometric returns with the balance we would calculate using the 4 percent annual arithmetic mean return from our earlier example. The ending value using the arithmetic mean return is EUR1.1249 ($=1.0000 \times 1.04^3$). This is much larger than the actual balance at the end of Year 3 of EUR0.8573.

In general, the arithmetic return is biased upward unless each of the underlying holding period returns are equal. The bias in arithmetic mean returns is particularly severe if holding period returns are a mix of both positive and negative returns, as in this example.

We will now look at three examples that calculate holding period returns over different time horizons.

EXAMPLE 2**Holding Period Return**

1. An investor purchased 100 shares of a stock for USD34.50 per share at the beginning of the quarter. If the investor sold all of the shares for USD30.50 per share after receiving a USD51.55 dividend payment at the end of the quarter, the investor's holding period return is *closest* to:

- A. -13.0 percent.
- B. -11.6 percent.
- C. -10.1 percent.

Solution:

C is correct. Applying Equation 2, the holding period return is -10.1 percent, calculated as follows:

$$R = (3,050 - 3,450 + 51.55)/3,450 = -10.1\%.$$

The holding period return comprised of a dividend yield of 1.49 percent ($= 51.55/3,450$) and a capital loss of -11.59 percent ($= -400/3,450$).

EXAMPLE 3**Holding Period Return**

1. An analyst obtains the following annual rates of return for a mutual fund, which are presented in Exhibit 3.

Exhibit 3: Mutual Fund Performance, 20X8–20X0

Year	Return (%)
20X8	14
20X9	-10
20X0	-2

The fund's holding period return over the three-year period is *closest* to:

- A. 0.18 percent.
- B. 0.55 percent.
- C. 0.67 percent.

Solution:

B is correct. The fund's three-year holding period return is 0.55 percent, calculated as follows:

$$R = [(1 + R_1) \times (1 + R_2) \times (1 + R_3)] - 1,$$

$$R = [(1 + 0.14)(1 - 0.10)(1 - 0.02)] - 1 = 0.0055 = 0.55\%.$$

EXAMPLE 4**Geometric Mean Return**

- An analyst observes the following annual rates of return for a hedge fund, which are presented in Exhibit 4.

Exhibit 4: Hedge Fund Performance, 20X8–20X0

Year	Return (%)
20X8	22
20X9	-25
20X0	11

The fund's geometric mean return over the three-year period is *closest* to:

- A. 0.52 percent.
- B. 1.02 percent.
- C. 2.67 percent.

Solution:

A is correct. Applying Equation 4, the fund's geometric mean return over the three-year period is 0.52 percent, calculated as follows:

$$\bar{R}_G = [(1 + 0.22)(1 - 0.25)(1 + 0.11)]^{(1/3)} - 1 = 1.0157^{(1/3)} - 1 = 0.0052 \\ = 0.52\%.$$

EXAMPLE 5**Geometric and Arithmetic Mean Returns**

- Consider the annual return data for the group of countries in Exhibit 5.

Exhibit 5: Annual Returns for Years 1 to 3 for Selected Countries' Stock Indexes

Index	52-Week Return (%)			Average 3-Year Return	
	Year 1	Year 2	Year 3	Arithmetic	Geometric
Country A	-15.6	-5.4	6.1	-4.97	-5.38
Country B	7.8	6.3	-1.5	4.20	4.12
Country C	5.3	1.2	3.5	3.33	3.32
Country D	-2.4	-3.1	6.2	0.23	0.15
Country E	-4.0	-3.0	3.0	-1.33	-1.38
Country F	5.4	5.2	-1.0	3.20	3.16
Country G	12.7	6.7	-1.2	6.07	5.91
Country H	3.5	4.3	3.4	3.73	3.73
Country I	6.2	7.8	3.2	5.73	5.72

Index	52-Week Return (%)			Average 3-Year Return	
	Year 1	Year 2	Year 3	Arithmetic	Geometric
Country J	8.1	4.1	-0.9	3.77	3.70
Country K	11.5	3.4	1.2	5.37	5.28

Calculate the arithmetic and geometric mean returns over the three years for the following three stock indexes: Country D, Country E, and Country F.

Solution:

The arithmetic mean returns are as follows:

	Annual Return (%)			Sum $\sum_{i=1}^3 R_i$	Arithmetic Mean Return (%)
	Year 1	Year 2	Year 3		
Country D	-2.4	-3.1	6.2	0.7	0.233
Country E	-4.0	-3.0	3.0	-4.0	-1.333
Country F	5.4	5.2	-1.0	9.6	3.200

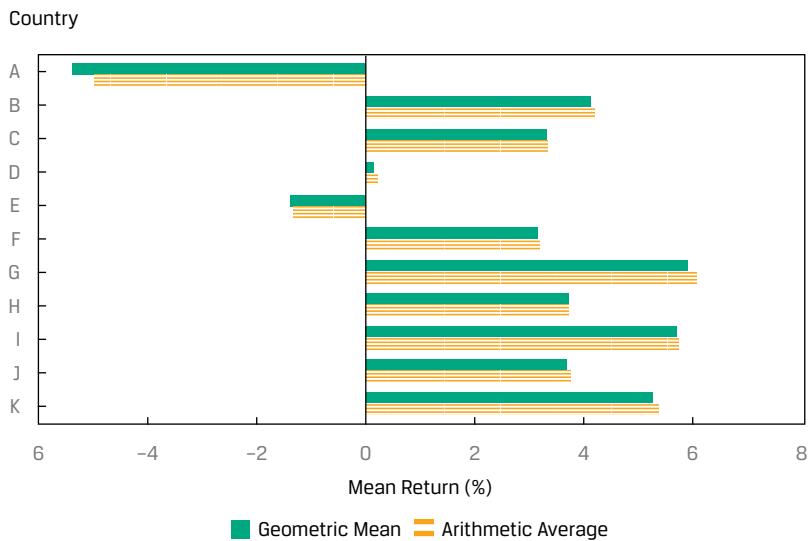
The geometric mean returns are as follows:

	1 + Return in Decimal Form			Product $\prod_t^T (1 + R_t)$	3rd root $\left[\prod_t^T (1 + R_t) \right]^{1/3}$	Geometric mean return (%)
	(1 + R_t)	Year 1	Year 2			
Country D	0.976	0.969	1.062	1.00438	1.00146	0.146
Country E	0.960	0.970	1.030	0.95914	0.98619	-1.381
Country F	1.054	1.052	0.990	1.09772	1.03157	3.157

In Example 5, the geometric mean return is less than the arithmetic mean return for each country's index returns. In fact, the geometric mean is always less than or equal to the arithmetic mean with one exception: the two means will be equal if there is no variability in the observations—that is, when all the observations in the series are the same.

In general, the difference between the arithmetic and geometric means increases with the variability within the sample; the more disperse the observations, the greater the difference between the arithmetic and geometric means. Casual inspection of the returns in Exhibit 5 and the associated graph of means in Exhibit 6 suggests a greater variability for Country A's index relative to the other indexes, and this is confirmed with the greater deviation of the geometric mean return (-5.38 percent) from the arithmetic mean return (-4.97 percent). How should the analyst interpret these results?

Exhibit 6: Arithmetic and Geometric Mean Returns for Country Stock Indexes, Years 1 to 3



The geometric mean return represents the growth rate or compound rate of return on an investment. One unit of currency invested in a fund tracking the Country B index at the beginning of Year 1 would have grown to $(1.078)(1.063)(0.985) = 1.128725$ units of currency, which is equal to 1 plus Country B's geometric mean return of 4.1189 percent compounded over three periods: $[1 + 0.041189]^3 = 1.128725$. This math confirms that the geometric mean is the compound rate of return. With its focus on the actual return of an investment over a multiple-period horizon, the geometric mean is of key interest to investors. The arithmetic mean return, focusing on average single-period performance, is also of interest. Both arithmetic and geometric means have a role to play in investment management, and both are often reported for return series.

For reporting historical returns, the geometric mean has considerable appeal because it is the rate of growth or return we would have to earn each year to match the actual, cumulative investment performance. Suppose we purchased a stock for EUR100 and two years later it was worth EUR100, with an intervening year at EUR200. The geometric mean of 0 percent is clearly the compound rate of growth during the two years, which we can confirm by compounding the returns: $[(1 + 1.00)(1 - 0.50)]^{1/2} - 1 = 0\%$. Specifically, the ending amount is the beginning amount times $(1 + R_G)^2$.

However, the arithmetic mean, which is $[100\% + -50\%]/2 = 25\%$ in the previous example, can distort our assessment of historical performance. As we noted, the arithmetic mean is always greater than or equal to the geometric mean. If we want to estimate the average return over a one-period horizon, we should use the arithmetic mean because the arithmetic mean is the average of one-period returns. If we want to estimate the average returns over more than one period, however, we should use the geometric mean of returns because the geometric mean captures how the total returns are linked over time.

The Harmonic Mean

The **harmonic mean**, X_H , is another measure of central tendency. The harmonic mean is appropriate in cases in which the variable is a rate or a ratio. The terminology "harmonic" arises from its use of a type of series involving reciprocals known as a harmonic series.

Rates of Return

Harmonic Mean Formula. The harmonic mean of a set of observations X_1, X_2, \dots, X_n is:

$$\bar{X}_H = \frac{n}{\sum_{i=1}^n (1/X_i)}, \quad (4)$$

with $X_i > 0$ for $i = 1, 2, \dots, n$.

The harmonic mean is the value obtained by summing the reciprocals of the observations,

$$\sum_{i=1}^n (1/X_i),$$

the terms of the form $1/X_i$, and then averaging their sum by dividing it by the number of observations, n , and, then finally, taking the reciprocal of that average,

$$\frac{n}{\sum_{i=1}^n (1/X_i)}.$$

The harmonic mean may be viewed as a special type of weighted mean in which an observation's weight is inversely proportional to its magnitude. For example, if there is a sample of observations of 1, 2, 3, 4, 5, 6, and 1,000, the harmonic mean is 2.8560. Compared to the arithmetic mean of 145.8571, we see the influence of the outlier (the 1,000) to be much less than in the case of the arithmetic mean. So, the harmonic mean is quite useful as a measure of central tendency in the presence of outliers.

The harmonic mean is used most often when the data consist of rates and ratios, such as P/Es. Suppose three peer companies have P/Es of 45, 15, and 15. The arithmetic mean is 25, but the harmonic mean, which gives less weight to the P/E of 45, is 19.3.

The harmonic mean is a relatively specialized concept of the mean that is appropriate for averaging ratios ("amount per unit") when the ratios are repeatedly applied to a fixed quantity to yield a variable number of units. The concept is best explained through an illustration. A well-known application arises in the investment strategy known as **cost averaging**, which involves the periodic investment of a fixed amount of money. In this application, the ratios we are averaging are prices per share at different purchase dates, and we are applying those prices to a constant amount of money to yield a variable number of shares. An illustration of the harmonic mean to cost averaging is provided in Example 6.

EXAMPLE 6**Cost Averaging and the Harmonic Mean**

- Suppose an investor invests EUR1,000 each month in a particular stock for $n = 2$ months. The share prices are EUR10 and EUR15 at the two purchase dates. What was the average price paid for the security?

Solution:

Purchase in the first month = EUR1,000/EUR10 = 100 shares.

Purchase in the second month = EUR1,000/EUR15 = 66.67 shares.

The investor purchased a total of 166.67 shares for EUR2,000, so the average price paid per share is EUR2,000/166.67 = EUR12.

The average price paid is in fact the harmonic mean of the asset's prices at the purchase dates. Using Equation 5, the harmonic mean price is $2/[(1/10) + (1/15)] = \text{EUR}12$. The value EUR12 is less than the arithmetic mean purchase price ($\text{EUR}10 + \text{EUR}15)/2 = \text{EUR}12.5$.

Because they use the same data but involve different progressions in their respective calculations, the arithmetic, geometric, and harmonic means are mathematically related to one another. We will not go into the proof of this relationship, but the basic result follows:

$$\text{Arithmetic mean} \times \text{Harmonic mean} = (\text{Geometric mean})^2.$$

Unless all the observations in a dataset are the same value, the harmonic mean is always less than the geometric mean, which, in turn, is always less than the arithmetic mean.

EXAMPLE 7

Calculating the Arithmetic, Geometric, and Harmonic Means for P/Es

Each year in December, a securities analyst selects her 10 favorite stocks for the next year. Exhibit 7 presents the P/Es, the ratio of share price to projected earnings per share (EPS), for her top 10 stock picks for the next year.

Exhibit 7: Analyst's 10 Favorite Stocks for Next Year

Stock	P/E
Stock 1	22.29
Stock 2	15.54
Stock 3	9.38
Stock 4	15.12
Stock 5	10.72
Stock 6	14.57
Stock 7	7.20
Stock 8	7.97
Stock 9	10.34
Stock 10	8.35

1. Calculate the arithmetic mean P/E for these 10 stocks.

Solution:

The arithmetic mean is calculated as:

$$121.48/10 = 12.1480.$$

2. Calculate the geometric mean P/E for these 10 stocks.

Solution:

The geometric mean P/E is calculated as:

$$\begin{aligned}\frac{P}{E_{Gi}} &= \sqrt[10]{\frac{P}{E_1} \times \frac{P}{E_2} \times \dots \times \frac{P}{E_9} \times \frac{P}{E_{10}}} \\ &= \sqrt[10]{22.29 \times 15.54 \dots \times 10.34 \times 8.35} \\ &= \sqrt[10]{38,016,128,040} = 11.4287.\end{aligned}$$

The geometric mean is 11.4287. This result can also be obtained as:

$$\frac{P}{E_{Gi}} = e^{\frac{\ln(22.29 \times 15.54 \dots \times 10.34 \times 8.35)}{10}} = e^{\frac{\ln(38,016,128,040)}{10}} = e^{24.3613/10} = 11.4287.$$

3. Calculate the harmonic mean P/E for the 10 stocks.

Solution:

The harmonic mean is calculated as:

$$\bar{X}_H = \frac{n}{\sum_{i=1}^n (1/X_i)},$$

$$\bar{X}_H = \frac{10}{\left(\frac{1}{22.29}\right) + \left(\frac{1}{15.54}\right) + \dots + \left(\frac{1}{10.34}\right) + \left(\frac{1}{8.35}\right)},$$

$$\bar{X}_H = 10/0.9247 = 10.8142.$$

In finance, the weighted harmonic mean is used when averaging rates and other multiples, such as the P/E ratio, because the harmonic mean gives equal weight to each data point, and reduces the influence of outliers.

These calculations can be performed using Excel:

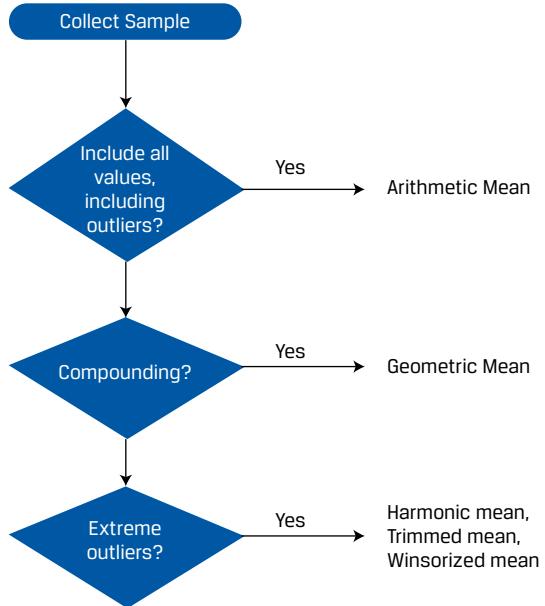
- To calculate the arithmetic mean or average return, the `=AVERAGE(return1, return2, ...)` function can be used.
- To calculate the geometric mean return, the `=GEOMEAN(return1, return2, ...)` function can be used.
- To calculate the harmonic mean return, the `=HARMEAN(return1, return2, ...)` function can be used.

In addition to arithmetic, geometric, and harmonic means, two other types of means can be used. Both the trimmed and the winsorized means seek to minimize the impact of outliers in a dataset. Specifically, the **trimmed mean** removes a small defined percentage of the largest and smallest values from a dataset containing our observation before calculating the mean by averaging the remaining observations.

A winsorized mean replaces the extreme observations in a dataset to limit the effect of the outliers on the calculations. The **winsorized mean** is calculated after replacing extreme values at both ends with the values of their nearest observations, and then calculating the mean by averaging the remaining observations.

However, the key question is: Which mean to use in what circumstances? The choice of which mean to use depends on many factors, as we describe in Exhibit 8:

- Are there outliers that we want to include?
- Is the distribution symmetric?
- Is there compounding?
- Are there extreme outliers?

Exhibit 8: Deciding Which Measure to Use**QUESTION SET**

A fund had the following returns over the past 10 years:

Exhibit 9: 10-Year Returns

Year	Return
1	4.5%
2	6.0%
3	1.5%
4	-2.0%
5	0.0%
6	4.5%
7	3.5%
8	2.5%
9	5.5%
10	4.0%

1. The arithmetic mean return over the 10 years is *closest* to:

- A. 2.97 percent.
- B. 3.00 percent.
- C. 3.33 percent.

Solution:

B is correct. The arithmetic mean return is calculated as follows:

$$\bar{R} = 30.0\%/10 = 3.0\%.$$

2. The geometric mean return over the 10 years is *closest* to:

- A. 2.94 percent.
- B. 2.97 percent.
- C. 3.00 percent.

Solution:

B is correct. The geometric mean return is calculated as follows:

$$\bar{R}_G = \sqrt[10]{(1 + 0.045) \times (1 + 0.06) \times \dots \times (1 + 0.055) \times (1 + 0.04)} - 1$$

$$\bar{R}_G = \sqrt[10]{1.3402338} - 1 = 2.9717\%.$$

MONEY-WEIGHTED AND TIME-WEIGHTED RETURN

4



compare the money-weighted and time-weighted rates of return and evaluate the performance of portfolios based on these measures

The arithmetic and geometric return computations do not account for the timing of cash flows into and out of a portfolio. For example, suppose an investor experiences the returns shown in Exhibit 2. Instead of only investing EUR1.0 at the start (Year 0) as was the case in Exhibit 2, suppose the investor had invested EUR10,000 at the start, EUR1,000 in Year 1, and EUR1,000 in Year 2. In that case, the return of –50 percent in Year 1 significantly hurts her given the relatively large investment at the start. Conversely, if she had invested only EUR100 at the start, the absolute effect of the –50 percent return on the total return is drastically reduced.

Calculating the Money Weighted Return

The **money-weighted return** accounts for the money invested and provides the investor with information on the actual return she earns on her investment. The money-weighted return and its calculation are similar to the internal rate of return and a bond's yield to maturity. Amounts invested are cash outflows from the investor's perspective and amounts returned or withdrawn by the investor, or the money that remains at the end of an investment cycle, is a cash inflow for the investor.

For example, assume that an investor invests EUR100 in a mutual fund at the beginning of the first year, adds another EUR950 at the beginning of the second year, and withdraws EUR350 at the end of the second year. The cash flows are presented in Exhibit 10.

Exhibit 10: Portfolio Balances across Three Years

Year	1	2	3
Balance from previous year	EUR0	EUR50	EUR1,000
New investment by the investor (cash inflow for the mutual fund) at the start of the year	100	950	0
Net balance at the beginning of year	100	1,000	1,000
Investment return for the year	-50%	35%	27%
Investment gain (loss)	-50	350	270
Withdrawal by the investor (cash outflow for the mutual fund) at the end of the year	0	-350	0
Balance at the end of year	EUR50	EUR1,000	EUR1,270

The **internal rate of return** is the discount rate at which the sum of present values of cash flows will equal zero. In general, the equation may be expressed as follows:

$$\sum_{t=0}^T \frac{CF_t}{(1 + IRR)^t} = 0, \quad (5)$$

where T is the number of periods, CF_t is the cash flow at time t , and IRR is the internal rate of return or the money-weighted rate of return.

A cash flow can be positive or negative; a positive cash flow is an inflow where money flows to the investor, whereas a negative cash flow is an outflow where money flows away from the investor. The cash flows are expressed as follows, where each cash inflow or outflow occurs at the end of each year. Thus, CF_0 refers to the cash flow at the end of Year 0 or beginning of Year 1, and CF_3 refers to the cash flow at end of Year 3 or beginning of Year 4. Because cash flows are being discounted to the present—that is, end of Year 0 or beginning of Year 1—the period of discounting CF_0 is zero.

$$CF_0 = -100$$

$$CF_1 = -950$$

$$CF_2 = +350$$

$$CF_3 = +1,270$$

$$\frac{CF_0}{(1 + IRR)^0} + \frac{CF_1}{(1 + IRR)^1} + \frac{CF_2}{(1 + IRR)^2} + \frac{CF_3}{(1 + IRR)^3} \\ = \frac{-100}{1} + \frac{-950}{(1 + IRR)^1} + \frac{+350}{(1 + IRR)^2} + \frac{+1270}{(1 + IRR)^3} = 0$$

$$IRR = 26.11\%$$

The investor's internal rate of return, or the money-weighted rate of return, is 26.11 percent, which tells the investor what she earned on the actual euros invested for the entire period on an annualized basis. This return is much greater than the arithmetic and geometric mean returns because only a small amount was invested when the mutual fund's return was -50 percent.

All the above calculations can be performed using Excel using the =IRR(values) function, which results in an IRR of 26.11 percent.

Money-Weighted Return for a Dividend-Paying Stock

Next, we'll illustrate calculating the money-weighted return for a dividend paying stock. Consider an investment that covers a two-year horizon. At time $t = 0$, an investor buys one share at a price of USD200. At time $t = 1$, he purchases an additional share at a price of USD225. At the end of Year 2, $t = 2$, he sells both shares at a price of USD235. During both years, the stock pays a dividend of USD5 per share. The $t = 1$ dividend is not reinvested. Exhibit 11 outlines the total cash inflows and outflows for the investment.

Exhibit 11: Cash Flows for a Dividend-Paying Stock

Time	Outflows
0	USD200 to purchase the first share
1	USD225 to purchase the second share
Time	Inflows
1	USD5 dividend received from first share (and not reinvested)
2	USD10 dividend (USD5 per share × 2 shares) received
2	USD470 received from selling two shares at USD235 per share

To solve for the money-weighted return, the first step is to group net cash flows by time. For this example, we have –USD200 for the $t = 0$ net cash flow, $-\text{USD}220 = -\text{USD}225 + \text{USD}5$ for the $t = 1$ net cash flow, and USD480 for the $t = 2$ net cash flow. After entering these cash flows, we use the spreadsheet's (such as Excel) or calculator's IRR function to find that the money-weighted rate of return is 9.39 percent.

$$\text{CF}_0 = -200$$

$$\text{CF}_1 = -220$$

$$\text{CF}_2 = +480$$

$$\frac{\text{CF}_0}{(1 + \text{IRR})^0} + \frac{\text{CF}_1}{(1 + \text{IRR})^1} + \frac{\text{CF}_2}{(1 + \text{IRR})^2} . \\ = \frac{-200}{1} + \frac{-220}{(1 + \text{IRR})^1} + \frac{480}{(1 + \text{IRR})^2} = 0$$

$$\text{IRR} = 9.39\%$$

All these calculations can be performed using Excel using the =IRR(values) function, which results in an IRR of 9.39 percent.

Now we take a closer look at what has happened to the portfolio during each of the two years.

In the first year, the portfolio generated a one-period holding period return of $(\text{USD}5 + \text{USD}225 - \text{USD}200)/\text{USD}200 = 15\%$. At the beginning of the second year, the amount invested is USD450, calculated as USD225 (per share price of stock) × 2 shares, because the USD5 dividend was spent rather than reinvested.

At the end of the second year, the proceeds from the liquidation of the portfolio are USD470 plus USD10 in dividends (as outlined in Exhibit 11). So, in the second year the portfolio produced a holding period return of $(\text{USD}10 + \text{USD}470 - \text{USD}450)/\text{USD}450 = 6.67\%$. The mean holding period return was $(15\% + 6.67\%)/2 = 10.84\%$.

The money-weighted rate of return, which we calculated as 9.39 percent, puts a greater weight on the second year's relatively poor performance (6.67 percent) than the first year's relatively good performance (15 percent), as more money was invested in the second year than in the first. That is the sense in which returns in this method of calculating performance are "money weighted."

Although the money-weighted return is an accurate measure of what the investor earned on the money invested, it is limited in its applicability to other situations. For example, it does not allow for a return comparison between different individuals or different investment opportunities. Importantly, two investors in the *same* mutual fund or with the same portfolio of underlying investments may have different money-weighted returns because they invested different amounts in different years.

EXAMPLE 8**Computation of Returns**

Ulli Lohrmann and his wife, Suzanne Lohrmann, are planning for retirement and want to compare the past performance of a few mutual funds they are considering for investment. They believe that a comparison over a five-year period would be appropriate. They gather information on a fund they are considering, the Rhein Valley Superior Fund, which is presented in Exhibit 12.

Exhibit 12: Rhein Valley Superior Fund Performance

Year	Assets under Management at the Beginning of Year (euros)	Annual Return (%)
1	30 million	15
2	45 million	-5
3	20 million	10
4	25 million	15
5	35 million	3

The Lohrmanns are interested in aggregating this information for ease of comparison with other funds.

Exhibit 13: Rhein Valley Superior Fund Annual Returns and Investments (euro millions)

Year	1	2	3	4	5
Balance from previous year	0	34.50	42.75	22.00	28.75
New investment by the investor (cash inflow for the Rhein fund)	30.00	10.50	0	3.00	6.25
Withdrawal by the investor (cash outflow for the Rhein fund)	0	0	-22.75	0	0
Net balance at the beginning of year	30.00	45.00	20.00	25.00	35.00
Investment return for the year	15%	-5%	10%	15%	3%
Investment gain (loss)	4.50	-2.25	2.00	3.75	1.05
Balance at the end of year	34.50	42.75	22.00	28.75	36.05

1. Compute the fund's holding period return for the five-year period.

Solution:

The five-year holding period return is calculated as:

$$R = (1 + R_1)(1 + R_2)(1 + R_3)(1 + R_4)(1 + R_5) - 1$$

$$R = (1.15)(0.95)(1.10)(1.15)(1.03) - 1 =$$

$$R = 0.4235 = 42.35\%.$$

2. Compute the fund's arithmetic mean annual return.

Solution:

The arithmetic mean annual return is calculated as:

$$\bar{R}_i = \frac{15\% - 5\% + 10\% + 15\% + 3\%}{5} = 7.60\%.$$

3. Compute the fund's geometric mean annual return. How does it compare with the arithmetic mean annual return?

Solution:

The geometric mean annual return can be computed as:

$$\bar{R}_{Gi} = \sqrt[5]{1.15 \times 0.95 \times 1.10 \times 1.15 \times 1.03} - 1,$$

$$\bar{R}_{Gi} = \sqrt[5]{1.4235} - 1 = 0.0732 = 7.32\%.$$

Thus, the geometric mean annual return is 7.32 percent, which is slightly less than the arithmetic mean return of 7.60 percent.

4. The Lohrmanns want to earn a minimum annual return of 5 percent. The annual returns and investment amounts are presented in Exhibit 13. Is the money-weighted annual return greater than 5 percent?

Solution:

To calculate the money-weighted rate of return, tabulate the annual returns and investment amounts to determine the cash flows, as shown in Exhibit 13:

$$CF_0 = -30.00, CF_1 = -10.50, CF_2 = +22.75, CF_3 = -3.00, CF_4 = -6.25, CF_5 = +36.05.$$

We can use the given 5 percent return to see whether or not the present value of the net cash flows is positive. If it is positive, then the money-weighted rate of return is greater than 5 percent, because a 5 percent discount rate could not reduce the present value to zero.

$$\frac{-30.00}{(1.05)^0} + \frac{-10.50}{(1.05)^1} + \frac{22.75}{(1.05)^2} + \frac{-3.00}{(1.05)^3} + \frac{-6.25}{(1.05)^4} + \frac{36.05}{(1.05)^5} = 1.1471.$$

Because the value is positive, the money-weighted rate of return is greater than 5 percent. The exact money-weighted rate of return (found by setting the above equation equal to zero) is 5.86 percent.

These calculations can be performed using Excel using the =IRR(cash flows) function, which results in an IRR of 5.86 percent.

Time-Weighted Returns

An investment measure that is not sensitive to the additions and withdrawals of funds is the time-weighted rate of return. The **time-weighted rate of return** measures the compound rate of growth of USD1 initially invested in the portfolio over a stated measurement period. For the evaluation of portfolios of publicly traded securities, the time-weighted rate of return is the preferred performance measure as it neutralizes the effect of cash withdrawals or additions to the portfolio, which are generally outside of the control of the portfolio manager.

Computing Time-Weighted Returns

To compute an exact time-weighted rate of return on a portfolio, take the following three steps:

1. Price the portfolio immediately prior to any significant addition or withdrawal of funds. Break the overall evaluation period into subperiods based on the dates of cash inflows and outflows.
2. Calculate the holding period return on the portfolio for each subperiod.
3. Link or compound holding period returns to obtain an annual rate of return for the year (the time-weighted rate of return for the year). If the investment is for more than one year, take the geometric mean of the annual returns to obtain the time-weighted rate of return over that measurement period.

Let us return to our dividend stock money-weighted example in the section, “Money-Weighted Return for a Dividend-Paying Stock” and calculate the time-weighted rate of return for that investor’s portfolio based on the information included in Exhibit 11. In that example, we computed the holding period returns on the portfolio, Step 2 in the procedure for finding the time-weighted rate of return. Given that the portfolio earned returns of 15 percent during the first year and 6.67 percent during the second year, what is the portfolio’s time-weighted rate of return over an evaluation period of two years?

We find this time-weighted return by taking the geometric mean of the two holding period returns, Step 3 in the previous procedure. The calculation of the geometric mean exactly mirrors the calculation of a compound growth rate. Here, we take the product of 1 plus the holding period return for each period to find the terminal value at $t = 2$ of USD1 invested at $t = 0$. We then take the square root of this product and subtract 1 to get the geometric mean return. We interpret the result as the annual compound growth rate of USD1 invested in the portfolio at $t = 0$. Thus, we have:

$$(1 + \text{Time-weighted return})^2 = (1.15)(1.0667)$$

$$\text{Time-weighted return} = \sqrt{(1.15)(1.0667)} - 1 = 10.76\%$$

The time-weighted return on the portfolio was 10.76 percent, compared with the money-weighted return of 9.39 percent, which gave larger weight to the second year’s return. We can see why investment managers find time-weighted returns more meaningful. If a client gives an investment manager more funds to invest at an unfavorable time, the manager’s money-weighted rate of return will tend to be depressed. If a client adds funds at a favorable time, the money-weighted return will tend to be elevated. The time-weighted rate of return removes these effects.

In defining the steps to calculate an exact time-weighted rate of return, we said that the portfolio should be valued immediately prior to any significant addition or withdrawal of funds. With the amount of cash flow activity in many portfolios, this task can be costly. We can often obtain a reasonable approximation of the time-weighted rate of return by valuing the portfolio at frequent, regular intervals, particularly if additions and withdrawals are unrelated to market movements.

The more frequent the valuation, the more accurate the approximation. Daily valuation is commonplace. Suppose that a portfolio is valued daily over the course of a year. To compute the time-weighted return for the year, we first compute each day’s holding period return. We compute 365 such daily returns, denoted R_1, R_2, \dots, R_{365} . We obtain the annual return for the year by linking the daily holding period returns in the following way: $(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_{365}) - 1$. If withdrawals and additions to the portfolio happen only at day’s end, this annual return is a precise time-weighted rate of return for the year. Otherwise, it is an approximate time-weighted return for the year.

If we have several years of data, we can calculate a time-weighted return for each year individually, as above. If R_i is the time-weighted return for year i , we calculate an annualized time-weighted return as the geometric mean of N annual returns, as follows:

$$R_{TW} = [(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_N)]^{1/N} - 1. \quad (6)$$

Example 9 illustrates the calculation of the time-weighted rate of return.

EXAMPLE 9

Time-Weighted Rate of Return

Strubeck Corporation sponsors a pension plan for its employees. It manages part of the equity portfolio in-house and delegates management of the balance to Super Trust Company. As chief investment officer of Strubeck, you want to review the performance of the in-house and Super Trust portfolios over the last four quarters. You have arranged for outflows and inflows to the portfolios to be made at the very beginning of the quarter. Exhibit 14 summarizes the inflows and outflows as well as the two portfolios' valuations. In Exhibit 11, the ending value is the portfolio's value just prior to the cash inflow or outflow at the beginning of the quarter. The amount invested is the amount each portfolio manager is responsible for investing.

Exhibit 14: Cash Flows for the In-House Strubeck Account and the Super Trust Account (US dollars)

	Quarter			
	1	2	3	4
Panel A: In-House Account				
Beginning value	4,000,000	6,000,000	5,775,000	6,720,000
Beginning of period inflow (outflow)	1,000,000	(500,000)	225,000	(600,000)
Amount invested	5,000,000	5,500,000	6,000,000	6,120,000
Ending value	6,000,000	5,775,000	6,720,000	5,508,000
Panel B: Super Trust Account				
Beginning value	10,000,000	13,200,000	12,240,000	5,659,200
Beginning of period inflow (outflow)	2,000,000	(1,200,000)	(7,000,000)	(400,000)
Amount invested	12,000,000	12,000,000	5,240,000	5,259,200
Ending value	13,200,000	12,240,000	5,659,200	5,469,568

1. Calculate the time-weighted rate of return for the in-house account.

Solution:

To calculate the time-weighted rate of return for the in-house account, we compute the quarterly holding period returns for the account and link them into an annual return. The in-house account's time-weighted rate of return is 27.01 percent, calculated as follows:

$$\begin{aligned}1Q HPR: r_1 &= (\text{USD}6,000,000 - \text{USD}5,000,000) / \text{USD}5,000,000 = 0.20 \\2Q HPR: r_2 &= (\text{USD}5,775,000 - \text{USD}5,500,000) / \text{USD}5,500,000 = 0.05 \\3Q HPR: r_3 &= (\text{USD}6,720,000 - \text{USD}6,000,000) / \text{USD}6,000,000 = 0.12 \\4Q HPR: r_4 &= (\text{USD}5,508,000 - \text{USD}6,120,000) / \text{USD}6,120,000 = -0.10\end{aligned}$$

$$R_{TW} = (1 + r_1)(1 + r_2)(1 + r_3)(1 + r_4) - 1,$$

$$R_{TW} = (1.20)(1.05)(1.12)(0.90) - 1 = 0.2701 \text{ or } 27.01\%.$$

2. Calculate the time-weighted rate of return for the Super Trust account.

Solution:

The account managed by Super Trust has a time-weighted rate of return of 26.02 percent, calculated as follows:

$$\begin{aligned}1Q HPR: r_1 &= (\text{USD}13,200,000 - \text{USD}12,000,000) / \text{USD}12,000,000 = 0.10 \\2Q HPR: r_2 &= (\text{USD}12,240,000 - \text{USD}12,000,000) / \text{USD}12,000,000 = 0.02 \\3Q HPR: r_3 &= (\text{USD}5,659,200 - \text{USD}5,240,000) / \text{USD}5,240,000 = 0.08 \\4Q HPR: r_4 &= (\text{USD}5,469,568 - \text{USD}5,259,200) / \text{USD}5,259,200 = 0.04\end{aligned}$$

$$R_{TW} = (1 + r_1)(1 + r_2)(1 + r_3)(1 + r_4) - 1,$$

$$R_{TW} = (1.10)(1.02)(1.08)(1.04) - 1 = 0.2602 \text{ or } 26.02\%.$$

The in-house portfolio's time-weighted rate of return is higher than the Super Trust portfolio's by 99 basis points. Note that 27.01 percent and 26.02 percent might be rounded to 27 percent and 26 percent, respectively. The impact of the rounding the performance difference (100 bp vs. 99 bp) may seem as trivial, yet it's impact on a large portfolio may be substantive.

Having worked through this exercise, we are ready to look at a more detailed case.

EXAMPLE 10

Time-Weighted and Money-Weighted Rates of Return Side by Side

Your task is to compute the investment performance of the Walbright Fund for the most recent year. The facts are as follows:

- On 1 January, the Walbright Fund had a market value of USD100 million.
- During the period 1 January to 30 April, the stocks in the fund generated a capital gain of USD10 million.
- On 1 May, the stocks in the fund paid a total dividend of USD2 million. All dividends were reinvested in additional shares.
- Because the fund's performance had been exceptional, institutions invested an additional USD20 million in Walbright on 1 May, raising assets under management to USD132 million ($\text{USD}100 + \text{USD}10 + \text{USD}2 + \text{USD}20$).
- On 31 December, Walbright received total dividends of USD2.64 million. The fund's market value on 31 December, not including the USD2.64 million in dividends, was USD140 million.
- The fund made no other interim cash payments during the year.

1. Compute the Walbright Fund's time-weighted rate of return.

Solution:

Because interim cash flows were made on 1 May, we must compute two interim total returns and then link them to obtain an annual return. Exhibit 15 lists the relevant market values on 1 January, 1 May, and 31 December, as well as the associated interim four-month (1 January to 1 May) and eight-month (1 May to 31 December) holding period returns.

Exhibit 15: Cash Flows for the Walbright Fund

1 January	Beginning portfolio value = USD100 million
1 May	Dividends received before additional investment = USD2 million
	Ending portfolio value = USD110 million
	Four-month holding period return: $R = \frac{USD2 + USD10}{USD100} = 12\%$
	New investment = USD20 million
	Beginning market value for last two-thirds of the year = USD132 million
31 December	Dividends received = USD2.64 million
	Ending portfolio value = USD140 million
	Eight-month holding period return: $R = \frac{USD2.64 + USD140 - USD132}{USD132} = 8.06\%$

Now we must geometrically link the four- and eight-month holding period returns to compute an annual return. We compute the time-weighted return as follows:

$$R_{TW} = 1.12 \times 1.0806 - 1 = 0.2103.$$

In this instance, we compute a time-weighted rate of return of 21.03 percent for one year. The four-month and eight-month intervals combine to equal one year. (Note: Taking the square root of the product 1.12×1.0806 would be appropriate only if 1.12 and 1.0806 each applied to one full year.)

2. Compute the Walbright Fund's money-weighted rate of return.

Solution:

To calculate the money-weighted return, we need to find the discount rate that sets the sum of the present value of cash inflows and outflows equal to zero. The initial market value of the fund and all additions to it are treated as cash outflows. (Think of them as expenditures.) Withdrawals, receipts, and the ending market value of the fund are counted as inflows. (The ending market value is the amount investors receive on liquidating the fund.) Because interim cash flows have occurred at four-month intervals, we must solve for the four-month internal rate of return. Exhibit 15 details the cash flows and their timing.

$$CF_0 = -100$$

$$CF_1 = -20$$

$$CF_2 = 0$$

$$CF_3 = 142.64$$

CF_0 refers to the initial investment of USD100 million made at the beginning of the first four-month interval on 1 January. CF_1 refers to the cash flows made at end of the first four-month interval or the beginning of the second four-month interval on 1 May. Those cash flows include a cash inflow of USD2 million for the dividend received and cash outflows of USD22 million for the dividend reinvested and additional investment, respectively. The second four-month interval had no cash flow so CF_2 is equal to zero. CF_3 refers to the cash inflows at the end of the third four-month interval. Those cash inflows include a USD2.64 million dividend received and the fund's terminal market value of USD140 million.

Using a spreadsheet or IRR-enabled calculator, we use -100, -20, 0, and USD142.64 for the $t = 0$, $t = 1$, $t = 2$, and $t = 3$ net cash flows, respectively. Using either tool, we get a four-month IRR of 6.28 percent.

$$\frac{CF_0}{(1 + IRR)^0} + \frac{CF_1}{(1 + IRR)^1} + \frac{CF_2}{(1 + IRR)^2} + \frac{CF_3}{(1 + IRR)^3} = 0$$

$$\frac{-100}{1} + \frac{-20}{(1 + IRR)^1} + \frac{0}{(1 + IRR)^2} + \frac{142.64}{(1 + IRR)^3} = 0$$

$$IRR = 6.28\%$$

The quick way to annualize this four-month return is to multiply it by 3. A more accurate way is to compute it on a compounded basis as: $(1.0628)^3 - 1 = 0.2005$ or 20.05 percent.

These calculations can also be performed using Excel using the =IRR(cash flows) function, which results in an IRR of 6.28 percent.

3. Interpret the differences between the Fund's time-weighted and money-weighted rates of return.

Solution:

In this example, the time-weighted return (21.03 percent) is greater than the money-weighted return (20.05 percent). The Walbright Fund's performance was relatively poorer during the eight-month period, when the fund had more money invested, than the overall performance. This fact is reflected in a lower money-weighted rate of return compared with the time-weighted rate of return, as the money-weighted return is sensitive to the timing and amount of withdrawals and additions to the portfolio.

The accurate measurement of portfolio returns is important to the process of evaluating portfolio managers. In addition to considering returns, however, analysts must also weigh risk. When we worked through Example 9, we stopped short of suggesting that in-house management was superior to Super Trust because it earned a higher time-weighted rate of return. A judgment as to whether performance was "better" or "worse" must include the risk dimension, which will be covered later in your study materials.

5

ANNUALIZED RETURN



calculate and interpret annualized return measures and continuously compounded returns, and describe their appropriate uses

The period during which a return is earned or computed can vary and often we have to annualize a return that was calculated for a period that is shorter (or longer) than one year. You might buy a short-term treasury bill with a maturity of three months, or you might take a position in a futures contract that expires at the end of the next quarter. How can we compare these returns?

In many cases, it is most convenient to annualize all available returns to facilitate comparison. Thus, daily, weekly, monthly, and quarterly returns are converted to annualized returns. Many formulas used for calculating certain values or prices also require all returns and periods to be expressed as annualized rates of return. For example, the most common version of the Black–Scholes option-pricing model requires annualized returns and periods to be in years.

Non-annual Compounding

Recall that interest may be paid semiannually, quarterly, monthly, or even daily. To handle interest payments made more than once a year, we can modify the present value formula as follows. Here, R_s is the quoted interest rate and equals the periodic interest rate multiplied by the number of compounding periods in each year. In general, with more than one compounding period in a year, we can express the formula for present value as follows:

$$PV = FV_N \left(1 + \frac{R_s}{m}\right)^{-mN}, \quad (7)$$

where

m = number of compounding periods per year,

R_s = quoted annual interest rate, and

N = number of years.

The formula in Equation 8 is quite similar to the simple present value formula. As we have already noted, present value and future value factors are reciprocals. Changing the frequency of compounding does not alter this result. The only difference is the use of the periodic interest rate and the corresponding number of compounding periods.

The following example presents an application of monthly compounding.

EXAMPLE 11

The Present Value of a Lump Sum with Monthly Compounding

The manager of a Canadian pension fund knows that the fund must make a lump-sum payment of CAD5 million 10 years from today. She wants to invest an amount today in a guaranteed investment contract (GIC) so that it will grow to the required amount. The current interest rate on GICs is 6 percent a year, compounded monthly.

1. How much should she invest today in the GIC?

Solution:

By applying Equation 8, the required present value is calculated as follow:

$$\begin{aligned}
 FV_N &= \text{CAD } 5,000,000 \\
 R_s &= 6\% = 0.06 \\
 m &= 12 \\
 R_s/m &= 0.06/12 = 0.005 \\
 N &= 10 \\
 mN &= 12(10) = 120 \\
 PV &= FV_N \left(1 + \frac{R_s}{m}\right)^{-mN} \\
 &= \text{CAD } 5,000,000(1.005)^{-120} \\
 &= \text{CAD } 5,000,000(0.549633) \\
 &= \text{CAD } 2,748,163.67
 \end{aligned}$$

In applying Equation 8, we use the periodic rate (in this case, the monthly rate) and the appropriate number of periods with monthly compounding (in this case, 10 years of monthly compounding, or 120 months).

Annualizing Returns

To annualize any return for a period shorter than one year, the return for the period must be compounded by the number of periods in a year. A monthly return is compounded 12 times, a weekly return is compounded 52 times, and a quarterly return is compounded 4 times. Daily returns are normally compounded 365 times. For an uncommon number of days, we compound by the ratio of 365 to the number of days.

If the weekly return is 0.2 percent, then the compound annual return is 10.95 percent (there are 52 weeks in a year):

$$\begin{aligned}
 R_{annual} &= (1 + R_{weekly})^{52} - 1 = (1 + 0.2\%)^{52} - 1 \\
 &= (1.002)^{52} - 1 = 0.1095 = 10.95\%
 \end{aligned}$$

If the return for 15 days is 0.4 percent, then the annualized return is 10.20 percent, assuming 365 days in a year:

$$\begin{aligned}
 R_{annual} &= (1 + R_{15})^{365/15} - 1 = (1 + 0.4\%)^{365/15} - 1 \\
 &= (1.004)^{365/15} - 1 = 0.1020 = 10.20\%
 \end{aligned}$$

A general equation to annualize returns is given, where c is the number of periods in a year. For a quarter, $c = 4$ and for a month, $c = 12$:

$$R_{annual} = (1 + R_{period})^c - 1. \quad (8)$$

How can we annualize a return when the holding period return is more than one year? For example, how do we annualize an 18-month holding period return? Because one year contains two-thirds of 18-month periods, $c = 2/3$ in the above equation. For example, an 18-month return of 20 percent can be annualized as follows:

$$R_{annual} = (1 + R_{18month})^{2/3} - 1 = (1 + 0.20)^{2/3} - 1 = 0.1292 = 12.92\%.$$

Similar expressions can be constructed when quarterly or weekly returns are needed for comparison instead of annual returns. In such cases, c is equal to the number of holding periods in a quarter or in a week. For example, assume that you want to convert daily returns to weekly returns or annual returns to weekly returns for comparison between weekly returns. To convert daily returns to weekly returns, $c = 5$, assume that there are five trading days in a week. However, daily return calculations can be annualized differently. For example, five can be used for trading-day-based calculations, giving approximately 250 trading days a year; seven can be used on calendar-day-based calculations. Specific methods used conform to specific business practices, market

Annualized Return

conventions, and standards. To convert annual returns to weekly returns, $c = 1/52$. The expressions for annual returns can then be rewritten as expressions for weekly returns as follows:

$$R_{\text{weekly}} = (1 + R_{\text{daily}})^5 - 1; R_{\text{weekly}} = (1 + R_{\text{annual}})^{1/52} - 1. \quad (9)$$

One major limitation of annualizing returns is the implicit assumption that returns can be repeated precisely, that is, money can be reinvested repeatedly while earning a similar return. This type of return is not always possible. An investor may earn a return of 5 percent during a week because the market rose sharply that week, but it is highly unlikely that he will earn a return of 5 percent every week for the next 51 weeks, resulting in an annualized return of 1,164.3 percent ($= 1.05^{52} - 1$). Therefore, it is important to annualize short-term returns with this limitation in mind.

EXAMPLE 12**Annualized Returns**

An analyst seeks to evaluate three securities she has held in her portfolio for different periods of time.

- Over the past 100 days, Security A has earned a return of 6.2 percent.
- Security B has earned 2 percent over the past four weeks.
- Security C has earned a return of 5 percent over the past three months.

1. Compare the relative performance of the three securities.

Solution:

To facilitate comparison, the three securities' returns need to be annualized:

- Annualized return for Security A: $R_{SA} = (1 + 0.062)^{365/100} - 1 = 0.2455 = 24.55\%$
- Annualized return for Security B: $R_{SB} = (1 + 0.02)^{52/4} - 1 = 0.2936 = 29.36\%$
- Annualized return for Security C: $R_{SC} = (1 + 0.05)^4 - 1 = 0.2155 = 21.55\%$

Security B generated the highest annualized return.

EXAMPLE 13**Exchange-Traded Fund Performance**

An investor is evaluating the returns of three recently formed exchange-traded funds. Selected return information on the exchange-traded funds (ETFs) is presented in Exhibit 16.

Exhibit 16: ETF Performance Information

ETF	Time Since Inception	Return Since Inception (%)
1	146 days	4.61
2	5 weeks	1.10
3	15 months	14.35

1. Which ETF has the highest annualized rate of return?

- A. ETF 1
- B. ETF 2
- C. ETF 3

Solution:

B is correct. The annualized rate of return for the three ETFs are as follows:

$$\text{ETF 1 annualized return} = (1.0461^{365/146}) - 1 = 11.93\%$$

$$\text{ETF 2 annualized return} = (1.0110^{52/5}) - 1 = 12.05\%$$

$$\text{ETF 3 annualized return} = (1.1435^{12/15}) - 1 = 11.32\%$$

Despite having the lowest value for the periodic rate, ETF 2 has the highest annualized rate of return because of the reinvestment rate assumption and the compounding of the periodic rate.

Continuously Compounded Returns

An important concept is the continuously compounded return associated with a holding period return, such as R_1 . The **continuously compounded return** associated with a holding period return is the natural logarithm of one plus that holding period return, or equivalently, the natural logarithm of the ending price over the beginning price (the price relative). Note that here we are using r to refer specifically to continuously compounded returns, but other textbooks and sources may use a different notation.

If we observe a one-week holding period return of 0.04, the equivalent continuously compounded return, called the one-week continuously compounded return, is $\ln(1.04) = 0.039221$; EUR1.00 invested for one week at 0.039221 continuously compounded gives EUR1.04, equivalent to a 4 percent one-week holding period return.

The continuously compounded return from t to $t + 1$ is

$$r_{t,t+1} = \ln(P_{t+1}/P_t) = \ln(1 + R_{t,t+1}). \quad (10)$$

For our example, an asset purchased at time t for a P_0 of USD30 and the same asset one period later, $t + 1$, has a value of P_1 of USD34.50 has a continuously compounded return given by $r_{0,1} = \ln(P_1/P_0) = \ln(1 + R_{0,1}) = \ln(\text{USD}34.50/\text{USD}30) = \ln(1.15) = 0.139762$.

Thus, 13.98 percent is the continuously compounded return from $t = 0$ to $t = 1$. The continuously compounded return is smaller than the associated holding period return. If our investment horizon extends from $t = 0$ to $t = T$, then the continuously compounded return to T is

$$r_{0,T} = \ln(P_T/P_0). \quad (11)$$

Applying the exponential function to both sides of the equation, we have $\exp(r_{0,T}) = \exp[\ln(P_T/P_0)] = P_T/P_0$, so

$$P_T = P_0 \exp(r_{0,T}).$$

We can also express P_T/P_0 as the product of price relatives:

$$P_T/P_0 = (P_T/P_{T-1})(P_{T-1}/P_{T-2}) \dots (P_1/P_0). \quad (12)$$

Taking logs of both sides of this equation, we find that the continuously compounded return to time T is the sum of the one-period continuously compounded returns:

$$r_{0,T} = r_{T-1,T} + r_{T-2,T-1} + \dots + r_{0,1}. \quad (13)$$

Using holding period returns to find the ending value of a USD1 investment involves the multiplication of quantities ($1 + \text{holding period return}$). Using continuously compounded returns involves addition (as shown in Equation 14), which is a desirable property of continuously compounded returns and which we will use throughout the curriculum.

OTHER MAJOR RETURN MEASURES AND THEIR APPLICATIONS

6

calculate and interpret major return measures and describe their appropriate uses

The statistical measures of return discussed in the previous section are generally applicable across a wide range of assets and time periods. Special assets, however, such as mutual funds, and other considerations, such as taxes or inflation, may require more specific return measures.

Although it is not possible to consider all types of special measures, we will discuss the effect of fees (gross versus net returns), taxes (pre-tax and after-tax returns), inflation (nominal and real returns), and the effect of **leverage**. Many investors use mutual funds or other external entities (i.e., investment vehicles) for investment. In those cases, funds charge management fees and expenses to the investors. Consequently, gross and net-of-fund-expense returns should also be considered. Of course, an investor may be interested in the net-of-expenses after-tax real return, which is in fact what an investor truly receives. We consider these additional return measures in the following sections.

Gross and Net Return

A gross return is the return earned by an asset manager prior to deductions for management expenses, custodial fees, taxes, or any other expenses that are not directly related to the generation of returns but rather related to the management and administration of an investment. These expenses are not deducted from the gross return because they may vary with the amount of assets under management or may vary because of the tax status of the investor. Trading expenses, however, such as commissions, *are* accounted for in (i.e., deducted from) the computation of gross return because trading expenses contribute directly to the return earned by the manager. Thus, gross return is an appropriate measure for evaluating and comparing the investment skill of asset managers because it does not include any fees related to the management and administration of an investment.

Net return is a measure of what the investment vehicle (e.g., mutual fund) has earned for the investor. Net return accounts for (i.e., deducts) all managerial and administrative expenses that reduce an investor's return. Because individual investors are most concerned about the net return (i.e., what they actually receive), small mutual funds with a limited amount of assets under management are at a disadvantage compared with the larger funds that can spread their largely fixed administrative expenses over a larger asset base. As a result, many small mutual funds waive part of the expenses to keep the funds competitive.

Pre-Tax and After-Tax Nominal Return

All return measures discussed up to this point are pre-tax nominal returns—that is, no adjustment has been made for taxes or inflation. In general, all returns are pre-tax nominal returns unless they are otherwise designated.

Many investors are concerned about the possible tax liability associated with their returns because taxes reduce the net return that they receive. Capital gains and income may be taxed differently, depending on the jurisdiction. Capital gains come in two forms: short-term capital gains and long-term capital gains. Long-term capital gains receive preferential tax treatment in a number of countries. Interest income is taxed as ordinary income in most countries. Dividend income may be taxed as ordinary income, may have a lower tax rate, or may be exempt from taxes depending on the country and the type of investor. The after-tax nominal return is computed as the total return minus any allowance for taxes on dividends, interest, and realized gains. Bonds issued at a discount to the par value may be taxed based on accrued gains instead of realized gains.

Because taxes are paid on realized capital gains and income, the investment manager can minimize the tax liability by selecting appropriate securities (e.g., those subject to more favorable taxation, all other investment considerations equal) and reducing trading turnover. Therefore, taxable investors evaluate investment managers based on the after-tax nominal return.

Real Returns

Previously this learning module approximated the relationship between the nominal rate and the real rate by the following relationship:

$$(1 + \text{nominal risk-free rate}) = (1 + \text{real risk-free rate})(1 + \text{inflation premium}).$$

This relationship can be extended to link the relationship between nominal and real returns. Specifically, the nominal return consists of a real risk-free rate of return to compensate for postponed consumption; inflation as loss of purchasing power; and a risk premium for assuming risk. Frequently, the real risk-free return and the risk premium are combined to arrive at the real “risky” rate and is simply referred to as the real return, or:

$$(1 + \text{real return}) = \frac{(1 + \text{real risk-free rate})(1 + \text{risk premium})}{1 + \text{inflation premium}}. \quad (14)$$

Real returns are particularly useful in comparing returns across time periods because inflation rates may vary over time. Real returns are also useful in comparing returns among countries when returns are expressed in local currencies instead of a constant investor currency and when inflation rates vary between countries (which are usually the case).

Finally, the after-tax real return is what the investor receives as compensation for postponing consumption and assuming risk after paying taxes on investment returns. As a result, the after-tax real return becomes a reliable benchmark for making investment decisions. Although it is a measure of an investor’s benchmark return, it is not commonly calculated by asset managers because it is difficult to estimate a general tax component applicable to all investors. For example, the tax component depends on an investor’s specific taxation rate (marginal tax rate), how long the investor holds an investment (long-term versus short-term), and the type of account the asset is held in (tax-exempt, tax-deferred, or normal).

EXAMPLE 14**Computation of Special Returns**

Let's return to Example 8. After reading this section, Mr. Lohrmann decided that he was not being fair to the fund manager by including the asset management fee and other expenses because the small size of the fund would put it at a competitive disadvantage. He learns that the fund spends a fixed amount of EUR500,000 every year on expenses that are unrelated to the manager's performance.

Mr. Lohrmann has become concerned that both taxes and inflation may reduce his return. Based on the current tax code, he expects to pay 20 percent tax on the return he earns from his investment. Historically, inflation has been around 2 percent and he expects the same rate of inflation to be maintained.

1. Estimate the annual gross return for the first year by adding back the fixed expenses.

Solution:

The gross return for the first year is higher by 1.67 percent (= EUR500,000 / EUR30,000,000) than the 15 percent investor return reported by the fund. Thus, the gross return for the first year is 16.67 percent (= 15% + 1.67%).

2. What is the net return that investors in the Rhein Valley Superior Fund earned during the five-year period?

Solution:

The investor return reported by the mutual fund is the net return of the fund after accounting for all direct and indirect expenses. The net return is also the pre-tax nominal return because it has not been adjusted for taxes or inflation. From Example 8, the net return for the five-year holding period was calculated as 42.35 percent.

3. What is the after-tax net return for the first year that investors earned from the Rhein Valley Superior Fund? Assume that all gains are realized at the end of the year and the taxes are paid immediately at that time.

Solution:

The net return earned by investors during the first year was 15 percent. Applying a 20 percent tax rate, the after-tax return that accrues to investors is 12 percent [= 15% - (0.20 × 15%)].

4. What is the after-tax real return that investors would have earned in the fifth year?

Solution:

The after-tax return earned by investors in the fifth year is 2.4 percent [= 3% - (0.20 × 3%)]. Inflation reduces the return by 2 percent so the after-tax real return earned by investors in the fifth year is 0.39 percent, as shown:

$$\frac{(1 + 2.40\%)}{(1 + 2.00\%)} - 1 = \frac{(1 + 0.0240)}{(1 + 0.0200)} - 1 = 1.0039 - 1 = 0.0039 = 0.39\%.$$

Note that taxes are paid before adjusting for inflation.

Leveraged Return

In the previous calculations, we have assumed that the investor's position in an asset is equal to the total investment made by an investor using his or her own money. This section differs in that the investor creates a leveraged position.

There are two ways of creating a claim on asset returns that are greater than the investment of one's own money. First, an investor may trade futures contracts in which the money required to take a position may be as little as 10 percent of the notional value of the asset. In this case, the leveraged return, the return on the investor's own money, is 10 times the actual return of the underlying security. Both the gains and losses are amplified by a factor of 10.

Investors can also invest more than their own money by borrowing money to purchase the asset. This approach is easily done in stocks and bonds, and very common when investing in real estate. If half (50 percent) of the money invested is borrowed, then the gross return to the investor is doubled, but the interest to be paid on borrowed money must be deducted to calculate the net return.

Using borrowed capital, debt, the size of the leveraged position increases by the additional, borrowed capital. If the total investment return earned on the leveraged portfolio, R_p , exceeds the borrowing cost on debt, r_D , taking on leverage increases the return on the portfolio. Denoting the return on a leveraged portfolio as R_L , then the return can be calculated as follows:

$$R_L = \frac{\text{Portfolio return}}{\text{Portfolio equity}} = \frac{[R_p \times (V_E + V_B) - (V_B \times r_D)]}{V_E} = R_p + \frac{V_B}{V_E}(R_p - r_D), \quad (15)$$

where V_E is the equity of the portfolio and V_B is the debt or borrowed funds. If $R_p < r_D$ then leverage decreases R_L .

For example, for a EUR10 million equity portfolio that generates an 8 percent total investment return, R_p , over one year and is financed 30 percent with debt at 5 percent, then the leveraged return, R_L , is:

$$\begin{aligned} R_L &= R_p + \frac{V_B}{V_E}(R_p - r_D) = 8\% + \frac{\text{EUR}3 \text{ million}}{\text{EUR}7 \text{ million}}(8\% - 5\%) = 8\% + 0.43 \times 3\% \\ &= 9.29\%. \end{aligned}$$

EXAMPLE 15

Return Calculations

An analyst observes the following historic asset class geometric returns:

Exhibit 17: Asset Class Geometric Return

Asset Class	Geometric Return (%)
Equities	8.0
Corporate Bonds	6.5
Treasury bills	2.5
Inflation	2.1

- The real rate of return for equities is closest to:

- A. 5.4 percent.
- B. 5.8 percent.

- C. 5.9 percent.

Solution:

B is correct. The real rate of return for equities is calculated as follows:

$$(1 + 0.080)/(1 + 0.0210) - 1 = 5.8\%.$$

-
2. The real rate of return for corporate bonds is *closest* to:

- A. 4.3 percent.

- B. 4.4 percent.

- C. 4.5 percent.

Solution:

A is correct. The real rate of return for corporate bonds is calculated as follows:

$$(1 + 0.065)/(1 + 0.0210) - 1 = 4.3\%.$$

-
3. The risk premium for equities is closest to:

- A. 5.4 percent.

- B. 5.5 percent.

- C. 5.6 percent.

Solution:

A is correct. The risk premium for equities is calculated as follows:

$$(1 + 0.080)/(1 + 0.0250) - 1 = 5.4\%.$$

-
4. The risk premium for corporate bonds is *closest* to:

- A. 3.5 percent.

- B. 3.9 percent.

- C. 4.0 percent.

Solution:

B is correct. The risk premium for corporate bonds is calculated as follows:

$$(1 + 0.0650)/(1 + 0.0250) - 1 = 3.9\%.$$

PRACTICE PROBLEMS

1. and a premium for:
 - A. maturity.
 - B. liquidity.
 - C. expected inflation.

2. Which of the following risk premiums is most relevant in explaining the difference in yields between 30-year bonds issued by the US Treasury and 30-year bonds issued by a small, private US corporate issuer?
 - A. Inflation
 - B. Maturity
 - C. Liquidity

3. Consider the following annual return for Fund Y over the past five years:

Exhibit 1: Five-Year Annual Returns

Year	Return (%)
Year 1	19.5
Year 2	-1.9
Year 3	19.7
Year 4	35.0
Year 5	5.7

The geometric mean return for Fund Y is *closest* to:

- A. 14.9 percent.
- B. 15.6 percent.
- C. 19.5 percent.

4. A portfolio manager invests EUR5,000 annually in a security for four years at the following prices:

Exhibit 1: Five-Year Purchase Prices

Year	Purchase Price (euros per unit)
Year 1	62.00
Year 2	76.00

Year	Purchase Price (euros per unit)
Year 3	84.00
Year 4	90.00

The average price is *best* represented as the:

- A. harmonic mean of EUR76.48.
 - B. geometric mean of EUR77.26.
 - C. arithmetic average of EUR78.00.
5. Which of the following statements regarding arithmetic and geometric means is correct?
- A. The geometric mean will exceed the arithmetic mean for a series with non-zero variance.
 - B. The geometric mean measures an investment's compound rate of growth over multiple periods.
 - C. The arithmetic mean measures an investment's terminal value over multiple periods.
6. A fund receives investments at the beginning of each year and generates returns for three years as follows:

Exhibit 1: Investments and Returns for Three Years

Year of Investment	Assets under Management at the Beginning of each year	Return during Year of Investment
1	USD1,000	15%
2	USD4,000	14%
3	USD45,000	-4%

Which return measure over the three-year period is negative?

- A. Geometric mean return
 - B. Time-weighted rate of return
 - C. Money-weighted rate of return
7. At the beginning of Year 1, a fund has USD10 million under management; it earns a return of 14 percent for the year. The fund attracts another net USD100 million at the start of Year 2 and earns a return of 8 percent for that year. The money-weighted rate of return of the fund is *most likely* to be:
- A. less than the time-weighted rate of return.
 - B. the same as the time-weighted rate of return.
 - C. greater than the time-weighted rate of return.
8. An investor is evaluating the returns of three recently formed ETFs. Selected

return information on the ETFs is presented in Exhibit 20:

Exhibit 1: Returns on ETFs

ETF	Time Since Inception	Return Since Inception (%)
1	125 days	4.25
2	8 weeks	1.95
3	16 months	17.18

Which ETF has the highest annualized rate of return?

- A. ETF 1
 - B. ETF 2
 - C. ETF 3
9. The price of a stock at $t = 0$ is USD208.25 and at $t = 1$ is USD186.75. The continuously compounded rate of return, $r_{1,T}$ for the stock from $t = 0$ to $t = 1$ is closest to:
- A. -10.90 percent.
 - B. -10.32 percent.
 - C. 11.51 percent.
10. A USD25 million equity portfolio is financed 20 percent with debt at a cost of 6 percent annual cost. If that equity portfolio generates a 10 percent annual total investment return, then the leveraged return is:
- A. 11.0 percent.
 - B. 11.2 percent.
 - C. 13.2 percent
11. An investment manager's gross return is:
- A. an after-tax nominal, risk-adjusted return.
 - B. the return earned by the manager prior to deduction of trading expenses.
 - C. an often used measure of an investment manager's skill because it does not include expenses related to management or administration.
12. The strategy of using leverage to enhance investment returns:
- A. amplifies gains but not losses.
 - B. doubles the net return if half of the invested capital is borrowed.
 - C. increases total investment return only if the return earned exceeds the borrowing cost.
13. At the beginning of the year, an investor holds EUR10,000 in a hedge fund. The investor borrowed 25 percent of the purchase price, EUR2,500, at an annual interest rate of 6 percent and expects to pay a 30 percent tax on the return she

earns from his investment. At the end of the year, the hedge fund reported the information in Exhibit 22:

Exhibit 1: Hedge Fund Investment

Gross return	8.46%
Trading expenses	1.10%
Managerial and administrative expenses	1.60%

The investor's after-tax return on the hedge fund investment is closest to:

- A. 3.60 percent.
- B. 3.98 percent.
- C. 5.00 percent.

SOLUTIONS

1. C is correct. The nominal risk-free rate is approximated as the sum of the real risk-free interest rate and an inflation premium.
2. C is correct. US Treasury bonds are highly liquid, whereas the bonds of small issuers trade infrequently and the interest rate includes a liquidity premium. This liquidity premium reflects the relatively high costs (including the impact on price) of selling a position. As the two bond issues have the same 30-year maturity, the observed difference in yields would not be solely explained by maturity. Further, the inflation premium embedded in the yield of both bonds is likely to be similar given they are both US-based bonds with the same maturity.
3. A is correct. The geometric mean return for Fund Y is calculated as follows:

$$\bar{R}_G = [(1 + 0.195) \times (1 - 0.019) \times (1 + 0.197) \times (1 + 0.350) \times (1 + 0.057)]^{(1/5)} - 1 \\ = 14.9\%.$$

4. A is correct. The harmonic mean is appropriate for determining the average price per unit as it gives equal weight to each data point and reduces the potential influence of outliers. It is calculated as follows:

$$\bar{X}_H = 4 / [(1/62.00) + (1/76.00) + (1/84.00) + (1/90.00)] = \text{EUR}76.48.$$

5. B is correct. The geometric mean compounds the periodic returns of every period, giving the investor a more accurate measure of the terminal value of an investment.
6. C is correct. The money-weighted rate of return considers both the timing and amounts of investments into the fund. To calculate the money-weighted rate of return, tabulate the annual returns and investment amounts to determine the cash flows.

Year	1	2	3
Balance from previous year	0	USD1,150	USD4,560
New investment	USD1,000	USD2,850	USD40,440
Net balance at the beginning of year	USD1,000	USD4,000	USD45,000
Investment return for the year	15%	14%	-4%
Investment gain (loss)	USD150	USD560	-USD1,800
Balance at the end of year	USD1,150	USD4,560	USD43,200

$$CF_0 = -\text{USD}1,000, CF_1 = -\text{USD}2,850, CF_2 = -\text{USD}40,440, CF_3 = +\text{USD}43,200.$$

$$CF_0 = -1,000$$

$$CF_1 = -2,850$$

$$CF_2 = -40,440$$

$$CF_3 = +43,200$$

$$\frac{CF_0}{(1 + IRR)^0} + \frac{CF_1}{(1 + IRR)^1} + \frac{CF_2}{(1 + IRR)^2} + \frac{CF_3}{(1 + IRR)^3} \\ = \frac{-1,000}{1} + \frac{-2,850}{(1 + IRR)^1} + \frac{-40,440}{(1 + IRR)^2} + \frac{43,200}{(1 + IRR)^3} = 0$$

Solving for *IRR* results in a value of *IRR* = -2.22 percent.

Note that A and B are incorrect because the time-weighted rate of return (TWR) of the fund is the same as the geometric mean return of the fund and is positive:

$$R_{TW} = \sqrt[3]{(1.15)(1.14)(0.96)} - 1 = 7.97\%.$$

7. A is correct. Computation of the money-weighted return, r , requires finding the discount rate that sums the present value of cash flows to zero. Because most of the investment came during Year 2, the money-weighted return will be biased toward the performance of Year 2 when the return was lower. The cash flows are as follows:

$$CF_0 = -10$$

$$CF_1 = -100$$

$$CF_2 = +120.31$$

The terminal value is determined by summing the investment returns for each period $[(10 \times 1.14 \times 1.08) + (100 \times 1.08)]$.

$$\frac{CF_0}{(1 + IRR)^0} + \frac{CF_1}{(1 + IRR)^1} + \frac{CF_2}{(1 + IRR)^2} \\ \frac{-10}{1} + \frac{-100}{(1 + IRR)^1} + \frac{120.31}{(1 + IRR)^2} = 0$$

This results in a value of $IRR = 8.53$ percent.

The time-weighted return of the fund is calculated as follows:

$$R_{TW} = \sqrt[2]{(1.14)(1.08)} - 1 = 10.96\%.$$

8. B is correct. The annualized rate of return for

$$\text{ETF 1 annualized return} = (1.0425^{365/125}) - 1 = 12.92\%$$

$$\text{ETF 2 annualized return} = (1.0195^{52/8}) - 1 = 13.37\%$$

$$\text{ETF 3 annualized return} = (1.1718^{12/16}) - 1 = 12.63\%$$

Despite having the lowest value for the periodic rate, ETF 2 has the highest annualized rate of return because of the reinvestment rate assumption and the compounding of the periodic rate.

9. A is correct. The continuously compounded return from $t = 0$ to $t = 1$ is $r_{0,1} = \ln(S_1/S_0) = \ln(186.75/208.25) = -0.10897 = -10.90\%$.

10. A is correct.

$$R_L = R_p + \frac{V_B}{V_E}(R_p - r_D) \\ = 10\% + \frac{\text{USD5 million}}{\text{USD20 million}}(10\% - 6\%) \\ = 10\% + 0.25 \times 4\% = 11.0\%.$$

11. C is correct. Gross returns are calculated on a pre-tax basis; trading expenses *are* accounted for in the computation of gross returns as they contribute directly to the returns earned by the manager. A is incorrect because investment managers' gross returns are pre-tax and not adjusted for risk. B is incorrect because managers' gross returns do reflect the deduction of trading expenses since they contribute directly to the return earned by the manager.

12. C is correct. The use of leverage can increase an investor's return if the total investment return earned on the leveraged investment exceeds the borrowing cost on debt. A is incorrect because leverage amplifies both gains and losses. B is incorrect because, if half of the invested capital is borrowed, then the investor's gross (not net) return would double.
13. C is correct. The first step is to compute the investor's net return from the hedge fund investment. The net return is the fund's gross return less managerial and administrative expenses of 1.60 percent, or $8.46\% - 1.60\% = 6.86\%$. Note that trading expenses are already reflected in the gross return, so they are not subtracted. The second step is to compute the investor's leveraged return (the investor borrowed EUR2,500 (25 percent) of the purchase), calculated as: follows

$$R_L = R_p + \frac{V_B}{V_E} (R_p - r_D)$$

$$R_L = 6.86\% + \frac{\text{EUR}2,500}{\text{EUR}7,500} (6.86\% - 6\%)$$

$$R_L = 6.86\% + 0.33 \times 0.86\% = 7.15\%.$$

The final step is to compute the after-tax return:

$$\text{After-tax return} = 7.15\% (1 - 0.30) = 5.00\%.$$

LEARNING MODULE

2

The Time Value of Money in Finance

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	calculate and interpret the present value (PV) of fixed-income and equity instruments based on expected future cash flows
<input type="checkbox"/>	calculate and interpret the implied return of fixed-income instruments and required return and implied growth of equity instruments given the present value (PV) and cash flows
<input type="checkbox"/>	explain the cash flow additivity principle, its importance for the no-arbitrage condition, and its use in calculating implied forward interest rates, forward exchange rates, and option values

INTRODUCTION

1

This learning module applies time value of money principles in valuing financial assets. The first lesson focuses on solving for the present value of expected future cash flows associated with bonds and stocks. In the second lesson, the focus shifts to solving for implied bond and stock returns given current prices. This includes solving for and interpreting implied growth rates associated with given stock prices. The final lesson introduces cash flow additivity, an important principle which ensures that financial asset prices do not allow investors to earn risk-free profits, illustrated with several examples. The material covered in this learning module provides an important foundation for candidates in understanding how financial assets are priced in markets.

LEARNING MODULE OVERVIEW



- The price of a bond is the sum of the present values of the bond's promised coupon payments and its par value. For discount bonds, the price reflects only the present value of the bond's par value.
- The value of a stock should reflect the sum of the present values of the stock's expected future dividends in perpetuity.
- Stock valuation models are classified by the expected growth pattern assumed for future dividends: (1) no growth, (2) constant growth, or (3) changing dividend growth.

- If a bond's price is known, the bond's implied return can be computed using the bond's price and its promised future cash flows.
- A stock's required return can be estimated given the stock's current price and assumptions about its expected future dividends and growth rates.
- A stock's implied dividend growth rate can be estimated given the stock's current price and assumptions about its expected future dividends and required return.
- If valuing two (or more) cash flow streams, the cash flow additivity principle allows for the cash flow streams to be compared (as long as the cash flows occur at the same point in time).
- Application of cash flow additivity allows for confirmation that asset prices are the same for economically equivalent assets (even if the assets have differing cash flow streams).
- Several real-world applications of cash flow additivity are used to illustrate no-arbitrage pricing.

2

TIME VALUE OF MONEY IN FIXED INCOME AND EQUITY

- calculate and interpret the present value (PV) of fixed-income and equity instruments based on expected future cash flows

The timing of cash flows associated with financial instruments affects their value, with cash inflows valued more highly the sooner they are received. The time value of money represents the trade-off between cash flows received today versus those received on a future date, allowing the comparison of the current or present value of one or more cash flows to those received at different times in the future. This difference is based upon an appropriate discount rate r as shown in the prior learning module, which varies based upon the type of instrument and the timing and riskiness of expected cash flows.

In general, the relationship between a current or present value (PV) and future value (FV) of a cash flow, where r is the stated discount rate per period and t is the number of compounding periods, is as follows:

$$FV_t = PV(1 + r)^t. \quad (1)$$

If the number of compounding periods t is very large, that is, $t \rightarrow \infty$, we compound the initial cash flow on a continuous basis as follows:

$$FV_t = PV e^{rt}. \quad (2)$$

Conversely, present values can be expressed in future value terms, which requires recasting Equation 1 as follows:

$$\begin{aligned} FV_t &= PV(1 + r)^t \\ PV &= FV_t \left[\frac{1}{(1 + r)^t} \right]. \\ PV &= FV_t (1 + r)^{-t} \end{aligned} \quad (3)$$

The continuous time equivalent expression of Equation 3 is as follows:

$$PV_t = FV e^{-r t}. \quad (4)$$

Fixed-Income Instruments and the Time Value of Money

Fixed-income instruments are debt instruments, such as a bond or a loan, that represent contracts under which an issuer borrows money from an investor in exchange for a promise of future repayment. The discount rate for fixed-income instruments is an interest rate, and the rate of return on a bond or loan is often referred to as its yield-to-maturity (YTM).

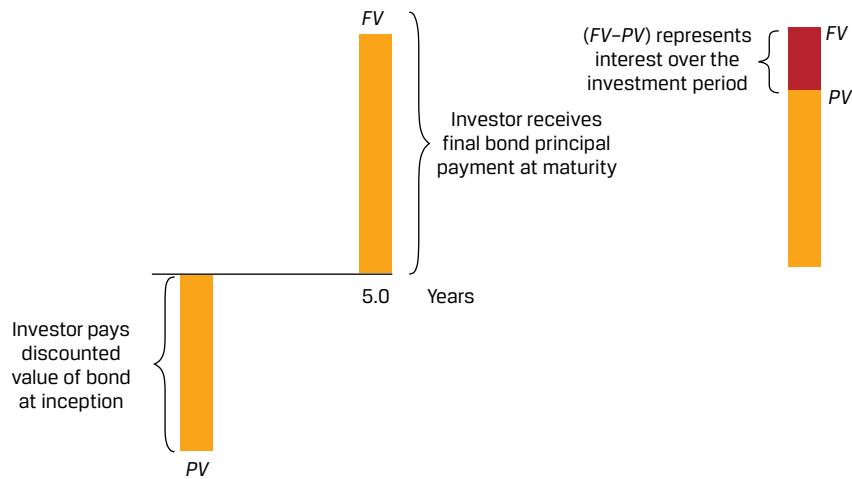
Cash flows associated with fixed-income instruments usually follow one of three general patterns:

- Discount: An investor pays an initial price (PV) for a bond or loan and receives a single principal cash flow (FV) at maturity. The difference ($FV - PV$) represents the interest earned over the life of the instrument.
- Periodic Interest: An investor pays an initial price (PV) for a bond or loan and receives interest cash flows (PMT) at pre-determined intervals over the life of the instrument, with the final interest payment and the principal (FV) paid at maturity.
- Level Payments: An investor pays an initial price (PV) and receives uniform cash flows at pre-determined intervals (A) through maturity which represent both interest and principal repayment.

Discount Instruments

The discount cash flow pattern is shown in Exhibit 1:

Exhibit 1: Discount Bond Cash Flows



The present value (PV) calculation for a discount bond with principal (FV) paid at time t with a market discount rate of r per period is:

$$PV(\text{Discount Bond}) = FV_t / (1 + r)^t. \quad (5)$$

The investor's sole source of return is the difference between the price paid (PV) and full principal (FV) received at maturity. This type of bond is often referred to as a **zero-coupon bond** given the lack of intermediate interest cash flows, which for bonds are generally referred to as coupons.

EXAMPLE 1

Discount Bonds under Positive and Negative Interest Rates

While most governments issue fixed coupon bonds with principal paid at maturity, for many government issuers such as the United States, United Kingdom, or India, investors buy and sell individual interest or principal cash flows separated (or stripped) from these instruments as discount bonds. Consider a single principal cash flow payable in 20 years on a Republic of India government bond issued when the YTM is 6.70 percent. For purposes of this simplified example, we use annual compounding, that is, t in Equation 5 is equal to the number of years until the cash flow occurs.

- What should an investor expect to pay for this discount bond per INR100 of principal?

Solution:

INR27.33

We solve for PV given r of 6.70 percent, $t = 20$, and FV_{20} of INR100 using Equation 5:

$$PV = \text{INR}100 / (1 + 0.067)^{20} = \text{INR}27.33.$$

We may also use the Microsoft Excel or Google Sheets PV function:

= PV (rate, nper, pmt, FV, type),

where:

rate = the market discount rate per period,

nper = the number of periods,

pmt = the periodic coupon payment (zero for a discount bond),

FV = future or face value, and

type = payments made at the end (0 as in this case) or beginning (1) of each period.

As a cash outflow (or price paid), the Excel PV solution has a negative sign, so:

$$PV = (27.33) = \text{PV} (0.067, 20, 0, 100, 0).$$

While the principal (FV) is a constant INR100, the price (PV) changes as both time passes, and interest rates change.

- If we assume that interest rates remain unchanged, what is the price (PV) of the bond in three years' time?

Solution:

INR33.21

Solve for PV by substituting $t = 17$ into the prior calculation using Equation 5:

$$PV = \text{INR}100 / (1 + 0.067)^{17} = \text{INR}33.21.$$

The INR5.88 price increase with a constant r represents implied interest earned over the three years. If the interest rate is positive, the PV generally rises (or accretes) over time to reach FV as time passes and t approaches zero.

3. Prices also change as interest rates change. Suppose after purchase at $t = 0$ we observe an immediate drop in the bond price to INR22.68224 per INR100 of principal. What is the implied interest rate on the discount bond?

Solution:

7.70 percent

Here we may solve for r in Equation 5 as follows:

$$\text{INR}22.68224 = \text{INR}100 / (1 + r)^{20}.$$

Rearranging Equation 5, we get:

$$r = 7.70 \text{ percent} = (100/22.68224)^{1/20} - 1.$$

Alternatively, we may use the Microsoft Excel or Google Sheets RATE function:

= RATE (nper, pmt, PV, FV, type, guess) using the same arguments as above, with *guess* as an optional estimate argument, which must be between 0 and 1 and defaults to 0.1, as follows:

$$7.70 \text{ percent} = \text{RATE} (20, 0, -22.68224, 100, 0, 0.1).$$

This shows that a 100 basis point (1.00 percent) increase in r results in an INR4.65 price decrease, underscoring the inverse relationship between price and YTM.

4. Now consider a bond issued at negative interest rates. In July 2016, Germany became the first eurozone country to issue 10-year sovereign bonds at a negative yield. If the German government bond annual YTM was -0.05 percent when issued, calculate the present value (PV) of the bond per EUR100 of principal (FV) at the time of issuance.

Solution:

EUR100.50

Solve for PV given r of -0.05 percent, $t = 10$, and FV_{10} of EUR100 using Equation 1:

$$PV = \text{EUR}100.50 = \text{EUR}100 / (1 - 0.0005)^{10},$$

or

$$PV = (100.50) = \text{PV} (-0.0005, 10, 0, 100, 0).$$

At issuance, this bond is priced at a **premium**, meaning that an investor purchasing the bond at issuance paid EUR0.50 above the future value expected at maturity, which is the principal.

5. Six years later, when German inflation reached highs not seen in decades and investors increased their required nominal rate of return, say we observed that the German government bond in question 4 is now trading at

a price (PV) of EUR95.72 per EUR100 principal. What is the YTM on this bond?

Solution:

1.10 percent

Use Equation 5 to solve for r given a PV of 95.72, FV of 100, and $t = 4$, we get:

$$\text{EUR95.72} = \text{EUR100} / (1 + r)^4; r = 1.10\%,$$

or

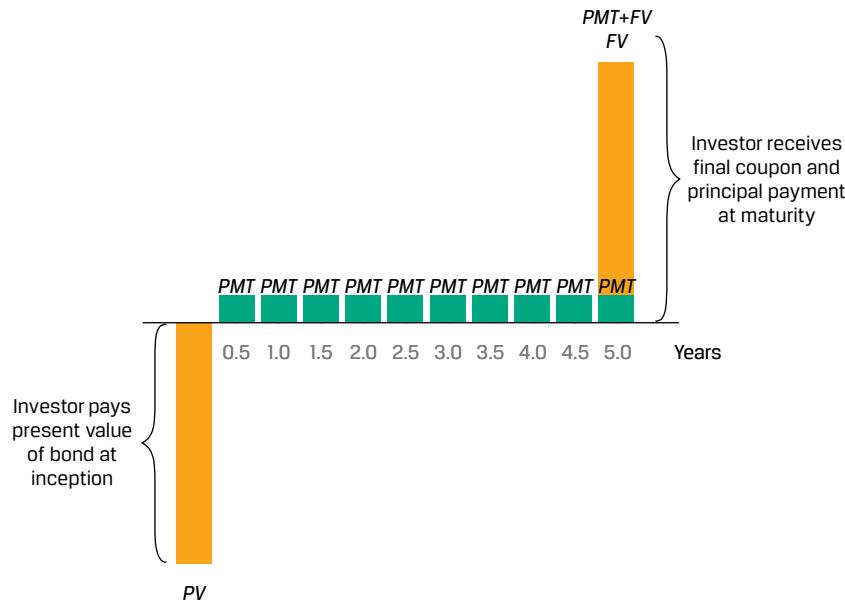
$$1.10\% = \text{RATE}(4, 0, -95.72, 100, 0, 0.1).$$

In the case of the Republic of India discount bond, a higher interest rate will reduce the bond's price. However, in contrast to the accreting price of a discount bond, the premium bond's price will decline (or amortize) over time to reach the EUR100 future value at maturity.

Coupon Instrument

A periodic cash flow pattern for fixed-income interest payments with principal repaid at maturity is shown in Exhibit 2. In this case, all the periodic cash flow payments are identical and occur on a semiannual basis.

Exhibit 2: Coupon Bond Cash Flows



Pricing a coupon bond extends the single cash flow calculation for a discount bond to a general formula for calculating a bond's price (PV) given the market discount rate on a coupon date, as follows:

$$PV(\text{Coupon Bond})$$

$$= PMT_1 / (1 + r)^1 + PMT_2 / (1 + r)^2 + \dots + (PMT_N + FV_N) / (1 + r)^N. \quad (6)$$

EXAMPLE 2**Hellenic Republic of Greece Annual Coupon Bond**

At the height of the COVID pandemic, the government of Greece issued a 2 percent annual coupon bond maturing in seven years.

1. If the observed YTM at issuance was 2.00 percent, what was the issuance price (PV) per EUR100 of principal?

Solution:

EUR100

Solve for PV using Equation 6 with $PMT_i = \text{EUR}2$, $r = 2.00\%$, and $FV = \text{EUR}100$:

$$PV = \text{EUR}100$$

$$= \frac{2}{1.02} + \frac{2}{1.02^2} + \frac{2}{1.02^3} + \frac{2}{1.02^4} + \frac{2}{1.02^5} + \frac{2}{1.02^6} + \frac{2}{1.02^7}.$$

We may solve this using the Microsoft Excel or Google Sheets PV function introduced earlier ($PV(0.02, 7, 2, 100, 0)$). The issuance price equals the principal ($PV = FV$). This relationship holds on a coupon date for any bond where the fixed periodic coupon is equal to the discount rate.

The present value of each cash flow may be solved using Equation 5. For example, the final EUR102 interest and principal cash flow in seven years is:

$$PV = \text{EUR}88.80 = \text{EUR}102/(1.02)^7.$$

The following table shows the present value of all bond cash flows:

Years/Periods	0	1	2	3	4	5	6	7
FV		2	2	2	2	2	2	102
PV	100.00	1.96	1.92	1.88	1.85	1.81	1.78	88.80

2. Next, let's assume that, exactly one year later, a sharp rise in Eurozone inflation drove the Greek bond's price lower to EUR93.091 (per EUR100 of principal). What would be the implied YTM expected by investors under these new market conditions?

Solution:

3.532 percent

In this case, we must solve for r using Equation 6, with PV equal to 93.091, as follows:

$$PV = 93.091 = 2/(1+r) + 2/(1+r)^2 + 2/(1+r)^3 + 2/(1+r)^4 + 102/(1+r)^5.$$

Here we may use the Microsoft Excel or Google Sheets RATE function (`RATE (5, 2, 93.091, 100, 0, 0.1)`) to solve for r of 3.532 percent.

Investors in fixed coupon bonds face a capital loss when investors expect a higher YTM.

The interest rate r used to discount all cash flows in Equation 6 is the bond's YTM, which is typically quoted on an annual basis. However, many bonds issued by public or private borrowers pay interest on a semiannual basis. In Example 3, we revisit the Republic of India bond from which a single cash flow was stripped in an earlier simplified example.

EXAMPLE 3**Republic of India Semiannual Coupon Bond**

Consider the 20-year Republic of India government bond from which the discount bond in Example 1 was separated (or stripped). The bond was issued at an annualized coupon rate of 6.70 percent and a YTM of 6.70 percent, and the coupon payments are semiannual.

1. Solve for the price of the bond at issuance.

Solution:

INR100

As coupon periods are semiannual, for a principal of INR100, $PMT = \text{INR}3.35 (=6.70/2)$ and the periodic discount rate is 3.35 percent ($=6.70\% / 2$), as follows:

$$PV = 3.35/(1.0335) + 3.35/(1.0335)^2 + \dots + 3.35/(1.0335)^{39} + 103.35/(1.0335)^{40}.$$

As shown in Example 2, because *the coupon rate is equal to the YTM*, we expect this bond to have a *PV* of INR100 at issuance.

2. What is the bond's price if the YTM immediately rises to 7.70 percent?

Solution:

INR89.88

We can solve for the *PV* as INR89.88 using Equation 6 with $r = 3.85\% (=7.70/2)$ or using the Microsoft Excel or Google Sheets PV function (PV (0.0385, 40, 3.35, 100, 0)). The first and final three cash flows are shown below:

		Years	0	0.5	1	1.5	19	19.5	20
r	Periods	0	1	2	3	38	39	40	
	FV			3.35	3.35	3.35	3.35	3.35	103.35
6.70%	PV	100.00	3.24	3.14	3.03	0.96	0.93	27.66	
7.70%	PV	89.88	3.23	3.11	2.99	0.80	0.77	22.81	

3. Recalculate the discount bond price for the final principal payment in 20 years from Example 1 using a 6.70 percent semiannual discount rate.

Solution:

INR26.77

Note that the *PV* calculation using the same annual discount rate for 40 semiannual periods will differ slightly using Equation 5, as follows:

$$PV = \text{INR}27.66 = (PMT_{40} + FV_{40})/(1+r/2)^{40},$$

$$PV(PMT_{40}) = \text{INR}0.90 = 3.35 / (1.0335)^{40},$$

$$PV(FV_{40}) = \text{INR}26.77 = 100 / (1.0335)^{40}.$$

Compounding on a semiannual basis for 40 periods, $PV(FV_{40})$ of 26.77 is less than the original *PV* of 27.33 using 20 annual periods from Example 1 (since $1/(1+r)^t > 1/(1+r/2)^{2t}$ when $r \geq 0$).

A **perpetual bond** is a less common type of coupon bond with no stated maturity date. Most perpetual bonds are issued by companies to obtain equity-like financing and often include redemption features. As $N \rightarrow \infty$ in Equation 6, we can simplify this to solve for the present value of a **perpetuity** (or perpetual fixed periodic cash flow without early redemption), where $r > 0$, as follows:

$$PV(\text{Perpetual Bond}) = PMT / r. \quad (7)$$

EXAMPLE 4

KB Financial Perpetual Bond

In 2020, KB Financial (the holding company for Kookmin Bank) issued KRW325 billion in perpetual bonds with a 3.30 percent quarterly coupon.

1. Calculate the bond's YTM if the market price was KRW97.03 (per KRW100).

Solution:

3.40 percent

Solve for r in Equation 7 given a PV of KRW97.03 and a periodic quarterly coupon (PMT) of KRW0.825 ($= (3.30\% \times \text{KRW}100)/4$):

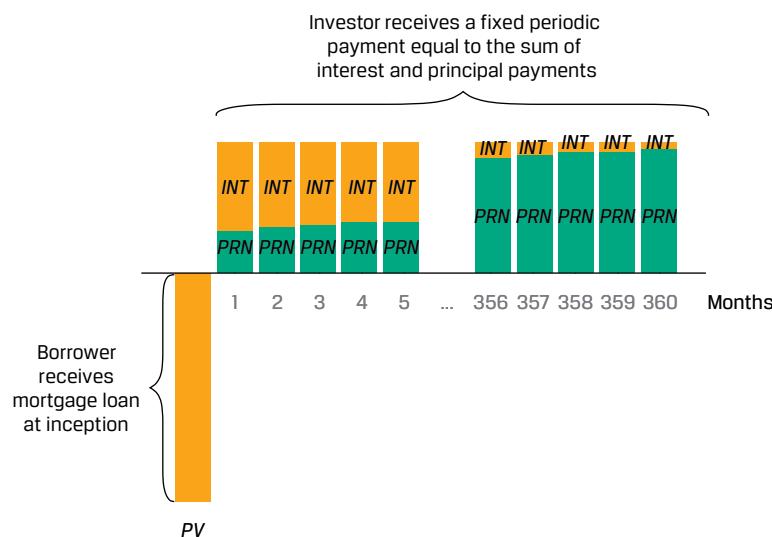
$$\text{KRW}97.03 = \text{KRW}0.825 / r ;$$

$r = 0.85\%$ per period, or 3.40% ($=0.85\% \times 4$) on an annualized basis.

Annuity Instruments

Examples of fixed-income instruments with level payments, which combine interest and principal cash flows through maturity, include fully amortizing loans such as mortgages and a fixed-income stream of periodic cash inflows over a finite period known as an annuity. Exhibit 3 illustrates an example of level monthly cash flows based upon a mortgage.

Exhibit 3: Mortgage Cash Flows



We may calculate the periodic annuity cash flow (A), which occurs at the end of each respective period, as follows:

$$A = \frac{r(PV)}{1 - (1 + r)^{-t}}, \quad (8)$$

where:

A = periodic cash flow,

r = market interest rate per period,

PV = present value or principal amount of loan or bond, and

t = number of payment periods.

EXAMPLE 5

Mortgage Cash Flows

An investor seeks a fixed-rate 30-year mortgage loan to finance 80 percent of the purchase price of USD1,000,000 for a residential building.

1. Calculate the investor's monthly payment if the annual mortgage rate is 5.25 percent.

Solution:

Solve for A using Equation 8 with $r = 0.4375\%$ ($= 5.25\%/12$), t of 360, and PV of USD800,000 ($= 80\% \times \text{USD}1,000,000$):

$$A = \text{USD}4,417.63 = \frac{0.4375\% (\text{USD}800,000)}{1 - (1 + 0.4375\%)^{-360}}.$$

The 360 level monthly payments consist of both principal and interest.

2. What is the principal amortization and interest breakdown of the first two monthly cash flows?

Solution:

Month 1 interest is USD3,500 and principal is USD917.63

Month 2 interest is USD3,495.99 and principal is USD921.64

Month 1:

Interest: $\text{USD}3,500 = \text{USD}800,000 \times 0.4375\% (= PV_0 \times r)$

Principal Amortization: $\text{USD}4,417.63 - \text{USD}3,500 = \text{USD}917.63$

Remaining Principal (PV_1): $\text{USD}799,082.37 = \text{USD}800,000 - \text{USD}917.63$

Month 2:

Interest: $\text{USD}3,495.99 = \text{USD}799,082.37 \times 0.4375\% (= PV_1 \times r)$

Principal Amortization: $\text{USD}4,417.63 - \text{USD}3,495.99 = \text{USD}921.64$

Remaining Principal (PV_2): $\text{USD}798,160.73 = \text{USD}799,082.37 - \text{USD}921.64$

Although the periodic mortgage payment is constant, the proportion of interest per payment declines while the principal amortization rises over

time. The following spreadsheet shows the first and final three monthly cash flows.

Month	Total Monthly Payment	Monthly Interest Payment	Monthly Principal Repayment	Remaining Principal
1	\$ 4,417.63	\$ 3,500.00	\$ 917.63	\$ 799,082.37
2	\$ 4,417.63	\$ 3,495.99	\$ 921.64	\$ 798,160.73
3	\$ 4,417.63	\$ 3,491.95	\$ 925.68	\$ 797,235.05
...				
358	\$ 4,417.63	\$ 57.48	\$ 4,360.15	\$ 8,777.61
359	\$ 4,417.63	\$ 38.40	\$ 4,379.23	\$ 4,398.39
360	\$ 4,417.63	\$ 19.24	\$ 4,398.39	\$ 0.00

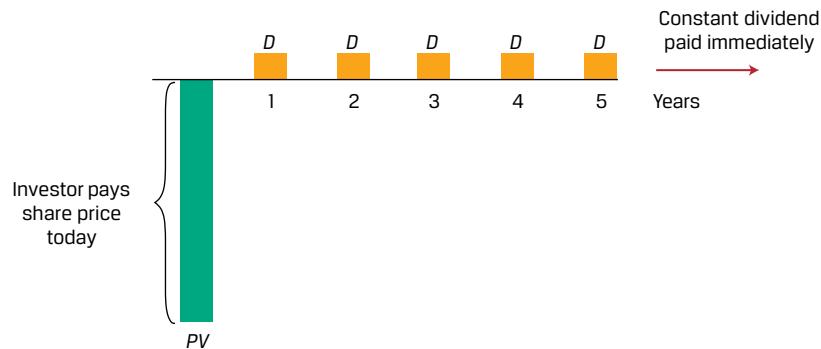
Equity Instruments and the Time Value of Money

Equity investments, such as preferred or common stock, represent ownership shares in a company which entitle investors to receive any discretionary cash flows in the form of dividends. Unlike fixed-income instruments, equity investments have no maturity date and are assumed to remain outstanding indefinitely, or until a company is sold, restructured, or liquidated. One way to value a company's shares is by discounting expected future cash flows using an expected rate of return (r). These cash flows include any periodic dividends received plus the expected price received at the end of an investment horizon.

Common assumptions associated with valuing equity instruments based upon dividend cash flows often follow one of three general approaches:

- Constant Dividends: An investor pays an initial price (PV) for a preferred or common share of stock and receives a fixed periodic dividend (D).
- Constant Dividend Growth Rate: An investor pays an initial price (PV) for a share of stock and receives an initial dividend in one period (D_{t+1}), which is expected to grow over time at a constant rate of g .
- Changing Dividend Growth Rate: An investor pays an initial price (PV) for a share of stock and receives an initial dividend in one period (D_{t+1}). The dividend is expected to grow at a rate that changes over time as a company moves from an initial period of high growth to slower growth as it reaches maturity.

The simplest case of a stock that is assumed to pay constant dividends in perpetuity is shown in Exhibit 4.

Exhibit 4: Equity Cash Flows with Constant Dividends

The price of a preferred or common share expected to pay a constant periodic dividend is an infinite series that simplifies to the formula for the present value of a perpetuity shown and is similar to the valuation of a perpetual bond that we encountered earlier. Specifically, the valuation in Equation 7:

$$PV_t = \sum_{i=1}^{\infty} \frac{D_t}{(1+r)^i}, \text{ and} \quad (9)$$

$$PV_t = \frac{D_t}{r}. \quad (10)$$

EXAMPLE 6**Constant Dividend Cash Flows**

Shipline PLC is a company that pays regular dividends of GBP1.50 per year, which are expected to continue indefinitely.

- What is Shipline's expected stock price if shareholders' required rate of return is 15 percent?

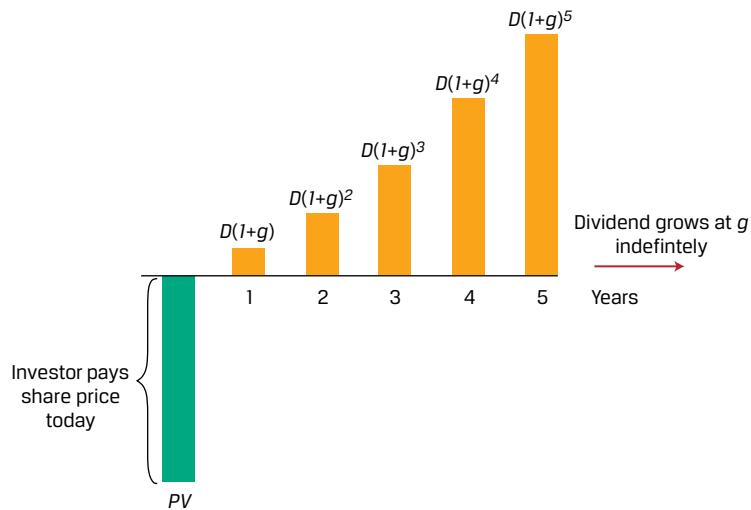
Solution:

GBP10

Solve for PV using Equation 10, where D is GBP1.50 and the rate of return per period r is 15 percent:

$$PV = \text{GBP}10.00 = \text{GBP}1.50 / 0.15.$$

Alternatively, a common equity forecasting approach is to assume a constant dividend growth rate (g) into perpetuity, as illustrated in Exhibit 5.

Exhibit 5: Equity Cash Flows with Constant Dividend Growth

If dividends grow at a rate of g per period and are paid at the end of each period, the next dividend (at time $t + 1$) may be shown as follows:

$$D_{t+1} = D_t (1 + g), \quad (11)$$

or generally in i periods as:

$$D_{t+i} = D_t (1 + g)^i. \quad (12)$$

If dividend cash flows continue to grow at g indefinitely, then we may rewrite Equation 10 as follows:

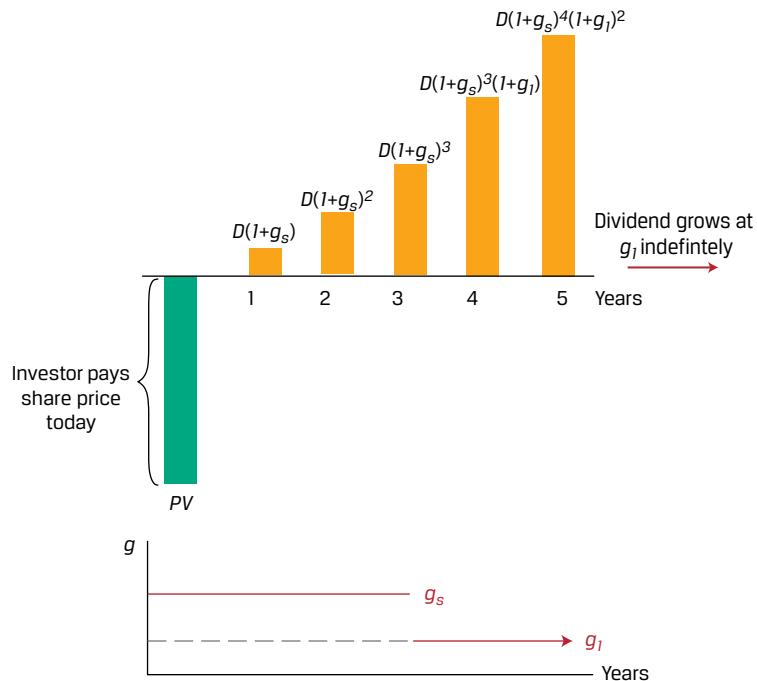
$$PV_t = \sum_{i=1}^{\infty} \frac{D_t (1 + g)^i}{(1 + r)^i}, \quad (13)$$

which simplifies to:

$$PV_t = \frac{D_t (1 + g)}{r - g} = \frac{D_{t+1}}{r - g}, \quad (14)$$

where $r - g > 0$.

An alternative to constant dividend growth is an assumption of changing growth. This is common for evaluating the share price of firms expected to experience an initial rapid rise in cash flow, followed by slower growth as a company matures. The simplest form of changing dividend growth is the two-stage model shown in Exhibit 6.

Exhibit 6: Equity Cash Flows with Two-Stage Dividend Growth


The example in Exhibit 6 shows a company with an initial higher short-term dividend growth of g_s for the first three years, followed by lower long-term growth (g_l , where $g_s > g_l$) indefinitely thereafter. If we generalize the initial growth phase to n periods followed by indefinite slower growth at g_l , we obtain a modified version of Equation 14 as follows:

$$PV_t = \sum_{i=1}^n \frac{D_t(1+g_s)^i}{(1+r)^i} + \sum_{j=n+1}^{\infty} \frac{D_{t+n}(1+g_l)^j}{(1+r)^j}. \quad (15)$$

Note that the second expression in Equation 15 involves constant growth starting in n periods, for which we can substitute the geometric series simplification:

$$PV_t = \sum_{i=1}^n \frac{D_t(1+g_s)^i}{(1+r)^i} + \frac{E(S_{t+n})}{(1+r)^n}, \quad (16)$$

where the stock value of the stock in n periods ($E(S_{t+n})$) is referred to as the **terminal value** and is equal to the following:

$$E(S_{t+n}) = \frac{D_{t+n+1}}{r - g_l}. \quad (17)$$

We revisit the Shipline PLC stock price example to evaluate the effects of constant dividend, constant dividend growth and changing dividend growth assumptions on a company's expected share price.

EXAMPLE 7
Constant and Changing Dividend Growth

Recall that based on a constant GBP1.50 annual dividend and required return of 15 percent, we showed Shipline PLC's expected stock price to be GBP10.00. Suppose instead that an investment analyst assumes that Shipline will grow its annual dividend by 6 percent per year indefinitely.

1. How does Shipline's expected share price change under the analyst's constant growth assumption?

Solution:

GBP17.67

Using Equation 14, solve for PV using D_{t+1} of GBP1.59 ($=\text{GBP}1.50 \times (1 + 0.06)$), r of 15%, and g of 6%:

$$PV = \text{GBP}1.59 / (15\% - 6\%) = \text{GBP}17.67.$$

Note that a higher growth rate g increases the PV by reducing the denominator ($r - g$).

2. How does Shipline's expected share price change if we assume instead an initial dividend growth of 6 percent over a three-year period followed by constant 2 percent dividend growth thereafter?

Solution:

GBP13.05

We may solve for Shipline's expected share price (PV) using Equations 15–17 and $D_t = 1.5$, $g_s = 6\%$, $n = 3$, $r = 15\%$, and $g_l = 2\%$, as follows:

$$PV_t = \sum_{i=1}^3 \frac{1.50(1 + 0.06)^i}{(1 + 0.15)^i} + \sum_{j=4}^{\infty} \frac{D_{t+3}(1 + 0.02)^j}{(1 + 0.15)^j}.$$

As a first step, we calculate the present value of dividends associated with the higher 6 percent growth rate over the first three years as shown in the first expression. As a second step, we calculate the present value of future dividends at a lower 2 percent growth rate for an indefinite period. The sum of these two steps is the expected share price.

Step 1 Solve for the first step as GBP3.832 with dividend cash flows for the initial growth period as shown in the following spreadsheet:

t	D	PV(D) at r%
0	1.500	
1	1.590	1.383
2	1.685	1.274
3	1.787	1.175
<i>PV(ST growth)</i>		3.832

Step 2 As shown in Equation 17, the second expression simplifies as follows:

$$\frac{E(S_4)}{(1 + r)^3}, \text{ with } E(S_4) = \frac{D_4}{r - g_l}.$$

We may solve for D_4 as GBP1.894 ($=1.787 \times 1.02 = D_3(1 + g_l)$) and the second expression to be GBP9.22 as follows:

$$\text{GBP}9.22 = \frac{1.894(0.15 - 0.02)}{(1.15)^3}.$$

Step 3 The sum of these two steps gives us an expected Shipline share price of GBP13.05 ($=3.83 + 9.22$).

In the following table, we show the sensitivity of Shipline's expected share price (PV) to changes in the long-term growth rate (g_l) after three years of 6 percent dividend growth:

Share Price	LT Growth Rate
11.66	0%
13.05	2%
14.94	4%
17.67	6%

3

IMPLIED RETURN AND GROWTH



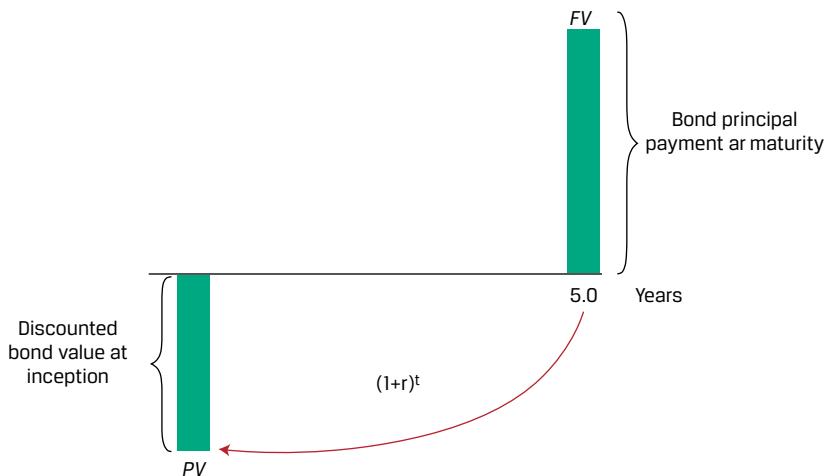
calculate and interpret the implied return of fixed-income instruments and required return and implied growth of equity instruments given the present value (PV) and cash flows

Lesson 1 addressed the time value of money trade-off between cash flows occurring today versus those in the future for certain fixed income and equity instruments. Market participants often face a situation in which both the present and future values of instruments or cash flows may be known. In this case it becomes possible to solve for the implied return or growth rate implied by the current price and features of the future cash flows. In this sense, solving for the implied growth or return provides a view of the market expectations that are incorporated into the market price of the asset.

Implied Return for Fixed-Income Instruments

Fixed-income instruments are characterized by contractual interest and principal cash flows. If we observe the present value (or price) and assume that all future cash flows occur as promised, then the discount rate (r) or yield-to-maturity (YTM) is a measure of implied return under these assumptions for the cash flow pattern.

In the case of a discount bond or instrument, recall that an investor receives a single principal cash flow (FV) at maturity, with $(FV - PV)$ representing the implied return, as shown in Exhibit 7.

Exhibit 7: Discount Bond Implied Return

We may rearrange Equation 5 from to solve for the implied periodic return earned over the life of the instrument (t periods):

$$r = \sqrt[t]{\frac{FV_t}{PV}} - 1 = \left(\frac{FV_t}{PV}\right)^{\frac{1}{t}} - 1. \quad (18)$$

EXAMPLE 8**Discount Bond Implied Return**

Recall from Example 1 that in 2016, German 10-year government bond investors faced a price of EUR100.50 per EUR100 principal and an annual YTM of -0.05 percent at issuance. That is, the German 10-year government bond was initially priced by the market to provide a negative return to the investor. Six years later, these bonds traded at a price (PV) of EUR95.72 per EUR100 principal.

- What was the initial investor's implied return on this bond over the six-year holding period?

Solution:

-0.81 percent

We can solve for an investor's annualized return (r) using Equation 18 and a PV of 100.5, FV of 95.72, and t of 6 as follows:

$$r = -0.81\% = \left(\frac{95.72}{100.5}\right)^{\frac{1}{6}} - 1.$$

Note that an investor who purchases the discount bond at issuance and receives EUR100 in 10 years expects an implied return equal to the issuance YTM of -0.05 percent. However, the EUR4.78 price decline for the first six years translates into an annualized return of -0.81 percent, which is below the initial YTM of -0.05 percent. This negative return is consistent with an expected decline in the price (PV) of a discount bond amid higher inflation.

2. What is the expected return of an investor who purchases the discount bond at EUR95.72 and holds it for the remaining four years?

Solution:

1.10 percent

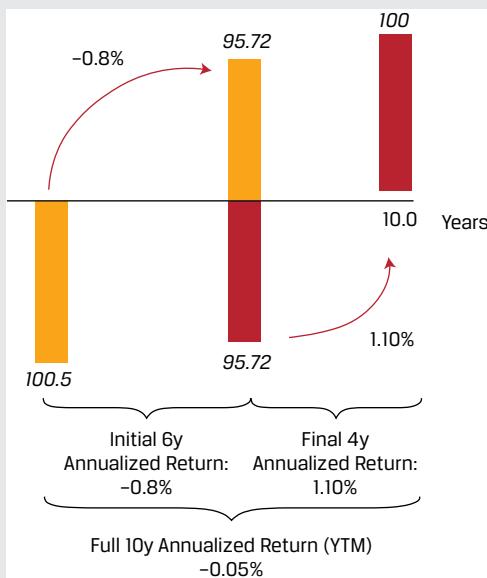
Here the current price of 95.72 in Equation 18 is now the PV , and the principal of 100 is FV , with $t = 4$:

$$r = 1.10\% = \left(\frac{100}{95.72}\right)^{\frac{1}{4}} - 1.$$

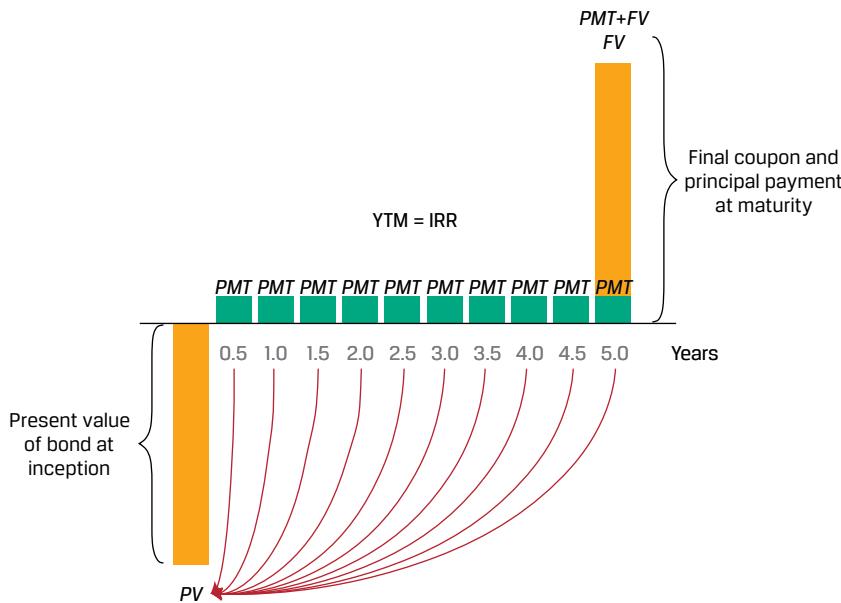
Note that r of 1.10 percent is equal to the YTM for the remaining four years we derived in Example 1. The cumulative returns across the two investors is equal to the initial -0.05 percent YTM, assuming no transaction costs as follows:

$$-0.05\% = [(1 - 0.0081)^6 \times (1 + 0.011)^4]^{\frac{1}{10}} - 1.$$

These relationships are summarized in the following diagram:



Unlike discount bonds, fixed-income instruments that pay periodic interest have cash flows throughout their life until maturity. The uniform discount rate (or internal rate of return) for all promised cash flows is the YTM, a single implied market discount rate for all cash flows regardless of timing as shown for a coupon bond in Exhibit 8.

Exhibit 8: Coupon Bond Implied Return

The YTM assumes an investor expects to receive all promised cash flows through maturity and reinvest any cash received at the same YTM. For coupon bonds, this involves periodic interest payments only, while for level payment instruments such as mortgages, the calculation assumes both interest and amortized principal may be invested at the same rate. Like other internal rates of return, the YTM cannot be solved using an equation, but it may be calculated using iteration with a spreadsheet or calculator, a process that solves for r in Equation 19, as follows:

$$PV(\text{Coupon Bond}) = \text{PMT}_1 / (1 + r)^1 + \text{PMT}_2 / (1 + r)^2 + \dots + (\text{PMT}_N + FV_N) / (1 + r)^N, \quad (19)$$

where FV equals a bond's principal and N is the number of periods to maturity.

The Microsoft Excel or Google Sheets YIELD function can be used to calculate YTM for fixed-income instruments with periodic interest and full principal payment at maturity:

```
= YIELD (settlement, maturity, rate, pr, redemption,
frequency, [basis])
```

where:

settlement = settlement date entered using the DATE function;

maturity = maturity date entered using the DATE function;

rate = semi-annual (or periodic) coupon;

pr = price per 100 face value;

redemption = future value at maturity;

frequency = number of coupons per year; and

[basis] = day count convention, typically 0 for US bonds (30/360 day count).

Example 9 illustrates the implied return on fixed income instruments with periodic interest.

EXAMPLE 9**Greek Coupon Bond Implied Return**

Recall from Example 2 that seven-year Greek government bonds issued in 2020 with a 2.00 percent annual coupon had a price of EUR93.091 per EUR100 principal two years later.

1. What is the implied two-year return for an investor able to reinvest periodic interest at the original YTM of 2.00 percent?

Solution:

-1.445 percent

Using Equation 18 as for the discount bond, we must first calculate the future value (FV) after two years including the future price of 93.091 and all cash flows reinvested to that date:

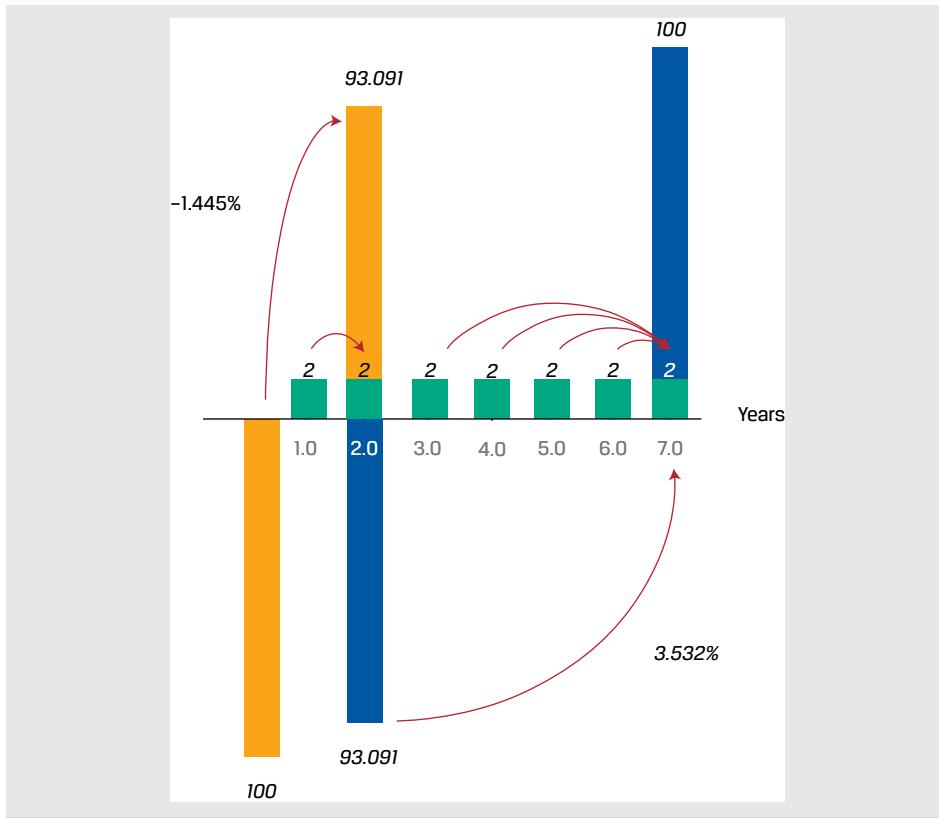
$$FV_2 = PMT_1(1 + r) + PMT_2 + PV_2$$

$$FV_2 = 97.13 = 2 \times (1.02) + 2 + 93.091$$

$$r = -1.445\% = \left(\frac{97.13}{100}\right)^{\frac{1}{2}} - 1.$$

This negative annualized return was due to a EUR6.91 (=100 – 93.091) capital loss which exceeded the periodic interest plus reinvestment proceeds of EUR4.04 (=2 × (1.02) + 2). The fall in the price of the bond to EUR93.091 in 2022 was a result of a rise in Eurozone inflation over the period.

For an investor purchasing the 2 percent coupon bond in 2022 at a price of EUR93.091, recall that we solved for an r (or YTM) of 3.532 percent. Note that this YTM calculation for the remaining five years assumes all cash flows can be reinvested at 3.532 percent through maturity. The rate(s) at which periodic interest can be reinvested is critical for the implied return calculation for coupon bonds as shown below. YTM is computed assuming that the bond is held to maturity.

Implied Return and Growth**Equity Instruments, Implied Return, and Implied Growth**

As noted in the discussion of calculating the present value of an equity investment, the price of a share of stock reflects not only the required return but also the growth of cash flows. If we begin with an assumption of constant growth of dividends from Equation 14, we can rearrange the formula as follows:

$$r - g = \frac{D_t(1+g)}{PV_t} = \frac{D_{t+1}}{PV_t}. \quad (20)$$

The left-hand side of Equation 20 simply reflects the difference between the required return and the constant growth rate, and the right-hand side is the dividend yield of the stock based on expected dividends over the next period. Thus, the implied return on a stock given its expected dividend yield and growth is given by Equation 21, as follows:

$$r = \frac{D_t(1+g)}{PV_t} + g = \frac{D_{t+1}}{PV_t} + g. \quad (21)$$

Simply put, if we assume that a stock's dividend grows at a constant rate in perpetuity, the stock's implied return is equal to its expected dividend yield plus the constant growth rate.

Alternatively, we may be interested in solving for a stock's implied growth rate, and this relation is given by Equation 22:

$$g = \frac{r * PV_t - D_t}{PV_t + D_t} = r - \frac{D_{t+1}}{PV_t}. \quad (22)$$

If a stock's next expected dividend is known, then we can calculate the implied growth by deducting its expected dividend yield from its required return.

EXAMPLE 10**Implied Return and Growth for a Stock**

Coca-Cola Company stock trades at a share price of USD63.00 and its annualized expected dividend per share during the next year is USD1.76.

1. If an analyst expects Coca-Cola's dividend per share to increase at a constant 4 percent per year indefinitely, calculate the required return expected by investors.

Solution:

6.79 percent

Using Equation 21,

$$r = \frac{1.76}{63} + 0.04 = 0.0679.$$

The required return expected for Coca-Cola stock is 6.79 percent given its current price, expected dividend, and expected dividend growth rate. Investor expectations of future stock returns are inferred by the combination of the current price, expected future cash flows and the cash flow growth rate. Suppose that, instead of attempting to estimate the required return, an investor wishes to determine an implied dividend growth rate. In this case, the investor must assume a future stock return, as in question 2.

2. If the analyst believes that Coca-Cola stock investors should expect a return of 7 percent, calculate the implied dividend growth rate for Coca-Cola.

Solution:

4.21 percent

Using Equation 22,

$$g = 0.07 - \frac{1.76}{63} = 0.0421.$$

The implied dividend growth rate for Coca-Cola stock is 4.21 percent given its expected return, price, and expected dividend. Given that a higher expected return is assumed in this question compared to the case in question 1, the result is a higher implied dividend growth rate to justify Coca-Cola's stock price of USD63.00.

Rather than comparing equity share prices directly in currency terms, a common practice is to compare ratios of share price to earnings per share, or the **price-to-earnings ratio**.

PRICE-TO-EARNINGS RATIO

Price-to-earnings ratio is a relative valuation metric that improves comparability by controlling for a known driver of value (earnings per share) as well as currency. It is analogous to expressing the price of real estate using a price per square meter.

A stock trading at a price-to-earnings ratio of 20 implies that its share price is 20 times its earnings per share and investors are willing to pay 20 times earnings per share for each share traded, which is more expensive than a stock trading at a price to earnings ratio of 10.

Price-to-earnings ratios not only are used for individual stocks but also are a valuation metric for stock indexes, such S&P 500, FTSE 100, or Nikkei 225. Here, the stock index value is divided by a weighted sum of the index constituents' earnings per share. This will be explored in depth later in the curriculum, but we can relate the price-to-earnings ratio to our earlier discussion of relating a stock's price (PV) to expected future cash flows to make some useful observations. First, recall Equation 14:

$$PV_t = \frac{D_t(1+g)}{r-g}.$$

We can divide both sides by E_t , earnings per share for period t , to obtain:

$$\frac{PV_t}{E_t} = \frac{\frac{D_t}{E_t} \times (1+g)}{r-g}. \quad (23)$$

The left-hand side of Equation 23 is the price-to-earnings ratio, whereas the first term in the numerator on the right is the proportion of earnings distributed to shareholders as dividends known as the **dividend payout ratio**.

Given a price-to-earnings ratio and dividend payout ratio, we can solve for either required return or implied dividend growth rate (given an assumption about the other). The required return is useful in understanding investor return expectations on a forward-looking basis. The implied constant growth rate is useful to compare with the company's expected growth rate and historical growth rate. For example, if the implied constant growth rate is 10 percent yet the analyst estimates that the company can only grow by 5 percent, the analyst may judge the shares to be overvalued.

In practice, the **forward price-to-earnings ratio** or ratio of its share price to an estimate of its next period ($t + 1$) earnings per share is commonly used. With it, we can simplify the previous equation as follows:

$$\frac{PV_t}{E_{t+1}} = \frac{D_{t+1}}{r-g}. \quad (24)$$

From Equation 24, we can see that forward price-to-earnings ratio is positively related to higher expected dividend payout ratio and higher expected growth but is negatively related to the required return.

EXAMPLE 11

Implied Return and Growth from Price to Earnings Ratio

- Suppose Coca-Cola stock trades at a forward price to earnings ratio of 28, its expected dividend payout ratio is 70 percent, and analysts believe that its dividend will grow at a constant rate of 4 percent per year. Calculate Coca-Cola's required return.

Solution:

6.50 percent

Using Equation 24,

$$28 = \frac{0.7}{r - 0.04}.$$

Solving this equation for r , Coca-Cola's required return is 6.50 percent.

Given the above result for Coca-Cola's required return, it should come as no surprise that if we instead assume that the required return on Coca-Cola stock is 6.50 percent, then we would find that Coca-Cola's implied growth rate is 4 percent, and this result is confirmed in question 2.

2. Suppose Coca-Cola stock trades at a forward price to earnings ratio of 28, its expected dividend payout ratio is 70 percent, and analysts believe that its required return is 6.50 percent. Calculate Coca-Cola's implied growth rate.

Solution:

4.00 percent

Using Equation 24,

$$28 = \frac{0.70}{0.065 - g}$$

Solving this equation for g , Coca-Cola's implied growth rate is 4.00 percent. In particular, Coca-Cola's return is expected to be 2.5 percent greater than its dividend growth rate (i.e., 6.50% – 4.00%).

As discussed earlier, the same principles apply to understanding required returns and growth rates on stock indexes.

3. A stock index is trading at a forward price to earnings ratio of 19. If the expected dividend payout ratio on the index is 60 percent, and equity investors expect an index rate of return of 8 percent, calculate the implied constant growth rate for the index.

Solution:

4.84 percent

Using Equation 24:

$$19 = \frac{0.60}{0.08 - g}$$

Solving for g , we find that investors expect the index's dividend growth rate to be 4.84 percent in the future. Alternatively, we could assume that investors expect a 4.84 percent growth rate and solve for r to calculate 8 percent as the required return. Thus, the index forward price-to-earnings ratio of 19 and an expected dividend payout ratio of 60 percent combine to reflect expectations of 8 percent return and 4.84 percent growth. Specifically, the required return exceeds the implied dividend growth rate by 3.16 percent (i.e., 8% – 4.84%).

An important point from the prior results is that equity prices, whether expressed simply as price or as a price-to-earnings ratio, reflect combined expectations about future returns and growth. Expectations of returns and growth are linked together by the difference between r and g . For example, the Coca-Cola example using a price-to-earnings ratio of 28 describes a situation in which investors must believe that the required return on Coca-Cola stock is 2.5 percent above its growth rate.

4. Suppose Coca-Cola stock trades at a forward price to earnings ratio of 28 and its expected dividend payout ratio is 70 percent. Analysts believe that Coca-Cola stock should earn a 9 percent return and that its dividends will grow by 4.50 percent per year indefinitely. Recommend a course of action for an investor interested in taking a position in Coca-Cola stock.

Solution:

Take a short position in Coca-Cola stock.

If we evaluate the above parameters using Equation 24, we can see that this results in an inequality.

$$28 > \frac{0.70}{0.09 - 0.045} = 15.56.$$

Coca-Cola's forward price-to-earnings ratio of 28 is much greater than 15.56, which is computed from the equation. Investor expectations of cash flow growth and return are inconsistent with Coca-Cola's forward price to earnings ratio. Specifically, an investor should consider a short position in Coca-Cola stock in the belief that its price should decline because its current price to earnings ratio is well above what its fundamentals imply. As shown in results for questions 1 and 2 in this example, expectations for the required return and growth rate must be such that $r - g = 2.50\%$ to justify a forward price-to-earnings ratio of 28 given the expected dividend payout ratio of 70 percent.

CASH FLOW ADDITIVITY

4

- explain the cash flow additivity principle, its importance for the no-arbitrage condition, and its use in calculating implied forward interest rates, forward exchange rates, and option values

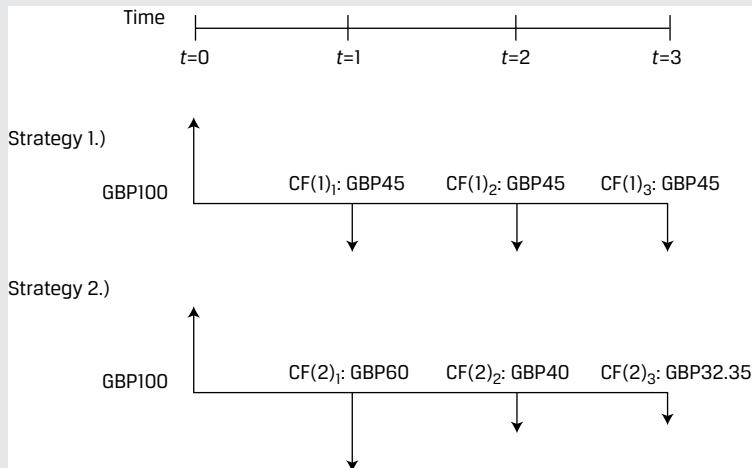
The time value of money trade-off between cash flows occurring today versus those in the future may be extended beyond pricing future cash flows today or calculating the implied return on a single instrument using the **cash flow additivity principle**. Under cash flow additivity, the present value of any future cash flow stream indexed at the same point equals the sum of the present values of the cash flows. This principle is important in ensuring that market prices reflect the condition of no arbitrage, or that no possibility exists to earn a riskless profit in the absence of transaction costs.

Let's begin with a basic example to demonstrate the cash flow additivity principle.

EXAMPLE 12

Basic Cash Flow Additivity

Let's assume that you have GBP100 to invest, and have two strategies from which to choose with the following cash flow streams as shown in Exhibit 9.

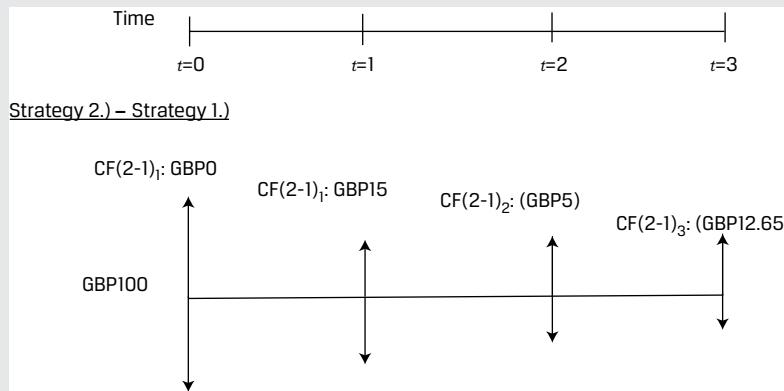
Exhibit 9: Two Investment Strategies


Your required return for both investment strategies is 10 percent per time period.

1. Recommend which investment strategy to choose.

Solution:

To make a recommendation between these two strategies, we need to establish which one has the higher present value as well as whether either has a positive present value. To accomplish these objectives, we can first compute the present values of both strategies and ensure that at least one has a positive present value in addition to comparing the present values of the two strategies. Alternatively, we can calculate the difference between the cash flows at each time period to effectively create a new set of cash flows. Both solution processes should yield an equivalent numeric result when comparing the two strategies, and this numeric equivalence represents the cash flow additivity principle. Let's first create the set of cash flows from the difference of the two strategies as shown in the diagram:



The present value of these net cash flows is zero, indicating equivalence, as follows:

$$PV = 0 + \frac{15}{1.10} + \frac{-5}{(1.10)^2} + \frac{-12.65}{(1.10)^3} = 0.$$

The conclusion from this analysis is that the two strategies are economically equivalent; therefore you should have no preference for one over the other. We must also determine whether these strategies are economically valuable. To do so, let's calculate the present value of each strategy individually.

For investment strategy 1, the present value is GBP11.91, as follows:

$$PV = -100 + \frac{45}{1.10} + \frac{45}{(1.10)^2} + \frac{45}{(1.10)^3} = 11.91.$$

For investment strategy 2, the present value is also GBP11.91, as follows:

$$PV = -100 + \frac{60}{1.10} + \frac{40}{(1.10)^2} + \frac{32.35}{(1.10)^3} = 11.91.$$

Both investment strategies are valuable in that the present value of the sums of their cash flows are positive. Since the present values are identical, subtracting the present value of strategy 1 from the present value of strategy 2 is zero (i.e., 11.91 – 11.91).

Either strategy is recommended as both are valuable and have equal present values.

In the following sections, we apply the cash flow additivity principle to three different economic situations and further illustrate the principle of no arbitrage by comparing two economically equivalent strategies in each situation.

Implied Forward Rates Using Cash Flow Additivity

Consider two risk-free discount bonds with different maturities as follows:

One-year bond: $r_1 = 2.50\%$

Two-year bond: $r_2 = 3.50\%$

A risk neutral investor seeking to earn a return on GBP100 over a two-year investment horizon has two possible strategies:

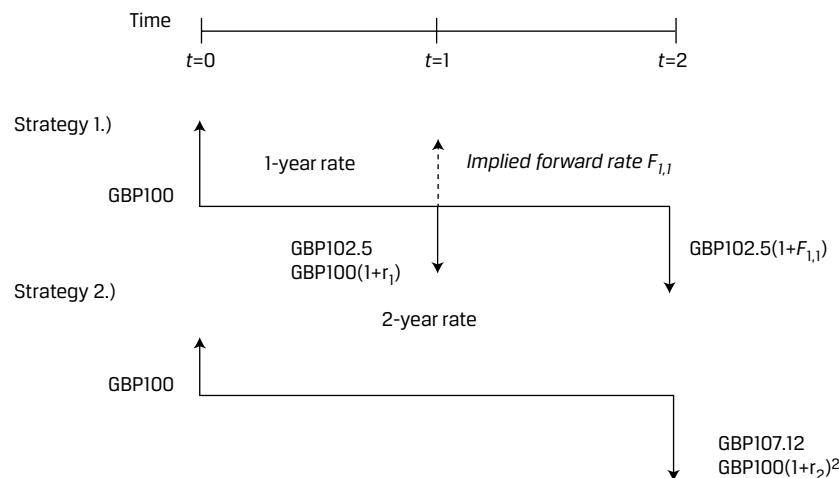
Investment strategy 1: Invest today for two years at a known annualized yield of 3.50 percent. Using Equation 5, we may solve for the future value in two years (FV_2) as follows:

$$\text{GBP100} = FV_2 / (1+r_2)^2;$$

$$FV_2 = \text{GBP107.12}.$$

Investment strategy 2: Invest today for one year at a known yield of 2.50 percent and reinvest in one year's time for one year at a rate of $F_{1,1}$ (the one-year forward rate starting in one year).

These two strategies are summarized in Exhibit 10:

Exhibit 10: Implied Forward Rate Example

Under the cash flow additivity principle, a risk-neutral investor would be indifferent between strategies 1 and 2 under the following condition:

$$FV_2 = PV_0 \times (1+r_2)^2 = PV_0 \times (1 + r_1)(1+F_{1,1}), \quad (25)$$

$$\text{GBP}107.12 = \text{GBP}100 (1.025)(1+F_{1,1}), \text{ and}$$

$$F_{1,1} = 4.51\%.$$

We can rearrange Equation 25 to solve for $F_{1,1}$ in general as follows:

$$F_{1,1} = (1 + r_2)^2 / (1 + r_1) - 1.$$

To illustrate why the one-year forward interest rate must be 4.51 percent, let's assume an investor could lock in a one-year rate of 5 percent starting in one year. The following arbitrage strategy generates a riskless profit:

1. Borrow GBP100 for two years at 3.50 percent and agree to pay GBP107.12 in two years.
2. Invest GBP100 for the first year at 2.50 percent and in year two at 5 percent, so receive GBP107.63 in two years.
3. The combination of these two strategies above yields a riskless profit of GBP0.51 in two years with zero initial investment.

This set of strategies illustrates that forward rates should be set such that investors cannot earn riskless arbitrage profits in financial markets. This is demonstrated by comparing two economically equivalent strategies (i.e., borrowing for two years at a two-year rate versus borrowing for two one-year horizons at two different rates). The forward interest rate $F_{1,1}$ may be interpreted as the breakeven one-year reinvestment rate in one year's time.

EXAMPLE 13**Forward Interest Rate Changes**

At its June 2022 meeting, the US Federal Reserve surprised markets by raising its target interest rate by a greater-than-expected 75 bps and suggested further increases in the future in response to sharply higher inflation. Exhibit 11 shows one-year and two-year US Treasury strip prices per USD100 at the end of May 2022 and after the Fed's decision in June:

Exhibit 11: One- and Two-Year US Treasury Strip Prices

Date	PV(1y)	PV(2y)
31 May 2022	98.028	95.109
15 June 2022	97.402	93.937

- Using the information in Exhibit 11, show the change in the breakeven one-year reinvestment in one year's time ($F_{1,1}$).

Solution:

62 bps or 0.62 percent

First, we use Equation 18 to solve for each market discount rate r . For example, in the case of the two-year discount bond on 15 June:

$$r = 3.177 \text{ percent} = \left(\frac{100}{93.937}\right)^{\frac{1}{2}} - 1.$$

The following table summarizes the respective discount bond rates:

Date	r_1	r_2
31 May 2022	2.012%	2.539%
15 June 2022	2.667%	3.177%

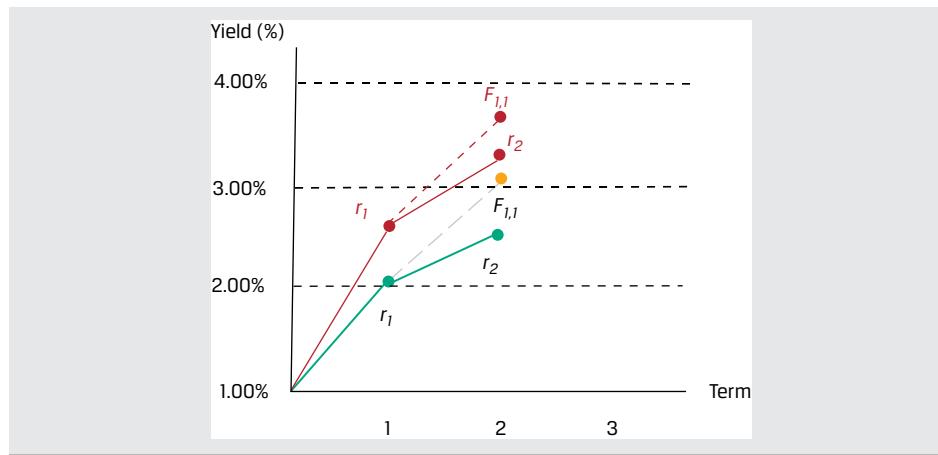
Solve for the respective forward rates ($F_{1,1}$) by rearranging Equation 25:

$$F_{1,1} = (1+r_2)^2/(1+r_1) - 1.$$

$$\underline{F_{1,1}(31 \text{ May})}: F_{1,1} = 3.069\% = (1 + 2.539\%)^2/(1 + 2.012\%) - 1.$$

$$\underline{F_{1,1}(15 \text{ June})}: F_{1,1} = 3.689\% = (1 + 3.177\%)^2/(1 + 2.667\%) - 1.$$

While the higher r_1 and r_2 in mid-June shows that investors have factored the immediate 75 bp increase into their nominal discount rate, the 62 bp (0.62 percent) increase in $F_{1,1}$ is a measure of expected future rate increases given that the one-year nominal rate of return starts in one year's time. Note also that in a rising rate environment, $F_{1,1} > r_2$, as shown in the following diagram comparing r_1 , r_2 and $F_{1,1}$ on 31 May (lower rates) to 15 June (higher rates).



Forward Exchange Rates Using No Arbitrage

We now extend the principle of cash flow additivity to an economic scenario involving different currencies.

Assume that you have USD1,000 to invest for six months. You are considering a riskless investment in either US or Japanese six-month government debt. Let's assume that the current exchange rate between Japanese yen (JPY) and US dollars (USD) is 134.40 (i.e., JPY134.40 = USD1). The six-month Japanese yen risk-free rate is assumed to be 0.05 percent, and the six-month US dollar risk-free rate is 2.00 percent. This example assumes continuous compounding.

Investment strategy 1: Invest USD1,000 in a six-month US Treasury bill

At time $t = 0$: Invest USD1,000 at the 2.00 percent US-dollar risk-free rate for six months.

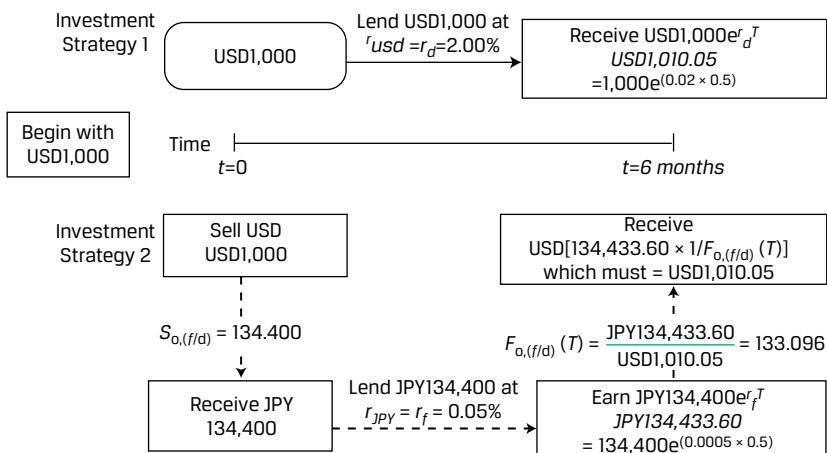
At time $t = T$ in six months: Receive USD1,010.05 ($= 1,000e^{(0.02 \times 0.5)}$).

Investment strategy 2: Convert the USD1,000 into Japanese yen at the current exchange rate of 134.40, invest in a six-month Japanese Treasury, and convert this known amount back into US dollars in six months at a six-month forward exchange rate set today of 133.096.

At time $t = 0$: Convert USD1,000 into JPY134,400. Lend the JPY134,400 at the 0.05 percent JPY risk-free rate for six months.

At time $t = T$ in six months: Receive JPY loan proceeds of 134,433.60 ($= 134,400e^{(0.0005 \times 0.5)}$). Exchange JPY loan proceeds for US dollars at the forward rate of 133.096 to receive USD1010.05 ($= 134,433.60 / 133.096$).

These two strategies are economically equivalent in that both involve investing USD1000 at $t = 0$ and receiving USD1010.05 in six months. The element that links these two strategies is the six-month forward exchange rate of 133.096 JPY/USD. If this forward rate is set above or below 133.096, an arbitrage opportunity would exist for investors converting between Japanese yen and US dollars. Exhibit 12 provides a visual layout of the two strategies.

Exhibit 12: No-Arbitrage Condition in the Foreign Exchange Market**EXAMPLE 14****Foreign Exchange Forward Rates in a Changing Interest Rate Environment**

Central banks responded differently to the sharp rise in inflation during 2022. For example, while the Fed raised rates by 75 bps to 1.75 percent in mid-June, the Bank of England opted for a more gradual approach, raising its benchmark rate by just 25 bps to 1.25 percent. Exhibit 13 compares one-year US Treasury strip rates (r_{USD}) from the prior example to UK gilt strip rates (r_{GBP}) over the same time frame:

Exhibit 13: Comparison of US Treasury Strip Rates

Date	r_{USD}	r_{GBP}
31 May 2022	2.012%	1.291%
15 June 2022	2.667%	1.562%

- If we assume the USD/GBP spot price of 1.2602 (or USD1.2602 per GBP1.00) from 31 May remains constant, how does the change in risk-free US dollars versus British pounds rates affect the one-year USD/GBP forward rate?

Solution:

The no-arbitrage USD/GBP forward rate increases from 1.2693 to 1.2742. Assume an investor has GBP1,000 to invest for one year in a British-pound or US-dollar risk-free discount bond. The US dollars needed to purchase GBP1 in one year must have a spot price equal to the discounted future price.

As of 31 May:

Domestic Strategy: Invest GBP1,000 at the 1.291 percent one-year British-pound risk-free rate to receive GBP1,012.99 ($= 1,000e^{(0.01291)}$) in one year.

Foreign Strategy: Convert GBP1,000 at USD/GBP1.2602 to receive USD1,260.20, which invested at the one-year US-dollar risk-free rate of 2.012 percent generates a return of USD1,285.81 ($= 1,260.20 e^{(0.02012)}$) in one year.

The no-arbitrage USD/GBP forward rate as of 31 May is therefore equal to 1.2693 ($= \text{USD}1,285.81 / \text{GBP}1,012.99$).

As of 15 June:

Domestic Strategy: Invest GBP1,000 at the 1.562 percent one-year British-pound risk-free rate to receive GBP1,015.74 ($= 1,000e^{(0.01562)}$) in one year.

Foreign Strategy: Convert GBP1,000 at 1.2602 to receive USD1,260.20, which invested at the one-year US-dollar risk-free rate of 2.667 percent returns USD1,294.27 ($= 1,260.20 e^{(0.02667)}$) in one year.

The no-arbitrage USD/GBP forward rate as of 15 June is equal to 1.2742 ($= \text{USD}1,294.26 / \text{GBP}1,015.74$).

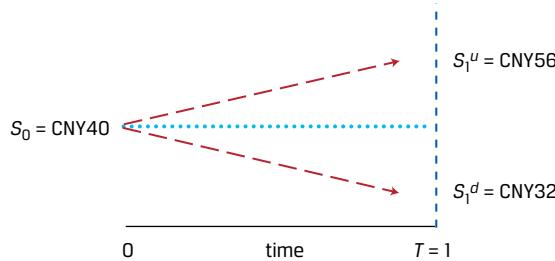
As of mid-June, an investor agreeing to exchange US dollars for British pounds in one year must deliver more US dollars (1.2742 versus 1.2693) in exchange for GBP1 than if the same contract had been entered at the end of May. The greater increase in r_{USD} widens the interest rate differential ($r_{\text{USD}} - r_{\text{GBP}}$), causing US dollars to depreciate on a forward basis versus British pounds over the period. Differently put, the expectation for US-dollar depreciation on a forward basis versus British pounds would require a higher US-dollar interest rate to attract investors to US dollars versus British pounds.

Option Pricing Using Cash Flow Additivity

Let's assume an asset has a current price of 40 Chinese yuan (i.e., CNY40). The asset is risky in that its price may rise 40 percent to CNY56 during the next time period or its price may fall 20 percent to CNY32 during the next time period.

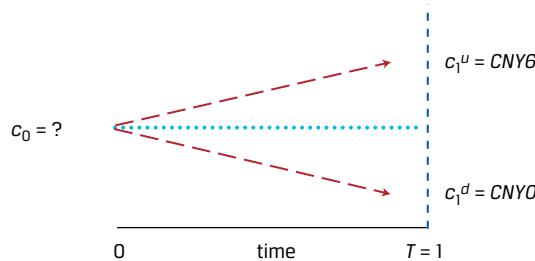
An investor wishes to sell a contract on the asset in which the buyer of the contract has the right, but not obligation, to buy the noted asset for CNY50 at the end of the next time period. We can apply the principle of cash flow additivity to establish no-arbitrage pricing for this contract.

The binomial tree in Exhibit 14 summarizes the two possible future outcomes of the asset:

Exhibit 14: One-Period Binomial Tree for the Asset's Price

The contract value under the two scenarios is as follows (shown visually in Exhibit 15):

- Price Increase: The 40 percent increase results in a contract value of CNY 6 ($=\text{CNY}56 - \text{CNY}50$), as the contract owner will choose to buy the asset at CNY50 and is able to sell it at a market price of CNY56.
- Price Decrease: The 20 percent decrease results in a contract value of zero (CNY0), as the contract owner will choose not to buy the asset at CNY50 when the market price is only CNY32.

Exhibit 15: One-Period Binomial Tree for the Contract's Price

Consider the position of the contract seller as follows:

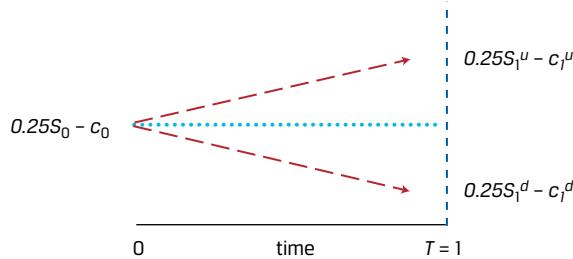
- At $t = 0$, the contract value is c_0 .
- At $t = 1$, the contract value is either CNY6 (if the underlying asset rises to CNY56) or CNY0 (if the underlying stock falls to CNY32).

The initial contract value c_0 is unknown and may be determined using cash flow additivity and no-arbitrage pricing. That is, the value of the contract and the underlying asset in each future scenario may be used to construct a risk-free portfolio, or a portfolio where the value is the same in both scenarios. For example, assume at $t = 0$ an investor creates a portfolio in which the contract is sold at a price of c_0 and 0.25 units of the underlying asset are purchased. This portfolio is called a replicating portfolio in that it is designed specifically to create a matching future cash flow stream to that of a risk-free asset. Denoting the value of the portfolio as V at $t = 0$ and at $t = 1$ under both the price increase and price decrease scenarios, we have the following:

- $V_0 = 0.25 \times 40 - c_0$,
- $V_1^U = 0.25 \times 56 - 6 = 8$, and
- $V_1^D = 0.25 \times 32 - 0 = 8$.

As can be seen, the value of the replicating portfolio is equal to CNY8 regardless of whether the price of the asset increases or decreases. Because of this, the portfolio is risk-free and can be discounted as a risk-free asset. The payoffs for these two scenarios are shown in Exhibit 16.

Exhibit 16: Call Option Replication



To solve for c_0 , we set the present value of the replicating portfolio ($0.25 \times 40 - c_0$) equal to the discounted future value of the risk-free payoff of CNY8 under each outcome as in Equation 5. Assume r of 5 percent:

$$V_0 = 0.25S_0 - c_0,$$

$$V_0 = \frac{V_1^u}{1+r} = \frac{V_1^d}{1+r}, \text{ and}$$

$$0.25 \times 40 - c_0 = \frac{\text{CNY8}}{1.05}.$$

We solve for c_0 as CNY2.38. Thus, CNY2.38 is a fair price for the seller of the contract to receive from the buyer of the contract.

As part of the example above, we assumed that the investor buys 0.25 units of the asset as part of the portfolio. This proportion of the underlying asset is known as the **hedge ratio** in option pricing. In fact, the example just completed is a simplified process for solving for the value of a call option (this material will be covered in more detail later in the curriculum). As in our prior examples, we compare different strategies using cash flow additivity to help solve for a no-arbitrage price for a financial instrument.

EXAMPLE 15

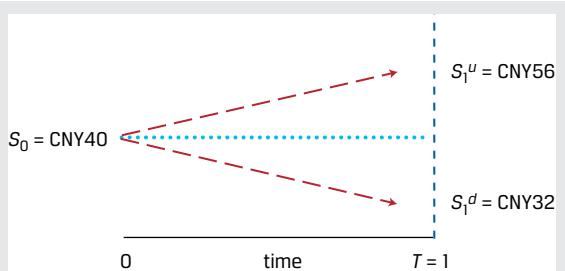
Put Option

A put option grants the owner the right, but not the obligation to sell a stock at a predetermined exercise price X . If we assume $X = \text{CNY}50$, then the put option value in one period (p_1) is equal to $\text{Max}(0, \text{CNY}50 - S_1)$.

- Using a one-period binomial tree model with the same prices in one year, initial stock price (S_0) of CNY40, and 5 percent discount rate ^{*}, create a risk-free portfolio replicating the put option with 0.75 units of the underlying asset and solve for the put option price (p_0).

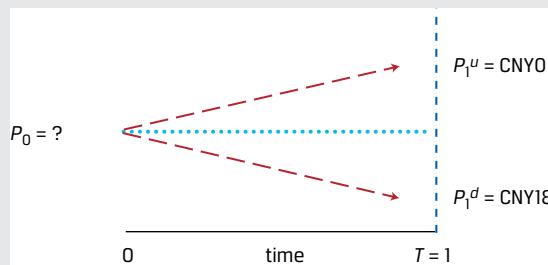
Solution:

The one-period binomial model is as follows:

Cash Flow Additivity

The put option value under the two scenarios is as follows:

- Price Increase (\$p_1^u\$): The 40 percent increase results in a put option value of CNY0, as the option owner will choose not to sell the stock at the exercise price.
- Price Decrease (\$p_1^d\$): The 20 percent decrease results in a put option value of CNY18 (\$\text{CNY}50 - \text{CNY}32\$), as the put option owner will choose to sell the stock with a market price of CNY32 at the exercise price of CNY50.



Unlike the call option, the put option increases in value as the stock price falls, so a risk-free portfolio combines a stock purchase with a purchased put option. Specifically, assume at \$t = 0\$, we sell a put option at a price of \$p_0\$ and purchase 0.75 units of the underlying stock. Portfolio values at \$t = 0\$ and \$t = 1\$ are as follows:

- \$V_0 = 0.75S_0 + p_0\$,
- \$V_1^u = 0.75S_1^u + p_1^u\$, and
- \$V_1^d = 0.75S_1^d + p_1^d\$,
where \$V_1^u = V_1^d\$.

The value of the risk-free portfolio in one period (\$V_1 = 0.75S_1 + p_1\$) under the two scenarios is:

- Price Increase (\$V_1^u\$): The portfolio value is CNY42 (\$=(0.75 \times \text{CNY}56) + \text{CNY}0\$).
- Price Decrease (\$V_1^d\$): The portfolio value is CNY42 (\$=(0.75 \times \text{CNY}32) + \text{CNY}18\$).

Solve for \$p_0\$ by setting the present value of the replicating portfolio (\$0.75S_0 + p_0\$) equal to the discounted future value of the payoff, \$S_0\$ is CNY40 and \$r\$ is 5 percent:

$$V_0 = 0.75S_0 + p_0,$$

$$V_0 = \frac{V_1^u}{1+r} = \frac{V_1^d}{1+r}, \text{ and}$$

$$p_0 = \frac{\text{CNY}42}{1.05} - 0.75(\text{CNY}40).$$

Solve for \$p_0\$ to be equal to CNY10. The higher put price is a result of the greater payoff of the put option versus the call option under the given pa-

parameters of the binomial model. The factors affecting option prices will be addressed in detail later in the curriculum.

PRACTICE PROBLEMS

1. Grupo Ignacia issued 10-year corporate bonds two years ago. The bonds pay an annualized coupon of 10.7 percent on a semiannual basis, and the current annualized YTM is 11.6 percent. The current price of Grupo Ignacia's bonds (per MXN100 of par value) is *closest* to:
 - A. MXN95.47.
 - B. MXN97.18.
 - C. MXN95.39.
2. Grey Pebble Real Estate seeks a fully amortizing fixed-rate five-year mortgage loan to finance 75 percent of the purchase price of a residential building that costs NZD5 million. The annual mortgage rate is 4.8 percent. The monthly payment for this mortgage would be *closest* to:
 - A. NZD70,424.
 - B. NZD93,899.
 - C. NZD71,781.
3. Mylandia Corporation pays an annual dividend to its shareholders, and its most recent payment was CAD2.40. Analysts following Mylandia expect the company's dividend to grow at a constant rate of 3 percent per year in perpetuity. Mylandia shareholders require a return of 8 percent per year. The expected share price of Mylandia is *closest* to:
 - A. CAD48.00.
 - B. CAD49.44.
 - C. CAD51.84.
4. Suppose Mylandia announces that it expects significant cash flow growth over the next three years, and now plans to increase its recent CAD2.40 dividend by 10 percent in each of the next three years. Following the 10 percent growth period, Mylandia is expected to grow its annual dividend by a constant 3 percent indefinitely. Mylandia's required return is 8 percent. Based upon these revised expectations, The expected share price of Mylandia stock is:
 - A. CAD49.98.
 - B. CAD55.84.
 - C. CAD59.71.
5. Consider a Swiss Confederation zero-coupon bond with a par value of CHF100, a remaining time to maturity of 12 years and a price of CHF89. In three years' time, the bond is expected to have a price of CHF95.25. If purchased today, the bond's expected annualized return is *closest* to:
 - A. 0.58 percent.
 - B. 1.64 percent.

- C. 2.29 percent.
6. Grupo Ignacia issued 10-year corporate bonds four years ago. The bonds pay an annualized coupon of 10.7 percent on a semiannual basis, and the current price of the bonds is MXN97.50 per MXN100 of par value. The YTM of the bonds is closest to:
- A. 11.28 percent.
 - B. 11.50 percent.
 - C. 11.71 percent.
7. Mylandia Corporation stock trades at CAD60.00. The company pays an annual dividend to its shareholders, and its most recent payment of CAD2.40 occurred yesterday. Analysts following Mylandia expect the company's dividend to grow at a constant rate of 3 percent per year. Mylandia's required return is:
- A. 8.00 percent.
 - B. 7.00 percent.
 - C. 7.12 percent.
8. An analyst observes the benchmark Indian NIFTY 50 stock index trading at a forward price-to-earnings ratio of 15. The index's expected dividend payout ratio in the next year is 50 percent, and the index's required return is 7.50 percent. If the analyst believes that the NIFTY 50 index dividends will grow at a constant rate of 4.50 percent in the future, which of the following statements is correct?
- A. The analyst should view the NIFTY 50 as overpriced.
 - B. The analyst should view the NIFTY 50 as underpriced.
 - C. The analyst should view the NIFTY 50 as fairly priced.
9. If you require an 8 percent return and must invest USD500,000, which of the investment opportunities in Exhibit 1 should you prefer?

Exhibit 1: Investment Opportunities

Cash flows (in thousands)	t = 0	t = 1	t = 2	t = 3
Opportunity 1	-500	195	195	195
Opportunity 2	-500	225	195	160.008

- A. Opportunity 1
 - B. Opportunity 2
 - C. Indifferent between the two opportunities.
10. Italian one-year government debt has an interest rate of 0.73 percent; Italian two-year government debt has an interest rate of 1.29 percent. The breakeven one-year reinvestment rate, one year from now is closest to:
- A. 1.01 percent.

- B. 1.11 percent.
- C. 1.85 percent.
11. The current exchange rate between the euro and US dollar is USD/EUR1.025. Risk-free interest rates for one year are 0.75 percent for the euro and 3.25 percent for the US dollar. The one-year USD/EUR forward rate that *best* prevents arbitrage opportunities is:
- A. USD/EUR1.051.
- B. USD/EUR1.025.
- C. USD/EUR0.975.
12. A stock currently trades at USD25. In one year, it will either increase in value to USD35 or decrease to USD15. An investor sells a call option on the stock, granting the buyer the right, but not the obligation, to buy the stock at USD25 in one year. At the same time, the investor buys 0.5 units of the stock. Which of the following statements about the value of the investor's portfolio at the end of one year is correct?
- A. The portfolio has a value of USD7.50 in both scenarios.
- B. The portfolio has a value of USD25 in both scenarios.
- C. The portfolio has a value of USD17.50 if the stock goes up and USD7.50 if the stock goes down.

SOLUTIONS

1. C is correct. The coupon payments are 5.35 ($=10.7/2$), the discount rate is 5.8 percent ($=11.6\%/2$) per period, and the number of periods is 16 ($=8\times2$). Using Equation 6, the calculation is as follows:

$$95.39 = \frac{5.35}{1.058} + \frac{5.35}{1.058^2} + \frac{5.35}{1.058^3} + \frac{5.35}{1.058^4} + \dots + \frac{5.35}{1.058^{14}} + \frac{5.35}{1.058^{15}} + \frac{105.35}{1.058^{16}}.$$

Alternatively, using the Microsoft Excel or Google Sheets PV function (`PV(0.058, 16, 5.35, 100, 0)`) also yields a result of MXN95.39.

A is incorrect. MXN95.47 is the result when incorrectly using coupon payments of 10.7, a discount rate of 11.6 percent, and 8 as the number of periods.

B is incorrect. MXN97.18 is the result when using the correct semiannual coupons and discount rate, but incorrectly using 8 as the number of periods.

2. A is correct. The present value of the mortgage is NZD3.75 million ($=0.75\times5,000,000$), the periodic discount rate is 0.004 ($=0.048/12$), and the number of periods is 60 ($=5\times12$). Using Equation 8,

$$A = \$70,424 = \frac{0.4\% (\text{NZD}3,750,000)}{1 - (1 + 0.4\%)^{-60}}.$$

Alternatively, the spreadsheet PMT function may be used with the inputs stated earlier.

B is incorrect. NZD93,899 is the result if NZD5 million is incorrectly used as the present value term.

C is incorrect. NZD71,781 is the result if the calculation is made using 4.8 percent as the rate and 5 as the number of periods, then the answer is divided by 12.

3. B is correct. Mylandia's next expected dividend is CAD2.472 ($=2.40\times1.03$), and using Equation 14,

$$PV_t = \frac{2.40(1 + 0.03)}{0.08 - 0.03} = \frac{2.472}{0.05} = 49.44.$$

4. C is correct. Following the first step, we observe the following expected dividends for Mylandia for the next three years:

In 1 year: $D_1 = \text{CAD}2.64$ ($=2.40\times1.10$)

In 2 years: $D_2 = \text{CAD}2.90$ ($=2.40\times1.10^2$)

In 3 years: $D_3 = \text{CAD}3.19$ ($=2.40\times1.10^3$)

The second step involves a lower 3 percent growth rate. At the end of year four, Mylandia's dividend (D_4) is expected to be CAD3.29 ($=2.40\times1.10^3\times1.03$). At this time, Mylandia's expected terminal value at the end of three years is CAD65.80 using Equation 17, as follows:

$$E(S_{t+n}) = \frac{3.29}{0.08 - 0.03} = 65.80.$$

Third, we calculate the sum of the present values of these expected dividends using Equation 16:

$$PV_t = \frac{2.64}{1.08} + \frac{2.90}{1.08^2} + \frac{3.19}{1.08^3} + \frac{65.80}{1.08^3} = 59.71.$$

5. C is correct. The FV of the bond is CHF95.25, the PV is CHF89, and the number of annual periods (t) is 3. Using Equation 18,

$$2.29 \text{ percent} = (92.25/89)^{(1/3)} - 1.$$

A is incorrect as the result is derived using t of 12. B is incorrect as this result is derived using a PV of CHF95.25 and an FV of 100.

6. A is correct. The YTM is calculated by solving for the RATE spreadsheet function with the following inputs: number of periods of 12 ($=6 \times 2$), coupon payments of 5.35 ($=10.7/2$), PV of -97.50, and FV of 100. The resulting solution for RATE of 5.64 percent is in semiannual terms, so multiply by 2 to calculate annualized YTM of 11.28 percent.

B is incorrect, as 11.50 percent is the result if number of periods used is eight, instead of 12. C is incorrect, as 11.71 percent is the result if the number of periods used is 6, instead of 12.

7. C is correct. We may solve for required return based upon the assumption of constant dividend growth using Equation 21:

$$r = \frac{2.40(1.03)}{60} + 0.03 = 0.0712.$$

B is incorrect as 7.00 percent is the result if we use the previous dividend of CAD2.40 instead of the next expected dividend. A is incorrect as 8.00 percent is simply the required return assumed from one of the Mylandia examples in Question Set 1 in which the price is solved to be a lower value.

8. B is correct. Using Equation 24, the previous input results in the following inequality:

$$15 < \frac{0.50}{0.075 - 0.045} = 16.67.$$

The above inequality implies that the analyst should view the NIFTY 50 as priced too low. The fundamental inputs into the equation imply a forward price to earnings ratio of 16.67 rather than 15. An alternative approach to answering the question would be to solve for implied growth using the observed forward price to earnings ratio of 15 and compare this to the analyst's growth expectations:

$$15 = \frac{0.50}{0.075 - g}.$$

Solving for g yields a result of 4.1667 percent. Since the analyst expects higher NIFTY 50 dividend growth of 4.50 percent, the index is viewed as underpriced.

9. Using cash flow additivity, compare the two opportunities by subtracting Opportunity 2 from Opportunity 1 yielding the following cash flows:

Opportunity 1 – Opportunity 2	0	-30	0	34.992
-------------------------------	---	-----	---	--------

Finding the present value of the above cash flows at 8 percent discount rate shows that both investment opportunities have the same present value. Thus, the two opportunities are economically identical, and there is no clear preference for one over the other.

10. C is correct. The one-year forward rate reflects the breakeven one-year reinvestment rate in one year, computed as follows:

$$F_{1,1} = (1+r_2)^2/(1+r_1) - 1,$$

$$F_{1,1} = (1.0129)^2/(1.0073) - 1 = 0.0185.$$

11. A is correct. To avoid arbitrage opportunities in exchanging euros and US dollars, investors must be able to lock in a one-year forward exchange rate of USD/

EUR1.051 today. The solution methodology is shown below.

In one year, a single unit of euro invested risk-free is worth EUR1.0075 ($=e^{0.0075}$).

In one year, a single unit of euro converted to US dollars and then invested risk-free is worth USD1.0589 ($=1.025 \times e^{0.0325}$).

To convert USD1.0589 into EUR1.0075 requires a forward exchange rate of USD/EUR1.051 ($=1.0589 / 1.0075$).

12. A is correct. Regardless of whether the stock increases or decreases in price, the investor's portfolio has a value of USD7.50 as follows:

If stock price goes to USD35, value = $0.5 \times 35 - 10 = 7.50$.

If stock price goes to USD15, value = $0.5 \times 15 - 0 = 7.50$.

If the stock price rises to USD35, the sold call option at USD25 has a value to the buyer of USD10, offsetting the rise in the stock price.

LEARNING MODULE

3

Statistical Measures of Asset Returns

by Pamela Peterson Drake, PhD, CFA, and Wu Jian, PhD.

Pamela Peterson Drake, PhD, CFA, is at James Madison University (USA). Jian Wu, PhD, is at State Street (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	calculate, interpret, and evaluate measures of central tendency and location to address an investment problem
<input type="checkbox"/>	calculate, interpret, and evaluate measures of dispersion to address an investment problem
<input type="checkbox"/>	interpret and evaluate measures of skewness and kurtosis to address an investment problem
<input type="checkbox"/>	interpret correlation between two variables to address an investment problem

INTRODUCTION

1

Data have always been a key input for securities analysis and investment management, allowing investors to explore and exploit an abundance of information for their investment strategies. While this data-rich environment offers potentially tremendous opportunities for investors, turning data into useful information is not so straightforward.

This module provides a foundation for understanding important concepts that are an indispensable part of the analytical tool kit needed by investment practitioners, from junior analysts to senior portfolio managers. These basic concepts pave the way for more sophisticated tools that will be developed as the quantitative methods topic unfolds, which are integral to gaining competencies in the investment management techniques and asset classes that are presented later in the CFA curriculum.

This learning module focuses on how to summarize and analyze important aspects of financial returns, including key measures of central tendency, dispersion, and the shape of return distributions—specifically, skewness and kurtosis. The learning module finishes with a graphical introduction to covariance and correlation between two variables, a key concept in constructing investment portfolios to achieve diversification across assets within a portfolio.

LEARNING MODULE OVERVIEW

- Sample statistics—such as measures of central tendency, dispersion, skewness, and kurtosis—help with investment analysis, particularly in making probabilistic statements about returns.
- Measures of central tendency include the mean, the median and the mode, and specify where data are centered.
- The arithmetic mean is the sum of the observations divided by the number of observations. It is the most frequently used measure of central tendency.
- The median is the value of the middle item of observations, or the mean of the values of the two middle items, when the items in a set are sorted into ascending or descending order. Since the median is not influenced by extreme values, it is most useful in the case of skewed distributions.
- The mode is the most frequently observed value and is the only measure of central tendency that can be used with nominal or categorical data. A distribution may be unimodal (one mode), bimodal (two modes), trimodal (three modes), or have even more modes.
- Quantiles, as the median, quartiles, quintiles, deciles, and percentiles, are location parameters that divide a distribution into halves, quarters, fifths, tenths, and hundredths, respectively.
- A box and whiskers plot illustrates the distribution of a set of observations. The “box” depicts the interquartile range, the difference between the first and the third quartile. The “whiskers” outside of the “box” indicate the others measures of dispersion.
- Dispersion measures, such as the range, mean absolute deviation (MAD), variance, standard deviation, target downside deviation, and coefficient of variation, describe the variability of outcomes around the arithmetic mean.
- The range is the difference between the maximum value and the minimum value of the dataset. The range has only a limited usefulness because it uses information from only two observations.
- The MAD for a sample is the average of the absolute deviations of observations from the mean.
- The variance is the average of the squared deviations around the mean, and the standard deviation is the positive square root of variance. In computing sample variance, s^2 , and sample standard deviation, s , the average squared deviation is computed using a divisor equal to the sample size minus 1.
- The target downside deviation, or target semideviation, is a measure of the risk of being below a given target.
- The coefficient of variation (CV) is the ratio of the standard deviation of a set of observations to their mean value. By expressing the magnitude of variation among observations relative to their average size, the CV allows for the direct comparisons of dispersion across different datasets. Reflecting the correction for scale, the CV is a scale-free measure, that is, it has no units of measurement.

- Skewness describes the degree to which a distribution is asymmetric about its mean. An asset return distribution with positive skewness has frequent small losses and a few extreme gains compared to a normal distribution. An asset return distribution with negative skewness has frequent small gains and a few extreme losses compared to a normal distribution. Zero skewness indicates a symmetric distribution of returns.
- Kurtosis measures the combined weight of the tails of a distribution relative to the rest of the distribution. A distribution with fatter tails than the normal distribution is referred to as fat-tailed (leptokurtic); a distribution with thinner tails than the normal distribution is referred to as thin-tailed (platykurtic). The kurtosis of a normal distribution is 3.
- The correlation coefficient measures the association between two variables. It is the ratio of covariance to the product of the two variables' standard deviations. A positive correlation coefficient indicates that the two variables tend to move together, whereas a negative coefficient indicates that the two variables tend to move in opposite directions. Correlation does not imply causation, simply association. Issues that arise in evaluating correlation include the presence of outliers and spurious correlation.

MEASURES OF CENTRAL TENDENCY AND LOCATION

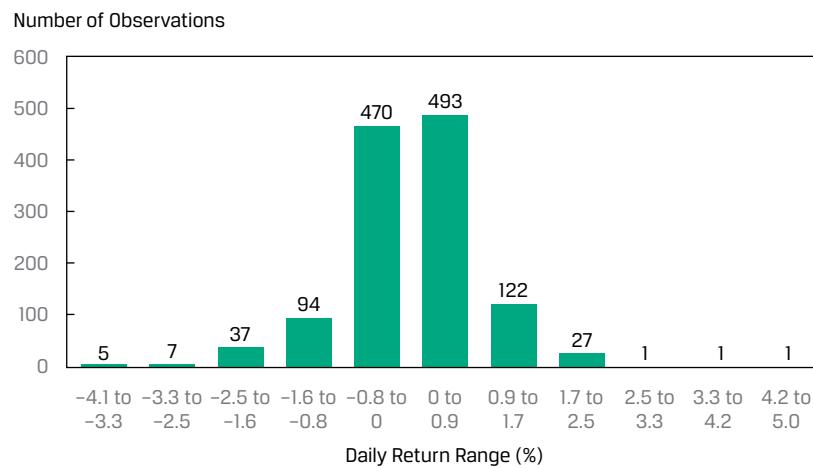
2



calculate, interpret, and evaluate measures of central tendency and location to address an investment problem

In this lesson, our focus is on measures of central tendency and other measures of location. A **measure of central tendency** specifies where the data are centered. For a return series, a measure of central tendency shows where the empirical distribution of returns is centered, essentially a measure of the “expected” return based on the observed sample. **Measures of location**, mean, the **median**, and the **mode** include not only measures of central tendency but also other measures that illustrate other aspects of the location or distribution of the data.

Frequency distributions, histograms, and contingency tables provide a convenient way to summarize a series of observations on an asset's returns as a first step toward describing the data. For example, a histogram for the frequency distribution of the daily returns for the fictitious EAA Equity Index over the past five years is shown in Exhibit 1.

Exhibit 1: Histogram of Daily Returns on the EAA Equity Index


Measures of Central Tendency

The Arithmetic Mean

Analysts and portfolio managers often want one number that describes a representative possible outcome of an investment decision. The arithmetic mean is one of the most frequently used measures of central tendency.

Arithmetic Mean. The **arithmetic mean** is the sum of the values of the observations in a dataset divided by the number of observations.

The Sample Mean

The sample mean is the arithmetic mean, or arithmetic average, computed for a sample.

Sample Mean Formula. The **sample mean** or average, \bar{X} (read “X-bar”), is the arithmetic mean value of a sample:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad (1)$$

where n is the number of observations in the sample.

A property and potential drawback of the arithmetic mean is its sensitivity to extreme values, or outliers. Because all observations are used to compute the mean and are given equal weight (i.e., importance), the arithmetic mean can be pulled sharply upward or downward by extremely large or small observations, respectively. The most common approach in this situation is to report the median, or middle value, in place of or in addition to the mean.

The Median

A second important measure of central tendency is the median.

Definition of Median. The median is the value of the middle item of a dataset that has been sorted into ascending or descending order. In an odd-numbered sample of n observations, the median is the value of the

observation that occupies the $(n + 1)/2$ position. In an even-numbered sample, we define the median as the mean of the values of the observations occupying the $n/2$ and $(n + 2)/2$ positions (the two middle observations).

Whether we use the calculation for an even- or odd-numbered sample, an equal number of observations lie above and below the median. A distribution has only one median. A potential advantage of the median is that, unlike the mean, outliers do not affect it.

The median, however, does not use all the information about the size of the observations; it focuses only on the relative position of the ranked observations. Calculating the median may also be more complex. Mathematicians express this disadvantage by saying that the median is less mathematically tractable than the mean.

The Mode

A third important measure of central tendency is the mode.

Definition of Mode. The mode is the most frequently occurring value in a dataset. A dataset can have more than one mode, or even no mode. When a dataset has a single value that is observed most frequently, its distribution is said to be **unimodal**. If a dataset has two most frequently occurring values, then it has two modes and its distribution is referred to as **bimodal**. When all the values in a dataset are different, the distribution has no mode because no value occurs more frequently than any other value.

Stock return data and other data from continuous distributions may not have a modal outcome. Exhibit 2 presents the frequency distribution of the daily returns for the EAA Equity Index over the past five years.

Exhibit 2: Frequency Distribution for Daily Returns of EAA Equity Index

Return Bin (%)	Absolute Frequency	Relative Frequency (%)	Cumulative Absolute Frequency	Cumulative Relative Frequency (%)
-5.0 to -4.0	1	0.08	1	0.08
-4.0 to -3.0	7	0.56	8	0.64
-3.0 to -2.0	23	1.83	31	2.46
-2.0 to -1.0	77	6.12	108	8.59
-1.0 to 0.0	470	37.36	578	45.95
0.0 to 1.0	555	44.12	1,133	90.06
1.0 to 2.0	110	8.74	1,243	98.81
2.0 to 3.0	13	1.03	1,256	99.84
3.0 to 4.0	1	0.08	1,257	99.92
4.0 to 5.0	1	0.08	1,258	100.00

A histogram for the frequency distribution of these daily returns was shown in Exhibit 1. The modal interval always has the highest bar in the histogram; in this case, the modal interval is 0.0 to 0.9 percent, and this interval has 493 observations out of a total of 1,258 observations.

Dealing with Outliers

In practice, although an extreme value or outlier in a financial dataset may represent a rare value in the population, it may also reflect an error in recording the value of an observation or an observation generated from a different population. After having checked and eliminated errors, we can address what to do with extreme values in the sample.

When dealing with a sample that has extreme values, there may be a possibility of transforming the variable (e.g., a log transformation) or of selecting another variable that achieves the same purpose. If, however, alternative model specifications or variable transformations are not possible, then three options exist for dealing with extreme values:

Option 1 Do nothing; use the data without any adjustment.

Option 2 Delete all the outliers.

Option 3 Replace the outliers with another value.

The first option is appropriate if the values are legitimate, correct observations, and it is important to reflect the whole of the sample distribution. Because outliers may contain meaningful information, excluding or altering these values may reduce valuable information. Further, because identifying a data point as extreme leaves it up to the judgment of the analyst, leaving in all observations eliminates that need to judge a value as extreme.

The second option excludes the extreme observations. One measure of central tendency in this case is the **trimmed mean**, which is computing an arithmetic mean after excluding a stated small percentage of the lowest and highest values. For example, a 5 percent trimmed mean discards the lowest 2.5 percent and the highest 2.5 percent of values and computes the mean of the remaining 95 percent of values. A trimmed mean is used in sports competitions when judges' lowest and highest scores are discarded in computing a contestant's score.

The third option involves substituting values for the extreme values. A measure of central tendency in this case is the **winsorized mean**. It is calculated after assigning one specified low value to a stated percentage of the lowest values in the dataset and one specified high value to a stated percentage of the highest values in the dataset. For example, a 95 percent winsorized mean sets the bottom 2.5 percent of values in the dataset equal to the value at or below which 2.5 percent of all the values lie (as will be seen shortly, this is called the "2.5th percentile" value) and the top 2.5 percent of values in the dataset equal to the value at or below which 97.5 percent of all the values lie (the "97.5th percentile" value).

Often comparing the statistical measures of datasets with outliers included and with outliers excluded can reveal important insights about the dataset. Such comparison can be particularly helpful when investors analyze the behavior of asset returns and rate, price, spread and volume changes.

In Example 1, we show the differences among these options for handling outliers using daily returns for the fictitious Euro-Asia-Africa (EAA) Equity Index in Exhibit 2.

EXAMPLE 1

Handling Outliers: Daily Returns to an Index

Using daily returns on the EAA Equity Index for the past five years, consisting of 1,258 trading days, we can see the effect of trimming and winsorizing the data:

Exhibit 3: Effect of Trimming and Winsorizing

	Arithmetic Mean	Trimmed Mean (Trimmed 5%)	Winsorized Mean (95%)
Mean	0.035%	0.048%	0.038%
Number of Observations	1,258	1,194	1,258

The trimmed mean eliminates the lowest 2.5 percent of returns, which in this sample is any daily return less than -1.934 percent, and it eliminates the highest 2.5 percent, which in this sample is any daily return greater than 1.671 percent. The result of this trimming is that the mean is calculated using 1,194 observations instead of the original sample's 1,258 observations.

The winsorized mean substitutes a return of -1.934 percent (the 2.5 percentile value) for any observation below -1.934 and substitutes a return of 1.671 percent (the 97.5 percentile value) for any observation above 1.671. The result in this case is that the trimmed and winsorized means are higher than the arithmetic mean, suggesting the potential evidence of significant negative returns in the observed daily return distribution.

Measures of Location

Having discussed measures of central tendency, we now examine an approach to describing the location of data that involves identifying values at or below which specified proportions of the data lie. For example, establishing that 25 percent, 50 percent, and 75 percent of the annual returns on a portfolio provides concise information about the distribution of portfolio returns. Statisticians use the word **quantile** as the most general term for a value at or below which a stated fraction of the data lies. In the following section, we describe the most commonly used quantiles—quartiles, quintiles, deciles, and percentiles—and their application in investments.

Quartiles, Quintiles, Deciles, and Percentiles

We know that the median divides a distribution of data in half. We can define other dividing lines that split the distribution into smaller sizes. **Quartiles** divide the distribution into quarters, **quintiles** into fifths, **deciles** into tenths, and **percentiles** into hundredths. The **interquartile range** (IQR) is the difference between the third quartile and the first quartile, or $IQR = Q_3 - Q_1$.

Example 2 illustrates the calculation of various quantiles for the daily return on the EAA Equity Index.

EXAMPLE 2

Percentiles, Quintiles, and Quartiles for the EAA Equity Index

Using the daily returns on the EAA Equity Index over the past five years and ranking them from lowest to highest daily return, we show the return bins from 1 (the lowest 5 percent) to 20 (the highest 5 percent) as follows:

Exhibit 4: EAA Equity Index Daily Returns Grouped by Size of Return

Bin	Cumulative Percentage of Sample Trading Days (%)	Daily Return (%) between		Number of Observations
		Lower Bound	Upper Bound	
1	5	-4.108	-1.416	63
2	10	-1.416	-0.876	63
3	15	-0.876	-0.629	63
4	20	-0.629	-0.432	63
5	25	-0.432	-0.293	63
6	30	-0.293	-0.193	63
7	35	-0.193	-0.124	62
8	40	-0.124	-0.070	63
9	45	-0.070	-0.007	63
10	50	-0.007	0.044	63
11	55	0.044	0.108	63
12	60	0.108	0.173	63
13	65	0.173	0.247	63
14	70	0.247	0.343	62
15	75	0.343	0.460	63
16	80	0.460	0.575	63
17	85	0.575	0.738	63
18	90	0.738	0.991	63
19	95	0.991	1.304	63
20	100	1.304	5.001	63

Because of the continuous nature of returns, it is not likely for a return to fall on the boundary for any bin other than the minimum (Bin = 1) and maximum (Bin = 20).

Using the data in Exhibit 4, complete the following tasks:

- Identify the 10th and 90th percentiles.

Solution:

The 10th and 90th percentiles correspond to the bins or ranked returns that include 10 percent and 90 percent of the daily returns, respectively. The 10th percentile corresponds to the return of -0.876 percent (and includes returns of that much and lower), and the 90th percentile corresponds to the return of 0.991 percent (and lower).

- Identify the first, second, and third quintiles.

Solution:

The first quintile corresponds to the lowest 20 percent of the ranked data, or -0.432 percent (and lower).

The second quintile corresponds to the lowest 40 percent of the ranked data, or -0.070 percent (and lower).

The third quintile corresponds to the lowest 60 percent of the ranked data, or 0.173 percent (and lower).

3. Identify the first and third quartiles.

Solution:

The first quartile corresponds to the lowest 25 percent of the ranked data, or -0.293 percent (and lower).

The third quartile corresponds to the lowest 75 percent of the ranked data, or 0.460 percent (and lower).

4. Identify the median.

Solution:

The median is the return for which 50 percent of the data lies on either side, which is 0.044 percent, the highest daily return in the 10th bin out of 20.

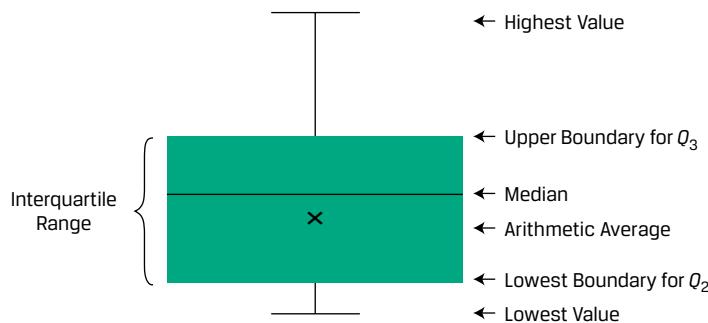
5. Calculate the interquartile range.

Solution:

The interquartile range is the difference between the third and first quartiles, 0.460 percent and -0.293 percent, or 0.753 percent.

One way to visualize the dispersion of data across quartiles is to use a diagram, such as a box and whisker chart. A **box and whisker plot** is shown in Exhibit 5. The “box” represents the lower bound of the second quartile and the upper bound of the third quartile, with the median or arithmetic average noted as a measure of central tendency of the entire distribution. The whiskers are the lines that run from the box and are bounded by the “fences,” which represent the lowest and highest values of the distribution.

Exhibit 5: Box and Whisker Plot



There are several variations for box and whisker displays. For example, for ease in detecting potential outliers, the fences of the whiskers may be a function of the interquartile range instead of the highest and lowest values like that in Exhibit 5.

In Exhibit 5, visually, the interquartile range is the height of the box and the fences are set at extremes. But another form of box and whisker plot typically uses 1.5 times the interquartile range for the fences. Thus, the upper fence is 1.5 times the interquartile range added to the upper bound of Q_3 , and the lower fence is 1.5 times the interquartile range subtracted from the lower bound of Q_2 . Observations beyond the fences (i.e., outliers) may also be displayed.

We can see the role of outliers in such a box and whisker plot using the EAA Equity Index daily returns, as shown in Exhibit 6.

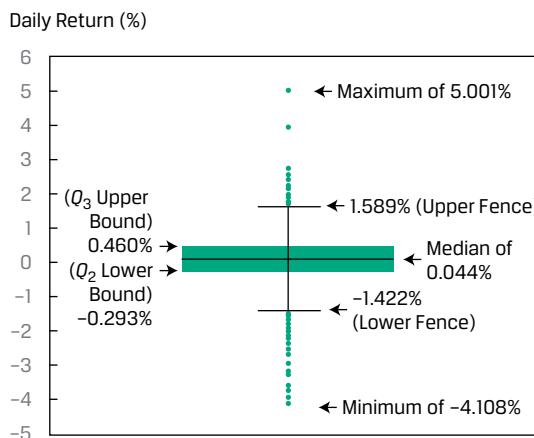
Exhibit 6: Box and Whisker Chart for EAA Equity Index Daily Returns

Exhibit 6 reveals the following:

- The maximum and minimum values of the distribution are 5.001 and -4.108, respectively, while the median (50th percentile) value is 0.044%.
- The interquartile range is 0.753 [= 0.460 – (–0.293)], and when multiplied by 1.5 and added to the Q_3 upper bound of 0.460 gives an upper fence of 1.589 [= $(1.5 \times 0.753) + 0.460$].
- The lower fence is determined in a similar manner, using the Q_1 lower bound, to be –1.422 [= $-(1.5 \times 0.753) + (-0.293)$].

As noted, any observation above (below) the upper (lower) fence is deemed to be an outlier.

Quantiles in Investment Practice

Quantiles are used in portfolio performance evaluation as well as in investment strategy development and research.

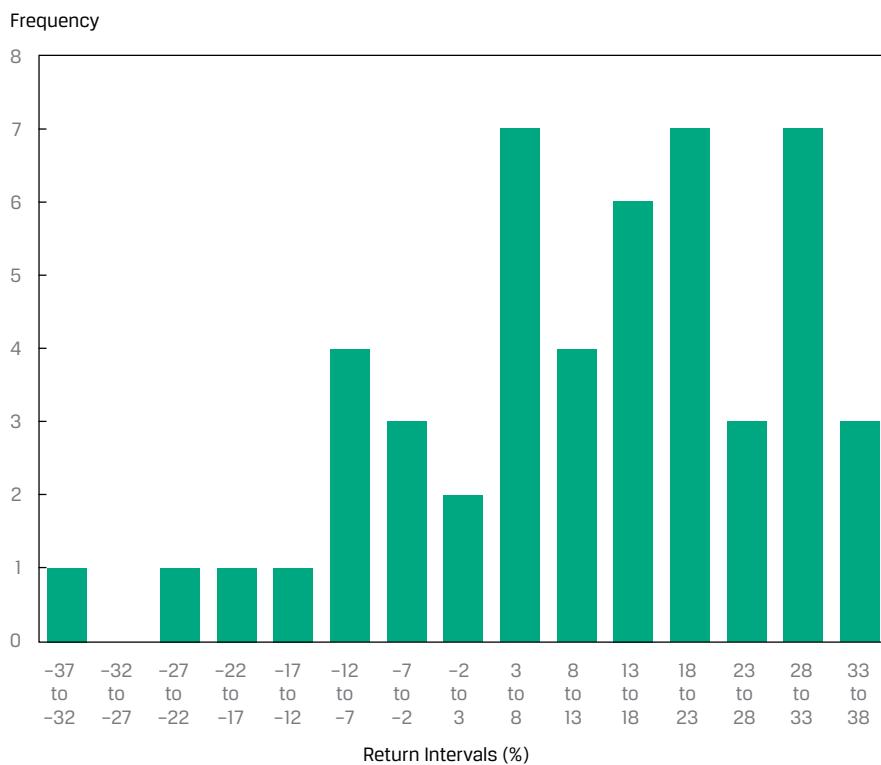
Investment analysts use quantiles to rank performance, such as the performance of assets, indexes, and portfolios. The performance of investment managers is often characterized in terms of the percentile in which their performance falls relative to the performance of their peer group of managers. The widely used Morningstar investment fund star rankings, for example, associate the number of stars with percentiles of performance metrics relative to similar-style investment funds.

Another key use of quantiles is in investment research. Dividing data into quantiles based on a specific objectively quantifiable characteristic, such as sales, market capitalization, or asset size allows analysts to evaluate the impact of that specific characteristic on a quantity of interest, such as asset returns, sales, growth, or valuation metrics. For instance, quantitatively driven investors often rank companies based on the market value of their equity, or their market capitalization, before sorting them into deciles. The first decile contains the portfolio of those companies with the smallest market values, usually called small capitalization companies. The tenth decile contains those companies with the largest market values, usually called large capitalization companies. Ranking companies by decile allows analysts to compare the absolute and relative performance of small market capitalization companies with large ones.

QUESTION SET

The histogram in Exhibit 7 shows a distribution of the annual returns for the S&P 500 Index for a 50-year period.

Exhibit 7: Annual Returns for the S&P 50 Index



1. The bin containing the median return is:

- A. 3 percent to 8 percent.
- B. 8 percent to 13 percent.
- C. 13 percent to 18 percent.

Solution:

C is correct. Because 50 data points are in the histogram, the median return would be the mean of the $50/2 = 25$ th and $(50 + 2)/2 = 26$ th positions. The sum of the return bin frequencies to the left of the 13 percent to 18 percent interval is 24. As a result, the 25th and 26th returns will fall in the 13 percent to 18 percent interval.

2. Based on Exhibit 7, the distribution would be *best* described as being:

- A. unimodal.
- B. bimodal.
- C. trimodal.

Solution:

C is correct. The mode of a distribution with data grouped in intervals is the interval with the highest frequency. The three intervals of 3 percent to 8

percent, 18 percent to 23 percent, and 28 percent to 33 percent all have the highest frequency of 7.

Consider the annual returns in Exhibit 8 for three portfolios for Portfolios P, Q and R. Portfolios P and R were created in Year 1, Portfolio Q was created in Year 2.

Exhibit 8: Annual Portfolio Returns

	Year 1 (%)	Year 2 (%)	Year 3 (%)	Year 4 (%)	Year 5 (%)
Portfolio P	-3.0	4.0	5.0	3.0	7.0
Portfolio Q	NA	-3.0	6.0	4.0	8.0
Portfolio R	1.0	-1.0	4.0	4.0	3.0

3. The median annual return for:

- A. Portfolio P is 4.5 percent.
- B. Portfolio Q is 4.0 percent.
- C. Portfolio R is higher than its arithmetic mean annual return.

Solution:

C is correct. The median of Portfolio R is 0.8 percent higher than the mean for Portfolio R. A is incorrect because the median annual return for Portfolio P is 4.0 percent. B is incorrect because the median annual return for Portfolio Q is 5.0 percent (midpoint of 4 percent and 6 percent).

4. The mode for Portfolio R is:

- A. 1.0 percent.
- B. 3.0 percent.
- C. 4.0 percent.

Solution:

C is correct. The mode is the most frequently occurring value in a dataset, which for Portfolio R is 4.0 percent.

A fund had the following returns over the past 10 years:

Exhibit 9: Fund Returns for 10 Years

Year	Return
1	4.5%
2	6.0%
3	1.5%
4	-2.0%
5	0.0%
6	4.5%
7	3.5%
8	2.5%

Year	Return
9	5.5%
10	4.0%

5. The fund's arithmetic mean return over the 10 years is *closest* to:

- A. 2.97 percent.
- B. 3.00 percent.
- C. 3.33 percent.

Solution:

B is correct. The sum of the returns is 30.0 percent, so the arithmetic mean is $30.0\%/10 = 3.0\%$.

6. The fund's geometric mean return over the 10 years is *closest* to:

- A. 2.94 percent.
- B. 2.97 percent.
- C. 3.00 percent.

Solution:

B is correct. The geometric mean return is calculated as follows:

$$\bar{R}_G = \sqrt[10]{(1 + 0.045) \times (1 + 0.06) \times \dots \times (1 + 0.055) \times (1 + 0.04)} - 1,$$

$$\bar{R}_G = \sqrt[10]{1.3402338} - 1 = 2.9717\%.$$

7. The harmonic mean return over the 10 years is *closest* to:

- A. 2.94 percent.
- B. 2.97 percent.
- C. 3.00 percent.

Solution:

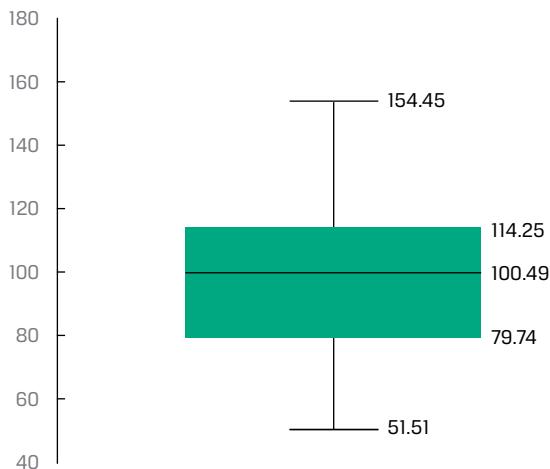
A is correct. The harmonic mean is calculated as follows:

$$\bar{X}_H = \frac{n}{\sum_{i=1}^n (1/(1+r_i))} - 1,$$

$$\bar{X}_H = \frac{10}{\left(\frac{1}{1.045}\right) + \left(\frac{1}{1.06}\right) + \dots + \left(\frac{1}{1.055}\right) + \left(\frac{1}{1.04}\right)} - 1,$$

$$\bar{X}_H = \left(\frac{10}{9.714}\right) - 1 = 2.9442\%.$$

Consider the box and whisker plot in Exhibit 10:

Exhibit 10: Box and Whisker Plot

8. The median is *closest* to:

- A. 34.51.
- B. 100.49.
- C. 102.98.

Solution:

B is correct. In a box and whisker plot, the “box” represents the lower bound of the second quartile and the upper bound of the third quartile, with the median or arithmetic average noted as a measure of central tendency of the entire distribution. The median is indicated within the box, which is 100.49.

9. The interquartile range is *closest* to:

- A. 13.76.
- B. 25.74.
- C. 34.51.

Solution:

C is correct. The interquartile range is the height of the box, which is the difference between 114.25 and 79.74, equal to 34.51.

3**MEASURES OF DISPERSION**

calculate, interpret, and evaluate measures of dispersion to address an investment problem

Few would disagree with the importance of expected return or mean return in investments: To understand an investment more completely, however, we also need to know how returns are dispersed around the mean. **Dispersion** is the variability around the central tendency. If mean return addresses reward, then dispersion addresses risk and uncertainty.

In this lesson, we examine the most common measures of dispersion: range, mean absolute deviation, variance, and standard deviation. These are all measures of **absolute dispersion**. Absolute dispersion is the amount of variability present without comparison to any reference point or benchmark.

The Range

We encountered range earlier when we discussed the construction of frequency distributions. It is the simplest of all the measures of dispersion.

Definition of Range. The **range** is the difference between the maximum and minimum values in a dataset:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}. \quad (2)$$

An alternative way to report the range is to specify both the maximum and minimum values. This alternative definition provides more information as the range is reported as “from Maximum Value to Minimum Value.”

One advantage of the range is ease of computation. A disadvantage is that the range uses only two pieces of information from the distribution. It cannot tell us how the data are distributed (i.e., the shape of the distribution). Because the range is the difference between the maximum and minimum values in the dataset, it is also sensitive to extremely large or small observations (“outliers”) that may not be representative of the distribution.

Mean Absolute Deviations

Measures of dispersion can be computed using all the observations in the distribution rather than just the highest and lowest. We could compute measures of dispersion as the arithmetic average of the deviations around the mean, but the problem is that deviations around the mean always sum to 0. Therefore, we need to find a way to address the problem of negative deviations canceling out positive deviations.

One solution is to examine the absolute deviations around the mean as in the **mean absolute deviation** (MAD).

MAD Formula. The MAD for a sample is:

$$\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}, \quad (3)$$

where \bar{X} is the sample mean, n is the number of observations in the sample, and the $| |$ indicate the absolute value of what is contained within these bars.

The MAD uses all of the observations in the sample and is thus superior to the range as a measure of dispersion. One technical drawback of MAD is that it is difficult to manipulate mathematically compared with the next measure we will introduce, sample variance.

Sample Variance and Sample Standard Deviation

A second approach to the problem of positive and negative deviations canceling out is to square them. Variance and standard deviation, which are based on squared deviations, are the two most widely used measures of dispersion. **Variance** is defined as the average of the squared deviations around the mean. **Standard deviation** is the square root of the variance.

Sample Variance

In investments, we often do not know the mean of a population of interest, so we estimate it using the mean from a sample drawn from the population. The corresponding measure of dispersion is the sample variance or standard deviation.

Sample Variance Formula. The **sample variance**, s^2 , is:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}, \quad (4)$$

where \bar{X} is the sample mean and n is the number of observations in the sample.

The variance calculation takes care of the problem of negative deviations from the mean canceling out positive deviations by squaring those deviations.

For the sample variance, by dividing by the sample size minus 1 (or $n - 1$) rather than n , we improve the statistical properties of the sample variance. The quantity $n - 1$ is also known as the number of degrees of freedom in estimating the population variance. To estimate the population variance with s^2 , we must first calculate the sample mean, which itself is an estimated parameter. Therefore, once we have computed the sample mean, there are only $n - 1$ independent pieces of information from the sample; that is, if you know the sample mean and $n - 1$ of the observations, you could calculate the missing sample observation.

Sample Standard Deviation

Variance is measured in squared units associated with the mean, and we need a way to return to those original units. Standard deviation, the square root of the variance, solves this problem and is more easily interpreted than the variance.

A useful property of the sample standard deviation is that, unlike sample variance, it is expressed in the same unit as the data itself. If the dataset is percentage of daily returns for an index, then both the average and the standard deviation of the dataset is in percentage terms, while the variance is in squared percentage of daily returns.

Sample Standard Deviation Formula. The **sample standard deviation**, s , is:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}, \quad (5)$$

where \bar{X} is the sample mean and n is the number of observations in the sample.

Because the standard deviation is a measure of dispersion about the arithmetic mean, we usually present the arithmetic mean and standard deviation together when summarizing data. When we are dealing with data that represent a time series of percentage changes, presenting the geometric mean (i.e., representing the compound rate of growth) is also very helpful.

Downside Deviation and Coefficient of Variation

An asset's variance or standard deviation of returns is often interpreted as a measure of the asset's risk. Variance and standard deviation of returns take account of returns above and below the mean, or upside and downside risks, respectively. However, investors are typically concerned only with **downside risk**—for example, returns below the mean or below some specified minimum target return. As a result, analysts have developed measures of downside risk.

Downside Deviation

In practice, we may be concerned with values of return (or another variable) below some level other than the mean. For example, if our return objective is 6.0 percent annually (our minimum acceptable return), then we may be concerned particularly with returns below 6.0 percent a year. The target downside deviation, also referred to as the **target semideviation**, is a measure of dispersion of the observations (here, returns) below a target—for example 6.0 percent. To calculate a sample target semideviation, we first specify the target. After identifying observations below the target, we find the sum of those squared negative deviations from the target, divide that sum by the total number of observations in the sample minus 1, and, finally, take the square root.

Sample Target Semideviation Formula. The target semideviation, s_{Target} is:

$$s_{\text{Target}} = \sqrt{\frac{\sum_{\text{for all } X_i \leq B} \frac{(X_i - B)^2}{n-1}}{n-1}}, \quad (6)$$

where B is the target and n is the total number of sample observations.

We illustrate this in Example 3.

EXAMPLE 3**Calculating Target Downside Deviation**

Consider the monthly returns on a portfolio as shown in Exhibit 11:

Exhibit 11: Monthly Portfolio Returns

Month	Return (%)
January	5
February	3
March	-1
April	-4
May	4
June	2
July	0
August	4
September	3
October	0
November	6
December	5

1. Calculate the target downside deviation when the target return is 3 percent.

Solution:

Month	Observation	Deviation from the 3% Target	Deviations below the Target	Squared Deviations below the Target
January	5	2	—	—
February	3	0	—	—

Month	Observation	Deviation from the 3% Target	Deviations below the Target	Squared Deviations below the Target
March	-1	-4	-4	16
April	-4	-7	-7	49
May	4	1	—	—
June	2	-1	-1	1
July	0	-3	-3	9
August	4	1	—	—
September	3	0	—	—
October	0	-3	-3	9
November	6	3	—	—
December	5	2	—	—
Sum				84

$$\text{Target semideviation} = \sqrt{\frac{84}{(12 - 1)}} = 2.7634\%.$$

2. If the target return were 4 percent, would your answer be different from that for question 1? Without using calculations, explain how would it be different?

Solution:

If the target return is higher, then the existing deviations would be larger and there would be several more values in the deviations and squared deviations below the target; so, the target semideviation would be larger.

How does the target downside deviation relate to the sample standard deviation? We illustrate the differences between the target downside deviation and the standard deviation in Example 4, using the data in Example 3.

EXAMPLE 4

Comparing the Target Downside Deviation with the Standard Deviation

1. Given the data in Example 3, calculate the sample standard deviation.

Solution:

Month	Observation	Deviation from the Mean	
		Mean	Squared Deviation
January	5	2.75	7.5625
February	3	0.75	0.5625
March	-1	-3.25	10.5625
April	-4	-6.25	39.0625
May	4	1.75	3.0625
June	2	-0.25	0.0625
July	0	-2.25	5.0625
August	4	1.75	3.0625

Month	Observation	Deviation from the	
		Mean	Squared Deviation
September	3	0.75	0.5625
October	0	-2.25	5.0625
November	6	3.75	14.0625
December	5	2.75	7.5625
Sum	27		96.2500

The sample standard deviation is $\sqrt{\frac{96.2500}{11}} = 2.958\%$.

2. Given the data in Example 3, calculate the target downside deviation if the target is 2 percent.

Solution:

Month	Observation	Deviation	Deviations	Squared
		from the 2% Target	below the Target	Deviations below the Target
January	5	3	—	—
February	3	1	—	—
March	-1	-3	-3	9
April	-4	-6	-6	36
May	4	2	—	—
June	2	0	—	—
July	0	-2	-2	4
August	4	2	—	—
September	3	1	—	—
October	0	-2	-2	4
November	6	4	—	—
December	5	3	—	—
Sum				53

The target semideviation with 2 percent target = $\sqrt{\frac{53}{11}} = 2.195$ percent.

3. Compare the standard deviation, the target downside deviation if the target is 2 percent, and the target downside deviation if the target is 3 percent.

Solution:

The standard deviation is based on the deviation from the mean, or $27\% / 12 = 2.25\%$. The standard deviation includes all deviations from the mean, not just those below it. This results in a sample standard deviation of 2.958 percent.

Considering just the four observations below the 2 percent target, the target semideviation is 2.195 percent. It is less than the sample standard deviation since target semideviation captures only the downside risk (i.e., deviations below the target). Considering target semideviation with a 3 percent target, there are now five observations below 3 percent, so the target semideviation is higher, at 2.763 percent.

Coefficient of Variation

We noted earlier that the standard deviation is more easily interpreted than variance because standard deviation uses the same units of measurement as the observations. We may sometimes find it difficult to interpret what standard deviation means in terms of the relative degree of variability of different sets of data, however, either because the datasets have markedly different means or because the datasets have different units of measurement. In this section, we explain a measure of relative dispersion, the coefficient of variation that can be useful in such situations. **Relative dispersion** is the amount of dispersion relative to a reference value or benchmark.

The coefficient of variation is helpful in such situations as that just described (i.e., datasets with markedly different means or different units of measurement).

Coefficient of Variation Formula. The **coefficient of variation** (CV) is the ratio of the standard deviation of a set of observations to their mean value:

$$CV = \frac{s}{\bar{X}}, \quad (7)$$

where s is the sample standard deviation and \bar{X} is the sample mean.

When the observations are returns, for example, the CV measures the amount of risk (standard deviation) per unit of reward (mean return). An issue that may arise, especially when dealing with returns, is that if \bar{X} is negative, the statistic is meaningless.

The CV may be stated as a multiple (e.g., two times) or as a percentage (e.g., 200 percent). Expressing the magnitude of variation among observations relative to their average size, the CV permits direct comparisons of dispersion across different datasets. Reflecting the correction for scale, the CV is a scale-free measure (i.e., it has no units of measurement).

We illustrate the usefulness of CV for comparing datasets with markedly different standard deviations using two hypothetical samples of companies in Example 5.

EXAMPLE 5

Coefficient of Variation of Returns on Assets

Suppose an analyst collects the return on assets (ROA), in percentage terms, for 10 companies for each of two industries:

Exhibit 12: Returns on Assets for Two Industries

Company	Industry A	Industry B
1	-5	-10
2	-3	-9
3	-1	-7
4	2	-3
5	4	1
6	6	3
7	7	5
8	9	18
9	10	20
10	11	22

These data can be represented graphically as shown in Exhibit 13:

Exhibit 13: Returns on Assets Depicted Graphically

Industry A



Industry B



1. Calculate the average ROA for each industry.

Solution:

The arithmetic mean ROA for both industries is the sum of the 10 observations, which in both cases is 40, divided by the 10 observations, or $40/10 = 4\%$.

2. Calculate the standard deviation of ROA for each industry.

Solution:

Using Equation 5, the standard deviation for Industry A is 5.60 and 12.12 for Industry B.

3. Calculate the coefficient of variation (CV) of the ROA for each industry.

Solution:

Using Equation 7, the CV for Industry A = $5.60/4 = 1.40$ and the CV for Industry B = $12.12/4 = 3.03$.

Though the two industries have the same arithmetic mean ROA, the dispersion is quite different—with Industry B's returns on assets being much more disperse than those of Industry A. The CV for these two industries reflects this, with Industry B having a larger CV. The interpretation is that the risk per unit of mean return is more than two times ($2.16 = 3.03/1.40$) greater for Industry B than Industry A.

QUESTION SET

Consider the annual MSCI World Index total returns for a 10-year period, as shown in Exhibit 14:

Exhibit 14: MSCI World Index Total Returns

Year 1	15.25%	Year 6	30.79%
Year 2	10.02%	Year 7	12.34%
Year 3	20.65%	Year 8	-5.02%

Year 4	9.57%	Year 9	16.54%
Year 5	-40.33%	Year 10	27.37%

1. For Years 6 through 10, the mean absolute deviation (MAD) of the MSCI World Index total returns is *closest* to:

- A. 10.20 percent.
- B. 12.74 percent.
- C. 16.40 percent.

Solution:

A is correct. The MAD is calculated as follows:

Step 1 Sum annual returns and divide by n to find the arithmetic mean (\bar{X}) of 16.40 percent.

Step 2 Calculate the absolute value of the difference between each year's return and the mean from Step 1. Sum the results and divide by n to find the MAD:

$$\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n},$$

$$\text{MAD} = \frac{50.98\%}{5} = 10.20\%.$$

These calculations are shown in the following table:

Year	Step 1		Step 2
	Return	$ X_i - \bar{X} $	
Year 6	30.79%	14.39%	
Year 7	12.34%	4.06%	
Year 8	-5.02%	21.42%	
Year 9	16.54%	0.14%	
Year 10	27.37%	10.97%	
Sum:	82.02%	50.98%	
n :	5	5	
\bar{X} :	16.40%	10.20%	

Annual returns and summary statistics for three funds are listed as follows:

Year	Annual Returns (%)		
	Fund ABC	Fund XYZ	Fund PQR
Year 1	-20.0	-33.0	-14.0
Year 2	23.0	-12.0	-18.0
Year 3	-14.0	-12.0	6.0
Year 4	5.0	-8.0	-2.0
Year 5	-14.0	11.0	3.0

Year	Annual Returns (%)		
	Fund ABC	Fund XYZ	Fund PQR
Mean	-4.0	-10.8	-5.0
Standard deviation	17.8	15.6	10.5

2. The fund with the mean absolute deviation (MAD) is Fund:

- A. ABC.
- B. XYZ.
- C. PQR.

Solution:

A is correct. The MAD of Fund ABC's returns is the highest among the three funds. Using Equation 3, the MAD for each fund is calculated as follows:

MAD for Fund ABC =

$$\frac{| -20 - (-4)| + | 23 - (-4)| + | -14 - (-4)| + | 5 - (-4)| + | -14 - (-4)|}{5} = 14.4\%.$$

MAD for Fund XYZ =

$$\frac{| -33 - (-10.8)| + | -12 - (-10.8)| + | -12 - (-10.8)| + | -8 - (-10.8)| + | 11 - (-10.8)|}{5} = 9.8\%.$$

MAD for Fund PQR =

$$\frac{| -14 - (-5)| + | -18 - (-5)| + | 6 - (-5)| + | -2 - (-5)| + | 3 - (-5)|}{5} = 8.8\%.$$

3. Consider the statistics in Exhibit 15 for Portfolio A and Portfolio B over the past 12 months:

Exhibit 15: Portfolio A and Portfolio B Statistics

	Portfolio A	Portfolio B
Average Return	3%	3%
Geometric Return	2.85%	?
Standard Deviation	4%	6%

4. The geometric mean return of Portfolio B is *most likely* to be:

- A. less than 2.85 percent.
- B. equal to 2.85 percent.
- C. greater than 2.85 percent.

Solution:

A is correct. The higher the dispersion of a distribution, the greater the difference between the arithmetic mean and the geometric mean.

4**MEASURES OF SHAPE OF A DISTRIBUTION**

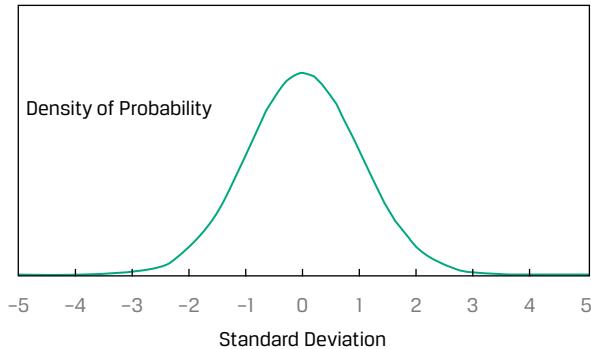
interpret and evaluate measures of skewness and kurtosis to address an investment problem

Mean and variance may not adequately describe an investment's distribution of returns. In calculations of variance, for example, the deviations around the mean are squared, so we do not know whether large deviations are likely to be positive or negative. We need to go beyond measures of central tendency, location, and dispersion to reveal other important characteristics of the distribution. One important characteristic of interest to analysts is the degree of symmetry in return distributions.

If a return distribution is symmetrical about its mean, each side of the distribution is a mirror image of the other. Thus, equal loss and gain intervals exhibit the same frequencies.

One of the most important distributions is the normal distribution, depicted in Exhibit 16. This symmetrical, bell-shaped distribution plays a central role in the mean–variance model of portfolio selection; it is also used extensively in financial risk management. The normal distribution has the following characteristics:

- Its mean, median, and mode are equal.
- It is completely described by two parameters—its mean and variance (or standard deviation).

Exhibit 16: The Normal Distribution

Other distributions may require more information than just the mean and variance to characterize their shape.

Skewness

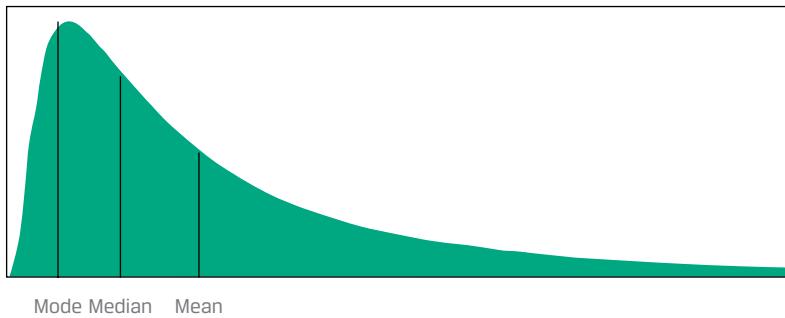
A distribution that is not symmetrical is termed **skewed**. A return distribution with positive skew has frequent small losses and a few extreme gains. A return distribution with negative skew has frequent small gains and a few extreme losses. Panel A

of Exhibit 17 illustrates a continuous positively skewed distribution, which has a long tail on its right side; Panel B illustrates a continuous negatively skewed distribution, which has a long tail on its left side.

Exhibit 17: Properties of Skewed Distributions

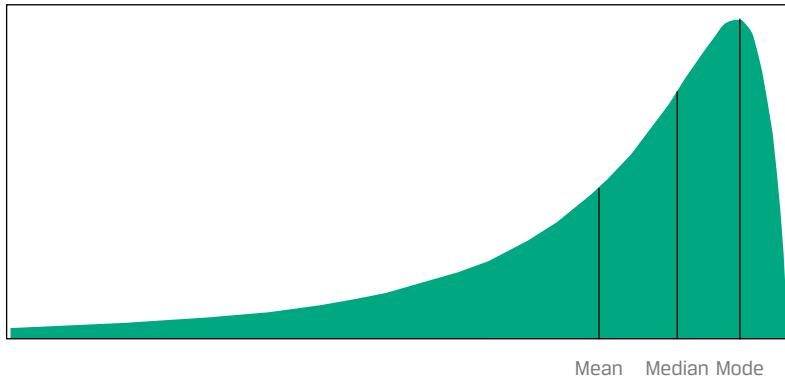
A. Positively Skewed

Density of Probability



B. Negatively Skewed

Density of Probability



For a continuous positively skewed unimodal distribution, the mode is less than the median, which is less than the mean. For the continuous negatively skewed unimodal distribution, the mean is less than the median, which is less than the mode. For a given expected return and standard deviation, investors should be attracted by a positive skew because the mean return lies above the median. Relative to the mean return, positive skew amounts to limited, though frequent, downside returns compared with somewhat unlimited, but less frequent, upside returns.

Skewness is the name given to a statistical measure of skew. (The word “skewness” is also sometimes used interchangeably for “skew.”) Like variance, skewness is computed using each observation’s deviation from its mean. **Skewness** (sometimes referred to as relative skewness) is computed as the average cubed deviation from the mean, standardized by dividing by the standard deviation cubed to make the measure free of scale.

Cubing, unlike squaring, preserves the sign of the deviations from the mean. If a distribution is positively skewed with a mean greater than its median, then more than half of the deviations from the mean are negative and less than half are positive. However, for the sum of the cubed deviations to be positive, the losses must be small

and likely and the gains less likely but more extreme. Therefore, if skewness is positive, the average magnitude of positive deviations is larger than the average magnitude of negative deviations.

The approximation for computing **sample skewness** when n is large (100 or more) is:

$$\text{Skewness} \approx \left(\frac{1}{n}\right) \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{s^3}. \quad (8)$$

As you will learn later in the curriculum, different investment strategies may introduce different types and amounts of skewness into returns.

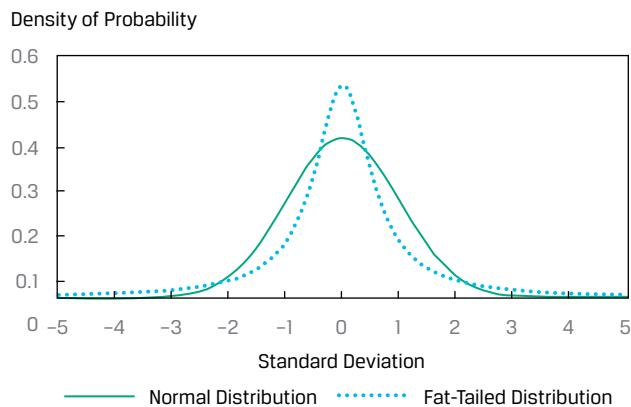
Kurtosis

Another way in which a return distribution might differ from a normal distribution is its relative tendency to generate large deviations from the mean. Most investors would perceive a greater chance of extremely large deviations from the mean as higher risk.

Kurtosis is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution—that is, the proportion of the total probability that is outside of, say, 2.5 standard deviations of the mean. A distribution that has fatter tails than the normal distribution is referred to as **leptokurtic** or **fat-tailed**; a distribution that has thinner tails than the normal distribution is referred to as being **platykurtic** or **thin-tailed**; and a distribution similar to the normal distribution as it concerns relative weight in the tails is called **mesokurtic**. A fat-tailed (thin-tailed) distribution tends to generate more frequent (less frequent) extremely large deviations from the mean than the normal distribution.

Exhibit 18 illustrates a distribution with fatter tails than the normal distribution. By construction, the fat-tailed and normal distributions in Exhibit 18 have the same mean, standard deviation, and skewness. Note that this fat-tailed distribution is more likely than the normal distribution to generate observations in the tail regions defined by the intersection of the distribution lines near a standard deviation of about ± 2.5 . This fat-tailed distribution is also more likely to generate observations that are near the mean, defined here as the region ± 1 standard deviation around the mean. However, to ensure probabilities sum to 1, this distribution generates fewer observations in the regions between the central region and the two tail regions.

Exhibit 18: Fat-Tailed Distribution Compared to the Normal Distribution



The calculation for kurtosis involves finding the average of deviations from the mean raised to the fourth power and then standardizing that average by dividing by the standard deviation raised to the fourth power. A normal distribution has kurtosis of 3.0, so a fat-tailed distribution has a kurtosis above 3.0 and a thin-tailed distribution has a kurtosis below 3.0.

Excess kurtosis is the kurtosis relative to the normal distribution. For a large sample size ($n = 100$ or more), **sample excess kurtosis** (K_E) is approximately as follows:

$$K_E \approx \left[\left(\frac{1}{n} \right) \sum_{i=1}^n (X_i - \bar{X})^4 \right] - 3. \quad (9)$$

As with skewness, this measure is free of scale. Many statistical packages report estimates of sample excess kurtosis, labeling this as simply “kurtosis.”

Excess kurtosis thus characterizes kurtosis relative to the normal distribution. A normal distribution has excess kurtosis equal to 0. A fat-tailed distribution has excess kurtosis greater than 0, and a thin-tailed distribution has excess kurtosis less than 0. A return distribution with positive excess kurtosis—a fat-tailed return distribution—has more frequent extremely large deviations from the mean than a normal distribution.

Exhibit 19: Summary of Kurtosis

If kurtosis is ...	then excess kurtosis is ...	Therefore, the distribution is ...	And we refer to the distribution as being ...
above 3.0	above 0	fatter-tailed than the normal distribution.	fat-tailed (leptokurtic)
equal to 3.0	equal to 0	similar in tails to the normal distribution.	mesokurtic
less than 3.0	less than 0	thinner-tailed than the normal distribution.	thin-tailed (platykurtic)

Most equity return series have been found to be fat-tailed. If a return distribution is fat-tailed and we use statistical models that do not account for that distribution, then we will underestimate the likelihood of very bad or very good outcomes. Example 6 revisits the EAA Equity Index from the earlier Example 1 and quantifies the shape of its return distribution.

EXAMPLE 6

Skewness and Kurtosis of EAA Equity Index Daily Returns

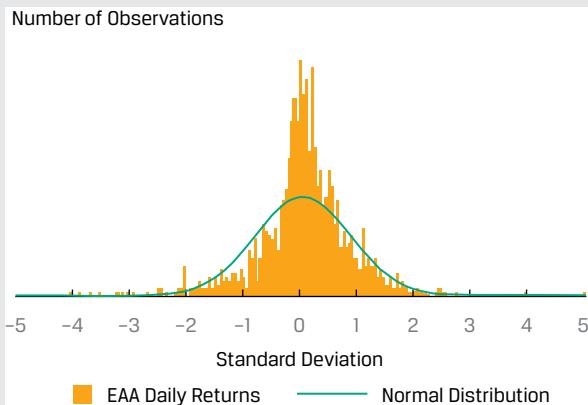
Consider the statistics in Exhibit 20 for the EAA Equity Index:

Exhibit 20: Properties of Skewed Distributions

Daily Return (%)	
Arithmetic mean	0.0347
Standard deviation	0.8341
Measure of Symmetry	
Skewness	-0.4260
Excess kurtosis	3.7962

The returns reflect negative skewness, which is illustrated in Exhibit 21 by comparing the distribution of the daily returns with a normal distribution with the same mean and standard deviation.

Exhibit 21: Negative Skewness



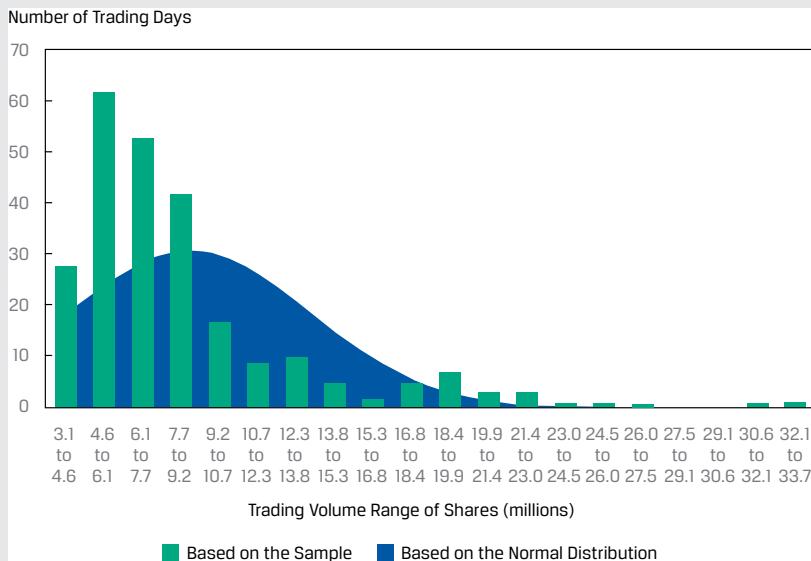
Using both the statistics and the graph, we see the following:

- The distribution is negatively skewed, as indicated by the negative calculated skewness of -0.4260 and the influence of observations below the mean of 0.0347 percent.
- The highest frequency of returns occurs within the 0.0 to 0.5 standard deviations from the mean (i.e., the mode is greater than the mean as the positive returns are offset by extreme negative deviations).
- The distribution is fat-tailed, as indicated by the positive excess kurtosis of 3.7962 . In Exhibit 21, we can see fat tails, a concentration of returns around the mean, and fewer observations in the regions between the central region and the two-tail regions.

To understand the trading liquidity of a stock, investors often look at the distribution of the daily trading volume for a stock. Analyzing the daily volume can provide insights about the interest in the stock, what factors may drive interest in the stock as well as whether the market can absorb a large trade in the stock. The latter may be of interest to investors interested in either establishing or exiting a large position in the particular stock.

INTERPRETING SKEWNESS AND KURTOSIS

Consider the daily trading volume for a stock for one year, as shown in Exhibit 22. In addition to the count of observations within each bin or interval, the number of observations anticipated based on a normal distribution (given the sample arithmetic average and standard deviation) is provided as well. The average trading volume per day for the stock during the year was 8.6 million shares, and the standard deviation was 4.9 million shares.

Exhibit 22: Histogram of Daily Trading Volume for a Stock for One Year


1. Would the distribution be characterized as being skewed? If so, what could account for this situation?

Solution:

The distribution appears to be skewed to the right, or positively skewed. This is likely due to: (1) no possible negative trading volume on a given trading day, so the distribution is truncated at zero; and (2) greater-than-typical trading occurring relatively infrequently, such as when there are company-specific announcements.

The actual skewness for this distribution is 2.1090, which supports this interpretation.

2. Does the distribution displays kurtosis? Explain.

Solution:

The distribution appears to have excess kurtosis, with a right-side fat tail and with maximum shares traded in the 4.6 to 6.1 million range, exceeding what is expected if the distribution was normally distributed. There are also fewer observations than expected between the central region and the tail.

The actual excess kurtosis for this distribution is 5.2151, which supports this interpretation.

QUESTION SET


1. An analyst calculates the excess kurtosis of a stock's returns as -0.75. From this information, the analyst should conclude that the distribution of returns is:
 - normally distributed.
 - fat-tailed compared to the normal distribution.

- C. thin-tailed compared to the normal distribution.

Solution:

C is correct. The distribution is thin-tailed relative to the normal distribution because the excess kurtosis is less than zero.

Use Exhibit 23 to answer questions 2–4.

An analyst examined a cross-section of annual returns for 252 stocks and calculated the following statistics:

Exhibit 23: Cross-Section of Annual Returns

Arithmetic Average	9.986%
Geometric Mean	9.909%
Variance	0.001723
Skewness	0.704
Excess Kurtosis	0.503

2. The coefficient of variation (CV) is closest to:

- A. 0.02.
- B. 0.42.
- C. 2.41.

Solution:

B is correct. The CV is the ratio of the standard deviation to the arithmetic average, or $\sqrt{0.001723}/0.09986 = 0.416$.

3. This distribution is best described as:

- A. negatively skewed.
- B. having no skewness.
- C. positively skewed.

Solution:

C is correct. The skewness is positive, so it is right-skewed (positively skewed).

4. Compared to the normal distribution, this sample's distribution is best described as having tails of the distribution with:

- A. less probability than the normal distribution.
- B. the same probability as the normal distribution.
- C. more probability than the normal distribution.

Solution:

C is correct. The excess kurtosis is positive, indicating that the distribution is "fat-tailed"; therefore, there is more probability in the tails of the distribution relative to the normal distribution.

CORRELATION BETWEEN TWO VARIABLES

5

- interpret correlation between two variables to address an investment problem

Scatter Plot

A **scatter plot** is a useful tool for displaying and understanding potential relationships between two variables. Suppose an analyst is interested in the relative performance of two sectors, information technology (IT) and utilities, compared to the market index over a specific five-year period. The analyst has obtained the sector and market index returns for each month over the five years under investigation. Exhibit 24 presents a scatterplot of returns for the IT sector index versus the S&P 500, and Exhibit 25 presents a scatterplot of returns for the utilities sector index versus the S&P 500.

Tight (loose) clustering signals a potentially stronger (weaker) relationship between the two variables.

Exhibit 24: Scatter Plot of Information Technology Sector Index Return versus S&P 500 Index Return

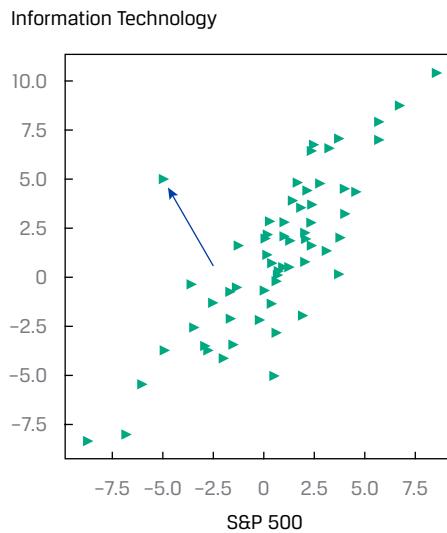
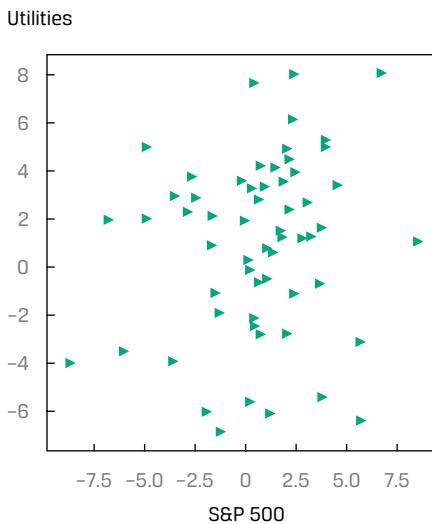


Exhibit 25: Scatter Plot of Utilities Sector Index Return versus S&P 500 Index Return



Despite their relatively straightforward construction, scatter plots convey valuable information. First, it is important to inspect for any potential association between the two variables. The pattern of the scatter plot may indicate no apparent relationship, a linear association, or a non-linear relationship. Furthermore, the strength of the association can be determined by how closely the data points are clustered around a line drawn across the observations.

Examining Exhibit 24 we can see the returns of the IT sector are highly positively associated with S&P 500 Index returns because the data points are tightly clustered along a positively sloped line. In contrast, Exhibit 25 tells a different story for relative performance of the utilities sector and S&P 500 index returns: The data points appear to be distributed in no discernable pattern, indicating no clear relationship among these variables.

Second, observing the data points located toward the ends of each axis, which represent the maximum or minimum values, provides a quick sense of the data range. Third, if a relationship among the variables is apparent, inspecting the scatter plot can help to spot extreme values (i.e., outliers). For example, an outlier data point is readily detected in Exhibit 24 as indicated by the arrow. Finding these extreme values and handling them with appropriate measures is an important part of the financial modeling process.

Scatter plots are a powerful tool for finding patterns between two variables, for assessing data range, and for spotting extreme values. In some situations, however, we need to inspect for pairwise associations among many variables—for example, when conducting feature selection from dozens of variables to build a predictive model.

Now that we have some understanding of sample variance and standard deviation, we can more formally consider the concept of correlation between two random variables that we previously explored visually in the scatter plots.

Covariance and Correlation

Correlation is a measure of the linear relationship between two random variables. The first step in considering how two variables vary together, however, is constructing their covariance.

Definition of Sample Covariance. The **sample covariance** (s_{XY}) is a measure of how two variables in a sample move together:

$$s_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}. \quad (10)$$

Equation 10 indicates that the sample covariance is the average value of the product of the deviations of observations on two random variables (X_i and Y_i) from their sample means. If the random variables are returns, the units would be returns squared. Also, note the use of $n - 1$ in the denominator, which ensures that the sample covariance is an unbiased estimate of population covariance.

Stated simply, covariance is a measure of the joint variability of two random variables. If the random variables vary in the same direction—for example, X tends to be above its mean when Y is above its mean, and X tends to be below its mean when Y is below its mean—then their covariance is positive. If the variables vary in the opposite direction relative to their respective means, then their covariance is negative.

The size of the covariance measure alone is difficult to interpret as it involves squared units of measure and so depends on the magnitude of the variables. This brings us to the normalized version of covariance, which is the correlation coefficient.

Definition of Sample Correlation Coefficient. The **sample correlation coefficient** is a standardized measure of how two variables in a sample move together. The sample correlation coefficient (r_{XY}) is the ratio of the sample covariance to the product of the two variables' standard deviations:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}. \quad (11)$$

Importantly, the correlation coefficient expresses the strength of the linear relationship between the two random variables.

Properties of Correlation

We now discuss the correlation coefficient, or simply correlation, and its properties in more detail:

1. Correlation ranges from -1 and $+1$ for two random variables, X and Y :

$$-1 \leq r_{XY} \leq +1.$$

2. A correlation of 0 , termed uncorrelated, indicates an absence of any linear relationship between the variables.
3. A positive correlation close to $+1$ indicates a strong positive linear relationship. A correlation of 1 indicates a perfect linear relationship.
4. A negative correlation close to -1 indicates a strong negative (i.e., inverse) linear relationship. A correlation of -1 indicates a perfect inverse linear relationship.

We return to scatter plots to illustrate correlation visually. In contrast to the correlation coefficient, which expresses the relationship between two data series using a single number, a scatter plot depicts the relationship graphically. Therefore, scatter plots are a very useful tool for the sensible interpretation of a correlation coefficient.

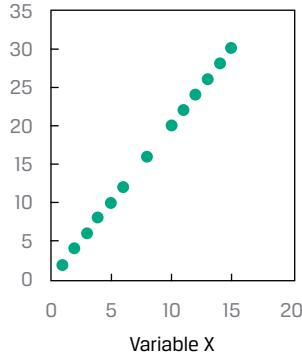
Exhibit 26 shows examples of scatter plots. Panel A shows the scatter plot of two variables with a correlation of $+1$. Note that all the points on the scatter plot in Panel A lie on a straight line with a positive slope. Whenever variable X increases by one unit, variable Y increases by two units. Because all of the points in the graph lie on a straight line, an increase of one unit in X is associated with an exact two-unit increase in Y , regardless of the level of X . Even if the slope of the line were different

(but positive), the correlation between the two variables would still be +1 as long as all the points lie on that straight line. Panel B shows a scatter plot for two variables with a correlation coefficient of -1. Once again, the plotted observations all fall on a straight line. In this graph, however, the line has a negative slope. As X increases by one unit, Y decreases by two units, regardless of the initial value of X .

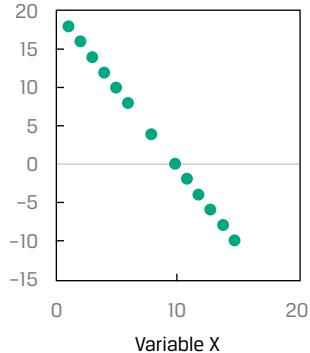
Exhibit 26: Scatter Plots Showing Various Degrees of Correlation

A. Variables With a Correlation of +1 **B. Variables With a Correlation of -1**

Variable Y

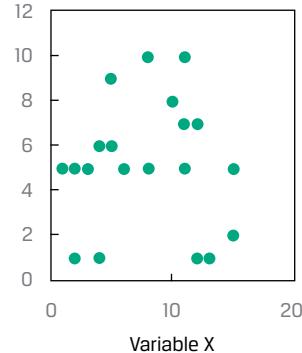


Variable Y



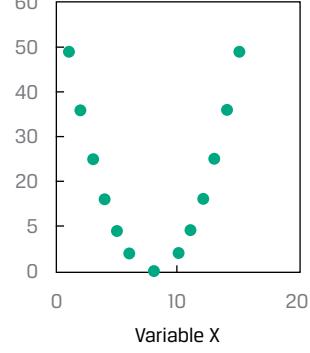
C. Variables With a Correlation of 0

Variable Y



D. Variables With a Strong Nonlinear Association

Variable Y



Panel C shows a scatter plot of two variables with a correlation of 0; they have no linear relation. This graph shows that the value of variable X tells us nothing about the value of variable Y . Panel D shows a scatter plot of two variables that have a non-linear relationship. Because the correlation coefficient is a measure of the linear association between two variables, it would not be appropriate to use the correlation coefficient in this case.

Limitations of Correlation Analysis

Exhibit 26 illustrates that correlation measures the linear association between two variables, but it may not always be reliable. Two variables can have a strong *nonlinear* relation and still have a very low correlation. A nonlinear relation between variables X and Y is shown in Panel D. Even though these two variables are perfectly associated, there is no linear association between them and hence no meaningful correlation.

Correlation may also be an unreliable measure when outliers are present in one or both variables. As we have seen, outliers are small numbers of observations at either extreme (small or large) of a sample. Correlation may be quite sensitive to outliers. In such a situation, we should consider whether it makes sense to exclude those outlier observations and whether they are noise or true information. We use judgment to determine whether those outliers contain information about the two variables' relationship, and should be included in the correlation analysis, or contain no information, and should be excluded. If they are to be excluded from the correlation analysis, as we have seen previously, outlier observations can be handled by trimming or winsorizing the dataset.

Importantly, keep in mind that correlation does not imply causation. Even if two variables are highly correlated, one does not necessarily cause the other in the sense that certain values of one variable bring about the occurrence of certain values of the other.

Moreover, with visualizations too, including scatter plots, we must be on guard against unconsciously making judgments about causal relationships that may or may not be supported by the data.

The term **spurious correlation** has been used to refer to:

- correlation between two variables that reflects chance relationships in a particular dataset;
- correlation induced by a calculation that mixes each of two variables with a third variable; and
- correlation between two variables arising not from a direct relation between them but from their relation to a third variable.

As an example of the chance relationship, consider the monthly US retail sales of beer, wine, and liquor and the atmospheric carbon dioxide levels from 2000 to 2018. The correlation is 0.824, indicating that a positive relation exists between the two. However, there is no reason to suspect that the levels of atmospheric carbon dioxide are related to the retail sales of beer, wine, and liquor.

As an example of the second type of spurious correlation, two variables that are uncorrelated may be correlated if divided by a third variable. For example, consider a cross-sectional sample of companies' dividends and total assets. While there may be a low correlation between these two variables, dividing each by market capitalization may increase the correlation.

As an example of the third type of spurious correlation, height may be positively correlated with the extent of a person's vocabulary, but the underlying relationships are between age and height and between age and vocabulary.

Investment professionals must be cautious in basing investment strategies on high correlations. Spurious correlations may suggest investment strategies that appear profitable but would not be, if implemented.

A further issue is that correlation does not tell the whole story about the data. Consider Anscombe's Quartet, discussed in Example 1, for which dissimilar graphs can be developed with variables that have the same mean, same standard deviation, and same correlation.

EXAMPLE 7**Anscombe's Quartet**

Francis Anscombe, a British statistician, developed datasets that illustrate why just looking at summary statistics (i.e., mean, standard deviation, and correlation) does not fully describe the data. He created four datasets (designated I, II, III, and IV), each with two variables, X and Y , such that:

- The X s in each dataset have the same mean and standard deviation, 9.00 and 3.32, respectively.
- The Y s in each dataset have the same mean and standard deviation, 7.50 and 2.03, respectively.
- The X s and Y s in each dataset have the same correlation of 0.82.

Exhibit 27: Summary Statistics

Observation	I		II		III		IV	
	X	Y	X	Y	X	Y	X	Y
1	10	8.04	10	9.14	10	7.46	8	6.6
2	8	6.95	8	8.14	8	6.77	8	5.8
3	13	7.58	13	8.74	13	12.74	8	7.7
4	9	8.81	9	8.77	9	7.11	8	8.8
5	11	8.33	11	9.26	11	7.81	8	8.5
6	14	9.96	14	8.1	14	8.84	8	7
7	6	7.24	6	6.13	6	6.08	8	5.3
8	4	4.26	4	3.1	4	5.39	19	13
9	12	10.8	12	9.13	12	8.15	8	5.6
10	7	4.82	7	7.26	7	6.42	8	7.9
11	5	5.68	5	4.74	5	5.73	8	6.9
<i>N</i>	11	11	11	11	11	11	11	11
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Standard deviation	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
Correlation	0.82		0.82		0.82		0.82	

While the X variable has the same values for I, II, and III in the quartet of datasets, the Y variables are quite different, creating different relationships. The four datasets are:

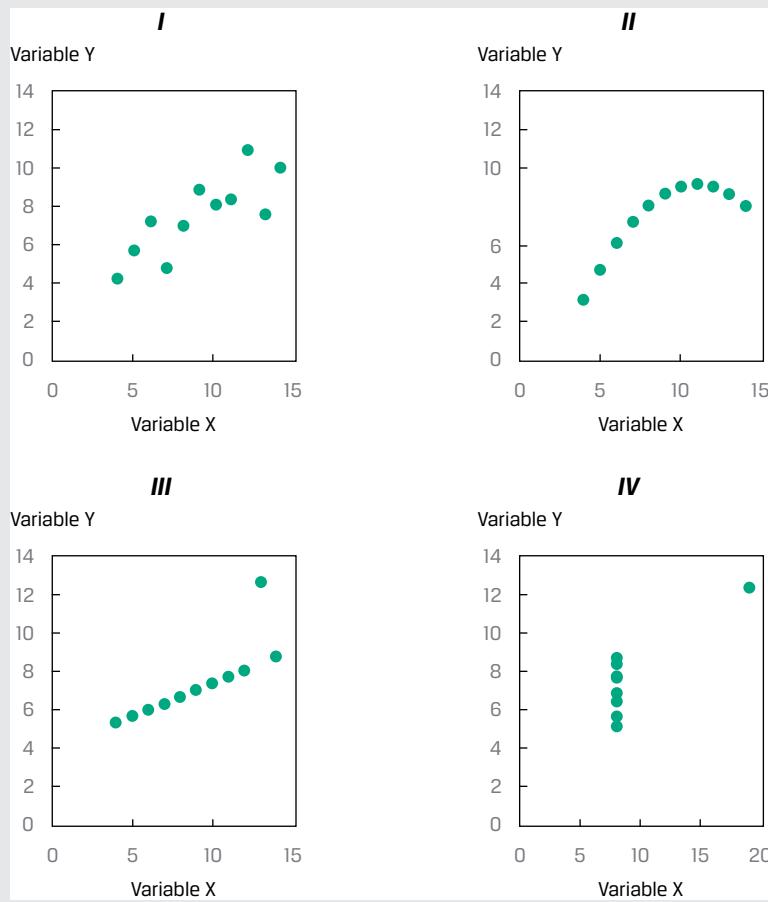
I An approximate linear relationship between X and Y .

II A curvilinear relationship between X and Y .

III A linear relationship except for one outlier.

IV A constant X with the exception of one outlier.

Depicting the quartet visually,

Exhibit 28: Visual Depiction

The bottom line? Knowing the means and standard deviations of the two variables, as well as the correlation between them, does not tell the entire story.

Source: Francis John Anscombe, "Graphs in Statistical Analysis," *American Statistician* 27 (February 1973): 17–21.

QUESTION SET

1. A correlation of 0.34 between two variables, X and Y , is best described as:

- A. changes in X causing changes in Y .
- B. a positive association between X and Y .
- C. a curvilinear relationship between X and Y .

Solution:

B is correct. The correlation coefficient is positive, indicating that the two series move together.

2. Which of the following is a potential problem with interpreting a correlation coefficient?

- A. Outliers
- B. Spurious correlation

- C. Both outliers and spurious correlation

Solution:

C is correct. Both outliers and spurious correlation are potential problems with interpreting correlation coefficients.

Use the information in Exhibit 29 to answer questions 3 and 4.

An analyst is evaluating the tendency of returns on the portfolio of stocks she manages to move along with bond and real estate indexes. She gathered monthly data on returns and the indexes:

Exhibit 29: Monthly Data on Returns and Indexes

	Returns (%)		
	Portfolio Returns	Bond Index Returns	Real Estate Index Returns
Arithmetic average	5.5	3.2	7.8
Standard deviation	8.2	3.4	10.3
	Portfolio Returns and Bond Index Returns	Portfolio Returns and Real Estate Index Returns	
Covariance	18.9		-55.9

3. Without calculating the correlation coefficient, the correlation of the portfolio returns and the bond index returns is *most likely* to be:

- A. negative.
- B. zero.
- C. positive.

Solution:

C is correct. The correlation coefficient is positive because the covariance is positive.

4. Without calculating the correlation coefficient, the correlation of the portfolio returns and the real estate index returns is:

- A. negative.
- B. zero.
- C. positive.

Solution:

A is correct. The correlation coefficient is negative because the covariance is negative.

5. Consider two variables, A and B. If variable A has a mean of -0.56, variable B has a mean of 0.23, and the covariance between the two variables is positive, the correlation between these two variables is:

- A. negative.
- B. zero.

C. positive.

Solution:

C is correct. The correlation coefficient must be positive because the covariance is positive. The fact that one or both variables have a negative mean does not affect the sign of the correlation coefficient.

PRACTICE PROBLEMS

The following information relates to questions 1–5

Consider the results of an analysis focusing on the market capitalizations of a sample of 100 firms:

Exhibit 1: Market Capitalization of a Sample of 100 Firms

Bin	Cumulative Percentage of Sample (%)	Market Capitalization (euro billions)		Number of Observations
		Lower Bound	Upper Bound	
1	5	0.28	15.45	5
2	10	15.45	21.22	5
3	15	21.22	29.37	5
4	20	29.37	32.57	5
5	25	32.57	34.72	5
6	30	34.72	37.58	5
7	35	37.58	39.90	5
8	40	39.90	41.57	5
9	45	41.57	44.86	5
10	50	44.86	46.88	5
11	55	46.88	49.40	5
12	60	49.40	51.27	5
13	65	51.27	53.58	5
14	70	53.58	56.66	5
15	75	56.66	58.34	5
16	80	58.34	63.10	5
17	85	63.10	67.06	5
18	90	67.06	73.00	5
19	95	73.00	81.62	5
20	100	81.62	96.85	5

1. The tenth percentile corresponds to observations in bin(s):
 - A. 2.
 - B. 1 and 2.
 - C. 19 and 20.

2. The second quintile corresponds to observations in bin(s):
 - A. 8.

- B.** 5, 6, 7, and 8.
- C.** 6, 7, 8, 9, and 10.
3. The fourth quartile corresponds to observations in bin(s):
- A.** 17.
- B.** 17, 18, 19, and 20.
- C.** 16, 17, 18, 19, and 20.
4. The median is *closest* to:
- A.** 44.86.
- B.** 46.88.
- C.** 49.40.
5. The interquartile range is *closest* to:
- A.** 20.76.
- B.** 23.62.
- C.** 25.52.
-
6. Exhibit 12 shows the annual MSCI World Index total returns for a 10-year period.

Exhibit 1: MSCI World Index Returns

Year 1	15.25%	Year 6	30.79%
Year 2	10.02%	Year 7	12.34%
Year 3	20.65%	Year 8	-5.02%
Year 4	9.57%	Year 9	16.54%
Year 5	-40.33%	Year 10	27.37%

The fourth quintile return for the MSCI World Index is *closest* to:

- A.** 20.65 percent.
- B.** 26.03 percent.
- C.** 27.37 percent.

**The following information relates to questions
7-9**

A fund had the following experience over the past 10 years:

Exhibit 1: Performance over 10 Years

Year	Return
1	4.5%
2	6.0%
3	1.5%
4	-2.0%
5	0.0%
6	4.5%
7	3.5%
8	2.5%
9	5.5%
10	4.0%

7. The fund's standard deviation of returns over the 10 years is *closest* to:
- A. 2.40 percent.
 - B. 2.53 percent.
 - C. 7.58 percent.
8. The target semideviation of the returns over the 10 years, if the target is 2 percent, is *closest* to:
- A. 1.42 percent.
 - B. 1.50 percent.
 - C. 2.01 percent.
9. Consider the mean monthly return and the standard deviation for three industry sectors, as shown in Exhibit 2:

Exhibit 2: Mean Monthly Return and Standard Deviations

Sector	Mean Monthly Return (%)	Standard Deviation of Return (%)
Utilities (UTIL)	2.10	1.23
Materials (MATR)	1.25	1.35
Industrials (INDU)	3.01	1.52

Based on the coefficient of variation (CV), the riskiest sector is:

- A. utilities.
- B. materials.
- C. industrials.

SOLUTIONS

1. B is correct. The tenth percentile corresponds to the lowest 10 percent of the observations in the sample, which are in bins 1 and 2.
2. B is correct. The second quintile corresponds to the second 20 percent of observations. The first 20 percent consists of bins 1 through 4. The second 20 percent of observations consists of bins 5 through 8.
3. C is correct. A quartile consists of 25 percent of the data, and the last 25 percent of the 20 bins are 16 through 20.
4. B is correct. The center of the 20 bins is represented by the market capitalization of the highest value of the 10th bin and the lowest value of the 11th bin, which is 46.88.
5. B is correct. The interquartile range is the difference between the lowest value in the second quartile and the highest value in the third quartile. The lowest value of the second quartile is 34.72, and the highest value of the third quartile is 58.34. Therefore, the interquartile range is $58.34 - 34.72 = 23.62$.
6. B is correct. Quintiles divide a distribution into fifths, with the fourth quintile occurring at the point at which 80 percent of the observations lie below it. The fourth quintile is equivalent to the 80th percentile. To find the y th percentile (P_y), we first must determine its location. The formula for the location (L_y) of a y th percentile in an array with n entries sorted in ascending order is $L_y = (n + 1) \times (y/100)$. In this case, $n = 10$ and $y = 80\%$, so

$$L_{80} = (10 + 1) \times (80/100) = 11 \times 0.8 = 8.8.$$

With the data arranged in ascending order (-40.33 percent, -5.02 percent, 9.57 percent, 10.02 percent, 12.34 percent, 15.25 percent, 16.54 percent, 20.65 percent, 27.37 percent, and 30.79 percent), the 8.8th position would be between the eighth and ninth entries, 20.65 percent and 27.37 percent, respectively. Using linear interpolation, $P_{80} = X_8 + (L_y - 8) \times (X_9 - X_8)$,

$$P_{80} = 20.65 + (8.8 - 8) \times (27.37 - 20.65)$$

$$= 20.65 + (0.8 \times 6.72) = 20.65 + 5.38$$

$$= 26.03 \text{ percent.}$$

7. B is correct. The fund's standard deviation of returns is calculated as follows:

Year	Return	Deviation from Mean	Deviation Squared
1	4.5%	0.0150	0.000225
2	6.0%	0.0300	0.000900
3	1.5%	-0.0150	0.000225
4	-2.0%	-0.0500	0.002500
5	0.0%	-0.0300	0.000900
6	4.5%	0.0150	0.000225
7	3.5%	0.0050	0.000025
8	2.5%	-0.0050	0.000025
9	5.5%	0.0250	0.000625

Year	Return	Deviation from Mean	Deviation Squared
10	4.0%	0.0100	0.000100
Mean	3.0%		
Sum			0.005750

The standard deviation is the square root of the sum of the squared deviations, divided by $n - 1$:

$$s = \sqrt{\frac{0.005750}{(10 - 1)}} = 2.5276\%.$$

8. B is correct. The target semideviation of the returns over the 10 years with a target of 2 percent is calculated as follows:

Year	Return	Deviation Squared below Target of 2%
1	4.5%	
2	6.0%	
3	1.5%	0.000025
4	-2.0%	0.001600
5	0.0%	0.000400
6	4.5%	
7	3.5%	
8	2.5%	
9	5.5%	
10	4.0%	
Sum		0.002025

The target semideviation is the square root of the sum of the squared deviations from the target, divided by $n - 1$:

$$s_{\text{Target}} = \sqrt{\frac{0.002025}{(10 - 1)}} = 1.5\%.$$

9. B is correct. The CV is the ratio of the standard deviation to the mean, where a higher CV implies greater risk per unit of return.

$$CV_{UTIL} = \frac{s}{X} = \frac{1.23\%}{2.10\%} = 0.59,$$

$$CV_{MATR} = \frac{s}{X} = \frac{1.35\%}{1.25\%} = 1.08,$$

$$CV_{INDU} = \frac{s}{X} = \frac{1.52\%}{3.01\%} = 0.51.$$

LEARNING MODULE

4

Probability Trees and Conditional Expectations

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA.

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Jacobs Levy Equity Management (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	calculate expected values, variances, and standard deviations and demonstrate their application to investment problems
<input type="checkbox"/>	formulate an investment problem as a probability tree and explain the use of conditional expectations in investment application
<input type="checkbox"/>	calculate and interpret an updated probability in an investment setting using Bayes' formula

INTRODUCTION

1

Investment decisions are made under uncertainty about the future direction of the economy, issuers, companies, and prices. This learning module presents probability tools that address many real-world problems involving uncertainty and applies to a variety of investment management applications.

Lesson 1 introduces the calculation of the expected value, variance, and standard deviation for a random variable. These are essential quantitative concepts in investment management. Lesson 2 introduces probability trees that help in visualizing the conditional expectations and the total probabilities for expected value.

When making investment decisions, analysts often rely on perspectives, which may be influenced by subsequent observations. Lesson 3 introduces Bayes' formula, a rational method to adjust probabilities with the arrival of new information. This method has wide business and investment applications.

LEARNING MODULE OVERVIEW



- The expected value of a random variable is a probability-weighted average of the possible outcomes of the random variable. For a random variable X , the expected value of X is denoted $E(X)$.
- The variance of a random variable is the expected value (the probability-weighted average) of squared deviations from the random variable's expected value $E(X)$: $\sigma^2(X) = E\{[X - E(X)]^2\}$, where $\sigma^2(X)$ stands for the variance of X .
- Standard deviation is the positive square root of variance. Standard deviation measures dispersion (as does variance), but it is measured in the same units as the variable.
- A probability tree is a means of illustrating the results of two or more independent events.
- A probability of an event given (conditioned on) another event is a conditional probability. The probability of an event A given an event B is denoted $P(A | B)$, and $P(A | B) = P(AB)/P(B)$, $P(B) \neq 0$.
- According to the total probability rule, if S_1, S_2, \dots, S_n are mutually exclusive and exhaustive scenarios or events, then $P(A) = P(A | S_1)P(S_1) + P(A | S_2)P(S_2) + \dots + P(A | S_n)P(S_n)$.
- Conditional expected value is $E(X | S) = P(X_1 | S)X_1 + P(X_2 | S)X_2 + \dots + P(X_n | S)X_n$ and has an associated conditional variance and conditional standard deviation.
- Bayes' formula is a method used to update probabilities based on new information.
- Bayes' formula is expressed as follows: Updated probability of event given the new information = [(Probability of the new information given event)/(Unconditional probability of the new information)] \times Prior probability of event.

2

EXPECTED VALUE AND VARIANCE



calculate expected values, variances, and standard deviations and demonstrate their application to investment problems

The expected value of a random variable is an essential quantitative concept in investments. Investors continually make use of expected values—in estimating the rewards of alternative investments, in forecasting earnings per share (EPS) and other corporate financial variables and ratios, and in assessing any other factor that may affect their financial position. The **expected value of a random variable** is the probability-weighted average of the possible outcomes of the random variable. For a random variable X , the expected value of X is denoted $E(X)$.

Expected value (e.g., expected stock return) looks either to the future, as a forecast, or to the “true” value of the mean (the population mean). We should distinguish expected value from the concepts of historical or sample mean. The sample mean also

summarizes in a single number a central value. However, the sample mean presents a central value for a particular set of observations as an equally weighted average of those observations. In sum, the contrast is forecast versus historical, or population versus sample.

An equation that summarizes the calculation of the expected value for a discrete random variable X is as follows:

$$E(X) = P(X_1)X_1 + P(X_2)X_2 + \dots + P(X_n)X_n = \sum_{i=1}^n P(X_i)X_i \quad (1)$$

where X_i is one of n possible outcomes of the discrete random variable X .

The expected value is our forecast. Because we are discussing random quantities, we cannot count on an individual forecast being realized (although we hope that, on average, forecasts will be accurate). It is important, as a result, to measure the risk we face. Variance and standard deviation measure the dispersion of outcomes around the expected value or forecast.

The **variance of a random variable** is the expected value (the probability-weighted average) of squared deviations from the random variable's expected value:

$$\sigma^2(X) = E[X - E(X)]^2. \quad (2)$$

The two notations for variance are $\sigma^2(X)$ and $\text{Var}(X)$.

Variance is a number greater than or equal to 0 because it is the sum of squared terms. If variance is 0, there is no dispersion or risk. The outcome is certain, and the quantity X is not random at all. Variance greater than 0 indicates dispersion of outcomes. Increasing variance indicates increasing dispersion, all else being equal.

The following equation summarizes the calculation of variance:

$$\begin{aligned} \sigma^2(X) &= P(X_1)[X_1 - E(X)]^2 + P(X_2)[X_2 - E(X)]^2 \\ &+ \dots + P(X_n)[X_n - E(X)]^2 = \sum_{i=1}^n P(X_i)[X_i - E(X)]^2, \end{aligned} \quad (3)$$

where X_i is one of n possible outcomes of the discrete random variable X .

Variance of X is a quantity in the squared units of X . For example, if the random variable is return in percent, variance of return is in units of percent squared. Standard deviation is easier to interpret than variance because it is in the same units as the random variable. **Standard deviation** is the square root of variance. If the random variable is return in percent, standard deviation of return is also in units of percent. In the following examples, when the variance of returns is stated as a percent or amount of money, to conserve space, we may suppress showing the unit squared.

The best way to become familiar with these concepts is to work examples.

EXAMPLE 1

BankCorp's Earnings per Share, Part 1

As part of your work as a banking industry analyst, you build models for forecasting earnings per share of the banks you cover. Today you are studying BankCorp. In Exhibit 1, you have recorded a probability distribution for BankCorp's EPS for the current fiscal year.

Exhibit 1: Probability Distribution for BankCorp's EPS

Probability	EPS (USD)
0.15	2.60
0.45	2.45
0.24	2.20

Probability	EPS (USD)
0.16	2.00
1.00	

1. What is the expected value of BankCorp's EPS for the current fiscal year?

Solution:

Following the definition of expected value, list each outcome, weight it by its probability, and sum the terms.

$$\begin{aligned} E(\text{EPS}) &= 0.15(\text{USD}2.60) + 0.45(\text{USD}2.45) + 0.24(\text{USD}2.20) + 0.16(\text{USD}2.00) \\ &= \text{USD}2.3405 \end{aligned}$$

The expected value of EPS is USD2.34.

2. Using the probability distribution of EPS from Exhibit 1, you want to measure the dispersion around your forecast. What are the variance and standard deviation of BankCorp's EPS for the current fiscal year?

Solution:

The order of calculation is always expected value, then variance, and then standard deviation. Expected value has already been calculated. Following the previous definition of variance, calculate the deviation of each outcome from the mean or expected value, square each deviation, weight (multiply) each squared deviation by its probability of occurrence, and then sum these terms.

$$\begin{aligned} \sigma^2(\text{EPS}) &= P(2.60)[2.60 - E(\text{EPS})]^2 + P(2.45)[2.45 - E(\text{EPS})]^2 \\ &\quad + P(2.20)[2.20 - E(\text{EPS})]^2 + P(2.00)[2.00 - E(\text{EPS})]^2 \\ &= 0.15(2.60 - 2.34)^2 + 0.45(2.45 - 2.34)^2 \\ &\quad + 0.24(2.20 - 2.34)^2 + 0.16(2.00 - 2.34)^2 \\ &= 0.01014 + 0.005445 + 0.004704 + 0.018496 = 0.038785 \end{aligned}$$

Standard deviation is the positive square root of 0.038785:

$$\sigma(\text{EPS}) = 0.038785^{1/2} = 0.196939, \text{ or approximately } 0.20.$$

3

PROBABILITY TREES AND CONDITIONAL EXPECTATIONS



formulate an investment problem as a probability tree and explain the use of conditional expectations in investment application

In investments, we make use of any relevant information available in making our forecasts. When we refine our expectations or forecasts, we are typically updating them based on new information or events; in these cases, we are using **conditional expected values**. The expected value of a random variable X given an event or scenario S is denoted $E(X | S)$. Suppose the random variable X can take on any one of n distinct

outcomes X_1, X_2, \dots, X_n (these outcomes form a set of mutually exclusive and exhaustive events). The expected value of X conditional on S is the first outcome, X_1 , times the probability of the first outcome given S , $P(X_1 | S)$, plus the second outcome, X_2 , times the probability of the second outcome given S , $P(X_2 | S)$, and so forth, as follows:

$$E(X | S) = P(X_1 | S)X_1 + P(X_2 | S)X_2 + \dots + P(X_n | S)X_n. \quad (4)$$

We will illustrate this equation shortly.

Parallel to the total probability rule for stating unconditional probabilities in terms of conditional probabilities, there is a principle for stating (unconditional) expected values in terms of conditional expected values. This principle is the **total probability rule for expected value**.

Total Probability Rule for Expected Value

The formula follows:

$$E(X) = E(X | S)P(S) + E(X | S^C)P(S^C), \quad (5)$$

where S^C is the “complement of S ,” which means event or scenario “ S ” does not occur.

$$E(X) = E(X | S_1)P(S_1) + E(X | S_2)P(S_2) + \dots + E(X | S_n)P(S_n), \quad (6)$$

where S_1, S_2, \dots, S_n are mutually exclusive and exhaustive scenarios or events.

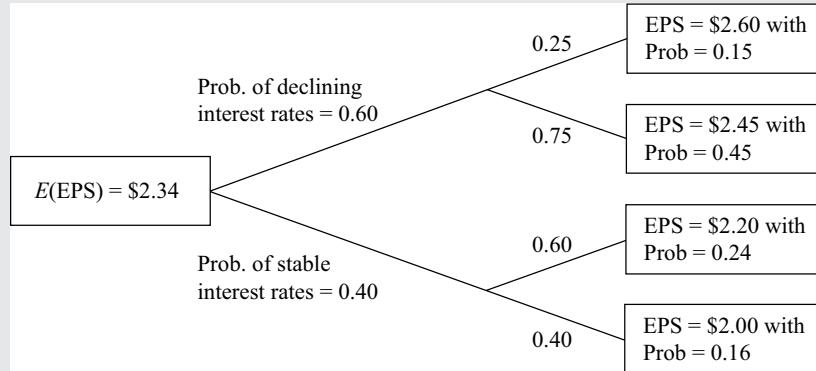
The general case, Equation 6, states that the expected value of X equals the expected value of X given Scenario 1, $E(X | S_1)$, times the probability of Scenario 1, $P(S_1)$, plus the expected value of X given Scenario 2, $E(X | S_2)$, times the probability of Scenario 2, $P(S_2)$, and so forth.

To use this principle, we formulate mutually exclusive and exhaustive scenarios that are useful for understanding the outcomes of the random variable. This approach was employed in developing the probability distribution of BankCorp’s EPS in Example 1, as we now discuss.

EXAMPLE 2

BankCorp’s Earnings per Share, Part 2

The earnings of BankCorp are interest rate sensitive, benefiting from a declining interest rate environment. Suppose there is a 0.60 probability that BankCorp will operate in a *declining interest rate environment* in the current fiscal year and a 0.40 probability that it will operate in a *stable interest rate environment* (assessing the chance of an increasing interest rate environment as negligible). If a *declining interest rate environment* occurs, the probability that EPS will be USD2.60 is estimated at 0.25, and the probability that EPS will be USD2.45 is estimated at 0.75. Note that 0.60, the probability of *declining interest rate environment*, times 0.25, the probability of USD2.60 EPS given a *declining interest rate environment*, equals 0.15, the (unconditional) probability of USD2.60 given in the table in Exhibit 1. The probabilities are consistent. Also, $0.60(0.75) = 0.45$, the probability of USD2.45 EPS given in Exhibit 1. The **probability tree diagram** in Exhibit 2 shows the rest of the analysis.

Exhibit 2: BankCorp's Forecasted EPS

A declining interest rate environment points us to the **node** of the tree that branches off into outcomes of USD2.60 and USD2.45. We can find expected EPS given a declining interest rate environment as follows, using Equation 6:

$$\begin{aligned}
 &E(\text{EPS} \mid \text{declining interest rate environment}) \\
 &= 0.25(\text{USD}2.60) + 0.75(\text{USD}2.45) \\
 &= \text{USD}2.4875
 \end{aligned}$$

If interest rates are stable,

$$\begin{aligned}
 &E(\text{EPS} \mid \text{stable interest rate environment}) = 0.60(\text{USD}2.20) + 0.40(\text{USD}2.00) \\
 &= \text{USD}2.12
 \end{aligned}$$

Once we have the new piece of information that interest rates are stable, for example, we revise our original expectation of EPS from USD2.34 downward to USD2.12. Now using the total probability rule for expected value,

$$\begin{aligned}
 &E(\text{EPS}) \\
 &= E(\text{EPS} \mid \text{declining interest rate environment})P(\text{declining interest rate environment}) \\
 &\quad + E(\text{EPS} \mid \text{stable interest rate environment})P(\text{stable interest rate environment})
 \end{aligned}$$

So, $E(\text{EPS}) = \text{USD}2.4875(0.60) + \text{USD}2.12(0.40) = \text{USD}2.3405$ or about USD2.34.

This amount is identical to the estimate of the expected value of EPS calculated directly from the probability distribution in Example 1. Just as our probabilities must be consistent, so too must our expected values, unconditional and conditional, be consistent; otherwise, our investment actions may create profit opportunities for other investors at our expense.

To review, we first developed the factors or scenarios that influence the outcome of the event of interest. After assigning probabilities to these scenarios, we formed expectations conditioned on the different scenarios. Then we worked backward to formulate an expected value as of today. In the problem just worked, EPS was the event of interest, and the interest rate environment was the factor influencing EPS.

We can also calculate the variance of EPS given each scenario:

$$\begin{aligned}
 &\sigma^2(\text{EPS} \mid \text{declining interest rate environment}) \\
 &= P(\text{USD}2.60 \mid \text{declining interest rate environment}) \\
 &\quad \times [\text{USD}2.60 - E(\text{EPS} \mid \text{declining interest rate environment})]^2 \\
 &\quad + P(\text{USD}2.45 \mid \text{declining interest rate environment})
 \end{aligned}$$

$$\begin{aligned} & \times [\text{USD}2.45 - E(\text{EPS} | \text{declining interest rate environment})]^2 \\ & = 0.25(\text{USD}2.60 - \text{USD}2.4875)^2 + 0.75(\text{USD}2.45 - \text{USD}2.4875)^2 = 0.004219 \end{aligned}$$

Similarly, $\sigma^2(\text{EPS} | \text{stable interest rate environment})$ is found to be equal to

$$= 0.60(\text{USD}2.20 - \text{USD}2.12)^2 + 0.40(\text{USD}2.00 - \text{USD}2.12)^2 = 0.0096$$

These are **conditional variances**, the variance of EPS given a *declining interest rate environment* and the variance of EPS given a *stable interest rate environment*. The relationship between unconditional variance and conditional variance is a relatively advanced topic. The main points are that (1) variance, like expected value, has a conditional counterpart to the unconditional concept; and (2) we can use conditional variance to assess risk given a particular scenario.

EXAMPLE 3

BankCorp's Earnings per Share, Part 3

Continuing with the BankCorp example, you focus now on BankCorp's cost structure. One model, a simple linear regression model, you are researching for BankCorp's operating costs is

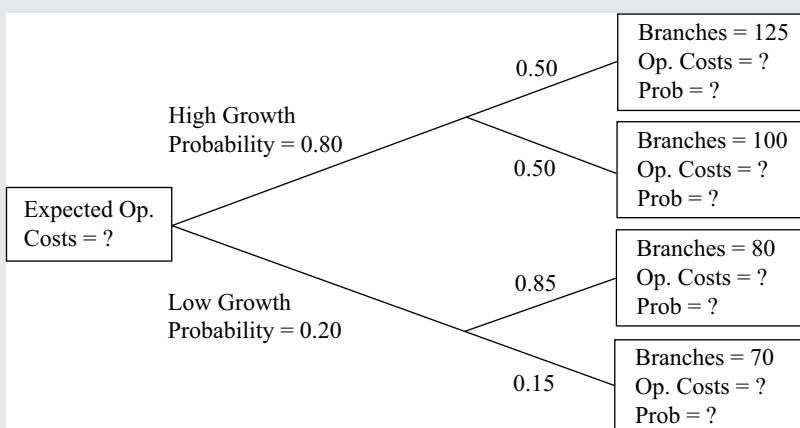
$$\hat{Y} = a + bX,$$

where \hat{Y} is a forecast of operating costs in millions of US dollars and X is the number of branch offices; and \hat{Y} represents the expected value of Y given X , or $E(Y | X)$. You interpret the intercept a as fixed costs and b as variable costs. You estimate the equation as follows:

$$\hat{Y} = 12.5 + 0.65X.$$

BankCorp currently has 66 branch offices, and the equation estimates operating costs as $12.5 + 0.65(66) = \text{USD}55.4$ million. You have two scenarios for growth, pictured in the tree diagram in Exhibit 3.

Exhibit 3: BankCorp's Forecasted Operating Costs



1. Compute the forecasted operating costs given the different levels of operating costs, using $\hat{Y} = 12.5 + 0.65X$. State the probability of each level of the number of branch offices. These are the answers to the questions in the terminal boxes of the tree diagram.

Solution:

Using $\hat{Y} = 12.5 + 0.65X$, from top to bottom, we have

Operating Costs	Probability
$\hat{Y} = 12.5 + 0.65(125) = \text{USD}93.75 \text{ million}$	$0.80(0.50) = 0.40$
$\hat{Y} = 12.5 + 0.65(100) = \text{USD}77.50 \text{ million}$	$0.80(0.50) = 0.40$
$\hat{Y} = 12.5 + 0.65(80) = \text{USD}64.50 \text{ million}$	$0.20(0.85) = 0.17$
$\hat{Y} = 12.5 + 0.65(70) = \text{USD}58.00 \text{ million}$	$0.20(0.15) = 0.03$
	Sum = 1.00

2. Compute the expected value of operating costs under the high growth scenario. Also calculate the expected value of operating costs under the low growth scenario.

Solution:

US dollar amounts are in millions.

$$\begin{aligned} E(\text{operating costs}|\text{high growth}) &= 0.50(\text{USD}93.75) + 0.50(\text{USD}77.50) \\ &= \text{USD}85.625 \end{aligned}$$

$$\begin{aligned} E(\text{operating costs}|\text{low growth}) &= 0.85(\text{USD}64.50) + 0.15(\text{USD}58.00) \\ &= \text{USD}63.525 \end{aligned}$$

3. Refer to the question in the initial box of the tree: What are BankCorp's expected operating costs?

Solution:

US dollar amounts are in millions.

$$\begin{aligned} E(\text{operating costs}) &= E(\text{operating costs}|\text{high growth})P(\text{high growth}) \\ &\quad + E(\text{operating costs}|\text{low growth})P(\text{low growth}) \\ &= 85.625(0.80) + 63.525(0.20) = 81.205 \end{aligned}$$

BankCorp's expected operating costs are USD81.205 million.

In this section, we have treated random variables, such as EPS, as standalone quantities. We have not explored how descriptors, such as expected value and variance of EPS, may be functions of other random variables. Portfolio return is one random variable that is clearly a function of other random variables, the random returns on the individual securities in the portfolio. To analyze a portfolio's expected return and variance of return, we must understand that these quantities are a function of characteristics of the individual securities' returns. Looking at the variance of portfolio return, we see that the way individual security returns move together or covary is key. We cover portfolio expected return, variance of return, and importantly, covariance and correlation in a separate learning module.

QUESTION SET



1. Suppose the prospects for recovering principal for a defaulted bond issue depend on which of two economic scenarios prevails. Scenario 1 has probability 0.75 and will result in recovery of USD0.90 per USD1 principal value with probability 0.45, or in recovery of USD0.80 per USD1 principal

value with probability 0.55. Scenario 2 has probability 0.25 and will result in recovery of USD0.50 per USD1 principal value with probability 0.85, or in recovery of USD0.40 per USD1 principal value with probability 0.15.

Using the data for Scenario 1 and Scenario 2, calculate the following:

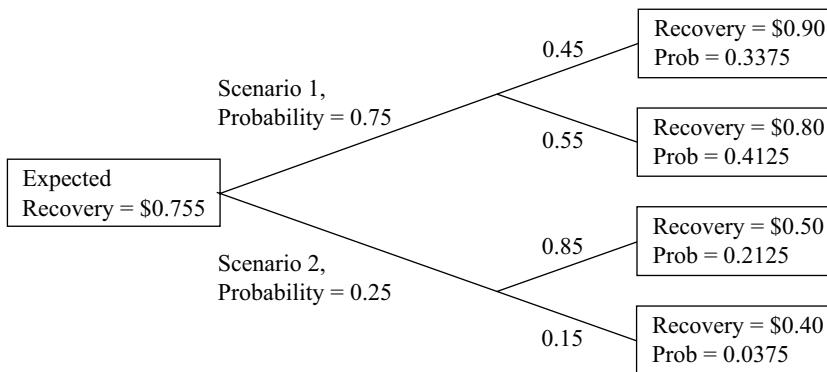
- Compute the expected recovery, given the first scenario.
- Compute the expected recovery, given the second scenario.
- Compute the expected recovery.
- Graph the information in a probability tree diagram.
- Compute the probability of each of the four possible recovery amounts: USD0.90, USD0.80, USD0.50, and USD0.40.

Solution:

- A. *Outcomes associated with Scenario 1:* With a 0.45 probability of a USD0.90 recovery per USD1 principal value, given Scenario 1, and with the probability of Scenario 1 equal to 0.75, the probability of recovering USD0.90 is $0.45(0.75) = 0.3375$. By a similar calculation, the probability of recovering USD0.80 is $0.55(0.75) = 0.4125$.

Outcomes associated with Scenario 2: With a 0.85 probability of a USD0.50 recovery per USD1 principal value, given Scenario 2, and with the probability of Scenario 2 equal to 0.25, the probability of recovering USD0.50 is $0.85(0.25) = 0.2125$. By a similar calculation, the probability of recovering USD0.40 is $0.15(0.25) = 0.0375$.

- $E(\text{recovery} | \text{Scenario 1}) = 0.45(\text{USD0.90}) + 0.55(\text{USD0.80}) = \text{USD0.845}$
- $E(\text{recovery} | \text{Scenario 2}) = 0.85(\text{USD0.50}) + 0.15(\text{USD0.40}) = \text{USD0.485}$
- $E(\text{recovery}) = 0.75(\text{USD0.845}) + 0.25(\text{USD0.485}) = \text{USD0.755}$



BAYES' FORMULA AND UPDATING PROBABILITY ESTIMATES

4



calculate and interpret an updated probability in an investment setting using Bayes' formula

A topic that is often useful in solving investment problems is Bayes' formula: what probability theory has to say about learning from experience.

Bayes' Formula

When we make decisions involving investments, we often start with viewpoints based on our experience and knowledge. These viewpoints may be changed or confirmed by new knowledge and observations. **Bayes' formula** is a rational method for adjusting our viewpoints as we confront new information. Bayes' formula and related concepts are used in many business and investment decision-making contexts.

Bayes' formula makes use of the total probability rule:

$$P(A) = \sum_n P(A \cap B_n). \quad (7)$$

To review, that rule expresses the probability of an event as a weighted average of the probabilities of the event, given a set of scenarios. Bayes' formula works in reverse; more precisely, it reverses the “given that” information. Bayes' formula uses the occurrence of the event to infer the probability of the scenario generating it. For that reason, Bayes' formula is sometimes called an inverse probability. In many applications, including those illustrating its use in this section, an individual is updating his/her beliefs concerning the causes that may have produced a new observation.

Bayes' Formula. Given a set of prior probabilities for an event of interest, if you receive new information, the rule for updating your probability of the event is as follows:

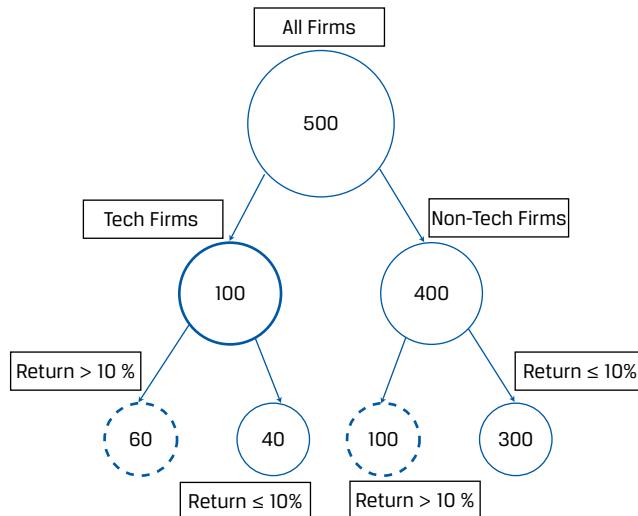
Updated probability of event given the new information

$$= \frac{\text{Probability of the new information given event}}{\text{Unconditional probability of the new information}} \times \text{Prior probability of event.}$$

In probability notation, this formula can be written concisely as follows:

$$P(\text{Event} | \text{Information}) = \frac{P(\text{Information} | \text{Event})}{P(\text{Information})} P(\text{Event}). \quad (8)$$

Consider the following example using frequencies—which may be more straightforward initially than probabilities—for illustrating and understanding Bayes' formula. Assume a hypothetical large-cap stock index has 500 member firms, of which 100 are technology firms, and 60 of these had returns of >10 percent, and 40 had returns of ≤10 percent. Of the 400 non-technology firms in the index, 100 had returns of >10 percent, and 300 had returns of ≤10 percent. The tree map in Exhibit 4 is useful for visualizing this example, which is summarized in the table in Exhibit 5.

Exhibit 4: Tree Map for Visualizing Bayes' Formula Using Frequencies

$$P(Tech | R > 10\%) = 60/(60+100)$$

Exhibit 5: Summary of Returns for Tech and Non-Tech Firms in Hypothetical Large-Cap Equity Index

Rate of Return (R)	Type of Firm in Stock Index		
	Non-Tech	Tech	Total
R > 10%	100	60	160
R ≤ 10%	300	40	340
Total	400	100	500

What is the probability a firm is a tech firm given that it has a return of >10 percent or $P(tech | R > 10\%)$? Looking at the frequencies in the tree map and in the table, we can see many empirical probabilities, such as the following:

- $P(tech) = 100 / 500 = 0.20$,
- $P(non-tech) = 400 / 500 = 0.80$,
- $P(R > 10\% | tech) = 60 / 100 = 0.60$,
- $P(R > 10\% | non-tech) = 100 / 400 = 0.25$,
- $P(R > 10\%) = 160 / 500 = 0.32$, and, finally,
- $P(tech | R > 10\%) = 60 / 160 = 0.375$.

This probability is the answer to our initial question.

Without looking at frequencies, let us use Bayes' formula to find the probability that a firm has a return of >10 percent and then the probability that a firm with a return of >10 percent is a tech firm, $P(tech | R > 10\%)$. First,

$$\begin{aligned} P(R > 10\%) \\ = P(R > 10\% | tech) \times P(tech) + P(R > 10\% | non-tech) \times P(non-tech) \end{aligned}$$

$$= 0.60 \times 0.20 + 0.25 \times 0.80 = 0.32.$$

Now we can implement the Bayes' formula answer to our question:

$$P(\text{tech} | R > 10\%) = \frac{P(R > 10\% | \text{tech}) \times P(\text{tech})}{P(R > 10\%)} = \frac{0.60 \times 0.20}{0.32} = 0.375.$$

The probability that a firm with a return of >10 percent is a tech firm is 0.375, which is impressive because the probability that a firm is a tech firm (from the whole sample) is only 0.20. In sum, it can be readily seen from the tree map and the underlying frequency data (Exhibit 4 and 5, respectively) or from the probabilities in Bayes' formula that 160 firms have $R > 10$ percent, and 60 of them are tech firms, so

$$P(\text{tech} | R > 10\%) = 60/160 = 0.375.$$

Users of Bayesian statistics do not consider probabilities (or likelihoods) to be known with certainty but believe that these should be subject to modification whenever new information becomes available. Our beliefs or probabilities are continually updated as new information arrives over time.

To further illustrate Bayes' formula, we work through an investment example that can be adapted to any actual problem. Suppose you are an investor in the stock of DriveMed, Inc. Positive earnings surprises relative to consensus EPS estimates often result in positive stock returns, and negative surprises often have the opposite effect. DriveMed is preparing to release last quarter's EPS result, and you are interested in which of these three events happened: *last quarter's EPS exceeded the consensus EPS estimate*, *last quarter's EPS exactly met the consensus EPS estimate*, or *last quarter's EPS fell short of the consensus EPS estimate*. This list of the alternatives is mutually exclusive and exhaustive.

On the basis of your own research, you write down the following **prior probabilities** (or priors, for short) concerning these three events:

- $P(\text{EPS exceeded consensus}) = 0.45$
- $P(\text{EPS met consensus}) = 0.30$
- $P(\text{EPS fell short of consensus}) = 0.25$

These probabilities are "prior" in the sense that they reflect only what you know now, before the arrival of any new information.

The next day, DriveMed announces that it is expanding factory capacity in Singapore and Ireland to meet increased sales demand. You assess this new information. The decision to expand capacity relates not only to current demand but probably also to the prior quarter's sales demand. You know that sales demand is positively related to EPS. So now it appears more likely that last quarter's EPS will exceed the consensus.

The question you have is, "In light of the new information, what is the updated probability that the prior quarter's EPS exceeded the consensus estimate?"

Bayes' formula provides a rational method for accomplishing this updating. We can abbreviate the new information as *DriveMed expands*. The first step in applying Bayes' formula is to calculate the probability of the new information (here: *DriveMed expands*), given a list of events or scenarios that may have generated it. The list of events should cover all possibilities, as it does here. Formulating these conditional probabilities is the key step in the updating process. Suppose your view, based on research of DriveMed and its industry, is

$$P(\text{DriveMed expands} | \text{EPS exceeded consensus}) = 0.75$$

$$P(\text{DriveMed expands} | \text{EPS met consensus}) = 0.20$$

$$P(\text{DriveMed expands} | \text{EPS fell short of consensus}) = 0.05$$

Conditional probabilities of an observation (here: *DriveMed expands*) are sometimes referred to as **likelihoods**. Again, likelihoods are required for updating the probability.

Next, you combine these conditional probabilities or likelihoods with your prior probabilities to get the unconditional probability for DriveMed expanding, $P(\text{DriveMed expands})$, as follows:

$$\begin{aligned}
 & P(\text{DriveMed expands}) \\
 &= P(\text{DriveMed expands} | \text{EPS exceeded consensus}) \times P(\text{EPS exceeded consensus}) \\
 &+ P(\text{DriveMed expands} | \text{EPS met consensus}) \times P(\text{EPS met consensus}) \\
 &+ P(\text{DriveMed expands} | \text{EPS fell short of consensus}) \times P(\text{EPS fell short of consensus}) \\
 &= 0.75(0.45) + 0.20(0.30) + 0.05(0.25) = 0.41, \text{ or } 41\%.
 \end{aligned}$$

This is the total probability rule in action. Now you can answer your question by applying Bayes' formula, Equation 8:

$$\begin{aligned}
 & P(\text{EPS exceeded consensus} | \text{DriveMed expands}) \\
 &= \frac{P(\text{DriveMed expands} | \text{EPS exceeded consensus})}{P(\text{DriveMed expands})} P(\text{EPS exceeded consensus}) \\
 &= (0.75/0.41)(0.45) = 1.829268(0.45) \\
 &= 0.823171
 \end{aligned}$$

Before DriveMed's announcement, you thought the probability that DriveMed would beat consensus expectations was 45 percent. On the basis of your interpretation of the announcement, you update that probability to 82.3 percent. This updated probability is called your **posterior probability** because it reflects or comes after the new information.

The Bayes' calculation takes the prior probability, which was 45 percent, and multiplies it by a ratio—the first term on the right-hand side of the equal sign. The denominator of the ratio is the probability that DriveMed expands, as you view it without considering (conditioning on) anything else. Therefore, this probability is unconditional. The numerator is the probability that DriveMed expands, if last quarter's EPS actually exceeded the consensus estimate. This last probability is larger than unconditional probability in the denominator, so the ratio (1.83 roughly) is greater than 1. As a result, your updated or posterior probability is larger than your prior probability. Thus, the ratio reflects the impact of the new information on your prior beliefs.

EXAMPLE 4

Inferring Whether DriveMed's EPS Met Consensus EPS

You are still an investor in DriveMed stock. To review the givens, your prior probabilities are $P(\text{EPS exceeded consensus}) = 0.45$, $P(\text{EPS met consensus}) = 0.30$, and $P(\text{EPS fell short of consensus}) = 0.25$. You also have the following conditional probabilities:

$$P(\text{DriveMed expands} | \text{EPS exceeded consensus}) = 0.75$$

$$P(\text{DriveMed expands} | \text{EPS met consensus}) = 0.20$$

$$P(\text{DriveMed expands} | \text{EPS fell short of consensus}) = 0.05$$

- What is your estimate of the probability $P(\text{EPS exceeded consensus} \mid \text{DriveMed expands})$?

Recall that you updated your probability that last quarter's EPS exceeded the consensus estimate from 45 percent to 82.3 percent after DriveMed announced it would expand. Now you want to update your other priors.

Update your prior probability that DriveMed's EPS met consensus.

Solution:

The probability is $P(\text{EPS met consensus} \mid \text{DriveMed expands}) =$

$$\frac{P(\text{DriveMed expands} \mid \text{EPS met consensus})}{P(\text{DriveMed expands})} P(\text{EPS met consensus})$$

The probability $P(\text{DriveMed expands})$ is found by taking each of the three conditional probabilities in the statement of the problem, such as $P(\text{DriveMed expands} \mid \text{EPS exceeded consensus})$; multiplying each one by the prior probability of the conditioning event, such as $P(\text{EPS exceeded consensus})$; and then adding the three products. The calculation is unchanged from the problem in the text above: $P(\text{DriveMed expands}) = 0.75(0.45) + 0.20(0.30) + 0.05(0.25) = 0.41$, or 41 percent. The other probabilities needed, $P(\text{DriveMed expands} \mid \text{EPS met consensus}) = 0.20$ and $P(\text{EPS met consensus}) = 0.30$, are givens. So

$$P(\text{EPS met consensus} \mid \text{DriveMed expands})$$

$$= [P(\text{DriveMed expands} \mid \text{EPS met consensus}) / P(\text{DriveMed expands})] P(\text{EPS met consensus})$$

$$= (0.20/0.41)(0.30) = 0.487805(0.30) = 0.146341$$

After taking account of the announcement on expansion, your updated probability that last quarter's EPS for DriveMed just met consensus is 14.6 percent compared with your prior probability of 30 percent.

- Update your prior probability that DriveMed's EPS fell short of consensus.

Solution:

$P(\text{DriveMed expands})$ was already calculated as 41 percent. Recall that $P(\text{DriveMed expands} \mid \text{EPS fell short of consensus}) = 0.05$ and $P(\text{EPS fell short of consensus}) = 0.25$ are givens.

$$P(\text{EPS fell short of consensus} \mid \text{DriveMed expands})$$

$$= [P(\text{DriveMed expands} \mid \text{EPS fell short of consensus}) /$$

$$P(\text{DriveMed expands})] P(\text{EPS fell short of consensus})$$

$$= (0.05/0.41)(0.25) = 0.121951(0.25) = 0.030488$$

As a result of the announcement, you have revised your probability that DriveMed's EPS fell short of consensus from 25 percent (your prior probability) to 3 percent.

3. Show that the three updated probabilities sum to 1. (Carry each probability to four decimal places.)

Solution:

The sum of the three updated probabilities is

$$\begin{aligned} P(\text{EPS exceeded consensus} | \text{DriveMed expands}) + P(\text{EPS met consensus} | \text{DriveMed expands}) + P(\text{EPS fell short of consensus} | \text{DriveMed expands}) \\ = 0.8232 + 0.1463 + 0.0305 = 1.000 \end{aligned}$$

The three events (*EPS exceeded consensus*, *EPS met consensus*, and *EPS fell short of consensus*) are mutually exclusive and exhaustive: One of these events or statements must be true, so the conditional probabilities must sum to 1. Whether we are talking about conditional or unconditional probabilities, whenever we have a complete set of distinct possible events or outcomes, the probabilities must sum to 1. This calculation serves to check your work.

4. Suppose, because of lack of prior beliefs about whether DriveMed would meet consensus, you updated on the basis of prior probabilities that all three possibilities were equally likely: $P(\text{EPS exceeded consensus}) = P(\text{EPS met consensus}) = P(\text{EPS fell short of consensus}) = 1/3$.

Solution:

Using the probabilities given in the question,

$$\begin{aligned} P(\text{DriveMed expands}) \\ = P(\text{DriveMed expands} | \text{EPS exceeded consensus})P(\text{EPS exceeded consensus}) + P(\text{DriveMed expands} | \text{EPS met consensus})P(\text{EPS met consensus}) + P(\text{DriveMed expands} | \text{EPS fell short of consensus})P(\text{EPS fell short of consensus}) \\ = 0.75(1/3) + 0.20(1/3) + 0.05(1/3) = 1/3 \end{aligned}$$

Not surprisingly, the probability of DriveMed expanding is one-third (1/3) because the decision maker has no prior beliefs or views regarding how well EPS performed relative to the consensus estimate.

Now we can use Bayes' formula to find $P(\text{EPS exceeded consensus} | \text{DriveMed expands}) = [P(\text{DriveMed expands} | \text{EPS exceeded consensus}) / P(\text{DriveMed expands})] P(\text{EPS exceeded consensus}) = [(0.75/(1/3)] (1/3) = 0.75$, or 75 percent. This probability is identical to your estimate of $P(\text{DriveMed expands} | \text{EPS exceeded consensus})$.

When the prior probabilities are equal, the probability of information given an event equals the probability of the event given the information. When a decision maker has equal prior probabilities (called **diffuse priors**), the probability of an event is determined by the information.

QUESTION SET

The following example shows how Bayes' formula is used in credit granting in cases in which the probability of payment given credit information is higher than the probability of payment without the information.

1. Jake Bronson is predicting the probability that consumer finance applicants granted credit will repay in a timely manner (i.e., their accounts will not

become “past due”). Using Bayes’ formula, he has structured the problem as follows:

$$P(\text{Event} | \text{Information}) = \frac{P(\text{Information} | \text{Event})}{P(\text{Information})} P(\text{Event}),$$

where the event (A) is “timely repayment” and the information (B) is having a “good credit report.”

Bronson estimates that the unconditional probability of receiving timely payment, $P(A)$, is 0.90 and that the unconditional probability of having a good credit report, $P(B)$, is 0.80. The probability of having a good credit report given that borrowers paid on time, $P(B | A)$, is 0.85.

What is the probability that applicants with good credit reports will repay in a timely manner?

- A.** 0.720
- B.** 0.944
- C.** 0.956

Solution:

The correct answer is C. The probability of timely repayment given a good credit report, $P(A | B)$, is

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A) = \frac{0.85}{0.80} \times 0.90 = 0.956$$

2. You have developed a set of criteria for evaluating distressed credits. Companies that do not receive a passing score are classed as likely to go bankrupt within 12 months. You gathered the following information when validating the criteria:

- Forty percent of the companies to which the test is administered will go bankrupt within 12 months: $P(\text{non-survivor}) = 0.40$.
- Fifty-five percent of the companies to which the test is administered pass it: $P(\text{pass test}) = 0.55$.
- The probability that a company will pass the test given that it will subsequently survive 12 months, is 0.85: $P(\text{pass test} | \text{survivor}) = 0.85$.

Using the information validating your criteria, calculate the following:

- A.** What is $P(\text{pass test} | \text{non-survivor})$?
- B.** Using Bayes’ formula, calculate the probability that a company is a survivor, given that it passes the test; that is, calculate $P(\text{survivor} | \text{pass test})$.
- C.** What is the probability that a company is a *non-survivor*, given that it fails the test?
- D.** Is the test effective?

Solution:

A. We can set up the equation using the total probability rule:

$$P(\text{pass test}) = P(\text{pass test} | \text{survivor}) P(\text{survivor})$$

$$+ P(\text{pass test}|\text{non-survivor})P(\text{non-survivor})$$

We know that $P(\text{survivor}) = 1 - P(\text{non-survivor}) = 1 - 0.40 = 0.60$. Therefore, $P(\text{pass test}) = 0.55 = 0.85(0.60) + P(\text{pass test} | \text{non-survivor})(0.40)$.

Thus, $P(\text{pass test} | \text{non-survivor}) = [0.55 - 0.85(0.60)]/0.40 = 0.10$.

B. We can calculate the probability that a company is a survivor as follows:

$$\begin{aligned} P(\text{survivor} | \text{pass test}) &= [P(\text{pass test} | \text{survivor})/P(\text{pass test})]P(\text{survivor}) \\ &= (0.85/0.55)0.60 = 0.927273 \end{aligned}$$

The information that a company passes the test causes you to update your probability that it is a survivor from 0.60 to approximately 0.927.

C. According to Bayes' formula, $P(\text{non-survivor} | \text{fail test}) = [P(\text{fail test} | \text{non-survivor})/P(\text{fail test})]P(\text{non-survivor}) = [P(\text{fail test} | \text{non-survivor})/0.45]0.40$.

We can set up the following equation to obtain $P(\text{fail test} | \text{non-survivor})$:

$$\begin{aligned} P(\text{fail test}) &= P(\text{fail test} | \text{non-survivor})P(\text{non-survivor}) \\ &\quad + P(\text{fail test} | \text{survivor})P(\text{survivor}) \\ 0.45 &= P(\text{fail test} | \text{non-survivor})0.40 + 0.15(0.60) \end{aligned}$$

where $P(\text{fail test} | \text{survivor}) = 1 - P(\text{pass test} | \text{survivor}) = 1 - 0.85 = 0.15$. So, $P(\text{fail test} | \text{non-survivor}) = [0.45 - 0.15(0.60)]/0.40 = 0.90$.

Using this result with the previous formula, we find $P(\text{non-survivor} | \text{fail test}) = [0.90/0.45]0.40 = 0.80$. Seeing that a company fails the test causes us to update the probability that it is a non-survivor from 0.40 to 0.80.

D. A company passing the test greatly increases our confidence that it is a survivor. A company failing the test doubles the probability that it is a non-survivor. Therefore, the test appears to be useful.

3. An analyst estimates that 20 percent of high-risk bonds will fail (go bankrupt). If she applies a bankruptcy prediction model, she finds that 70 percent of the bonds will receive a “good” rating, implying that they are less likely to fail. Of the bonds that failed, only 50 percent had a “good” rating. Using Bayes’ formula, what is the predicted probability of failure given a “good” rating? (Hint, let $P(A)$ be the probability of failure, $P(B)$ be the probability of a “good” rating, $P(B | A)$ be the likelihood of a “good” rating given failure, and $P(A | B)$ be the likelihood of failure given a “good” rating.)

- A. 5.7 percent
- B. 14.3 percent
- C. 28.6 percent

Solution:

B is correct. With Bayes’ formula, the probability of failure given a “good” rating is

$$P(A|B) = \frac{P(B|A)}{P(B)}P(A)$$

where

$$P(A) = 0.20 = \text{probability of failure}$$

$$P(B) = 0.70 = \text{probability of a “good” rating}$$

$$P(B | A) = 0.50 = \text{probability of a “good” rating given failure}$$

With these estimates, the probability of failure given a “good” rating is

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A) = \frac{0.50}{0.70} \times 0.20 = 0.143$$

If the analyst uses the bankruptcy prediction model as a guide, the probability of failure declines from 20 percent to 14.3 percent.

4. In a typical year, 5 percent of all CEOs are fired for “performance” reasons. Assume that CEO performance is judged according to stock performance and that 50 percent of stocks have above-average returns or “good” performance. Empirically, 30 percent of all CEOs who were fired had “good” performance. Using Bayes’ formula, what is the probability that a CEO will be fired given “good” performance? (Hint, let $P(A)$ be the probability of a CEO being fired, $P(B)$ be the probability of a “good” performance rating, $P(B | A)$ be the likelihood of a “good” performance rating given that the CEO was fired, and $P(A | B)$ be the likelihood of the CEO being fired given a “good” performance rating.)

- A. 1.5 percent
- B. 2.5 percent
- C. 3.0 percent

Solution:

C is correct. With Bayes’ formula, the probability of the CEO being fired given a “good” rating is

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A)$$

where

$P(A) = 0.05$ = probability of the CEO being fired

$P(B) = 0.50$ = probability of a “good” rating

$P(B | A) = 0.30$ = probability of a “good” rating given that the CEO is fired

With these estimates, the probability of the CEO being fired given a “good” rating is

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A) = \frac{0.30}{0.50} \times 0.05 = 0.03$$

Although 5 percent of all CEOs are fired, the probability of being fired given a “good” performance rating is 3 percent.

PRACTICE PROBLEMS

1. An analyst developed two scenarios with respect to the recovery of USD100,000 principal from defaulted loans:

Scenario	Probability of Scenario (%)	Amount Recovered (USD)	Probability of Amount (%)
1	40	50,000	60
		30,000	40
2	60	80,000	90
		60,000	10

The amount of the expected recovery is *closest* to which of the following?

- A. USD36,400.
 - B. USD55,000.
 - C. USD63,600.
2. The probability distribution for a company's sales is:

Probability	Sales (USD, millions)
0.05	70
0.70	40
0.25	25

The standard deviation of sales is *closest* to which of the following?

- A. USD9.81 million.
- B. USD12.20 million.
- C. USD32.40 million.

SOLUTIONS

1. C is correct. If Scenario 1 occurs, the expected recovery is 60% (USD50,000) + 40% (USD30,000) = USD42,000, and if Scenario 2 occurs, the expected recovery is 90% (USD80,000) + 10% (USD60,000) = USD78,000. Weighting by the probability of each scenario, the expected recovery is 40% (USD42,000) + 60% (USD78,000) = USD63,600. Alternatively, first calculating the probability of each amount occurring, the expected recovery is $(40\%)(60\%)(\text{USD}50,000) + (40\%)(40\%)(\text{USD}30,000) + (60\%)(90\%)(\text{USD}80,000) + (60\%)(10\%)(\text{USD}60,000) = \text{USD}63,600$.
2. A is correct. The analyst must first calculate expected sales as $0.05 \times \text{USD}70 + 0.70 \times \text{USD}40 + 0.25 \times \text{USD}25 = \text{USD}3.50 \text{ million} + \text{USD}28.00 \text{ million} + \text{USD}6.25 \text{ million} = \text{USD}37.75 \text{ million}$.

After calculating expected sales, we can calculate the variance of sales:

$$\begin{aligned}\sigma^2(\text{Sales}) &= P(\text{USD}70)[\text{USD}70 - E(\text{Sales})]^2 + P(\text{USD}40)[\text{USD}40 - E(\text{Sales})]^2 + P(\text{USD}25) \\ &\quad [\text{USD}25 - E(\text{Sales})]^2 \\ &= 0.05(\text{USD}70 - 37.75)^2 + 0.70(\text{USD}40 - 37.75)^2 + 0.25(\text{USD}25 - 37.75)^2 \\ &= \text{USD}52.00 \text{ million} + \text{USD}3.54 \text{ million} + \text{USD}40.64 \text{ million} \\ &= \text{USD}96.18 \text{ million}.\end{aligned}$$

The standard deviation of sales is thus $\sigma = (\text{USD}96.18)^{1/2} = \text{USD}9.81 \text{ million}$.

LEARNING MODULE

5

Portfolio Mathematics

by Richard A. DeFusco, PhD, CFA, Dennis W. McLeavey, DBA, CFA, Jerald E. Pinto, PhD, CFA, and David E. Runkle, PhD, CFA.

Richard A. DeFusco, PhD, CFA, is at the University of Nebraska-Lincoln (USA). Dennis W. McLeavey, DBA, CFA, is at the University of Rhode Island (USA). Jerald E. Pinto, PhD, CFA, is at CFA Institute (USA). David E. Runkle, PhD, CFA, is at Jacobs Levy Equity Management (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	calculate and interpret the expected value, variance, standard deviation, covariances, and correlations of portfolio returns
<input type="checkbox"/>	calculate and interpret the covariance and correlation of portfolio returns using a joint probability function for returns
<input type="checkbox"/>	define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion

INTRODUCTION

1

Modern portfolio theory makes frequent use of the idea that investment opportunities can be evaluated using expected return as a measure of reward and variance of return as a measure of risk. In Lesson 1, we will develop an understanding of portfolio return and risk metrics. The forecast expected return and variance of return are functions of the returns on the individual portfolio holdings. To begin, the expected return on a portfolio is a weighted average of the expected returns on the securities in the portfolio. When we have estimated the expected returns on the individual securities, we immediately have portfolio expected return. Lesson 2 focuses on forecasting certain portfolio metrics, such as correlations and covariances by looking at the risk and return on the individual components of a portfolio. Lesson 3 introduces various portfolio risk metrics widely used in portfolio management.

LEARNING MODULE OVERVIEW



- A portfolio's variance measures its expected investment risk and is defined as $\sigma^2(R_p) = E\{[R_p - E(R_p)]^2\}$. A portfolio's expected return ($E(R_p)$) is a weighted average of the expected returns (R_1 to R_n) on the component securities using their respective proportions of the portfolio in currency units as weights (w_1 to w_n):

$$\begin{aligned}E(R_p) &= E(w_1 R_1 + w_2 R_2 + \dots + w_n R_n) \\&= w_1 E(R_1) + w_2 E(R_2) + \dots + w_n E(R_n)\end{aligned}$$

- Portfolio variance is affected by both the risk of the individual component assets and their combined risks together as measured by their covariance, which is defined as

$$\sigma^2(R_p) = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \text{Cov}(R_i, R_j).$$

- Covariance of returns can be negative (an average negative relationship between returns), zero if returns on the assets are unrelated, or positive (an average positive relationship between returns). Correlation, like covariance, measures linear association and ranges between -1 (strongly inverse) to $+1$ (strongly direct), with 0 indicating no relationship.
- The covariance of portfolio returns can be estimated using a joint probability function of random variables. Defined on variables X and Y , as $P(X,Y)$, which gives the probability of joint occurrences of their values. For example, $P(X=3, Y=2)$, is the probability X equals 3 and Y equals 2.
- A formula for computing the covariance between random variables R_A and R_B , such as the different assets of a portfolio, is

$$\text{Cov}(R_A, R_B) = \sum_i \sum_j P(R_{A,i}, R_{B,j}) (R_{A,i} - E R_A)(R_{B,j} - E R_B).$$

The value is derived by summing all possible deviation cross-products weighted by the appropriate joint probability.

- The joint probability function simplifies for independent variables, defined for two random variables X and Y if and only if $P(X,Y) = P(X)P(Y)$. The expected value of the product of both independent and uncorrelated random variables is the product of their expected values.
- An application of normal distribution theory to practical investment problems involves safety-first rules. These focus on reducing the shortfall risk, defined as portfolio value (or portfolio return) falling below some minimum acceptable level over some time horizon,
- The safety-first ratio is defined as $(SFRatio = [E(R_p) - R_L]/\sigma_p)$, where $E(R_p)$ is expected portfolio return, R_L is a predetermined minimum threshold level for a variable of interest like portfolio return, and σ_p is portfolio standard deviation. When R_L is the risk-free rate, the safety-first ratio is equivalent to the Sharpe ratio.
- Roy's safety-first criterion states that the optimal portfolio minimizes the probability that portfolio return, R_p , will fall below R_L . For a portfolio with a given safety-first ratio (SFRatio), the probability that its return will be less than R_L is $\text{Normal}(-SFRatio)$, and the safety-first

optimal portfolio has the lowest such probability. The criterion is implemented by first calculating each potential portfolio's SFRatio and then choosing the portfolio with the highest SFRatio.

PORTRFOLIO EXPECTED RETURN AND VARIANCE OF RETURN

2

- calculate and interpret the expected value, variance, standard deviation, covariances, and correlations of portfolio returns

The **expected return on the portfolio** ($E(R_p)$) is a weighted average of the expected returns (R_1 to R_n) on the component securities using their respective proportions of the portfolio in currency units as weights (w_1 to w_n):

$$\begin{aligned} E(R_p) &= E(w_1 R_1 + w_2 R_2 + \dots + w_n R_n) \\ &= w_1 E(R_1) + w_2 E(R_2) + \dots + w_n E(R_n) \end{aligned} \quad (1)$$

Suppose we have estimated expected returns on assets in the three-asset portfolio shown in Exhibit 1.

Exhibit 1: Weights and Expected Returns of Sample Portfolio

Asset Class	Weight	Expected Return (%)
S&P 500	0.50	13
US long-term corporate bonds	0.25	6
MSCI EAFE	0.25	15

We calculate the expected return on the portfolio as 11.75 percent:

$$\begin{aligned} E(R_p) &= w_1 E(R_1) + w_2 E(R_2) + w_3 E(R_3) \\ &= 0.50(13\%) + 0.25(6\%) + 0.25(15\%) = 11.75\% \end{aligned}$$

Here we are interested in portfolio variance of return as a measure of investment risk. Accordingly, portfolio variance is as follows:

$$\sigma^2(R_p) = E\{[R_p - E(R_p)]^2\}. \quad (2)$$

This is expected variance or variance in a forward-looking sense. To implement this definition of portfolio variance, we use information about the individual assets in the portfolio, but we also need the concept of covariance. To avoid notational clutter, we write ER_p for $E(R_p)$.

Covariance

Given two random variables R_i and R_j , the **covariance** between R_i and R_j is as follows:

$$\text{Cov}(R_i, R_j) = E[(R_i - ER_i)(R_j - ER_j)]. \quad (3)$$

Alternative notations are $\sigma(R_i, R_j)$ and σ_{ij} . Equation 3 states that the covariance between two random variables is the probability-weighted average of the cross-products of each random variable's deviation from its own expected value. The previous measure is the population covariance and is forward-looking. The sample covariance between two random variables R_i and R_j , based on a sample of past data of size n is as follows:

$$\text{Cov}(R_i, R_j) = \sum_{n=1}^n (R_{i,t} - E\bar{R}_i)(R_{j,t} - E\bar{R}_j) / (n - 1). \quad (4)$$

Start with the definition of variance for a three-asset portfolio and see how it decomposes into three variance terms and six covariance terms. Dispensing with the derivation, the result is Equation 5:

$$\begin{aligned} \sigma^2(R_p) &= E[(R_p - ER_p)^2] \\ &= E\{[w_1 R_1 + w_2 R_2 + w_3 R_3 - (w_1 R_1 + w_2 R_2 + w_3 R_3)]^2\} \\ &= E\{[w_1 R_1 + w_2 R_2 + w_3 R_3 - w_1 ER_1 - w_2 ER_2 - w_3 ER_3]^2\}. \\ &= w_1^2 \sigma^2(R_1) + w_1 w_2 \text{Cov}(R_1, R_2) + w_1 w_3 \text{Cov}(R_1, R_3) \\ &\quad + w_1 w_2 \text{Cov}(R_1, R_2) + w_2^2 \sigma^2(R_2) + w_2 w_3 \text{Cov}(R_2, R_3) \\ &\quad + w_1 w_3 \text{Cov}(R_1, R_3) + w_2 w_3 \text{Cov}(R_2, R_3) + w_3^2 \sigma^2(R_3). \end{aligned} \quad (5)$$

Noting that the order of variables in covariance does not matter, for example, $\text{Cov}(R_2, R_1) = \text{Cov}(R_1, R_2)$, and that diagonal variance terms $\sigma^2(R_1)$, $\sigma^2(R_2)$, and $\sigma^2(R_3)$ can be expressed as $\text{Cov}(R_1, R_1)$, $\text{Cov}(R_2, R_2)$, and $\text{Cov}(R_3, R_3)$, respectively, the most compact way to state Equation 5 is

$$\sigma^2(R_p) = \sum_{i=1}^3 \sum_{j=1}^3 w_i w_j \text{Cov}(R_i, R_j).$$

Moreover, this expression generalizes for a portfolio of any size n to

$$\sigma^2(R_p) = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \text{Cov}(R_i, R_j). \quad (6)$$

Equation 6 shows that individual variances of return constitute part, but not all, of portfolio variance. The three variances are outnumbered by the six covariance terms off the diagonal. If there are 20 assets, there are 20 variance terms and $20(20) - 20 = 380$ off-diagonal covariance terms. A first observation is that as the number of holdings increases, covariance becomes increasingly important, all else equal.

The covariance terms capture how the co-movements of returns affect aggregate portfolio variance. From the definition of covariance, we can establish two essential observations about covariance.

1. We can interpret the sign of covariance as follows:

- Covariance of returns is negative if, when the return on one asset is above its expected value, the return on the other asset tends to be below its expected value (an average inverse relationship between returns).
- Covariance of returns is 0 if returns on the assets are unrelated.

Covariance of returns is positive when the returns on both assets tend to be on the same side (above or below) their expected values at the same time (an average positive relationship between returns). The covariance of a random variable with itself (*own covariance*) is its own variance: $\text{Cov}(R, R) = E\{[RE(R)][RE(R)]\} = E\{[RE(R)]^2\} = \sigma^2(R)$.

Exhibit 2 summarizes the inputs for portfolio expected return (Panel A) and variance of return (Panel B). A complete list of the covariances constitutes all the statistical data needed to compute portfolio variance of return as shown in the covariance matrix in Panel B.

Exhibit 2: Inputs to Portfolio Expected Return and Variance**A. Inputs to Portfolio Expected Return**

Asset	A	B	C
	$E(R_A)$	$E(R_B)$	$E(R_C)$

B. Covariance Matrix: The Inputs to Portfolio Variance of Return

Asset	A	B	C
A	$\text{Cov}(R_A R_A, R_A R_A)$	$\text{Cov}(R_A R_B, R_B R_A)$	$\text{Cov}(R_A R_C, R_C R_A)$
B	$\text{Cov}(R_B R_A, R_A R_B)$	$\text{Cov}(R_B R_B, R_B R_B)$	$\text{Cov}(R_B R_C, R_C R_B)$
C	$\text{Cov}(R_C R_A, R_A R_C)$	$\text{Cov}(R_C R_B, R_B R_C)$	$\text{Cov}(R_C R_C, R_C R_C)$

With three assets, the covariance matrix has $3^2 = 3 \times 3 = 9$ entries, but the diagonal terms, the variances (bolded in Exhibit 2), are treated separately from the off-diagonal terms. So, there are $9 - 3 = 6$ covariances, excluding variances. But $\text{Cov}(R_B R_A) = \text{Cov}(R_A R_B)$, $\text{Cov}(R_C R_A) = \text{Cov}(R_A R_C)$, and $\text{Cov}(R_C R_B) = \text{Cov}(R_B R_C)$. The covariance matrix below the diagonal is the mirror image of the covariance matrix above the diagonal, so you only need to use one (i.e., either below or above the diagonal). As a result, there are only $6/2 = 3$ distinct covariance terms to estimate. In general, for n securities, there are $n(n - 1)/2$ distinct covariances and n variances to estimate.

Suppose we have the covariance matrix shown in Exhibit 3 with returns expressed as a percentage. The table entries are shown as return percentages squared ($\%^2$). The terms $38\%^2$ and $400\%^2$ are 0.0038 and 0.0400, respectively, stated as decimals; the correct usage of percents and decimals leads to identical answers.

Exhibit 3: Covariance Matrix

	S&P 500	US Long-Term Corporate Bonds	MSCI EAFE
S&P 500	400	45	189
US long-term corporate bonds	45	81	38
MSCI EAFE	189	38	441

Taking Equation 5 and grouping variance terms together produces the following:

$$\begin{aligned}
 \sigma^2(R_p) &= w_1^2 \sigma^2(R_1) + w_2^2 \sigma^2(R_2) + w_3^2 \sigma^2(R_3) + 2w_1 w_2 \text{Cov}(R_1, R_2) \\
 &\quad + 2w_1 w_3 \text{Cov}(R_1, R_3) + 2w_2 w_3 \text{Cov}(R_2, R_3) \\
 &= (0.50)^2(400) + (0.25)^2(81) + (0.25)^2(441) \\
 &\quad + 2(0.50)(0.25)(45) + 2(0.50)(0.25)(189) \\
 &\quad + 2(0.25)(0.25)(38) \\
 &= 100 + 5.0625 + 27.5625 + 11.25 + 47.25 + 4.75 = 195.875.
 \end{aligned}$$

The variance is 195.875. Standard deviation of return is $195.875^{1/2} = 14\%$. To summarize, the portfolio has an expected annual return of 11.75 percent and a standard deviation of return of 14 percent.

Looking at the first three terms in the calculation above, their sum (100 + 5.0625 + 27.5625) is 132.625, the contribution of the individual variances to portfolio variance. If the returns on the three assets were independent, covariances would be 0 and the standard deviation of portfolio return would

be $132.625^{1/2} = 11.52$ percent as compared to 14 percent before, so a less risky portfolio. If the covariance terms were negative, then a negative number would be added to 132.625, so portfolio variance and risk would be even smaller, while expected return would not change. For the same expected portfolio return, the portfolio has less risk. This risk reduction is a diversification benefit, meaning a risk-reduction benefit from holding a portfolio of assets. The diversification benefit increases with decreasing covariance. This observation is a key insight of modern portfolio theory. This insight is even more intuitively stated when we can use the concept of correlation.

Correlation

The **correlation** between two random variables, R_i and R_j , is defined as follows:

$$\rho(R_i, R_j) = \text{Cov}(R_i, R_j)/[\sigma(R_i)\sigma(R_j)]. \quad (7)$$

Alternative notations are $\text{Corr}(R_i, R_j)$ and ρ_{ij} .

The above definition of correlation is forward-looking because it involves dividing the forward-looking covariance by the product of forward-looking standard deviations. Frequently, covariance is substituted out using the relationship $\text{Cov}(R_i, R_j) = \rho(R_i, R_j) \sigma(R_i)\sigma(R_j)$. Like covariance, the correlation coefficient is a measure of linear association. However, the division in the definition makes correlation a pure number (without a unit of measurement) and places bounds on its largest and smallest possible values, which are +1 and -1, respectively.

If two variables have a strong positive linear relation, then their correlation will be close to +1. If two variables have a strong negative linear relation, then their correlation will be close to -1. If two variables have a weak linear relation, then their correlation will be close to 0. Using the previous definition, we can state a correlation matrix from data in the covariance matrix alone. Exhibit 4 shows the correlation matrix.

Exhibit 4: Correlation Matrix of Returns

	S&P 500	US Long-Term Corporate Bonds	MSCI EAFE
S&P 500	1.00	0.25	0.45
US long-term corporate bonds	0.25	1.00	0.20
MSCI EAFE	0.45	0.20	1.00

For example, from Exhibit 3, we know the covariance between long-term bonds and MSCI EAFE is 38. The standard deviation of long-term bond returns is $81^{1/2} = 9$ percent, that of MSCI EAFE returns is $441^{1/2} = 21$ percent, from diagonal terms in Exhibit 3. The correlation $\rho(R_{\text{long-term bonds}}, R_{\text{EAFE}})$ is $38/[(9\%)(21\%)] = 0.201$, rounded to 0.20. The correlation of the S&P 500 with itself equals 1: The calculation is its own covariance divided by its standard deviation squared.

EXAMPLE 1

Portfolio Expected Return and Variance of Return with Varying Portfolio Weights

Anna Cintara is constructing different portfolios from the following two stocks:

Exhibit 5: Description of Two-Stock Portfolio

	Stock 1	Stock 2
Expected return	4%	8%
Standard deviation	6%	15%
Current portfolio weights	0.40	0.60
Correlation between returns	0.30	

1. Calculate the covariance between the returns on the two stocks.

Solution:

The correlation between two stock returns is $\rho(R_i, R_j) = \text{Cov}(R_i, R_j)/[\sigma(R_i)\sigma(R_j)]$, so the covariance is $\text{Cov}(R_i, R_j) = \rho(R_i, R_j)\sigma(R_i)\sigma(R_j)$. For these two stocks, the covariance is $\text{Cov}(R_1, R_2) = \rho(R_1, R_2)\sigma(R_1)\sigma(R_2) = 0.30(6)(15) = 27$.

2. What is the portfolio expected return and standard deviation if Cintara puts 100 percent of her investment in Stock 1 ($w_1 = 1.00$ and $w_2 = 0.00$)? What is the portfolio expected return and standard deviation if Cintara puts 100 percent of her investment in Stock 2 ($w_1 = 0.00$ and $w_2 = 1.00$)?

Solution:

If the portfolio is 100 percent invested in Stock 1, the portfolio has an expected return of 4 percent and a standard deviation of 6 percent. If the portfolio is 100 percent invested in Stock 2, the portfolio has an expected return of 8 percent and a standard deviation of 15 percent.

3. What are the portfolio expected return and standard deviation using the current portfolio weights?

Solution:

For the current 40/60 portfolio, the expected return is

$$E(R_p) = w_1 E(R_1) + (1 - w_1) E(R_2) = 0.40(4\%) + 0.60(8\%) = 6.4\%$$

The portfolio variance and standard deviation are as follows:

$$\begin{aligned}\sigma^2(R_p) &= w_1^2 \sigma^2(R_1) + w_2^2 \sigma^2(R_2) + 2w_1 w_2 \text{Cov}(R_1, R_2) \\ &= (0.40)^2(36) + (0.60)^2(225) + 2(0.40)(0.60)(27) \\ &= 5.76 + 81 + 12.96 = 99.72\end{aligned}$$

$$\sigma(R_p) = 99.72^{1/2}$$

4. Calculate the expected return and standard deviation of the portfolios when w_1 goes from 0.00 to 1.00 in 0.10 increments (and $w_2 = 1 - w_1$). Place the results (stock weights, portfolio expected return, and portfolio standard deviation) in a table, and then sketch a graph of the results with the standard deviation on the horizontal axis and expected return on the vertical axis.

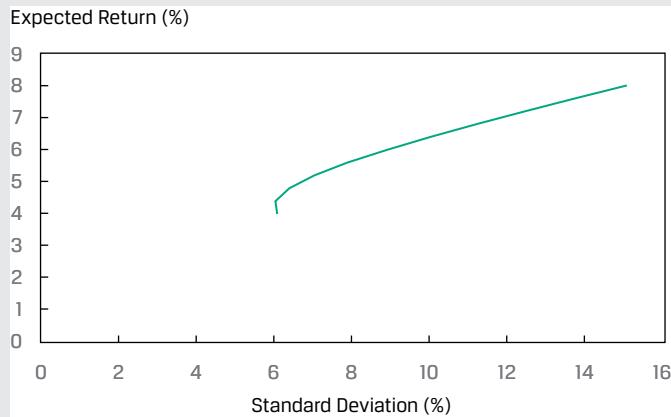
Solution:

The portfolio expected returns, variances, and standard deviations for the different sets of portfolio weights are given in the following table. Three of the rows are already computed in the solutions to 2 and 3, and the other

rows are computed using the same expected return, variance, and standard deviation formulas as in the solution to 3:

Stock 1 weight	Stock 2 weight	Expected return (%)	Variance (% ²)	Standard deviation (%)
1.00	0.00	4.00	36.00	6.00
0.90	0.10	4.40	36.27	6.02
0.80	0.20	4.80	40.68	6.38
0.70	0.30	5.20	49.23	7.02
0.60	0.40	5.60	61.92	7.87
0.50	0.50	6.00	78.75	8.87
0.40	0.60	6.40	99.72	9.99
0.30	0.70	6.80	124.83	11.17
0.20	0.80	7.20	154.08	12.41
0.10	0.90	7.60	187.47	13.69
0.00	1.00	8.00	225.00	15.00

The graph of the expected return and standard deviation follows:



QUESTION SET



1. US and Spanish bonds returns measured in the same currency units have standard deviations of 0.64 and 0.56, respectively. If the correlation between the two bonds is 0.24, the covariance of returns is *closest* to:

- A. 0.086.
- B. 0.335.
- C. 0.390.

Solution:

A is correct. The covariance is the product of the standard deviations and correlation using the formula $\text{Cov}(\text{US bond returns}, \text{Spanish bond returns}) = \sigma(\text{US bonds}) \times \sigma(\text{Spanish bonds}) \times \rho(\text{US bond returns}, \text{Spanish bond returns}) = 0.64 \times 0.56 \times 0.24 = 0.086$.

2. The covariance of returns is positive when the returns on two assets tend to:
- A. have the same expected values.

- B.** be above their expected value at different times.
C. be on the same side of their expected value at the same time.

Solution:

C is correct. The covariance of returns is positive when the returns on both assets tend to be on the same side (above or below) their expected values at the same time, indicating an average positive relationship between returns.

3. Which of the following correlation coefficients indicates the weakest linear relationship between two variables?

- A.** -0.67
B. -0.24
C. 0.33

Solution:

B is correct. Correlations near +1 exhibit strong positive linearity, whereas correlations near -1 exhibit strong negative linearity. A correlation of 0 indicates an absence of any linear relationship between the variables. The closer the correlation is to 0, the weaker the linear relationship.

4. An analyst develops the following covariance matrix of returns:

	Hedge Fund	Market Index
Hedge fund	256	110
Market index	110	81

The correlation of returns between the hedge fund and the market index is closest to:

- A.** 0.005.
B. 0.073.
C. 0.764.

Solution:

C is correct. The correlation between two random variables R_i and R_j is defined as $\rho(R_i, R_j) = \text{Cov}(R_i, R_j)/[\sigma(R_i)\sigma(R_j)]$. Using the subscript i to represent hedge funds and the subscript j to represent the market index, the standard deviations are $\sigma(R_i) = 256^{1/2} = 16$ and $\sigma(R_j) = 81^{1/2} = 9$. Thus, $\rho(R_i, R_j) = \text{Cov}(R_i, R_j)/[\sigma(R_i)\sigma(R_j)] = 110/(16 \times 9) = 0.764$.

5. All else being equal, as the correlation between two assets approaches +1.0, the diversification benefits:

- A.** decrease.
B. stay the same.
C. increase.

Solution:

A is correct. As the correlation between two assets approaches +1, diversification benefits decrease. In other words, an increasingly positive correlation indicates an increasingly strong positive linear relationship and fewer diversification benefits.

6. Given a portfolio of five stocks, how many unique covariance terms, excluding variances, are required to calculate the portfolio return variance?

- A. 10
- B. 20
- C. 25

Solution:

A is correct. A covariance matrix for five stocks has $5 \times 5 = 25$ entries. Subtracting the 5 diagonal variance terms results in 20 off-diagonal entries. Because a covariance matrix is symmetrical, only 10 entries are unique ($20/2 = 10$).

7. Which of the following statements is *most* accurate? If the covariance of returns between two assets is 0.0023, then the:

- A. assets' risk is near zero.
- B. asset returns are unrelated.
- C. asset returns have a positive relationship.

Solution:

C is correct. The covariance of returns is positive when the returns on both assets tend to be on the same side (above or below) their expected values at the same time.

8. A two-stock portfolio includes stocks with the following characteristics:

	Stock 1	Stock 2
Expected return	7%	10%
Standard deviation	12%	25%
Portfolio weights	0.30	0.70
Correlation		0.20

What is the standard deviation of portfolio returns?

- A. 14.91 percent
- B. 18.56 percent
- C. 21.10 percent

Solution:

B is correct. The covariance between the returns for the two stocks is $\text{Cov}(R_1, R_2) = \rho(R_1, R_2) \sigma(R_1) \sigma(R_2) = 0.20 (12) (25) = 60$. The portfolio variance is

$$\begin{aligned}\sigma^2(R_p) &= w_1^2 \sigma^2(R_1) + w_2^2 \sigma^2(R_2) + 2 w_1 w_2 \text{Cov}(R_1, R_2) \\ &= (0.30)^2 (12)^2 + (0.70)^2 (25)^2 + 2(0.30)(0.70)(60) \\ &= 12.96 + 306.25 + 25.2 = 344.41.\end{aligned}$$

The portfolio standard deviation is

$$\sigma^2(R_p) = 344.41^{1/2} = 18.56\%.$$

9. Lena Hunziger has designed the following three-asset portfolio:

	Asset 1	Asset 2	Asset 3
Expected return	5%	6%	7%
Portfolio weight	0.20	0.30	0.50

Variance-Covariance Matrix			
	Asset 1	Asset 2	Asset 3
Asset 1	196	105	140
Asset 2	105	225	150
Asset 3	140	150	400

Hunziger estimated the portfolio return to be 6.3 percent. What is the portfolio standard deviation?

- A. 13.07 percent
- B. 13.88 percent
- C. 14.62 percent

Solution:

C is correct. For a three-asset portfolio, the portfolio variance is

$$\begin{aligned}\sigma^2(R_p) &= w_1^2 \sigma^2(R_1) + w_2^2 \sigma^2(R_2) + w_3^2 \sigma^2(R_3) + 2w_1 w_2 \text{Cov}(R_1, R_2) \\ &\quad + 2w_1 w_3 \text{Cov}(R_1, R_3) + 2w_2 w_3 \text{Cov}(R_2, R_3) \\ &= (0.20)^2(196) + (0.30)^2(225) + (0.50)^2(400) + 2(0.20)(0.30)(105) \\ &\quad + 2(0.20)(0.50)(140) + 2(0.30)(0.50)(150) \\ &= 7.84 + 20.25 + 100 + 12.6 + 28 + 45 = 213.69.\end{aligned}$$

The portfolio standard deviation is

$$\sigma^2(R_p) = 213.69^{1/2} = 14.62\%.$$

FORECASTING CORRELATION OF RETURNS: COVARIANCE GIVEN A JOINT PROBABILITY FUNCTION

3

- calculate and interpret the covariance and correlation of portfolio returns using a joint probability function for returns

How do we estimate return covariance and correlation? Frequently, we make forecasts on the basis of historical covariance or use other methods such as a market model regression based on historical return data. We can also calculate covariance using the **joint probability function** of the random variables, if that can be estimated. The joint probability function of two random variables X and Y , denoted $P(X, Y)$, gives the probability of joint occurrences of values of X and Y . For example, $P(X=3, Y=2)$, is the probability that X equals 3 and Y equals 2.

Suppose that the joint probability function of the returns on BankCorp stock (R_A) and the returns on NewBank stock (R_B) has the simple structure given in Exhibit 6.

**Exhibit 6: Joint Probability Function of BankCorp and NewBank Returns
(Entries Are Joint Probabilities)**

	$R_B = 20\%$	$R_B = 16\%$	$R_B = 10\%$
$R_A = 25\%$	0.20	0	0
$R_A = 12\%$	0	0.50	0
$R_A = 10\%$	0	0	0.30

The expected return on BankCorp stock is $0.20(25\%) + 0.50(12\%) + 0.30(10\%) = 14\%$. The expected return on NewBank stock is $0.20(20\%) + 0.50(16\%) + 0.30(10\%) = 15\%$. The joint probability function above might reflect an analysis based on whether banking industry conditions are good, average, or poor. Exhibit 7 presents the calculation of covariance.

Exhibit 7: Covariance Calculations

Banking Industry Condition	Deviations BankCorp	Deviations NewBank	Product of Deviations	Probability of Condition	Probability-Weighted Product
Good	25–14	20–15	55	0.20	11
Average	12–14	16–15	-2	0.50	-1
Poor	10–14	10–15	20	0.30	6
					$\text{Cov}(R_A, R_B) = 16$

Note: Expected return for BankCorp is 14% and for NewBank, 15%.

The first and second columns of numbers show, respectively, the deviations of BankCorp and NewBank returns from their mean or expected value. The next column shows the product of the deviations. For example, for good industry conditions, $(25 - 14)(20 - 15) = 11(5) = 55$. Then, 55 is multiplied or weighted by 0.20, the probability that banking industry conditions are good: $55(0.20) = 11$. The calculations for average and poor banking conditions follow the same pattern. Summing up these probability-weighted products, we find $\text{Cov}(R_A, R_B) = 16$.

A formula for computing the covariance between random variables R_A and R_B is

$$\text{Cov}(R_A, R_B) = \sum_i \sum_j P(R_{A,i}, R_{B,j})(R_{A,i} - ER_A)(R_{B,j} - ER_B). \quad (8)$$

The formula tells us to sum all possible deviation cross-products weighted by the appropriate joint probability.

Next, we take note of the fact that when two random variables are independent, their joint probability function simplifies.

Two random variables X and Y are **independent** if and only if $P(X,Y) = P(X)P(Y)$.

For example, given independence, $P(3,2) = P(3)P(2)$. We multiply the individual probabilities to get the joint probabilities. *Independence* is a stronger property than *uncorrelatedness* because correlation addresses only linear relationships. The following condition holds for independent random variables and, therefore, also holds for uncorrelated random variables.

The expected value of the product of uncorrelated random variables is the product of their expected values.

$$E(XY) = E(X)E(Y) \text{ if } X \text{ and } Y \text{ are uncorrelated.}$$

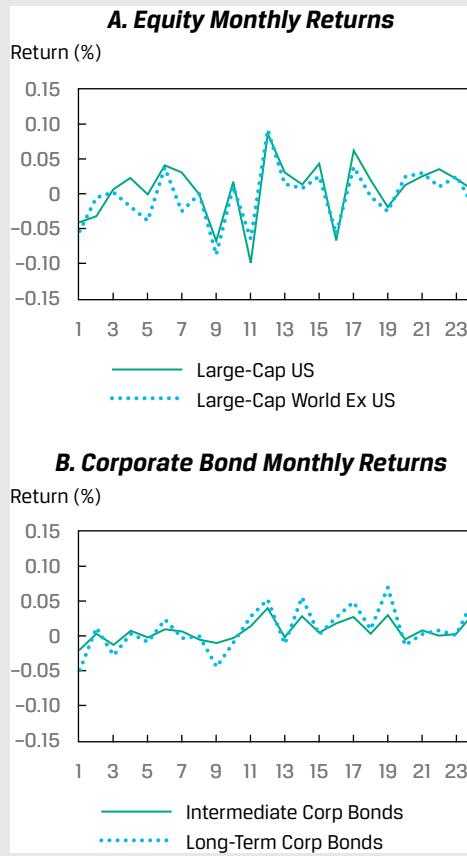
Many financial variables, such as revenue (price times quantity), are the product of random quantities. When applicable, the previous rule simplifies the calculation of the expected value of a product of random variables.

EXAMPLE 2

Covariances and Correlations of Security Returns

Isabel Vasquez is reviewing the correlations between four of the asset classes in her company portfolio. In Exhibit 8, she plots 24 recent monthly returns for large-cap US stocks versus for large-cap world ex-US stocks (Panel 1) and the 24 monthly returns for intermediate-term corporate bonds versus long-term corporate bonds (Panel 2). Vasquez presents the returns, variances, and covariances in decimal form instead of percentage form. Note the different ranges of their vertical axes (Return %).

Exhibit 8: Monthly Returns for Four Asset Classes



1. Selected data for the four asset classes are shown in Exhibit 9.

Exhibit 9: Selected Data for Four Asset Classes

Asset Classes	Large-Cap US Equities	World (ex US) Equities	Intermediate Corp Bonds	Long-Term Corp Bonds
Variance	0.001736	0.001488	0.000174	0.000699
Standard deviation	0.041668	0.038571	0.013180	0.026433
Covariance		0.001349		0.000318
Correlation		0.87553		0.95133

Vasquez noted, as shown in Exhibit 9, that although the two equity classes had much greater variances and covariance than the two bond classes, the correlation between the two equity classes was lower than the correlation between the two bond classes. She also noted that long-term bonds were more volatile (higher variance) than intermediate-term bonds; however, long- and intermediate-term bond returns still had a high correlation.

4

PORTFOLIO RISK MEASURES: APPLICATIONS OF THE NORMAL DISTRIBUTION

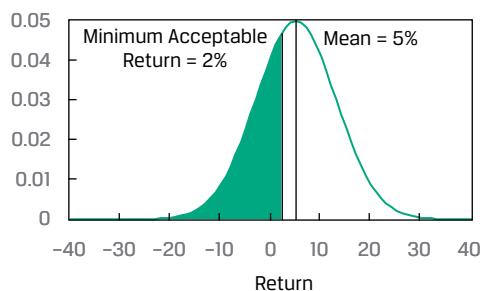
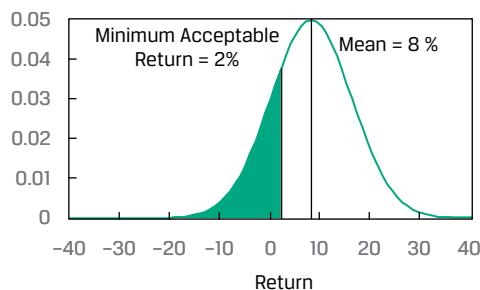
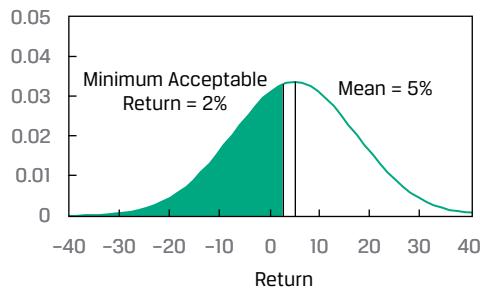
- define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion

Modern portfolio theory (MPT) often involves valuing investment opportunities using mean return and variance of return measures. In economic theory, **mean-variance analysis** holds exactly when investors are risk averse; when they choose investments to maximize expected utility or satisfaction; and when either (assumption 1) returns are normally distributed or (assumption 2) investors have quadratic utility functions (a concept used in economics for a mathematical representation of risk and return trade-offs). Mean-variance analysis, however, can still be useful—that is, it can hold approximately—when either assumption 1 or 2 is violated. Because practitioners prefer to work with observables, such as returns, the proposition that returns are at least approximately normally distributed has played a key role in much of MPT.

To illustrate this concept, assume an investor is saving for retirement. Although her goal is to earn the highest real return possible, she believes that the portfolio should at least achieve real capital preservation over the long term. Assuming a long-term expected inflation rate of 2 percent, the minimum acceptable return would be 2 percent. Exhibit 10 compares three investment alternatives in terms of their expected returns and standard deviation of returns. The probability of falling below 2 percent is calculated on basis of the assumption of normally distributed returns. In Exhibit 10, we see that Portfolio II, which combines the highest expected return and the lowest volatility, has the lowest probability of earning less than 2 percent (or equivalently, the highest probability of earning at least 2 percent). This also can be seen in Panel B, which shows that Portfolio II has the smallest shaded area to the left of 2 percent (the probability of earning less than the minimum acceptable return).

Exhibit 10: Probability of Earning a Minimum Acceptable Return**Panel A: Alternative Portfolio Characteristics**

Portfolio	I	II	III
Expected return	5%	8%	5%
Standard deviation of return	8%	8%	12%
Probability of earning $< 2\% [P(x < 2)]$	37.7%	24.6%	41.7%
Probability of earning $\geq 2\% [P(x \geq 2)]$	62.3%	75.4%	58.3%

Panel B: Likelihoods of Attaining Minimal Acceptable Return**A. Portfolio I****B. Portfolio II****C. Portfolio III**

Mean-variance analysis generally considers risk symmetrically in the sense that standard deviation captures variability both above and below the mean. An alternative approach evaluates only downside risk. We discuss one such approach, safety-first rules, because they provide an excellent illustration of the application of normal distribution theory to practical investment problems. **Safety-first rules** focus on **shortfall**.

risk, the risk that portfolio value (or portfolio return) will fall below some minimum acceptable level over some time horizon. The risk that the assets in a defined benefit plan will fall below plan liabilities is an example of a shortfall risk.

Suppose an investor views any return below a level of R_L as unacceptable. Roy's safety-first criterion (Roy 1952) states that the optimal portfolio minimizes the probability that portfolio return, R_p , will fall below the threshold level, R_L . That is, the investor's objective is to choose a portfolio that minimizes $P(R_p < R_L)$. When portfolio returns are normally distributed, we can calculate $P(R_p < R_L)$ using the number of standard deviations that R_L lies below the expected portfolio return, $E(R_p)$. The portfolio for which $E(R_p) - R_L$ is largest relative to standard deviation minimizes $P(R_p < R_L)$. Therefore, if returns are normally distributed, the safety-first optimal portfolio *maximizes* the safety-first ratio (SFRatio), as follows:

$$\text{SFRatio} = [E(R_p) - R_L]/\sigma_p \quad (9)$$

The quantity $E(R_p) - R_L$ is the distance from the mean return to the shortfall level. Dividing this distance by σ_p gives the distance in units of standard deviation. When choosing among portfolios using Roy's criterion (assuming normality), follow these two steps:

1. Calculate each portfolio's SFRatio.
2. Choose the portfolio with the highest SFRatio.

For a portfolio with a given safety-first ratio, the probability that its return will be less than R_L is $N(-\text{SFRatio})$, and the safety-first optimal portfolio has the lowest such probability. For example, suppose an investor's threshold return, R_L , is 2 percent. He is presented with two portfolios. Portfolio 1 has an expected return of 12 percent, with a standard deviation of 15 percent. Portfolio 2 has an expected return of 14 percent, with a standard deviation of 16 percent. The SFRatios, using Equation 9, are $0.667 = (12 - 2)/15$ and $0.75 = (14 - 2)/16$ for Portfolios 1 and 2, respectively. For the superior Portfolio 2, the probability that portfolio return will be less than 2 percent is $N(-0.75) = 1 - N(0.75) = 1 - 0.7734 = 0.227$, or about 23 percent, assuming that portfolio returns are normally distributed.

You may have noticed the similarity of the SFRatio to the Sharpe ratio. If we substitute the risk-free rate, R_F , for the critical level R_L , the SFRatio becomes the Sharpe ratio. The safety-first approach provides a new perspective on the Sharpe ratio: When we evaluate portfolios using the Sharpe ratio, the portfolio with the highest Sharpe ratio is the one that minimizes the probability that portfolio return will be less than the risk-free rate (given a normality assumption).

EXAMPLE 3

The Safety-First Optimal Portfolio for a Client

You are researching asset allocations for a client in Canada with a CAD800,000 portfolio. Although her investment objective is long-term growth, at the end of a year, she may want to liquidate CAD30,000 of the portfolio to fund educational expenses. If that need arises, she would like to be able to take out the CAD30,000 without invading the initial capital of CAD800,000. Exhibit 11 shows three alternative allocations.

Exhibit 11: Mean and Standard Deviation for Three Allocations (in Percent)

Allocation	A	B	C
Expected annual return	25	11	14
Standard deviation of return	27	8	20

Address these questions (assume normality for Questions 2 and 3):

- Given the client's desire not to invade the CAD800,000 principal, what is the shortfall level, R_L ? Use this shortfall level to answer question 2.

Solution:

Because CAD30,000/CAD800,000 is 3.75 percent, for any return less than 3.75 percent the client will need to invade principal if she takes out CAD30,000. So, $R_L = 3.75\%$.

- According to the safety-first criterion, which of the three allocations is the best?

(Hint, to decide which of the three allocations is safety-first optimal, select the alternative with the highest ratio $[E(R_p) - R_L]/\sigma_p$)

- A. $0.787037 = (25 - 3.75)/27$
- B. $0.90625 = (11 - 3.75)/8$
- C. $0.5125 = (14 - 3.75)/20$

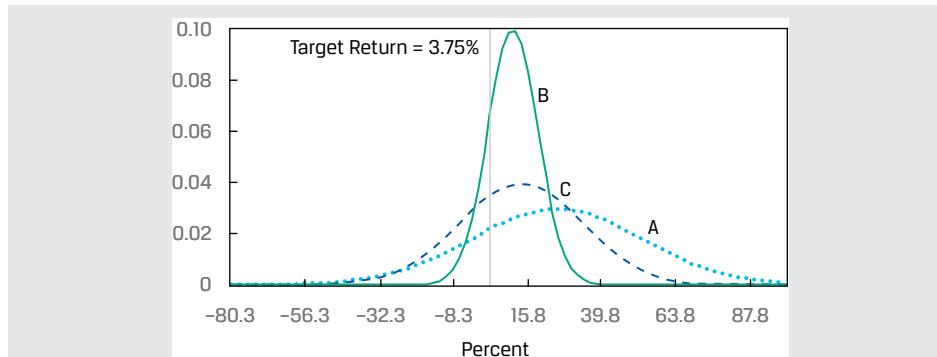
Solution:

B is correct. Allocation B, with the largest ratio (0.90625), is the best alternative according to the safety-first criterion.

- What is the probability that the return on the safety-first optimal portfolio will be less than the shortfall level?

Solution:

To answer this question, note that $P(R_B < 3.75) = \text{Normal}(-0.90625)$. We can round 0.90625 to 0.91 for use with tables of the standard normal CDF. First, we calculate $\text{Normal}(-0.91) = 1 - \text{Normal}(0.91) = 1 - 0.8186 = 0.1814$, or about 18.1 percent. Using a spreadsheet function for the standard normal CDF on -0.90625 without rounding, we get 0.182402, or about 18.2 percent. The safety-first optimal portfolio has a roughly 18 percent chance of not meeting a 3.75 percent return threshold. This can be seen in the following graphic, in which Allocation B has the smallest area under the distribution curve to the left of 3.75 percent.



Several points are worth noting. First, if the inputs were slightly different, we could get a different ranking. For example, if the mean return on B were 10 percent rather than 11 percent, Allocation A would be superior to B. Second, if meeting the 3.75 percent return threshold were a necessity rather than a wish, CAD830,000 in one year could be modeled as a liability. Fixed-income strategies, such as cash flow matching, could be used to offset or immunize the CAD830,000 quasi-liability.

In many investment contexts besides Roy's safety-first criterion, we use the normal distribution to estimate a probability. Another arena in which the normal distribution plays an important role is financial risk management. Financial institutions, such as investment banks, security dealers, and commercial banks, have formal systems to measure and control financial risk at various levels, from trading positions to the overall risk for the firm. Two mainstays in managing financial risk are value at risk (VaR) and stress testing/scenario analysis. **Stress testing** and **scenario analysis** refer to a set of techniques for estimating losses in extremely unfavorable combinations of events or scenarios. **Value at risk** (VaR) is a money measure of the minimum value of losses expected over a specified time period (e.g., a day, a quarter, or a year) at a given level of probability (often 0.05 or 0.01). Suppose we specify a one-day time horizon and a level of probability of 0.05, which would be called a 95 percent one-day VaR. If this VaR equaled EUR5 million for a portfolio, there would be a 0.05 probability that the portfolio would lose EUR5 million or more in a single day (assuming our assumptions were correct). One of the basic approaches to estimating VaR, the variance–covariance or analytical method, assumes that returns follow a normal distribution.

QUESTION SET



A client has a portfolio of common stocks and fixed-income instruments with a current value of GBP1,350,000. She intends to liquidate GBP50,000 from the portfolio at the end of the year to purchase a partnership share in a business. Furthermore, the client would like to be able to withdraw the GBP50,000 without reducing the initial capital of GBP1,350,000. The following table shows four alternative asset allocations.

Mean and Standard Deviation for Four Allocations (in Percent)

	A	B	C	D
Expected annual return	16	12	10	9
Standard deviation of return	24	17	12	11

1. Address the following questions (assume normality for Parts B and C):

- A. Given the client's desire not to invade the GBP1,350,000 principal, what is the shortfall level, R_L ? Use this shortfall level to answer Question 2.
- B. According to the safety-first criterion, which of the allocations is the best?
- C. What is the probability that the return on the safety-first optimal portfolio will be less than the shortfall level, R_L ?

Solution:

- A. Because GBP50,000/GBP1,350,000 is 3.7 percent, for any return less than 3.7 percent the client will need to invade principal if she takes out GBP50,000. So $R_L = 3.7$ percent.
- B. To decide which of the allocations is safety-first optimal, select the alternative with the highest ratio $[E(R_P) - R_L]/\sigma_P$:

$$\text{Allocation 1 } 0.5125 = (16 - 3.7)/24.$$

$$\text{Allocation 2 } 0.488235 = (12 - 3.7)/17.$$

$$\text{Allocation 3 } 0.525 = (10 - 3.7)/12.$$

$$\text{Allocation 4 } 0.481818 = (9 - 3.7)/11.$$

Allocation C, with the largest ratio (0.525), is the best alternative according to the safety-first criterion.

- C. To answer this question, note that $P(R_C < 3.7) = N(0.037 - 0.10)/0.12 = \text{Normal}(-0.525)$. By using Excel's NORM.S.DIST() function, we get $\text{NORM.S.DIST}((0.037 - 0.10)/0.12) = 29.98\%$, or about 30 percent. The safety-first optimal portfolio has a roughly 30 percent chance of not meeting a 3.7 percent return threshold.

2. A client holding a GBP2,000,000 portfolio wants to withdraw GBP90,000 in one year without invading the principal. According to Roy's safety-first criterion, which of the following portfolio allocations is optimal?

	Allocation A	Allocation B	Allocation C
Expected annual return	6.5%	7.5%	8.5%
Standard deviation of returns	8.35%	10.21%	14.34%

- A. Allocation A
- B. Allocation B
- C. Allocation C

Solution:

B is correct. Allocation B has the highest safety-first ratio. The threshold return level, R_L , for the portfolio is $\text{GBP90,000}/\text{GBP}2,000,000 = 4.5$ percent; thus, any return less than $R_L = 4.5\%$ will invade the portfolio principal. To compute the allocation that is safety-first optimal, select the alternative with the highest ratio:

$$\frac{[E(R_P) - R_L]}{\sigma_P}.$$

$$\text{Allocation A} = \frac{6.5 - 4.5}{8.35} = 0.240.$$

$$\text{Allocation B} = \frac{7.5 - 4.5}{10.21} = 0.294.$$

$$\text{Allocation C} = \frac{8.5 - 4.5}{14.34} = 0.279.$$

PRACTICE PROBLEMS

1. An analyst produces the following joint probability function for a foreign index (FI) and a domestic index (DI).

	$R_{DI} = 30\%$	$R_{DI} = 25\%$	$R_{DI} = 15\%$
$R_{FI} = 25\%$	0.25		
$R_{FI} = 15\%$		0.50	
$R_{FI} = 10\%$			0.25

The covariance of returns on the foreign index and the returns on the domestic index is *closest* to:

- A. 26.39.
- B. 26.56.
- C. 28.12.

SOLUTIONS

1. B is correct. The covariance is 26.56, calculated as follows. First, expected returns are

$$E(R_{FI}) = (0.25 \times 25) + (0.50 \times 15) + (0.25 \times 10)$$

$$= 6.25 + 7.50 + 2.50 = 16.25 \text{ and}$$

$$E(R_{DI}) = (0.25 \times 30) + (0.50 \times 25) + (0.25 \times 15)$$

$$= 7.50 + 12.50 + 3.75 = 23.75.$$

Covariance is

$$\text{Cov}(R_{FI}, R_{DI}) = \sum_i \sum_j P(R_{FI,i}, R_{DI,j})(R_{FI,i} - E R_{FI})(R_{DI,j} - E R_{DI})$$

$$= 0.25[(25 - 16.25)(30 - 23.75)] + 0.50[(15 - 16.25)(25 - 23.75)] + \\ 0.25[(10 - 16.25)(15 - 23.75)]$$

$$= 13.67 + (-0.78) + 13.67 = 26.56.$$

LEARNING MODULE

6

Simulation Methods

by Kobor Adam, PhD, CFA.

Adam Kobor, PhD, CFA, at New York University Investment Office (USA)

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	explain the relationship between normal and lognormal distributions and why the lognormal distribution is used to model asset prices when using continuously compounded asset returns
<input type="checkbox"/>	describe Monte Carlo simulation and explain how it can be used in investment applications
<input type="checkbox"/>	describe the use of bootstrap resampling in conducting a simulation based on observed data in investment applications

INTRODUCTION

1

The understanding and application of probability distributions is a critical component of forecasting financial variables and asset prices. This learning module provides a foundation for understanding important concepts related to probability distributions. Regarding the application of probability distributions, this learning module explains how to construct and interpret a Monte Carlo simulation analysis. Bootstrapping, with some similarities to Monte Carlo simulations, is also demonstrated to illustrate the use and application of this statistical sampling approach.

LEARNING MODULE OVERVIEW



- The lognormal distribution is widely used for modeling the probability distribution of financial asset prices because the distribution is bounded from below by 0 as asset prices and usually describes accurately the statistical distribution properties of financial assets prices. Lognormal distribution is typically skewed to the right.
- Continuously compounded returns play a role in many asset pricing models, as well as in risk management.

- Monte Carlo simulation is widely used to estimate risk and return in investment applications. Specifically, it is commonly used to value securities with complex features, such as embedded options, where no analytic pricing formula is available
- A Monte Carlo simulation generates a large number of random samples from a specified probability distribution or a series of distributions to obtain the likelihood of a range of results.
- Bootstrapping mimics the process of performing random sampling from a population to construct the sampling distribution by treating the randomly drawn sample as if it were the population.
- Because a random sample offers a good representation of the population, bootstrapping can simulate sampling from the population by sampling from the observed sample.

2

LOGNORMAL DISTRIBUTION AND CONTINUOUS COMPOUNDING



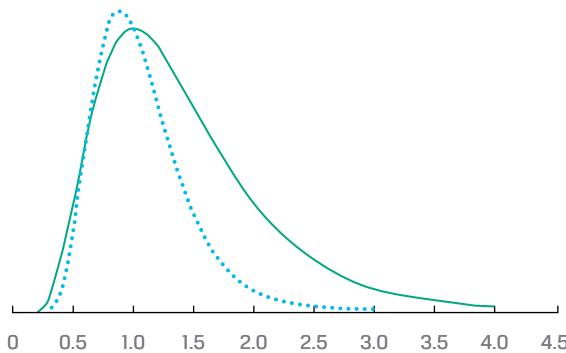
explain the relationship between normal and lognormal distributions and why the lognormal distribution is used to model asset prices when using continuously compounded asset returns

The Lognormal Distribution

Closely related to the normal distribution, the lognormal distribution is widely used for modeling the probability distribution of share and other asset prices. For example, the lognormal distribution appears in the Black–Scholes–Merton option pricing model. The Black–Scholes–Merton model assumes that the price of the asset underlying the option is lognormally distributed.

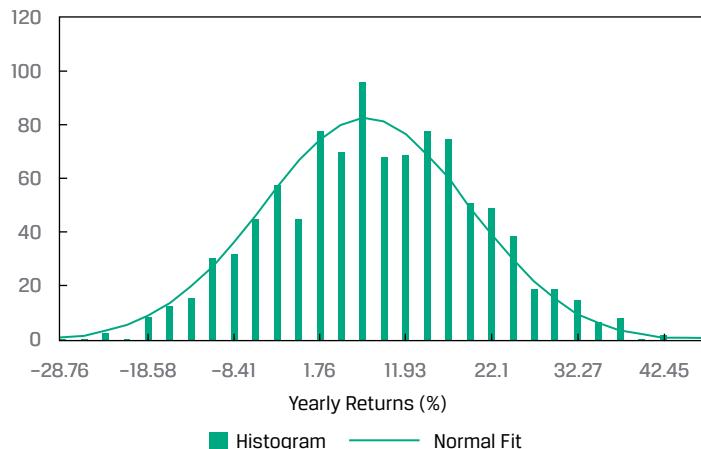
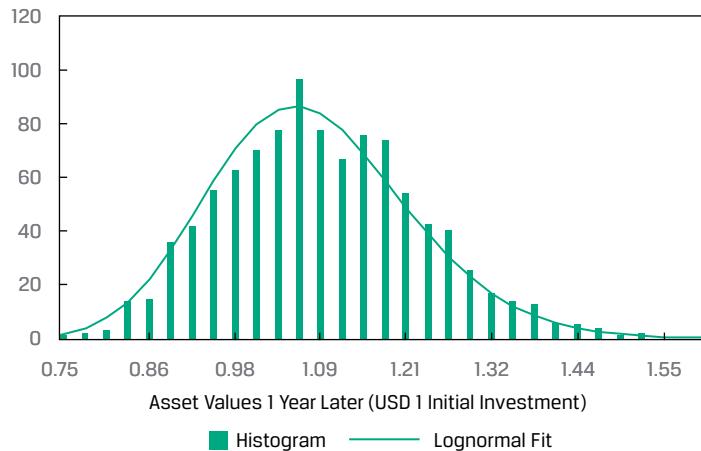
A random variable Y follows a lognormal distribution if its natural logarithm, $\ln Y$, is normally distributed. The reverse is also true: If the natural logarithm of random variable Y , $\ln Y$, is normally distributed, then Y follows a lognormal distribution. If you think of the term lognormal as “the log is normal,” you will have no trouble remembering this relationship.

The two most noteworthy observations about the lognormal distribution are that it is bounded below by 0 and it is skewed to the right (it has a long right tail). Note these two properties in the graphs of the probability density functions (pdfs) of two lognormal distributions in Exhibit 1. Asset prices are bounded from below by 0. In practice, the lognormal distribution has been found to be a usefully accurate description of the distribution of prices for many financial assets. However, the normal distribution is often a good approximation for returns. For this reason, both distributions are very important for finance professionals.

Exhibit 1: Two Lognormal Distributions

Like the normal distribution, the lognormal distribution is completely described by two parameters. Unlike many other distributions, a lognormal distribution is defined in terms of the parameters of a *different* distribution. The two parameters of a lognormal distribution are the mean and standard deviation (or variance) of its associated normal distribution: the mean and variance of $\ln Y$, given that Y is lognormal. So, we must keep track of two sets of means and standard deviations (or variances): (1) the mean and standard deviation (or variance) of the associated normal distribution (these are the parameters) and (2) the mean and standard deviation (or variance) of the lognormal variable itself.

To illustrate this relationship, we simulated 1,000 scenarios of yearly asset returns, assuming that returns are normally distributed with 7 percent mean and 12 percent standard deviation. For each scenario i , we converted the simulated continuously compounded returns (r_i) to future asset prices with the formula $\text{Price}(1 \text{ year later})_i = \text{USD}1 \times \exp(r_i)$, where \exp is the exponential function and assuming that the asset's price is USD1 today. In Exhibit 2, Panel A shows the distribution of the simulated returns together with the fitted normal pdf, whereas Panel B shows the distribution of the corresponding future asset prices together with the fitted lognormal pdf. Again, note that the lognormal distribution of future asset prices is bounded below by 0 and has a long right tail.

Exhibit 2: Simulated Returns (Normal PDF) and Asset Prices (Lognormal PDF)
A. Normal PDF

B. Lognormal PDF


The expressions for the mean and variance of the lognormal variable are challenging. Suppose a normal random variable X has expected value μ and variance σ^2 . Define $Y = \exp(X)$. Remember that the operation indicated by $\exp(X)$ or e^X (where $e \approx 2.7183$) is the opposite operation from taking logs. Because $\ln Y = \ln [\exp(X)] = X$ is normal (we assume X is normal), Y is lognormal. What is the expected value of $Y = \exp(X)$? A guess might be that the expected value of Y is $\exp(\mu)$. The expected value is actually $\exp(\mu + 0.50\sigma^2)$, which is larger than $\exp(\mu)$ by a factor of $\exp(0.50\sigma^2) > 1$. To get some insight into this concept, think of what happens if we increase σ^2 . The distribution spreads out; it can spread upward, but it cannot spread downward past 0. As a result, the center of its distribution is pushed to the right: The distribution's mean increases.

The expressions for the mean and variance of a lognormal variable are summarized below, where μ and σ^2 are the mean and variance of the associated normal distribution (refer to these expressions as needed, rather than memorizing them):

- Mean (μ_L) of a lognormal random variable = $\exp(\mu + 0.50\sigma^2)$.

- Variance (σ_L^2) of a lognormal random variable = $\exp(2\mu + \sigma^2) \times [\exp(\sigma^2) - 1]$.

Continuously Compounded Rates of Return

We now explore the relationship between the distribution of stock return and stock price. In this section, we show that if a stock's continuously compounded return is normally distributed, then future stock price is necessarily lognormally distributed. Furthermore, we show that stock price may be well described by the lognormal distribution even when continuously compounded returns do not follow a normal distribution. These results provide the theoretical foundation for using the lognormal distribution to model asset prices.

Showing that the stock price at some future time T , P_T , equals the current stock price, P_0 , multiplied by e raised to power $r_{0,T}$, the continuously compounded return from 0 to T :

$$P_T = P_0 \exp(r_{0,T}).$$

We showed in an earlier lesson that $r_{0,T}$, the continuously compounded return to time T , is the sum of the one-period continuously compounded returns, as follows:

$$r_{0,T} = r_{T-1,T} + r_{T-2,T-1} + \dots + r_{0,1}. \quad (1)$$

If these shorter-period returns are normally distributed, then $r_{0,T}$ is normally distributed (given certain assumptions) or approximately normally distributed (not making those assumptions). As P_T is proportional to the log of a normal random variable, P_T is lognormal.

A key assumption in many investment applications is that returns are **independently and identically distributed** (i.i.d.). Independence captures the proposition that investors cannot predict future returns using past returns. Identical distribution captures the assumption of stationarity, a property implying that the mean and variance of return do not change from period to period.

Assume that the one-period continuously compounded returns (such as $r_{0,1}$) are i.i.d. random variables with mean μ and variance σ^2 (but making no normality or other distributional assumption). Then,

$$E(r_{0,T}) = E(r_{T-1,T}) + E(r_{T-2,T-1}) + \dots + E(r_{0,1}) = \mu T, \quad (2)$$

(we add up μ for a total of T times), and

$$\sigma^2(r_{0,T}) = \sigma^2 T \quad (3)$$

(as a consequence of the independence assumption).

The variance of the T holding period continuously compounded return is T multiplied by the variance of the one-period continuously compounded return; also, $\sigma(r_{0,T}) = \sigma\sqrt{T}$. If the one-period continuously compounded returns on the right-hand side of Equation 1 are normally distributed, then the T holding period continuously compounded return, $r_{0,T}$, is also normally distributed with mean μT and variance $\sigma^2 T$. This is because a linear combination of normal random variables is also a normal random variable.

Even if the one-period continuously compounded returns are not normal, their sum, $r_{0,T}$, is approximately normal according to the central limit theorem. Now compare $P_T = P_0 \exp(r_{0,T})$ to $Y = \exp(X)$, where X is *normal* and Y is lognormal (as we discussed previously). Clearly, we can model future stock price P_T as a lognormal random variable because $r_{0,T}$ should be at least approximately *normal*. This assumption of normally distributed returns is the basis in theory for the lognormal distribution as a model for the distribution of prices of shares and other financial assets.

Continuously compounded returns play a role in many asset pricing models, as well as in risk management. **Volatility** measures the standard deviation of the continuously compounded returns on the underlying asset; by convention, it is stated as an annualized measure. In practice, we often estimate volatility using a historical series of continuously compounded daily returns. We gather a set of daily holding period returns, convert them into continuously compounded daily returns and then compute the standard deviation of the continuously compounded daily returns and annualize that number using Equation 3.

Annualizing is typically done based on 250 days in a year, the approximate number of business days that financial markets are typically open for trading. Thus, if daily volatility were 0.01, we would state volatility (on an annual basis) as $0.01 \sqrt{250} = 0.1581$. Example 1 illustrates the estimation of volatility for the shares of Astra International.

EXAMPLE 1

Volatility of Share Price

Suppose you are researching Astra International (Indonesia Stock Exchange: ASII) and are interested in Astra's price action in a week in which international economic news had significantly affected the Indonesian stock market. You decide to use volatility as a measure of the variability of Astra shares during that week. Exhibit 3 shows closing prices during that week.

Exhibit 3: Astra International Daily Closing Prices

Day	Closing Price (Indonesian rupiah, IDR)
Monday	6,950
Tuesday	7,000
Wednesday	6,850
Thursday	6,600
Friday	6,350

Use the data provided to do the following:

1. Estimate the volatility of Astra shares. (Annualize volatility on the basis of 250 trading days in a year.)

Solution:

First, calculate the continuously compounded daily returns; then, find their standard deviation in the usual way. In calculating sample variance, to get sample standard deviation, the divisor is sample size minus 1.

$$\ln(7,000/6,950) = 0.007168.$$

$$\ln(6,850/7,000) = -0.021661.$$

$$\ln(6,600/6,850) = -0.037179.$$

$$\ln(6,350/6,600) = -0.038615.$$

$$\text{Sum} = -0.090287.$$

$$\text{Mean} = -0.022572.$$

$$\text{Variance} = 0.000452.$$

$$\text{Standard deviation} = 0.021261.$$

The standard deviation of continuously compounded daily returns is 0.021261. Equation 3 states that $\hat{\sigma}(r_{0,T}) = \hat{\sigma}\sqrt{T}$. In this example, $\hat{\sigma}$ is the sample standard deviation of one-period continuously compounded returns. Thus, $\hat{\sigma}$ refers to 0.021261. We want to annualize, so the horizon T corresponds to one year. Because $\hat{\sigma}$ is in days, we set T equal to the number of trading days in a year (250).

Therefore, we find that annualized volatility for Astra stock that week was 33.6 percent, calculated as $0.021261\sqrt{250} = 0.336165$.

2. Calculate an estimate of the expected continuously compounded annual return for Astra.

Solution:

Note that the sample mean, -0.022572 (from the Solution to 1), is a sample estimate of the mean, μ , of the continuously compounded one-period or daily returns. The sample mean can be translated into an estimate of the expected continuously compounded annual return using Equation 2, $\hat{\mu}_T = -0.022572(250)$ (using 250 to be consistent with the calculation of volatility).

3. Discuss why it may not be prudent to use the sample mean daily return to estimate the expected continuously compounded annual return for Astra.

Solution:

Four daily return observations are far too few to estimate expected returns. Further, the variability in the daily returns overwhelms any information about expected return in a series this short.

4. Identify the probability distribution for Astra share prices if continuously compounded daily returns follow the normal distribution.

Solution:

Astra share prices should follow the lognormal distribution if the continuously compounded daily returns on Astra shares follow the normal distribution.

We have shown that the distribution of stock price is lognormal, given certain assumptions. Earlier we gave bullet-point expressions for the mean and variance of a lognormal random variable. In the context of a stock price, the $\hat{\mu}$ and $\hat{\sigma}^2$ in these expressions would refer to the mean and variance of the T horizon, not the one-period, continuously compounded returns compatible with the horizon of P_T .

3

MONTE CARLO SIMULATION



describe Monte Carlo simulation and explain how it can be used in investment applications

After gaining an understanding of probability distributions used to characterize asset prices and asset returns, we explore a technique called **Monte Carlo simulation** in which probability distributions play an integral role. A characteristic of Monte Carlo simulation is the generation of a very large number of random samples from a specified probability distribution or distributions to obtain the likelihood of a range of results.

Monte Carlo simulation is widely used to estimate risk and return in investment applications. In this setting, we simulate the portfolio's profit and loss performance for a specified time horizon, either on an asset-by-asset basis or an aggregate, portfolio basis. Repeated trials within the simulation, each trial involving a draw of random observations from a probability distribution, produce a simulated frequency distribution of portfolio returns from which performance and risk measures are derived.

Another important use of Monte Carlo simulation in investments is as a tool for valuing complex securities for which no analytic pricing formula is available. For other securities, such as mortgage-backed securities with complex embedded options, Monte Carlo simulation is also an important modeling resource. Because we control the assumptions when we carry out a simulation, we can run a model for valuing such securities through a Monte Carlo simulation to examine the model's sensitivity to a change in key assumptions.

To understand the technique of Monte Carlo simulation, we present the process as a series of steps; these can be viewed as providing an overview rather than a detailed recipe for implementing a Monte Carlo simulation in its many varied applications.

To illustrate the steps, we use Monte Carlo simulation to value an option, **contingent claim**, whose value is based on some other underlying security. For this option, no analytic pricing formula is available. For our purposes, the value of this contingent claim (an Asian option), equals the difference between the underlying stock price at that maturity and the *average* stock price during the life of the contingent claim or USD 0, whichever is greater. For instance, if the final underlying stock price is USD 34 and the average value over the life of the claim is USD 31, the value of the contingent claim at its maturity is USD 3 (the greater of USD 34 – USD 31 = USD 3 and USD 0).

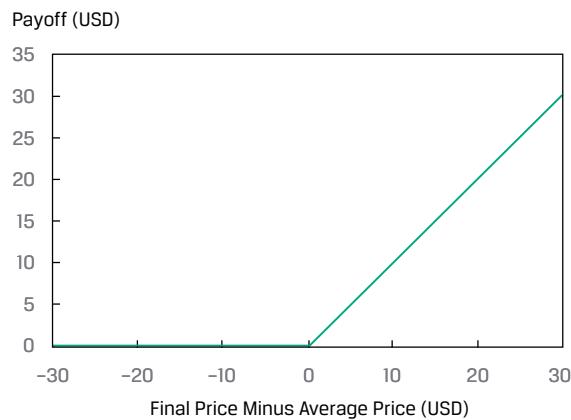
Assume that the maturity of the claim is one year from today; we will simulate stock prices in monthly steps over the next 12 months and will generate 1,000 scenarios to value this claim. The payoff diagram of this contingent claim security is depicted in Panel A of Exhibit 4, a histogram of simulated average and final stock prices is shown in Panel B, and a histogram of simulated payoffs of the contingent claim is presented in Panel C.

The payoff diagram (Panel A) is a snapshot of the contingent claim at maturity. If the stock's final price is less than or equal to its average over the life of the contingent claim, then the payoff would be zero. However, if the final price exceeds the average price, the payoff is equal to this difference. Panel B shows histograms of the simulated

final and average stock prices. Note that the simulated final price distribution is wider than the simulated average price distribution. Also, note that the contingent claim's value depends on the difference between the final and average stock prices, which cannot be directly inferred from these histograms.

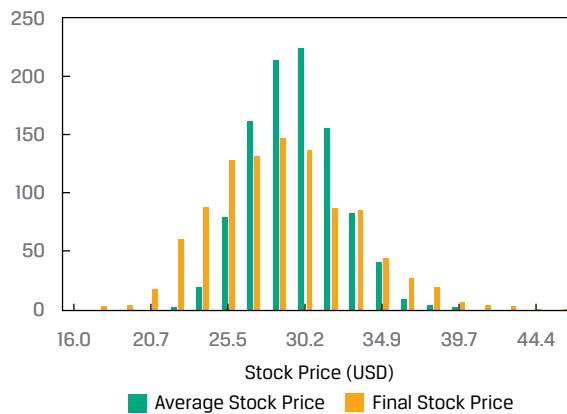
Exhibit 4: Payoff Diagram, Histogram of Simulated Average, and Final Stock Prices, and Histogram of Simulated Payoffs for Contingent Claim

A. Contingent Claim Payoff Diagram



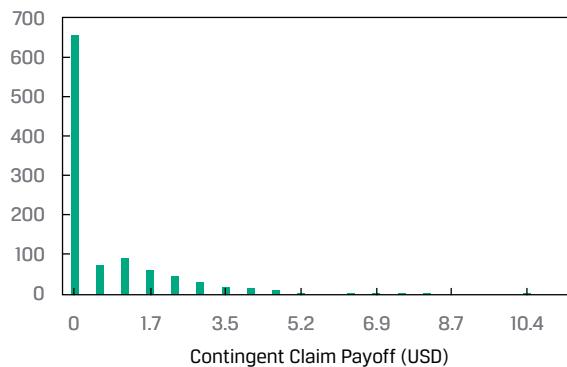
B. Histogram of Simulated Average and Final Stock Prices

Number of Trials



C. Histogram of Simulated Contingent Claim Payoffs

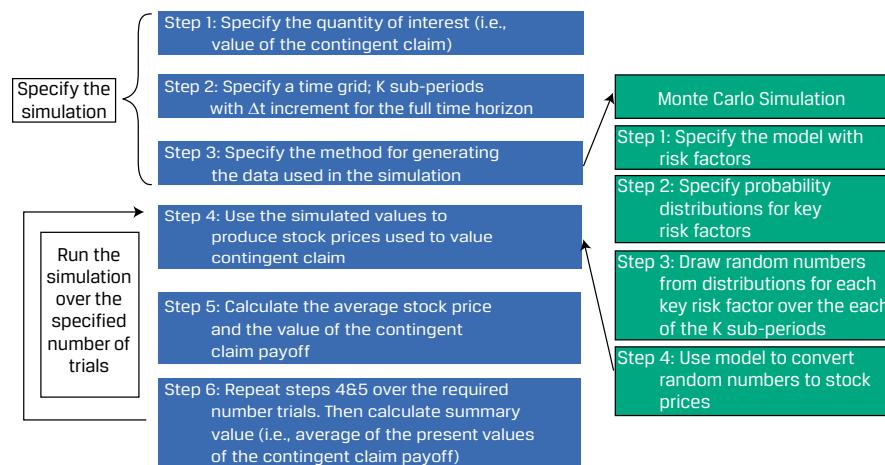
Number of Trials



Finally, Panel C shows the histogram of the contingent claim's simulated payoffs. In 654 of 1,000 total trials, the final stock price was less than or equal to the average price, so in 65.4 percent of the trials the contingent claim paid off zero. In the remaining 34.6 percent of the trials, however, the claim paid the positive difference between the final and average prices, with the maximum payoff being USD 11.

The process flowchart in Exhibit 5 shows the steps for implementing the Monte Carlo simulation for valuing this contingent claim. Steps 1 through 3 of the process describe specifying the simulation; Steps 4 through 6 describe running the simulation.

Exhibit 5: Steps in Implementing the Monte Carlo Simulation



The mechanics of implementing the Monte Carlo simulation for valuing the contingent claim using the six-step process are described as follows:

1. Specify the quantity of interest in terms of underlying variables. The quantity of interest is the contingent claim value, and the underlying variable is the stock price. Then, specify the starting value(s) of the underlying variable(s).

We use C_{iT} to represent the value of the claim at maturity, T . The subscript i in C_{iT} indicates that C_{iT} is a value resulting from the i th **simulation trial**, each simulation trial involving a drawing of random values (an iteration of Step 4).

2. Specify a time grid. Take the horizon in terms of calendar time and split it into a number of subperiods—say, K in total. Calendar time divided by the number of subperiods, K , is the time increment, Δt . In our example, calendar time is one year and K is 12, so Δt equals one month.
3. Specify the method for generating the data used in the simulation. This step will require that distributional assumptions be made for the key risk factors that drive the underlying variables. For example, stock price is the underlying variable for the contingent claim, so we need a model for stock price movement, effectively a period return. We choose the following model for changes in stock price, where Z_k stands for the standard normal random variable:

$$\Delta \text{Stock price} = (\mu \times \text{Prior stock price} \times \Delta t) + (\sigma \times \text{Prior stock price} \times Z_k).$$

The term Z_k is the key risk factor in the simulation. Through our choice of μ (mean) and σ (standard deviation), we control the distribution of the stock price variable. Although this example has one key risk factor, a given simulation may have multiple key risk factors.

Then, using a computer program or spreadsheet function, draw K random values of each risk factor. In our example, the spreadsheet function would produce a draw of K ($= 12$) values of the standard normal variable Z_k : $Z_1, Z_2, Z_3, \dots, Z_K$. We will discuss generating standard normal random numbers (or, in fact, random numbers with any kind of distribution) after describing the sequence of simulation steps.

4. Use the simulated values to produce stock prices used to value the contingent claim. This step will convert the standard normal random numbers generated in Step 3 into stock price changes (Δ Stock price) by using the model of stock price dynamics from Step 3. The result is K observations on possible changes in stock price over the K subperiods (remember, $K = 12$). An additional calculation is needed to convert those changes into a sequence of K stock prices, with the initial stock price as the starting value over the K subperiods. This is an important step: we rely on the distributional assumptions of the Monte Carlo simulation to randomly create a very large number of stock price processes.
5. Calculate the average stock price and the value of the contingent claim. This calculation produces the average stock price during the life of the contingent claim (the sum of K stock prices divided by K). Then, compute the value of the contingent claim at maturity, C_{iT} , and then calculate its present value, C_{i0} , by discounting this terminal value using an appropriate interest rate as of today. (The subscript i in C_{i0} stands for the i th simulation trial, as it does in C_{iT} .) We have now completed one simulation trial.
6. Repeat steps 4 and 5 over the required number of trials. Iteratively, go back to Step 4 until the specified number of trials, I , is completed. Finally, produce summary values and statistics for the simulation. The quantity of interest in our example is the mean value of C_{i0} for the total number of simulation trials ($I = 1,000$). This mean value is the Monte Carlo estimate of the value of our contingent claim.

In Example 2, we continue with the application of Monte Carlo simulation to value another type of contingent claim.

EXAMPLE 2

Valuing a Lookback Contingent Claim Using Monte Carlo Simulation

1. A standard lookback contingent claim on a stock has a value at maturity equal to (Value of the stock at maturity – Minimum value of stock during the life of the claim prior to maturity) or USD 0, whichever is greater. If the minimum value reached prior to maturity was USD 20.11 and the value of

the stock at maturity is USD 23, for example, the contingent claim is worth USD 23 – USD 20.11 = USD 2.89.

How might you use Monte Carlo simulation in valuing a lookback contingent claim?

Solution:

We previously described how to use Monte Carlo simulation to value a certain type of contingent claim. Just as we can calculate the average value of the stock over a simulation trial to value that claim, for a lookback contingent claim, we can also calculate the minimum value of the stock over a simulation trial. Then, for a given simulation trial, we can calculate the terminal value of the claim, given the minimum value of the stock for the simulation trial. We can then discount this terminal value back to the present to get the value of the claim today ($t = 0$). The average of these $t = 0$ values over all simulation trials is the Monte Carlo simulated value of the lookback contingent claim.

Finally, note that Monte Carlo simulation is a complement to analytical methods. It provides only statistical estimates, not exact results. Analytical methods, where available, provide more insight into cause-and-effect relationships. However, as financial product innovations proceed, the applications for Monte Carlo simulation in investment management continue to grow.

QUESTION SET



1. Define Monte Carlo simulation and explain its use in investment management.

Solution:

A Monte Carlo simulation generates a large number of random samples from a specified probability distribution (or distributions) to represent the role of risk in the system. Monte Carlo simulation is widely used to estimate risk and return in investment applications. In this setting, we simulate the portfolio's profit and loss performance for a specified time horizon. Repeated trials within the simulation produce a simulated frequency distribution of portfolio returns from which performance and risk measures are derived. Another important use of Monte Carlo simulation in investments is as a tool for valuing complex securities for which no analytic pricing formula is available. It is also an important modeling resource for securities with complex embedded options.

2. Compared with analytical methods, what are the strengths and weaknesses of using Monte Carlo simulation for valuing securities?

Solution:

- *Strengths:* Monte Carlo simulation can be used to price complex securities for which no analytic expression is available, particularly European-style options.
- *Weaknesses:* Monte Carlo simulation provides only statistical estimates, not exact results. Analytic methods, when available, provide more insight into cause-and-effect relationships than does Monte Carlo simulation.

3. A Monte Carlo simulation can be used to:

- A. directly provide precise valuations of call options.
- B. simulate a process from historical records of returns.
- C. test the sensitivity of a model to changes in assumptions—for example, on distributions of key variables.

Solution:

C is correct. A characteristic feature of Monte Carlo simulation is the generation of a large number of random samples from a specified probability distribution or distributions to represent the role of risk in the system. Therefore, it is very useful for investigating the sensitivity of a model to changes in assumptions—for example, on distributions of key variables.

4. A limitation of Monte Carlo simulation is:

- A. its failure to do “what if” analysis.
- B. that it requires historical records of returns.
- C. its inability to independently specify cause-and-effect relationships.

Solution:

C is correct. Monte Carlo simulation is a complement to analytical methods. Monte Carlo simulation provides statistical estimates and not exact results. Analytical methods, when available, provide more insight into cause-and-effect relationships.

BOOTSTRAPPING

4



describe the use of bootstrap resampling in conducting a simulation based on observed data in investment applications

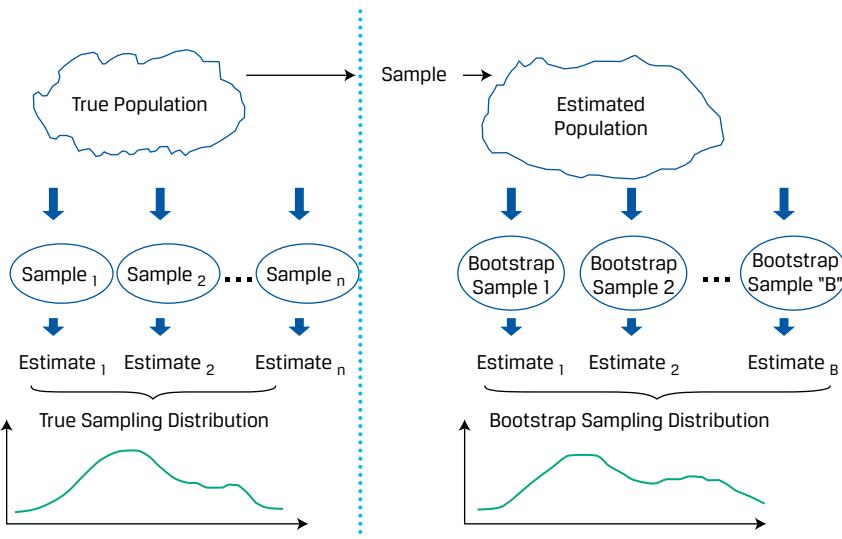
Earlier, we demonstrated how to find the standard error of the sample mean, which can be computed based on the central limit theorem. We now introduce a computational tool called **resampling**, which repeatedly draws samples from the original observed data sample for the statistical inference of population parameters. **Bootstrap**, one of the most popular resampling methods, uses computer simulation for statistical inference without using an analytical formula such as a z -statistic or t -statistic.

The idea behind bootstrap is to mimic the process of performing random sampling from a population to construct the sampling distribution. The difference lies in the fact that we have no knowledge of what the population looks like, except for a sample with size n drawn from the population. Because a random sample offers a good representation of the population, we can simulate sampling from the population by sampling from the observed sample. In other words, the bootstrap mimics the process by treating the randomly drawn sample as if it were the population.

Both the bootstrap and the Monte Carlo simulation build on repetitive sampling. Bootstrapping resamples a dataset as the true population, and infers from the sampling statistical distribution parameter values (i.e., mean, variance, skewness, and kurtosis) for the population. Monte Carlo simulation builds on generating random data with certain known statistical distribution of parameter values.

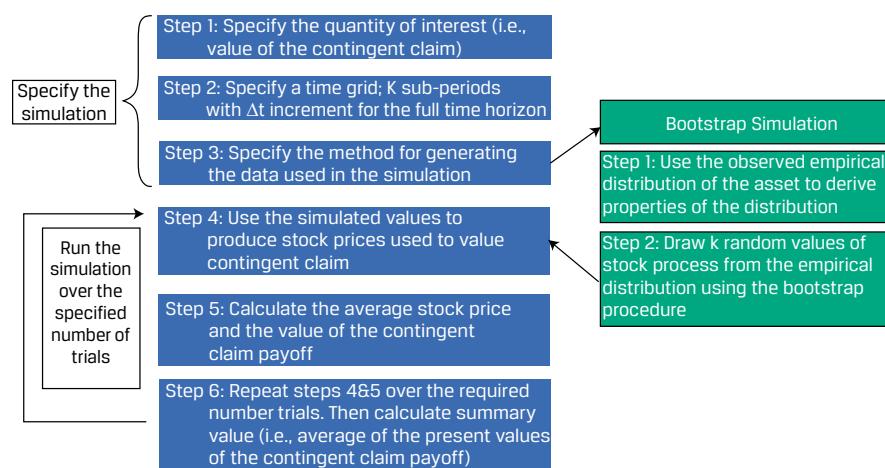
The right-hand side of Exhibit 6 illustrates the process. In bootstrap, we repeatedly draw samples from the original sample, and each resample is of the same size as the original sample. Note that each item drawn is replaced for the next draw (i.e., the identical element is put back into the group so that it can be drawn more than once). Although some items may appear several times in the resamples, other items may not appear at all.

Exhibit 6: Bootstrap Resampling



The mechanics of implementing the simulation for valuing the contingent claim used in the previous lesson using the bootstrap differ only in the source of the random variable used. Instead of being drawn from a probability distribution, under bootstrapping, the random variable is drawn from the sample as described in the discussion related to Exhibit 6. Exhibit 7 shows the steps for the bootstrap process highlighting the differences between the bootstrap process and the Monte Carlo simulation from Exhibit 6.

Exhibit 7: Steps in Implementing Simulation using Bootstrapping



The steps in using bootstrap to simulate the contingent claim are then (with the first two steps and the last three steps) the same for bootstrapping as they are for the Monte Carlo simulation:

1. Specify the quantity of interest in terms of underlying variables. The quantity of interest is the contingent claim value, and the underlying variable is the stock price. Then, specify the starting value(s) of the underlying variable(s).

We use C_{iT} to represent the value of the claim at maturity, T . The subscript i in C_{iT} indicates that C_{iT} is a value resulting from the i th simulation trial, each simulation trial involving a drawing of random values (an iteration of Step 4).

2. Specify a time grid that is consistent with the periodicity of the sample observations. Take the horizon in terms of calendar time and split it into a number of subperiods—say, K in total. Calendar time divided by the number of subperiods, K , is the time increment, Δt . In the example, calendar time was one year, and K is 12, so Δt equals one month.
3. Specify the method for generating the data used in the simulation. In our example, stock price is the underlying variable for the contingent claim, so we use the observed changes in stock price as our empirical distribution. We use the observed historical behavior of stock price processes: price changes or price returns.
4. Use the simulated values to produce stock prices used to value the contingent claim. Using a computer program or spreadsheet function, draw K random values of stock process from the empirical distribution using the bootstrap procedure. Then, convert the stock price changes (Δ Stock price) from Step 3 into the stock price dynamics. The calculation is necessary to convert those changes into a sequence of K stock prices, with the initial stock price as the starting value over the K subperiods. This is an important step: we rely on the distribution of the bootstrapped trials drawn from observed, historical stock price processes.
5. Calculate the average stock price and the value of the contingent claim. Another calculation produces the average stock price during the life of the contingent claim (the sum of K stock prices divided by K). Then, compute the value of the contingent claim at maturity, C_{iT} , and then calculate its present value, C_{i0} , by discounting this terminal value using an appropriate interest rate as of today. (The subscript i in C_{i0} stands for the i th simulation trial, as it does in C_{iT} .) We have now completed one simulation trial.
6. Repeat steps 4 and 5 over the required number of trials. Iteratively, go back to Step 4 until the specified number of trials, I , is completed. Finally, produce summary values and statistics for the simulation. The quantity of interest in our example is the mean value of C_{i0} for the total number of bootstrapping runs ($I = 1,000$). This mean value is the bootstrap estimate of the value of our contingent claim based on the observed empirical distribution.

Again, note that bootstrap simulation is a complement to analytical methods. It provides only statistical estimates based on the empirical distribution created by the bootstrapping process from observed, historical prices and price processes; these are not exact results. Analytical methods, where available, provide more insight into cause-and-effect relationships.

QUESTION SET

1. What are the main strengths and weaknesses of bootstrapping?

Solution:*Strengths:*

- Bootstrapping is simple to perform.
- Bootstrapping offers a good representation of the statistical features of the population and can simulate sampling from the population by sampling from the observed sample.

Weaknesses:

- Bootstrapping provides only statistical estimates, not exact results.

PRACTICE PROBLEMS

1. The weekly closing prices of Mordice Corporation shares are as follows:

Exhibit 1: Mordice Corporation Shares

Date	Closing Price (euros)
1 August	112
8 August	160
15 August	120

The continuously compounded return of Mordice Corporation shares for the period August 1 to August 15 is *closest* to:

- A. 6.90 percent.
 - B. 7.14 percent.
 - C. 8.95 percent.
2. In contrast to normal distributions, lognormal distributions:
 - A. are skewed to the left.
 - B. have outcomes that cannot be negative.
 - C. are more suitable for describing asset returns than asset prices.
 3. The lognormal distribution is a more accurate model for the distribution of stock prices than the normal distribution because stock prices are:
 - A. symmetrical.
 - B. unbounded.
 - C. non-negative.
 4. Analysts performing bootstrap:
 - A. seek to create statistical inferences of population parameters from a single sample.
 - B. repeatedly draw samples of the same size, with replacement, from the original population.
 - C. must specify probability distributions for key risk factors that drive the underlying random variables.

SOLUTIONS

1. A is correct. The continuously compounded return of an asset over a period is equal to the natural log of the asset's price change during the period. In this case, $\ln(120/112) = 6.90\%$.

Note that the continuously compounded return from period 0 to period T is the sum of the incremental one-period continuously compounded returns, which in this case are weekly returns. Specifically:

Week 1 return: $\ln(160/112) = 35.67\%$.

Week 2 return: $\ln(120/160) = -28.77\%$.

Continuously compounded return = $35.67\% + -28.77\% = 6.90\%$.

2. B is correct. By definition, lognormal random variables cannot have negative values (bounded below by 0) and have distributions that are skewed to the right.
3. C is correct. A lognormal distributed variable has a lower bound of zero. The lognormal distribution is also right skewed, which is a useful property in describing asset prices.
4. A is correct. Bootstrapping through random sampling generates the observed variable from a random sampling with unknown population parameters. The analyst does not know the true population distribution, but through sampling can infer the population parameters from the randomly generated sample. B is incorrect because, when performing bootstrap, the analyst repeatedly draws samples from the original sample and not population, where each individual resample has the same size as the original sample and each item drawn is replaced for the next draw. C is incorrect because, when performing bootstrap, analysts simply use the observed empirical distribution of the underlying variables. In a Monte Carlo simulation, in contrast, the analyst would specify probability distributions for key risk factors that drive the underlying variables.

LEARNING MODULE

7

Estimation and Inference

by Wu Jian, PhD.

Jian Wu, PhD, is at State Street (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling and their implications for sampling error in an investment problem
<input type="checkbox"/>	explain the central limit theorem and its importance for the distribution and standard error of the sample mean
<input type="checkbox"/>	describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

INTRODUCTION

1

In this Learning Module, we present the various methods for obtaining information on a population (all members of a specified group) through samples (part of the population). The information on a population that we seek usually concerns the value of a **parameter**, a quantity computed from or used to describe a population of data. In Lesson 1 we introduce sampling, which we use a sample to estimate a parameter; we make use of sample statistics. A statistic is a quantity computed from or used to describe a sample of data.

Supposing that a sample is representative of the underlying population, how can the analyst assess the sampling error in estimating a population parameter? In Lesson 2, the Central Limit Theorem helps us understand the sampling distribution of the sample mean in many of the estimation problems we face. This provides guidance on how closely a sample mean can be expected to match its underlying population mean, allowing an analyst to use the sampling distribution to assess the accuracy of the sample and test hypotheses about the underlying parameter. Lesson 3 covers various resampling approaches.

LEARNING MODULE OVERVIEW

- Of the two types of sampling methods, probability sampling includes simple random sampling and stratified random sampling, and non-probability sampling includes convenience sampling and judgmental sampling. Probability sampling involves equal chance of sample selection, while non-probability sampling has a significant risk of being non-representative.
- Sampling error is the difference between the observed value of a statistic and the quantity it is intended to estimate as a result of using subsets of the population.
- Non-probability sampling methods rely not on a fixed selection process but instead on a researcher's sample selection capabilities. Its advantages include quick and low-cost data collection, and can apply expert judgment for efficient sample selection.
- The Central Limit Theorem is defined as follows: Given a population described by any probability distribution with mean μ and finite variance σ^2 , the sampling distribution of the sample mean \bar{X} computed from random samples of size n from this population will be approximately normal with mean μ (the population mean) and variance σ^2/n (the population variance divided by n) when the sample size n is large.
- The standard error of the sample mean is an important quantity in applying the central limit theorem in practice. It is typically estimated using the square root of the sample variance, calculated as follows:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}.$$

- The central limit theorem shows that when sampling from any distribution, the sample mean distribution will have these two properties when the sample size is large: (1) the distribution of the sample mean \bar{X} will be approximately normal, and (2) the mean of the distribution of \bar{X} will be equal to the mean of the population from which samples are drawn.
- Bootstrap, a popular resampling method which repeatedly draws samples of the same size as the original sample, uses computer simulation for statistical inference without using an analytical formula such as a z-statistic or t-statistic. It can be used as a simple but powerful method for any complicated estimators such as the standard error of a sample mean.
- Bootstrap has potential advantages in accuracy. Given these advantages, it can be applied widely in finance, such as for historical simulations in asset allocation or in gauging an investment strategy's performance against a benchmark.
- Jackknife is another resampling technique with samples selected by taking the original observed data sample and leaving out one observation at a time from the set (and not replacing it). Jackknife is often used to reduce the bias of an estimator, and other applications include finding the standard error and confidence interval of an estimator.

SAMPLING METHODS

2



compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling and their implications for sampling error in an investment problem

We take samples for one of two reasons. In some cases, we cannot possibly examine every member of the population. In other cases, examining every member of the population would not be economically efficient. Thus, savings of time and money are two primary factors that cause an analyst to use sampling to answer a question about a population.

There are two types of sampling methods: **probability sampling** and **non-probability sampling**. Probability sampling gives every member of the population an equal chance of being selected. Hence it can create a sample that is representative of the population. In contrast, non-probability sampling depends on factors other than probability considerations, such as a sampler's judgment or the convenience to access data. Consequently, there is a significant risk that non-probability sampling might generate a non-representative sample. In general, all else being equal, probability sampling can yield more accuracy and reliability compared with non-probability sampling.

We first focus on probability sampling, particularly the widely used **simple random sampling** and **stratified random sampling**. We then turn our attention to non-probability sampling.

Simple Random Sampling

Suppose a wireless equipment analyst wants to know how much major customers will spend on average for equipment during the coming year. One strategy is to survey the population of wireless equipment customers and inquire what their purchasing plans are. Surveying all companies, however, would be very costly in terms of time and money.

Alternatively, the analyst can collect a representative sample of companies and survey them about upcoming wireless equipment expenditures. In this case, the analyst will compute the sample mean expenditure, \bar{X} , a statistic. This strategy has a substantial advantage over polling the whole population because it can be accomplished more quickly and at lower cost.

Sampling, however, introduces error. The error arises because not all of the companies in the population are surveyed. The analyst who decides to sample is trading time and money for sampling error.

When an analyst chooses to sample, they must formulate a sampling plan. A **sampling plan** is the set of rules used to select a sample. The basic type of sample from which we can draw statistically sound conclusions about a population is the simple random sample.

A **simple random sample** is a subset of a larger population created in such a way that each element of the population has an equal probability of being selected to the subset.

The procedure of drawing a sample to satisfy the definition of a simple random sample is called **simple random sampling**. Simple random sampling is particularly useful when data in the population is homogeneous—that is, the characteristics of the data or observations (e.g., size or region) are broadly similar. If this condition is not satisfied, other types of sampling may be more appropriate.

Systematic sampling can be used when we cannot code (or even identify) all the members of a population. With systematic sampling, we select every k th member until we have a sample of the desired size. The sample that results from this procedure should be approximately random.

Suppose the wireless equipment analyst polls a random sample of wireless equipment customers to determine the average equipment expenditure. The derived sample mean will provide the analyst with an estimate of the population mean expenditure. The mean obtained from the sample this way will differ from the population mean that we are trying to estimate. It is subject to error. An important part of this error is known as sampling error, which comes from sampling variation and occurs because we have data on only a subset of the population.

Sampling error is the difference between the observed value of a statistic and the quantity it is intended to estimate as a result of using subsets of the population.

A random sample reflects the properties of the population in an unbiased way, and sample statistics, such as the sample mean, computed on the basis of a random sample are valid estimates of the underlying population parameters. Thus a sample statistic is a random variable. In other words, not only do the original data from the population have a distribution but so does the sample statistic. This distribution is the statistic's sampling distribution.

Sampling distribution of a statistic is the distribution of all the distinct possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population.

In the case of the sample mean, for example, we refer to the "sampling distribution of the sample mean" or the distribution of the sample mean. We will have more to say about sampling distributions later in this text. Next, we look at another sampling method that is useful in investment analysis.

Stratified Random Sampling

The simple random sampling method just discussed may not be the best approach in all situations. One frequently used alternative is stratified random sampling.

In **stratified random sampling**, the population is divided into subpopulations (strata) based on one or more classification criteria. Simple random samples are then drawn from each stratum in sizes proportional to the relative size of each stratum in the population. These samples are then pooled to form a stratified random sample.

In contrast to simple random sampling, stratified random sampling guarantees that population subdivisions of interest are represented in the sample. Another advantage is that estimates of parameters produced from stratified sampling have greater precision—that is, smaller variance or dispersion—than estimates obtained from simple random sampling.

Bond indexing is one area in which stratified sampling is frequently applied. **Indexing** is an investment strategy in which an investor constructs a portfolio to mirror the performance of a specified index. In pure bond indexing, also called the full-replication approach, the investor attempts to fully replicate an index by owning all the bonds in the index in proportion to their market value weights. Many bond indexes consist of thousands of issues, however, so pure bond indexing is difficult to implement. In addition, transaction costs would be high because many bonds do not have liquid markets.

Although a simple random sample could be a solution to the cost problem, the sample would probably not match the index's major risk factors, such as interest rate sensitivity. Because the major risk factors of fixed-income portfolios are well known and quantifiable, stratified sampling offers a more effective approach. In this approach, we divide the population of index bonds into groups of similar duration (interest rate sensitivity), cash flow distribution, sector, credit quality, and call exposure. We refer

to each group as a stratum or cell (a term frequently used in this context). Then, we choose a sample from each stratum proportional to the relative market weighting of the stratum in the index to be replicated.

EXAMPLE 1**Bond Indexes and Stratified Sampling**

Suppose you are the manager of a portfolio of bonds indexed to the Bloomberg Barclays US Government/Credit Index, meaning that the portfolio returns should be similar to those of the index. You are exploring several approaches to indexing, including a stratified sampling approach. You first distinguish among agency bonds, US Treasury bonds, and investment-grade corporate bonds. For each of these three groups, you define 10 maturity intervals—1 to 2 years, 2 to 3 years, 3 to 4 years, 4 to 6 years, 6 to 8 years, 8 to 10 years, 10 to 12 years, 12 to 15 years, 15 to 20 years, and 20 to 30 years—and also separate the bonds with coupons (annual interest rates) of 6 percent or less from the bonds with coupons of more than 6 percent.

1. How many cells or strata does this sampling plan entail?

Solution:

We have 3 issuer classifications, 10 maturity classifications, and 2 coupon classifications. So, in total, this plan entails $3(10)(2) = 60$ different strata or cells.

2. If you use this sampling plan, what is the minimum number of issues the indexed portfolio can have?

Solution:

One cannot have less than 1 issue for each cell, so the portfolio must include at least 60 issues.

3. Suppose that in selecting among the securities that qualify for selection within each cell, you apply a criterion concerning the liquidity of the security's market. Is the sample obtained random? Explain your answer.

Solution

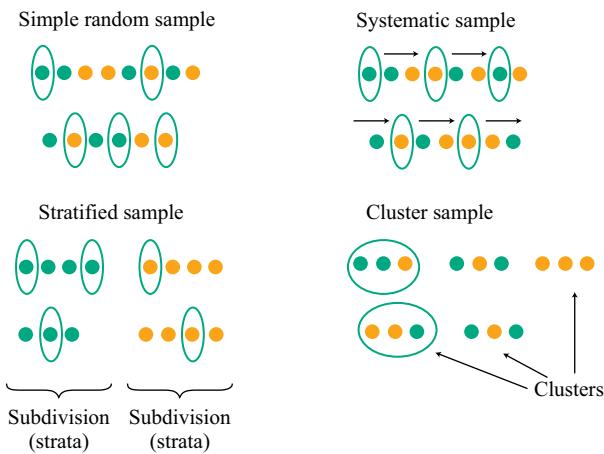
Applying any additional criteria to the selection of securities for the cells, not every security that might be included has an equal probability of being selected. There is no proportionality in the selection, and as a result, the sampling is not random. In practice, indexing using stratified sampling usually does not strictly involve random sampling because the selection of bond issues within cells is subject to various additional criteria. Because the purpose of sampling in this application is not to make an inference about a population parameter but rather to index a portfolio, lack of randomness is not in itself a problem in this application of stratified sampling.

Cluster Sampling

Another sampling method, **cluster sampling**, also requires the division or classification of the population into subpopulation groups, called clusters. In this method, the population is divided into clusters, each of which is essentially a mini-representation of the entire populations. Then certain clusters are chosen as a whole using simple

random sampling. If all the members in each sampled cluster are sampled, this sample plan is referred to as one-stage cluster sampling. If a subsample is randomly selected from each selected cluster, then the plan is referred as two-stage cluster sampling. Exhibit 1 (bottom-right panel) shows how cluster sampling works and how it compares with the other probability sampling methods.

Exhibit 1: Probability Sampling



A major difference between cluster and stratified random samples is that in cluster sampling, a whole cluster is regarded as a sampling unit and only sampled clusters are included. In stratified random sampling, however, all the strata are included and only specific elements within each stratum are then selected as sampling units.

Cluster sampling is commonly used for broad market surveys, and the most popular version identifies clusters based on geographic parameters. For example, a research institute is looking to survey if individual investors in the United States are bullish, bearish, or neutral on the stock market. It would be impossible to carry out the research by surveying all the individual investors in the country. The two-stage cluster sampling is a good solution in this case. In the first stage, a researcher can group the population by states and all the individual investors of each state represent a cluster. A handful of the clusters are then randomly selected. At the second stage, a simple random sample of individual investors is selected from each sampled cluster to conduct the survey.

Compared with other probability sampling methods, given equal sample size, cluster sampling usually yields lower accuracy because a sample from a cluster might be less representative of the entire population. Its major advantage, however, is offering the most time-efficient and cost-efficient probability sampling plan for analyzing a vast population.

Non-Probability Sampling

Non-probability sampling methods rely not on a fixed selection process but instead on a researcher's sample selection capabilities. We introduce two major types of non-probability sampling methods here.

- **Convenience Sampling:** In this method, an element is selected from the population based on whether or not it is accessible to a researcher or on how easy it is for a researcher to access the element. The samples are not

necessarily representative of the entire population, and hence the level of sampling accuracy could be limited. The advantage of **convenience sampling** is that data can be collected quickly at a low cost. In situations such as the preliminary stage of research or in circumstances subject to cost constraints, convenience sampling is often used as a time-efficient and cost-effective sampling plan for a small-scale pilot study before testing a large-scale and more representative sample.

- **Judgmental Sampling:** This sampling process involves selectively handpicking elements from the population based on a researcher's knowledge and professional judgment. Sample selection under **judgmental sampling** can be affected by the bias of the researcher and might lead to skewed results that do not represent the whole population. In circumstances where there is a time constraint, however, or when the specialty of researchers is critical to select a more representative sample than by using other probability or non-probability sampling methods, judgmental sampling allows researchers to go directly to the target population of interest. For example, when auditing financial statements, seasoned auditors can apply their sound judgment to select accounts or transactions that can provide sufficient audit coverage. Example 2 illustrates an application of these sampling methods.

EXAMPLE 2

Demonstrating the Power of Sampling

To demonstrate the power of sampling, we conduct two sampling experiments on a large dataset. The full dataset is the “population,” representing daily returns of the fictitious Euro-Asia-Africa (EAA) Equity Index. This dataset spans a five-year period and consists of 1,258 observations of daily returns with a minimum value of -4.1 percent and a maximum value of 5.0 percent.

First, we calculate the mean daily return of the EAA Equity Index (using the population).

By taking the average of all the data points, the mean of the entire daily return series is computed as 0.035 percent.

First Experiment: Random Sampling

The sample size m is set to 5, 10, 20, 50, 100, 200, 500, and 1,000. At each sample size, we run random sampling multiple times ($N = 100$) to collect 100 samples to compute mean absolute error. The aim is to compute and plot the mean error versus the sample size.

For a given sample size m , we use a random sampling procedure to compute mean absolute error in order to measure sampling error.

By applying this procedure, we compute mean absolute errors for eight different sample sizes: $m = 5, 10, 20, 50, 100, 200, 500$, and 1000.

Second Experiment: Stratified Random Sampling

We now conduct stratified random sampling by dividing daily returns into groups by year. The sample size m is again set to 5, 10, 20, 50, 100, 200, 500, and 1,000. At each sample size, we run random sampling multiple times ($N = 100$) to collect 100 samples to compute mean absolute error.

We follow the same steps as before, except for the first step. Rather than running a simple random sampling, we conduct stratified random sampling—that is, randomly selecting subsamples of equal number from daily return groups by year to generate a full sample. For example, for a sample of 50, 10 data points are randomly selected from daily returns of each year from 2014 to 2018, respectively. Exhibit 2 summarizes the results.

Exhibit 2: Mean Absolute Errors under Different Sampling Procedures

Mean Absolute Error of Random Sampling

Sample size	5	10	20	50	100	200	500	1,000
Mean absolute error	0.297%	0.218%	0.163%	0.091%	0.063%	0.039%	0.019%	0.009%

Mean Absolute Error of Stratified Random Sampling

Sample size	5	10	20	50	100	200	500	1,000
Mean absolute error	0.294%	0.205%	0.152%	0.083%	0.071%	0.051%	0.025%	0.008%

Under both random sampling and stratified sampling mean absolute errors quickly shrink as sample size increases. Stratified sampling produces smaller mean absolute errors as it more accurately reflects the character of the population, but this difference shrinks—and in this case actually expands—as the sample size increases.

Exhibit 2 also indicates that a minimum sample size is needed to limit sample error and achieve a certain level of accuracy. After a certain size, however, there is little incremental benefit from adding more observations (200 to 400 in this case).

Sampling from Different Distributions

In practice, in addition to selecting appropriate sampling methods, we also need to be careful when sampling from a population that is not under one single distribution. Example 3 illustrates the problems that can arise when sampling from more than one distribution.

EXAMPLE 3

Calculating Sharpe Ratios: One or Two Years of Quarterly Data

Analysts often use the Sharpe ratio to evaluate the performance of a managed portfolio. The Sharpe ratio is the average return in excess of the risk-free rate divided by the standard deviation of returns. This ratio measures the return of a fund or a security above the risk-free rate (the excess return) earned per unit of standard deviation of return.

To compute the Sharpe ratio, suppose that an analyst collects eight quarterly excess returns (i.e., total return in excess of the risk-free rate). During the first year, the investment manager of the portfolio followed a low-risk strategy, and during the second year, the manager followed a high-risk strategy. For each of these years, the analyst also tracks the quarterly excess returns of some benchmark against which the manager will be evaluated. For each of the two years, the Sharpe ratio for the benchmark is 0.21. Exhibit 3 gives the calculation of the Sharpe ratio of the portfolio.

Exhibit 3: Calculation of Sharpe Ratios: Low-Risk and High-Risk Strategies

Quarter/Measure	Year 1 Excess Returns	Year 2 Excess Returns
Quarter 1	-3%	-12%
Quarter 2	5	20
Quarter 3	-3	-12
Quarter 4	5	20
Quarterly average	1%	4%
Quarterly standard deviation	4.62%	18.48%

$$\text{Sharpe ratio} = 0.22 = 1/4.62 = 4/18.48$$

For the first year, during which the manager followed a low-risk strategy, the average quarterly return in excess of the risk-free rate was 1 percent with a standard deviation of 4.62 percent. The Sharpe ratio is thus $1/4.62 = 0.22$. The second year's results mirror the first year except for the higher average return and volatility. The Sharpe ratio for the second year is $4/18.48 = 0.22$. The Sharpe ratio for the benchmark is 0.21 during the first and second years. Because larger Sharpe ratios are better than smaller ones (providing more return per unit of risk), the manager appears to have outperformed the benchmark.

Now, suppose the analyst believes a larger sample to be superior to a small one. She thus decides to pool the two years together and calculate a Sharpe ratio based on eight quarterly observations. The average quarterly excess return for the two years is the average of each year's average excess return. For the two-year period, the average excess return is $(1 + 4)/2 = 2.5\%$ per quarter. The standard deviation for all eight quarters measured from the sample mean of 2.5 percent is 12.57 percent. The portfolio's Sharpe ratio for the two-year period is now $2.5/12.57 = 0.199$; the Sharpe ratio for the benchmark remains 0.21. Thus, when returns for the two-year period are pooled, the manager appears to have provided less return per unit of risk than the benchmark and less when compared with the separate yearly results.

The problem with using eight quarters of return data is that the analyst has violated the assumption that the sampled returns come from the same population. As a result of the change in the manager's investment strategy, returns in Year 2 followed a different distribution than returns in Year 1. Clearly, during Year 1, returns were generated by an underlying population with lower mean and variance than the population of the second year. Combining the results for the first and second years yielded a sample that was representative of no population. Because the larger sample did not satisfy model assumptions, any conclusions the

analyst reached based on the larger sample are incorrect. For this example, she was better off using a smaller sample than a larger sample because the smaller sample represented a more homogeneous distribution of returns.

QUESTION SET



An analyst is studying research and development (R&D) spending by pharmaceutical companies around the world. She considers three sampling methods for understanding a company's level of R&D. Method 1 is to simply use all the data available to her from an internal database that she and her colleagues built while researching several dozen representative stocks in the sector. Method 2 involves relying on a commercial database provided by a data vendor. She would select every fifth pharmaceutical company on the list to pull the data. Method 3 is to first divide pharmaceutical companies in the commercial database into three groups according to the region where a company is headquartered (e.g., Asia, Europe, or North America) and then randomly select a subsample of companies from each group, with the sample size proportional to the size of its associated group in order to form a complete sample.

1. Method 1 is an example of:

- A. simple random sampling.
- B. stratified random sampling.
- C. convenience sampling.

Solution:

C is correct. The analyst selects the data from the internal database because they are easy and convenient to access.

2. Method 2 is an example of:

- A. judgmental sampling.
- B. systematic sampling.
- C. cluster sampling.

Solution:

B is correct. The sample elements are selected with a fixed interval ($k = 5$) from the large population provided by data vendor.

3. Method 3 is an example of:

- A. simple random sampling.
- B. stratified random sampling.
- C. cluster sampling.

Solution:

B is correct. The population of pharmaceutical companies is divided into three strata by region to perform random sampling individually.

4. Perkiomen Kinzua, a seasoned auditor, is auditing last year's transactions for Conemaugh Corporation. Unfortunately, Conemaugh had a very large number of transactions, and Kinzua is under a time constraint to finish the audit.

He decides to audit only the small subset of the transaction population that is of interest and to use sampling to create that subset.

The most appropriate sampling method for Kinzua to use is:

- A. judgmental sampling.
- B. systematic sampling.
- C. convenience sampling.

Solution:

A is correct. With judgmental sampling, Kinzua will use his knowledge and professional judgment as a seasoned auditor to select transactions of interest from the population. This approach will allow Kinzua to create a sample that is representative of the population and that will provide sufficient audit coverage. Judgmental sampling is useful in cases that have a time constraint or in which the specialty of researchers is critical to select a more representative sample than by using other probability or non-probability sampling methods. Judgment sampling, however, entails the risk that Kinzua is biased in his selections, leading to skewed results that are not representative of the whole population.

CENTRAL LIMIT THEOREM AND INFERENCE

3



explain the central limit theorem and its importance for the distribution and standard error of the sample mean

Earlier we presented a wireless equipment analyst who decided to sample in order to estimate mean planned capital expenditures by the customers of wireless equipment vendors. Supposing that the sample is representative of the underlying population, how can the analyst assess the sampling error in estimating the population mean?

The sample mean is itself a random variable with a probability distribution called the statistic's sampling distribution. To estimate how closely the sample mean can be expected to match the underlying population mean, the analyst needs to understand the sampling distribution of the mean. The central limit theorem helps us understand the sampling distribution of the mean in many of the estimation problems we face.

The Central Limit Theorem

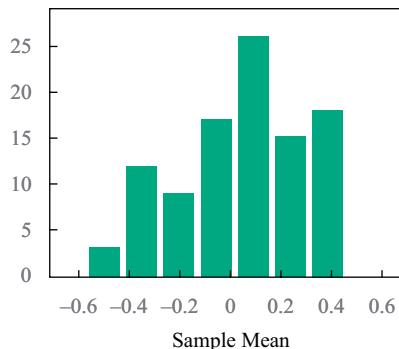
To explain the central limit theorem, we will revisit the daily returns of the fictitious Euro-Asia-Africa Equity Index shown earlier. The dataset (the population) consists of daily returns of the index over a five-year period. The 1,258 return observations have a population mean of 0.035 percent.

We conduct four different sets of random sampling from the population. We first draw a random sample of 10 daily returns and obtain a sample mean. We repeat this exercise 99 more times, drawing a total of 100 samples of 10 daily returns. We plot the sample mean results in a histogram, as shown in the top left panel of Exhibit 4. We then repeat the process with a larger sample size of 50 daily returns. We draw 100 samples of 50 daily returns and plot the results (the mean returns) in the histogram shown in the top-right panel of Exhibit 4. We then repeat the process for sample sizes of 100 and 300 daily returns, respectively, again drawing 100 samples in each case.

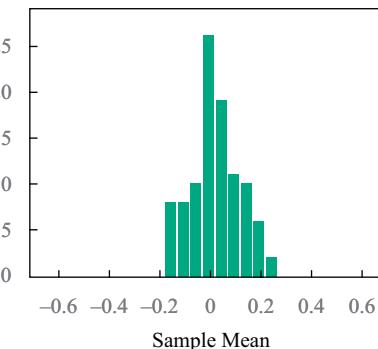
These results appear in the bottom-left and bottom-right panels of Exhibit 4. Looking at all four panels together, we observe that the larger the sample size, the more closely the histogram follows the shape of normal distribution.

Exhibit 4: Sampling Distribution with Increasing Sample Size

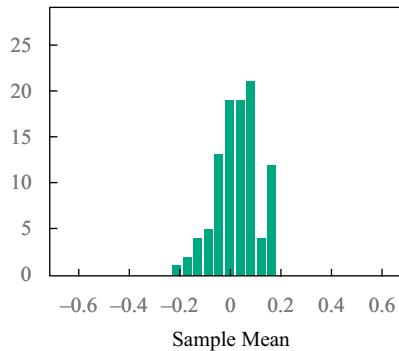
A. Sample Size $n = 10$



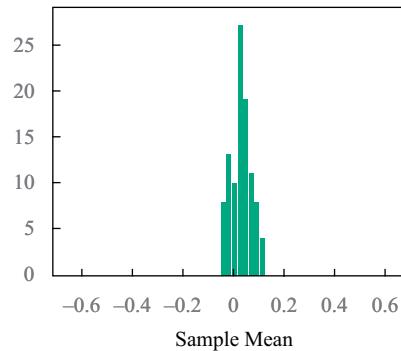
B. Sample Size $n = 50$



C. Sample Size $n = 100$



D. Sample Size $n = 300$



The results of this exercise show that as we increase the size of a random sample, the distribution of the sample means tends towards a normal distribution and the sampling error of the sample mean is reduced. This is a significant outcome and brings us to the central limit theorem concept, one of the most practically useful theorems in probability theory. It has important implications for how we construct confidence intervals and test hypotheses. Formally, the **central limit theorem** is stated as follows:

- **Central Limit Theorem.** Given a population described by any probability distribution having mean μ and finite variance σ^2 , the sampling distribution of the sample mean \bar{X} computed from random samples of size n from this population will be approximately normal with mean μ (the population mean) and variance σ^2/n (the population variance divided by n) when the sample size n is large.

Consider what the expression σ^2/n signifies. Variance (σ^2) stays the same, but as n increases, the size of the fraction decreases. This suggests that it becomes progressively less common to obtain a sample mean that is far from the true population mean with progressively larger sample sizes.

The central limit theorem allows us to make quite precise probability statements about the population mean by using the sample mean, *regardless of the population distribution* (so long as it has finite variance), because the sample mean follows an approximate normal distribution for large-size samples. The obvious question is, “When is a sample’s size large enough that we can assume the sample mean is normally distributed?” In general, when sample size n is greater than or equal to 30, we can assume that the sample mean is approximately normally distributed. When the underlying population is very non-normal, a sample size well in excess of 30 may be required for the normal distribution to be a good description of the sampling distribution of the mean.

Standard Error of the Sample Mean

The central limit theorem states that the variance of the distribution of the sample mean is σ^2/n . The positive square root of variance is standard deviation. The standard deviation of a sample statistic is known as the standard error of the statistic. The standard error of the sample mean is an important quantity in applying the central limit theorem in practice.

- **Definition of the Standard Error of the Sample Mean.** For sample mean \bar{X} calculated from a sample generated by a population with standard deviation σ , the standard error of the sample mean is given by one of two expressions:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}, \quad (1)$$

when we know σ , the population standard deviation, or by

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}, \quad (2)$$

when we do not know the population standard deviation and need to use the sample standard deviation, s , to estimate it.

In practice, we almost always need to use Equation 2. The estimate of s is given by the square root of the sample variance, s^2 , calculated as follows:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}. \quad (3)$$

Note that although the standard error is the standard deviation of the sampling distribution of the parameter, “standard deviation” in general and “standard error” are two distinct concepts, and the terms are not interchangeable.

Simply put, standard deviation measures the dispersion of the data from the mean, whereas standard error measures how much inaccuracy of a population parameter estimate comes from sampling. The contrast between standard deviation and standard error reflects the distinction between data description and inference. If we want to draw conclusions about how spread out the data are, standard deviation is the statistic to use. If we want to find out how precise the estimate of a population parameter from sampled data is relative to its true value, standard error is the statistic to use.

In another learning module we will see how the sample mean and its standard error are used in hypothesis testing to make probability statements about the population mean. To summarize, the central limit theorem tells us that when we sample from any distribution, the distribution of the sample mean will have the following properties as long as our sample size is large:

- The distribution of the sample mean \bar{X} will be approximately normal.
- The mean of the distribution of \bar{X} will be equal to the mean of the population from which the samples are drawn.

- The variance of the distribution of \bar{X} will be equal to the variance of the population divided by the sample size.

QUESTION SET



A research analyst makes two statements about repeated random sampling:

- Statement 1 When repeatedly drawing large samples from datasets, the sample means are approximately normally distributed.
- Statement 2 The underlying population from which samples are drawn must be normally distributed in order for the sample mean to be normally distributed.

1. Which of the following best describes the validity of the analyst's statements?

- A.** Statement 1 is false; Statement 2 is true.
- B.** Both statements are true.
- C.** Statement 1 is true; Statement 2 is false.

Solution:

C is correct. According to the central limit theorem, Statement 1 is true. Statement 2 is false because the underlying population does not need to be normally distributed in order for the sample mean to be normally distributed.

2. Although he knows security returns are not independent, a colleague makes the claim that because of the central limit theorem, if we diversify across a large number of investments, the portfolio standard deviation will eventually approach zero as n becomes large. Is he correct?

Solution:

No. First the conclusion on the limit of zero is wrong; second, the support cited for drawing the conclusion (i.e., the central limit theorem) is not relevant in this context.

3. Why is the central limit theorem important?

Solution:

In many instances, the distribution that describes the underlying population is not normal or the distribution is not known. The central limit theorem states that if the sample size is large, regardless of the shape of the underlying population, the distribution of the sample mean is approximately normal. Therefore, even in these instances, we can still construct confidence intervals (and conduct tests of inference) as long as the sample size is large (generally $n \geq 30$).

4. What is wrong with the following statement of the central limit theorem?

Central Limit Theorem. "If the random variables $X_1, X_2, X_3, \dots, X_n$ are a random sample of size n from any distribution with finite mean μ and variance

σ^2 , then the distribution of \bar{X} will be approximately normal, with a standard deviation of σ/\sqrt{n} .

Solution:

The statement makes the following mistakes:

- Given the conditions in the statement, the distribution of \bar{X} will be approximately normal only for large sample sizes.
- The statement omits the important element of the central limit theorem that the distribution of \bar{X} will have mean μ .

5. Peter Biggs wants to know how growth managers performed last year. Biggs assumes that the population cross-sectional standard deviation of growth manager returns is 6 percent and that the returns are independent across managers.

- A. How large a random sample does Biggs need if he wants the standard deviation of the sample means to be 1 percent?
- B. How large a random sample does Biggs need if he wants the standard deviation of the sample means to be 0.25 percent?

Solution:

A. The standard deviation or standard error of the sample mean is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Substituting in the values for $\sigma_{\bar{X}}$ and σ , we have $1\% = 6\% / \sqrt{n}$, or $\sqrt{n} = 6$. Squaring this value, we get a random sample of $n = 36$.

B. As in Part A, the standard deviation of sample mean is $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Substituting in the values for $\sigma_{\bar{X}}$ and σ , we have $0.25\% = 6\% / \sqrt{n}$, or $\sqrt{n} = 24$. Squaring this value, we get a random sample of $n = 576$, which is substantially larger than for Part A of this question.

BOOTSTRAPPING AND EMPIRICAL SAMPLING DISTRIBUTIONS

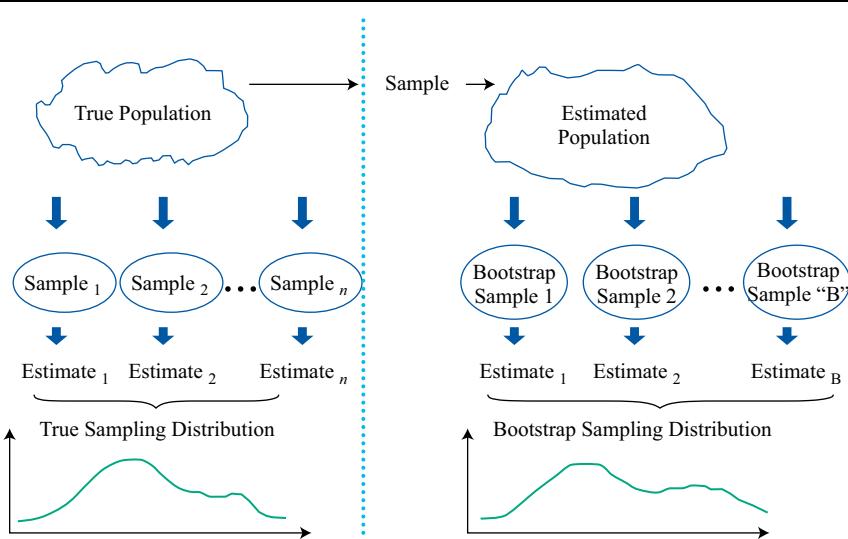
4



describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

We demonstrated how to find the standard error of the sample mean, based on the central limit theorem. We return to the computational tool called **resampling**, which repeatedly draws samples from the original observed data sample for the statistical inference of population parameters. **Bootstrap**, one of the most popular resampling methods, uses computer simulation for statistical inference without using an analytical formula such as a *z*-statistic or *t*-statistic.

In bootstrap, we repeatedly draw samples from the original sample, and each resample is of the same size as the original sample. Note that each item drawn is replaced for the next draw (i.e., the identical element is put back into the group so that it can be drawn more than once). Assuming we are looking to find the standard error of sample mean, we take many resamples and then compute the mean of each resample. Note that although some items may appear several times in the resamples, other items may not appear at all.

Exhibit 5: Bootstrap Resampling

Subsequently, we construct a sampling distribution with these resamples. The bootstrap sampling distribution (right-hand side of Exhibit 5) will approximate the true sampling distribution. We estimate the standard error of the sample mean using Equation 4. Note that to distinguish the foregoing resampling process from other types of resampling, it is often called model-free resampling or non-parametric resampling.

$$s_{\bar{X}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2}, \quad (4)$$

where:

$s_{\bar{X}}$ = the estimate of the standard error of the sample mean,

B = the number of resamples drawn from the original sample,

$\hat{\theta}_b$ = the mean of a resample, and

$\bar{\theta}$ = the mean across all the resample means.

Bootstrap is one of the most powerful and widely used tools for statistical inference. As we have explained, it can be used to estimate the standard error of a sample mean. Similarly, bootstrap can be used to find the standard error or construct confidence intervals for the statistic of other population parameters, such as the median, which does not apply to the previously discussed methodologies. Compared with conventional statistical methods, bootstrap does not rely on an analytical formula to estimate the distribution of the estimators. It is a simple but powerful method for any complicated estimators and particularly useful when no analytical formula is available. In addition, bootstrap has potential advantages in accuracy. Given these advantages, bootstrap can be applied widely in finance, such as for historical simulations in asset allocation or in gauging an investment strategy's performance against a benchmark.

EXAMPLE 4**Bootstrap Resampling Illustration**

Exhibit 6 displays a set of 12 monthly returns of a rarely traded stock, shown in Column A. Our aim is to calculate the standard error of the sample mean. Using the bootstrap resampling method, a series of bootstrap samples, labelled as “resamples” (with replacement) are drawn from the sample of 12 returns. Notice how some of the returns from data sample in Column A feature more than once in some of the resamples (e.g., 0.055 features twice in Resample 1).

Exhibit 6: Rarely Traded Stock, 12 Monthly Returns

Column A	Resample 1	Resample 2	Resample 3	Resample 1,000	
-0.096	0.055	-0.096	-0.033	...	-0.072
-0.132	-0.033	0.055	-0.132	...	0.255
-0.191	0.255	0.055	-0.157	...	0.055
-0.096	-0.033	-0.157	0.255	...	0.296
0.055	0.255	-0.096	-0.132	...	0.055
-0.053	-0.157	-0.053	-0.191	...	-0.096
-0.033	-0.053	-0.096	0.055	...	0.296
0.296	-0.191	-0.132	0.255	...	-0.132
0.055	-0.132	-0.132	0.296	...	0.055
-0.072	-0.096	0.055	-0.096	...	-0.096
0.255	0.055	-0.072	0.055	...	-0.191
-0.157	-0.157	-0.053	-0.157	...	0.055
Sample mean	-0.019	-0.060	0.001	...	0.040

Drawing 1,000 such samples, we obtain 1,000 sample means. The mean across all resample means is -0.01367 . The sum of squares of the differences between each sample mean and the mean across all resample means ($\sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2$) is 1.94143 . Using Equation 4, we calculate an estimate of the standard error of the sample mean:

$$s_{\bar{X}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2} = \sqrt{\frac{1}{999} \times 1.94143} = 0.04408.$$

Jackknife is another resampling technique for statistical inference of population parameters. Unlike bootstrap, which repeatedly draws samples with replacement, jackknife samples are selected by taking the original observed data sample and leaving out one observation at a time from the set (and not replacing it). Jackknife method is often used to reduce the bias of an estimator, and other applications include finding the standard error and confidence interval of an estimator. According to its computation procedure, we can conclude that jackknife produces similar results for every run, whereas bootstrap usually gives different results because bootstrap resamples are randomly drawn. For a sample of size n , jackknife usually requires n repetitions, whereas with bootstrap, we are left to determine how many repetitions are appropriate.

QUESTION SET

1. An analyst in a real estate investment company is researching the housing market of the Greater Boston area. From a sample of collected house sale price data in the past year, she estimates the median house price of the area. To find the standard error of the estimated median, she is considering two options:

Option 1: The standard error of the sample median can be given by $\frac{s}{\sqrt{n}}$, where s denotes the sample standard deviation and n denotes the sample size.

Option 2: Apply the bootstrap method to construct the sampling distribution of the sample median, and then compute the standard error using Equation 7.

Which of the following statements is accurate?

- A. Option 1 is suitable to find the standard error of the sample median.
- B. Option 2 is suitable to find the standard error of the sample median.
- C. Both options are suitable to find the standard error of the sample median.

Solution:

B is correct. Option 1 is valid for estimating the standard error of the sample mean but not for that of the sample median, which is not based on the given formula. Thus, both A and C are incorrect. The bootstrap method is a simple way to find the standard error of an estimator even if no analytical formula is available or it is too complicated.

Having covered many of the fundamental concepts of sampling and estimation, we now focus on sampling issues of special concern to analysts. The quality of inferences depends on the quality of the data as well as on the quality of the sampling plan used. Financial data pose special problems, and sampling plans frequently reflect one or more biases. The next section examines these issues.

2. Otema Chi has a spreadsheet with 108 monthly returns for shares in Marunou Corporation. He writes a software program that uses bootstrap resampling to create 200 resamples of this Marunou data by sampling with replacement. Each resample has 108 data points. Chi's program calculates the mean of each of the 200 resamples, and then it calculates that the mean of these 200 resample means is 0.0261. The program subtracts 0.0261 from each of the 200 resample means, squares each of these 200 differences, and adds the squared differences together. The result is 0.835. The program then calculates an estimate of the standard error of the sample mean.

The estimated standard error of the sample mean is closest to:

- A. 0.0115
- B. 0.0648

C. 0.0883

Solution:

B is correct. The estimate of the standard error of the sample mean with bootstrap resampling is calculated as follows:

$$s_X = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2} = \sqrt{\frac{1}{200-1} \sum_{b=1}^{200} (\hat{\theta}_b - 0.0261)^2} = \sqrt{0.004196}$$

$$s_X = 0.0648$$

PRACTICE PROBLEMS

1. Which one of the following statements is true about non-probability sampling?
 - A. There is significant risk that the sample is not representative of the population.
 - B. Every member of the population has an equal chance of being selected for the sample.
 - C. Using judgment guarantees that population subdivisions of interest are represented in the sample.
2. The *best* approach for creating a stratified random sample of a population involves:
 - A. drawing an equal number of simple random samples from each subpopulation.
 - B. selecting every k th member of the population until the desired sample size is reached.
 - C. drawing simple random samples from each subpopulation in sizes proportional to the relative size of each subpopulation.
3. A population has a non-normal distribution with mean μ and variance σ^2 . The sampling distribution of the sample mean computed from samples of large size from that population will have:
 - A. the same distribution as the population distribution.
 - B. its mean approximately equal to the population mean.
 - C. its variance approximately equal to the population variance.
4. A sample mean is computed from a population with a variance of 2.45. The sample size is 40. The standard error of the sample mean is *closest* to:
 - A. 0.039.
 - B. 0.247.
 - C. 0.387.
5. Compared with bootstrap resampling, jackknife resampling:
 - A. is done with replacement.
 - B. usually requires that the number of repetitions is equal to the sample size.
 - C. produces dissimilar results for every run because resamples are randomly drawn.

SOLUTIONS

1. A is correct. Because non-probability sampling is dependent on factors other than probability considerations, such as a sampler's judgment or the convenience to access data, there is a significant risk that non-probability sampling might generate a non-representative sample.
2. C is correct. Stratified random sampling involves dividing a population into subpopulations based on one or more classification criteria. Then, simple random samples are drawn from each subpopulation in sizes proportional to the relative size of each subpopulation. These samples are then pooled to form a stratified random sample.
3. B is correct. Given a population described by any probability distribution (normal or non-normal) with finite variance, the central limit theorem states that the sampling distribution of the sample mean will be approximately normal, with the mean approximately equal to the population mean, when the sample size is large.
4. B is correct. Taking the square root of the known population variance to determine the population standard deviation (σ) results in

$$\sigma = \sqrt{2.45} = 1.565.$$

The formula for the standard error of the sample mean (σ_X), based on a known sample size (n), is

$$\sigma_X = \frac{\sigma}{\sqrt{n}}.$$

Therefore,

$$\sigma_X = \frac{1.565}{\sqrt{40}} = 0.247.$$

5. B is correct. For a sample of size n , jackknife resampling usually requires n repetitions. In contrast, with bootstrap resampling, we are left to determine how many repetitions are appropriate.

LEARNING MODULE

8

Hypothesis Testing

by Pamela Peterson Drake, PhD, CFA.

Pamela Peterson Drake, PhD, CFA, is at James Madison University (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	explain hypothesis testing and its components, including statistical significance, Type I and Type II errors, and the power of a test.
<input type="checkbox"/>	construct hypothesis tests and determine their statistical significance, the associated Type I and Type II errors, and power of the test given a significance level
<input type="checkbox"/>	compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test

INTRODUCTION

1

Hypothesis testing is covered extensively in the pre-read. This learning module builds on that coverage and assumes a functional understanding of the topic gained there or elsewhere.

Lesson 1 summarizes the hypothesis testing process by exemplifying its use in finance and investment management. Lesson 2 brings forward the impact of errors in the hypothesis testing process. Lesson 3 introduces nonparametric tests and their applications in investment management.

LEARNING MODULE OVERVIEW



- The steps in testing a hypothesis are as follows:
 1. State the hypotheses.
 2. Identify the appropriate test statistic and its probability distribution.
 3. Specify the significance level.
 4. State the decision rule.
 5. Collect the data and calculate the test statistic.
 6. Make a decision.

- A test statistic is a quantity, calculated using a sample, whose value is the basis for deciding whether to reject or not reject the null hypothesis. We compare the computed value of the test statistic to a critical value for the same test statistic to determine whether to reject or not reject the null hypothesis.
- In reaching a statistical decision, two possible errors can be made: reject a true null hypothesis (a Type I error, or false positive), or fail to reject a false null hypothesis (a Type II error, or false negative).
- The level of significance of a test is the probability of a Type I error when conducting a hypothesis test. The standard approach to hypothesis testing involves specifying a level of significance (i.e., the probability of a Type I error). The complement of the level of significance is the confidence level.
- For hypothesis tests concerning the population mean of a normally distributed population with an unknown variance, the theoretically correct test statistic is the t -statistic.
- To test whether the observed difference between two means is statistically significant, the analyst must first decide whether the samples are independent or dependent (related). If the samples are independent, a test concerning differences between means is employed. If the samples are dependent, a test of mean differences (paired comparisons test) is employed.
- To determine whether the difference between two population means from normally distributed populations with unknown but equal variances, the appropriate test is a t -test based on pooling the observations of the two samples to estimate the common but unknown variance. This test is based on an assumption of independent samples.
- In tests concerning two means based on two samples that are not independent, the data are often arranged in paired observations and a test of mean differences (a paired comparisons test) is conducted. When the samples are from normally distributed populations with unknown variances, the appropriate test statistic is t -distributed.
- In tests concerning the variance of a single normally distributed population, the test statistic is chi-square with $n - 1$ degrees of freedom, where n is sample size.
- For tests concerning differences between the variances of two normally distributed populations based on two random, independent samples, the appropriate test statistic is based on an F -test (the ratio of the sample variances). The degrees of freedom for this F -test are $n_1 - 1$ and $n_2 - 1$, where n_1 corresponds to the number of observations in the calculation of the numerator, and n_2 is the number of observations in the calculation of the denominator of the F -statistic.
- A parametric test is a hypothesis test concerning a population parameter, or a hypothesis test based on specific distributional assumptions. In contrast, a nonparametric test either is not concerned with a parameter or makes minimal assumptions about the population from which the sample was taken.
- A nonparametric test is primarily used when data do not meet distributional assumptions, when there are outliers, when data are given in ranks, or when the hypothesis we are addressing does not concern a parameter.

HYPOTHESIS TESTS FOR FINANCE

2

- explain hypothesis testing and its components, including statistical significance, Type I and Type II errors, and the power of a test.

We use **hypothesis testing** to make decisions using data. Hypothesis testing is part of statistical inference, the process of making judgments about a larger group (a population) based on a smaller group of observations (a sample).

In hypothesis testing, we test to determine whether a sample statistic is likely to come from a population with the hypothesized value of the population parameter.

The concepts and tools of hypothesis testing provide an objective means to gauge whether the available evidence supports the hypothesis. After applying a statistical test, we should have a clearer idea of the probability that a hypothesis is true or not, although our conclusion always stops short of certainty.

The main focus of this lesson is on the framework of hypothesis testing and tests concerning mean and variance, two measures frequently used in investments.

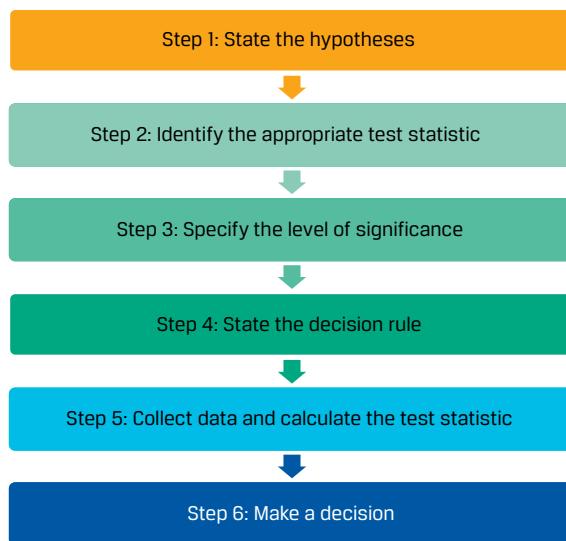
The Process of Hypothesis Testing

Hypothesis testing is part of the branch of statistics known as statistical inference. In statistical inference, there is estimation and hypothesis testing. Estimation involves point estimates and interval estimates. Consider a sample mean, which is a point estimate, that we can use to form a confidence interval. In hypothesis testing, the focus is examining how a sample statistic informs us about a population parameter. A **hypothesis** is a statement about one or more populations that we test using sample statistics.

The process of hypothesis testing begins with the formulation of a theory to organize and explain observations. We judge the correctness of the theory by its ability to make accurate predictions—for example, to predict the results of new observations. If the predictions are correct, we continue to maintain the theory as a possibly correct explanation of our observations. Risk plays a role in the outcomes of observations in finance, so we can only try to make unbiased, probability-based judgments about whether the new data support the predictions. Statistical hypothesis testing fills that key role of testing hypotheses when there is uncertainty. When an analyst correctly formulates the question into a testable hypothesis and carries out a test of hypotheses, the use of well-established scientific methods supports the conclusions and decisions made on the basis of this test.

We organize this introduction to hypothesis testing around the six steps presented in Exhibit 1, which illustrate the standard approach to hypothesis testing.

Exhibit 1: The Process of Hypothesis Testing



Stating the Hypotheses

For each hypothesis test, we always state two hypotheses: the **null hypothesis** (or null), designated H_0 , and the **alternative hypothesis**, designated H_a . The null hypothesis is a statement concerning a population parameter or parameters considered to be true unless the sample we use to conduct the hypothesis test gives convincing evidence that the null hypothesis is false. In fact, the null hypothesis is what we want to reject. If there is sufficient evidence to indicate that the null hypothesis is not true, we reject it in favor of the alternative hypothesis.

Importantly, the null and alternative hypotheses are stated in terms of population parameters, and we use sample statistics to test these hypotheses.

Second, the null and alternative hypotheses must be mutually exclusive and collectively exhaustive; in other words, all possible values are contained in either the null or the alternative hypothesis.

Identify the Appropriate Test Statistic and Distribution

A test statistic is a value calculated on the basis of a sample that, when used in conjunction with a decision rule, is the basis for deciding whether to reject the null hypothesis.

The focal point of our statistical decision is the value of the test statistic. The test statistic that we use depends on what we are testing.

Following the identification of the appropriate test statistic, we must be concerned with the distribution of the test statistic. We show examples of test statistics, and their corresponding distributions, in Exhibit 2.

Exhibit 2: Test Statistics and Their Distribution

What We Want to Test	Test Statistic	Probability Distribution of the Statistic	Degrees of Freedom
Test of a single mean	$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	<i>t</i> -distributed	$n - 1$
Test of the difference in means	$t = \frac{(\bar{X}_{d1} - \bar{X}_{d2}) - (\mu_{d1} - \mu_{d2})}{\sqrt{\frac{s_p^2}{n_{d1}} + \frac{s_p^2}{n_{d2}}}}$	<i>t</i> -distributed	$n_1 + n_2 - 2$
Test of the mean of differences	$t = \frac{\bar{d} - \mu_{d0}}{\frac{s_d}{\sqrt{n_d}}}$	<i>t</i> -distributed	$n - 1$
Test of a single variance	$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2}$	Chi-square distributed	$n - 1$
Test of the difference in variances	$F = \frac{s_{Before}^2}{s_{After}^2}$	<i>F</i> -distributed	$n_1 - 1, n_2 - 1$
Test of a correlation	$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}$	<i>t-Distributed</i>	$n - 2$
Test of independence (categorical data)	$\chi^2 = \sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$	<i>Chi-square distributed</i>	$(r - 1)(c - 1)$

Note: μ_0 , μ_{d0} , and σ_0^2 denote hypothesized values of the mean, mean difference, and variance, respectively. The \bar{x} , \bar{d} , s^2 , s , and r denote for a sample the mean, mean of the differences, variance, standard deviation, and correlation, respectively, with subscripts indicating the sample, if appropriate. The sample size is indicated as n , and the subscript indicates the sample, if appropriate. O_{ij} and E_{ij} are observed and expected frequencies, respectively, with r indicating the number of rows and c indicating the number of columns in the contingency table.

Specify the Level of Significance

The level of significance reflects how much sample evidence we require to reject the null hypothesis. The required standard of proof can change according to the nature of the hypotheses and the seriousness of the consequences of making a mistake.

There are four possible outcomes when we test a null hypothesis, as shown in Exhibit 3. A **Type I error** is a false positive (reject when the null is true), whereas a **Type II error** is a false negative (fail to reject when the null is false).

Exhibit 3: Correct and Incorrect Decisions in Hypothesis Testing

Decision	H_0 True	H_0 False
Fail to reject H_0	Correct decision: Do not reject a true null hypothesis.	Type II error: Fail to reject a false null hypothesis. False negative
	Type I error: Reject a true null hypothesis.	Correct decision: Reject a false null hypothesis.
Reject H_0	False positive	

When we make a decision in a hypothesis test, we run the risk of making either a Type I or a Type II error. As shown in Exhibit 3, these errors are mutually exclusive: If we mistakenly reject the true null, we can only be making a Type I error; if we mistakenly fail to reject the false null, we can only be making a Type II error.

The probability of a Type I error is denoted by the lowercase Greek letter alpha, α . This probability is also known as the **level of significance** of the test, and its complement, $(1 - \alpha)$, is the **confidence level**. For example, a level of significance of 5 percent for a test means that there is a 5 percent probability of rejecting a true null hypothesis and corresponds to the 95 percent confidence level.

Controlling the probabilities of the two types of errors involves a trade-off. All else equal, if we decrease the probability of a Type I error by specifying a smaller significance level (say, 1 percent rather than 5 percent), we increase the probability of making a Type II error because we will reject the null less frequently, including when it is false. Both Type I and Type II errors are risks of being wrong. Whether to accept more of one type versus the other depends on the consequences of the errors, such as costs. This trade-off weighs the impact of errors we are willing to accept and if so, at what cost. The only way to reduce the probabilities of both types of errors simultaneously is to increase the sample size, n .

Whereas the significance level of a test is the probability of incorrectly rejecting the true null, the **power of a test** is the probability of *correctly* rejecting the null—that is, the probability of rejecting the null when it is false. The power of a test is, in fact, the complement of the Type II error. The probability of a Type II error is often denoted by the lowercase Greek letter beta, β . We can classify the different probabilities in Exhibit 4 to reflect the notation that is often used.

Exhibit 4: Probabilities Associated with Hypothesis Testing Decisions

Decision	H_0 True	H_0 False
Fail to reject H_0	$1 - \alpha$	β
Reject H_0	α	$1 - \beta$

State the Decision Rule

The fourth step in hypothesis testing is stating the decision rule: When do we reject the null hypothesis, and when do we not? The action we take is based on comparing the calculated sample test statistic with a specified value or values, which are referred to as **critical values**.

The critical value or values we choose are based on the level of significance and the probability distribution associated with the test statistic. If we find that the calculated value of the test statistic is more extreme than the critical value or values, then we reject the null hypothesis; we say the result is **statistically significant**. Otherwise, we fail to reject the null hypothesis; there is not sufficient evidence to reject the null hypothesis. Recall that the smallest level of significance at which the null hypothesis can be rejected is the **p-value**, the area in the probability distribution outside the calculated test statistic.

QUESTION SET



- Willco is a manufacturer in a mature cyclical industry. During the most recent industry cycle, its net income averaged USD30 million per year with a standard deviation of USD10 million ($n = 6$ observations). Management claims that Willco's performance during the most recent cycle results

from new approaches and that Willco's profitability will exceed the average of USD24 million per year observed in prior cycles.

- A. With μ as the population value of mean annual net income, formulate null and alternative hypotheses consistent with testing Willco management's claim.
- B. Assuming that Willco's net income is at least approximately normally distributed, identify the appropriate test statistic and calculate the degrees of freedom.
- C. Based on a critical value of 2.015, determine whether to reject the null hypothesis.

Solution:

- A. We often set up the "hoped for" or "suspected" condition as the alternative hypothesis. Here, that condition is that the population value of Willco's mean annual net income exceeds USD24 million. Thus, we have $H_0: \mu \leq 24$ versus $H_a: \mu > 24$.
- B. Given that net income is normally distributed with unknown variance, the appropriate test statistic is a t-statistic with $n - 1 = 6 - 1 = 5$ degrees of freedom.
- C. We reject the null if the calculated t-statistic is greater than 2.015. The calculated t-statistic is

$$t = \frac{30 - 24}{10/\sqrt{6}} = 1.4697.$$

- D. Because the calculated test statistic does not exceed 2.015, we fail to reject the null hypothesis. There is not sufficient evidence to indicate that the mean net income is greater than USD24 million.
2. All else equal, is specifying a smaller significance level in a hypothesis test likely to increase the probability of a:

Type I error?	Type II error?
A. No	No
B. No	Yes
C. Yes	No

Solution:

B is correct. Specifying a smaller significance level decreases the probability of a Type I error (rejecting a true null hypothesis) but increases the probability of a Type II error (not rejecting a false null hypothesis). As the level of significance decreases, the null hypothesis is less frequently rejected.

3. For each of the following hypothesis tests concerning the population mean, μ , state the conclusion regarding the test of the hypotheses.
- A. $H_0: \mu = 10$ versus $H_a: \mu \neq 10$, with a calculated t-statistic of 2.05 and critical t-values of ± 1.984 .
 - B. $H_0: \mu \leq 10$ versus $H_a: \mu > 10$, with a calculated t-statistic of 2.35 and a critical t-value of +1.679
 - C. $H_0: \mu = 10$ versus $H_a: \mu \neq 10$, with a calculated t-statistic of 2.05, a p-value of 4.6352%, and a level of significance of 5%.

- D. $H_0: \mu \leq 10$ versus $H_a: \mu > 10$, with a 2% level of significance and a calculated test statistic with a p -value of 3%.

Solution:

We make the decision either by comparing the calculated test statistic with the critical values or by comparing the p -value for the calculated test statistic with the level of significance.

- A. Reject the null hypothesis because the calculated test statistic of 2.05 is outside the bounds of the critical values of ± 1.984 .
- B. Reject the null hypothesis because the calculated test statistic of 2.35 is outside the bounds of the critical value of +1.679.
- C. The p -value corresponding to the calculated test statistic of 4.6352% is less than the level of significance (5%), so we reject the null hypothesis.
- D. We fail to reject because the p -value for the calculated test statistic of 3% is greater than the level of significance (2%).

3

TESTS OF RETURN AND RISK IN FINANCE

- construct hypothesis tests and determine their statistical significance, the associated Type I and Type II errors, and power of the test given a significance level

Hypothesis tests concerning return and risk are among the most common in finance. The sampling distribution of the mean, when the population standard deviation is unknown, is t -distributed, and when the population standard deviation is known, it is normally distributed, or z -distributed. Since the population standard deviation is unknown in almost all cases, we will focus on the use of a t -distributed test statistic.

TEST OF A SINGLE MEAN: RISK AND RETURN CHARACTERISTICS OF AN EQUITY MUTUAL FUND

Suppose you are analyzing Sendar Equity Fund,. During the past 24 months, it has achieved a mean monthly return of 1.50%, with a sample standard deviation of monthly returns of 3.60 percent. Given its level of market risk and according to a pricing model, this mutual fund was expected to have earned a 1.10 percent mean monthly return during that time period.

Assuming returns are normally distributed, are the actual results consistent with a population mean monthly return of 1.10 percent?

Formulate and test a hypothesis that the fund's performance was different than the mean return of 1.1 percent inferred from the pricing model. Use a 5 percent level of significance.

Exhibit 5: Test of a Single Mean

Step 1	State the hypotheses.	$H_0: \mu = 1.1\%$ versus $H_a: \mu \neq 1.1\%$
Step 2	Identify the appropriate test statistic.	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ with $24 - 1 = 23$ degrees of freedom.

Step 3	Specify the level of significance.	$\alpha = 5\%$ (two tailed).
Step 4	State the decision rule.	Critical t -values = ± 2.069 . Reject the null if the calculated t -statistic is less than -2.069 , and reject the null if the calculated t -statistic is greater than $+2.069$.
		Excel
		<i>Lower: T.INV(0.025, 23)</i>
		<i>Upper: T.INV(0.975, 23)</i>
		R qt(c(.025, .975), 23)
		Python from scipy.stats import t <i>Lower: t.ppf(.025, 23)</i> <i>Upper: t.ppf(.975, 23)</i>
Step 5	Calculate the test statistic.	$t = \frac{1.5 - 1.1}{3.6/\sqrt{24}} = 0.54433$
Step 6	Make a decision.	Fail to reject the null hypothesis because the calculated t -statistic falls between the two critical values. There is not sufficient evidence to indicate that the population mean monthly return is different from 1.10%.

Test the hypothesis using the 95 percent confidence interval.

The 95 percent confidence interval is $\bar{X} \pm \text{Critical value}\left(\frac{s}{\sqrt{n}}\right)$, so

$$\begin{aligned} & \{1.5 - 2.069(3.6/\sqrt{24}), 1.5 + 2.069(3.6/\sqrt{24})\} \\ & \{1.5 - 1.5204, 1.5 + 1.5204\} \\ & \{-0.0204, 3.0204\} \end{aligned}$$

The hypothesized value of 1.1 percent is within the bounds of the 95 percent confidence interval, so we fail to reject the null hypothesis.

TEST OF A SINGLE VARIANCE: RISK CHARACTERISTICS OF AN EQUITY MUTUAL FUND

Suppose we want to use the observed sample variance of the fund to test whether the true variance of the fund is less than some trigger level, say 4 percent. Performing a test of a population variance requires specifying the hypothesized value of the variance. We can formulate hypotheses concerning whether the variance is equal to a specific value or whether it is greater than or less than a hypothesized value:

One-sided alternative (left tail): $H_0: \sigma^2 \geq \sigma_0^2$ versus $H_a: \sigma^2 < \sigma_0^2$.

Note that the fund's variance is less than the trigger level $\sigma_0^2 = 4\%$ if the null hypothesis is rejected in favor of the alternative hypothesis. In tests concerning the variance of a single normally distributed population, we make use of a chi-square test statistic, denoted χ^2 .

You continue with your analysis of Sendar Equity Fund, a midcap growth fund that has been in existence for only 24 months. During this period, Sendar Equity achieved a mean monthly return of 1.50 percent and a standard deviation of monthly returns of 3.60 percent. Using a 5 percent level of significance, test whether the standard deviation of returns is less than 4 percent. Recall that the standard deviation is the square root of the variance, hence a standard deviation of 4 percent or 0.04, is a variance of 0.0016.

Exhibit 6: Test of Single Variance

Step 1	State the hypotheses.	$H_0: \sigma^2 \geq 16$ versus $H_a: \sigma^2 < 16$
Step 2	Identify the appropriate test statistic.	$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2}$
Step 3	Specify the level of significance.	5%
Step 4	State the decision rule.	With $24 - 1 = 23$ degrees of freedom, the critical value is 13.09051. We reject the null hypothesis if the calculated χ^2 statistic is less than 13.09051. Excel CHISQ.INV(0.05, 23) R qchisq(.05, 23) Python <pre>from scipy.stats import chi2 chi2.ppf(.05, 23)</pre>
Step 5	Calculate the test statistic.	$\chi^2 = \frac{(24 - 1)12.96}{16} = 18.63000$
Step 6	Make a decision.	Fail to reject the null hypothesis because the calculated χ^2 statistic is greater than the critical value. There is insufficient evidence to indicate that the variance is less than 16% (or, equivalently, that the standard deviation is less than 4%).

TEST CONCERNING DIFFERENCES BETWEEN MEANS WITH INDEPENDENT SAMPLES

We often want to know whether a mean value—for example, a mean return—differs for two groups. Is an observed difference due to chance or to different underlying values for the mean? We test this by drawing a sample from each group. When it is reasonable to believe that the samples are from populations that are approximately normally distributed and that the samples are also independent of each other, we use the test of the difference in the means. We may assume that population variances are equal or unequal. However, our focus in discussing the test of the difference of means is using the assumption that the population variances are equal. In the calculation of the test statistic, we combine the observations from both samples to obtain a pooled estimate of the common population variance.

Let μ_1 and μ_2 represent, respectively, the population means of the first and second populations, respectively. Most often we want to test whether the population means are equal or whether one is larger than the other. Thus, we formulate the following hypotheses:

$$H_0: \mu_1 - \mu_2 = 0 \text{ versus } H_a: \mu_1 - \mu_2 \neq 0.$$

EXAMPLE 1**Comparison of Returns on the ACE High Yield Index for Two Periods**

Suppose we want to test whether the returns of the ACE High Yield Total Return Index, shown in Exhibit 7, are different for two different time periods, Period 1 and Period 2.

Exhibit 7: Descriptive Statistics for ACE High Yield Total Return Index for Periods 1 and 2

	Period 1	Period 2
Mean	0.01775%	0.01134%
Standard deviation	0.31580%	0.38760%
Sample size	445 days	859 days

Note that these periods are of different lengths and the samples are independent; that is, there is no pairing of the days for the two periods.

1. Is there a difference between the mean daily returns in Period 1 and in Period 2, using a 5% level of significance?

Solution:

Step 1	State the hypotheses.	$H_0: \mu_{Period1} = \mu_{Period2}$ versus $H_a: \mu_{Period1} \neq \mu_{Period2}$
Step 2	Identify the appropriate test statistic.	$t = \frac{(\bar{X}_{Period1} - \bar{X}_{Period2}) - (\mu_{Period1} - \mu_{Period2})}{\sqrt{\frac{s_p^2}{n_{Period1}} + \frac{s_p^2}{n_{Period2}}}},$ where $s_p^2 = \frac{(n_{Period1} - 1)s_{Period1}^2 + (n_{Period2} - 1)s_{Period2}^2}{n_{Period1} + n_{Period2} - 2}$ with $445 + 859 - 2 = 1,302$ degrees of freedom.
Step 3	Specify the level of significance.	$\alpha = 5\%$
Step 4	State the decision rule.	Critical t -values = ± 1.962 Reject the null if the calculated t -statistic is less than -1.962 , and reject the null if the calculated t -statistic is greater than $+1.962$.
		Excel
		<i>Lower: T.INV(0.025, 1302)</i>
		<i>Upper: T.INV(0.975, 1302)</i>
		R <i>qt(c(.025, .975), 1302)</i>
		Python <i>from scipy.stats import t</i>
		<i>Lower: t.ppf(.025, 1302)</i>
		<i>Upper: t.ppf(.975, 1302)</i>
Step 5	Calculate the test statistic.	$s_p^2 = \frac{(445 - 1)0.09973 + (859 - 1)0.15023}{445 + 859 - 2} = 0.1330$ $t = \frac{(0.01775 - 0.01134) - 0}{\sqrt{\frac{0.1330}{445} + \frac{0.1330}{859}}} = \frac{0.0064}{0.0213} = 0.3009.$
Step 6	Make a decision.	Fail to reject the null because the calculated t -statistic falls within the bounds of the two critical values. We conclude that there is insufficient evidence to indicate that the returns are different for the two time periods.

Test Concerning Differences between Means with Dependent Samples

When we compare two independent samples, we use a t -distributed test statistic that uses the difference in the means and a pooled variance. An assumption for the validity of those tests is that the samples are independent—that is, unrelated to each other. When we want to conduct tests on two means based on samples that we believe are dependent, we use the **test of the mean of the differences** (a paired comparisons test).

How is this test of paired differences different from the test of the difference in means in independent samples? The test of paired comparisons is more powerful than the test of the difference in the means because by using the common element (such as the same periods or companies), we eliminate the variation between the samples that could be caused by something other than what we are testing.

EXAMPLE 2

A Comparison of the Returns of Two Indexes

Suppose we want to compare the returns of the ACE High Yield Index with those of the ACE BBB Index. We collect data over the same 1,304 days for both indexes and calculate their means and standard deviations as shown in Exhibit 8.

Exhibit 8: Mean and Standard Deviations for the ACE High Yield Index and the ACE BBB Index

	ACE High Yield Index (%)	ACE BBB Index (%)	Difference (%)
Mean return	0.0157	0.0135	-0.0021
Standard deviation	0.3157	0.3645	0.3622

- Using a 5 percent level of significance, is the mean of the differences is different from zero?

Solution:

Step 1	State the hypotheses.	$H_0: \mu_{d0} = 0$ versus $H_a: \mu_{d0} \neq 0$
Step 2	Identify the appropriate test statistic.	$t = \frac{\bar{d} - \mu_{d0}}{s_d}$
Step 3	Specify the level of significance.	5%
Step 4	State the decision rule.	With $1,304 - 1 = 1,303$ degrees of freedom, the critical values are ± 1.962 . We reject the null hypothesis if the calculated t -statistic is less than -1.962 or greater than $+1.962$.

Excel

Lower: T.INV(0.025, 1303)

Upper: T.INV(0.975, 1303)

R *qt(c(.025, .975), 1303)*

Python *from scipy.stats import t*

Lower: t.ppf(.025, 1303)

Upper: t.ppf(.975, 1303)

Step 5	Calculate the test statistic.	$\bar{d} = -0.0021\%$ $s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{0.3622}{\sqrt{1,304}} = 0.01003\%$ $t = \frac{-0.00210 - 0}{0.01003} = -0.20937$
Step 6	Make a decision.	Fail to reject the null hypothesis because the calculated t -statistic falls within the bounds of the two critical values. There is insufficient evidence to indicate that the mean of the differences of returns is different from zero.C

Test Concerning the Equality of Two Variances

There are many instances in which we want to compare the volatility of two samples, in which case we can test for the equality of two variances. Examples include comparisons of baskets of securities against indexes or benchmarks, as well as comparisons of volatility in different periods.

EXAMPLE 3

Volatility and Regulation

You are investigating whether the population variance of returns on a stock market index changed after a change in market regulation. The first 418 weeks occurred before the regulation change, and the second 418 weeks occurred after the regulation change. You gather the data in Exhibit 9 for 418 weeks of returns both before and after the change in regulation. You have specified a 5 percent level of significance.

Exhibit 9: Index Returns and Variances before and after the Market Regulation Change

	N	Mean Weekly Return (%)	Variance of Returns
Before regulation change	418	0.250	4.644
After regulation change	418	0.110	3.919

- Are the variance of returns different before the regulation change versus after the regulation change?

Solution:

Step 1	State the hypotheses.	$H_0: \sigma_{Before}^2 = \sigma_{After}^2$ versus $H_a: \sigma_{Before}^2 \neq \sigma_{After}^2$
Step 2	Identify the appropriate test statistic.	$F = \frac{s_{Before}^2}{s_{After}^2}$
Step 3	Specify the level of significance.	5%

Step 4	State the decision rule.	With $418 - 1 = 417$ and $418 - 1 = 417$ degrees of freedom, the critical values are 0.82512 and 1.21194. Reject the null if the calculated F -statistic is less than 0.82512 or greater than 1.21194.
Excel		<i>Left side: F.INV(0.025, 417, 417)</i> <i>Right side: F.INV(0.975, 417, 417)</i>
R		<i>qf(c(.025, .975), 417, 417)</i>
Python		<i>from scipy.stats import f</i> <i>Left side: f.ppf(.025, 417, 417)</i> <i>Right side: f.ppf(.975, 417, 417)</i>
Step 5	Calculate the test statistic.	$F = \frac{4.644}{3.919} = 1.18500$
Step 6	Make a decision.	Fail to reject the null hypothesis since the calculated F -statistic falls within the bounds of the two critical values. There is not sufficient evidence to indicate that the weekly variances of returns are different in the periods before and after the regulation change.

2. Is the variance of returns greater before the regulation change versus after the regulation change?

Solution:

Step 1	State the hypotheses.	$H_0: \sigma_{Before}^2 \leq \sigma_{After}^2$ versus $H_a: \sigma_{Before}^2 > \sigma_{After}^2$
Step 2	Identify the appropriate test statistic.	$F = \frac{s_{Before}^2}{s_{After}^2}$
Step 3	Specify the level of significance.	5%
Step 4	State the decision rule.	With $418 - 1 = 417$ and $418 - 1 = 417$ degrees of freedom, the critical value is 1.17502. We reject the null hypothesis if the calculated F -statistic is greater than 1.17502.
		Excel <i>F.INV(0.95, 417, 417)</i> R <i>qf(.95, 417, 417)</i>
		Python <i>from scipy.stats import f</i> <i>f.ppf(.95, 417, 417)</i>
Step 5	Calculate the test statistic.	$F = \frac{4.644}{3.919} = 1.18500$
Step 6	Make a decision.	Reject the null hypothesis since the calculated F -statistic is greater than 1.17502. There is sufficient evidence to indicate that the weekly variances of returns before the regulation change are greater than the variances after the regulation change.

QUESTION SET



Investment analysts often use earnings per share (EPS) forecasts. One test of forecasting quality is the zero-mean test, which states that optimal forecasts should have a mean forecasting error of zero. The forecasting error is the difference between the predicted value of a variable and the actual value of the variable.

You have collected data (shown in Exhibit 10) for two analysts who cover two different industries: Analyst A covers the telecom industry; Analyst B covers automotive parts and suppliers.

Exhibit 10: Test of Return and Risk
Performance in Forecasting Quarterly Earnings per Share

	Number of Forecasts	Mean Forecast Error (Predicted – Actual)	Standard Deviation of Forecast Errors
Analyst A	10	0.05	0.10
Analyst B	15	0.02	0.09

Critical t -values:

Degrees of Freedom	Area in the Right-Side Rejection Area	
	$p = 0.05$	$p = 0.025$
8	1.860	2.306
9	1.833	2.262
10	1.812	2.228
11	1.796	2.201
12	1.782	2.179
13	1.771	2.160
14	1.761	2.145
15	1.753	2.131
16	1.746	2.120
17	1.740	2.110
18	1.734	2.101
19	1.729	2.093
20	1.725	2.086
21	1.721	2.080
22	1.717	2.074
23	1.714	2.069
24	1.711	2.064
25	1.708	2.060
26	1.706	2.056
27	1.703	2.052

- With μ as the population mean forecasting error, formulate null and alternative hypotheses for a zero-mean test of forecasting quality.

- For Analyst A, determine whether to reject the null at the 0.05 level of significance.

- B.** For Analyst B, determine whether to reject the null at the 0.05 level of significance.

Solution:

$H_0: \mu = 0$ versus $H_a: \mu \neq 0$.

- A.** The t-test is based on $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ with $n - 1 = 10 - 1 = 9$ degrees of freedom. At the 0.05 significance level, we reject the null if the calculated t-statistic is outside the bounds of ± 2.262 (from the table for 9 degrees of freedom and 0.025 in the right side of the distribution). For Analyst A, we have a calculated test statistic of

$$t = \frac{0.05 - 0}{0.10/\sqrt{10}} = 1.58114.$$

We, therefore, fail to reject the null hypothesis at the 0.05 level.

- B.** For Analyst B, the t-test is based on t with $15 - 1 = 14$ degrees of freedom. At the 0.05 significance level, we reject the null if the calculated t-statistic is outside the bounds of ± 2.145 (from the table for 14 degrees of freedom). The calculated test statistic is

$$t = \frac{0.02 - 0}{0.09/\sqrt{15}} = 0.86066.$$

Because 0.86066 is within the range of ± 2.145 , we fail to reject the null at the 0.05 level.

2. Reviewing the EPS forecasting performance data for Analysts A and B, you want to investigate whether the larger average forecast errors of Analyst A relative to Analyst B are due to chance or to a higher underlying mean value for Analyst A. Assume that the forecast errors of both analysts are normally distributed and that the samples are independent.

- A.** Formulate null and alternative hypotheses consistent with determining whether the population mean value of Analyst A's forecast errors (μ_1) are larger than Analyst B's (μ_2).
- B.** Identify the test statistic for conducting a test of the null hypothesis formulated in Part A.
- C.** Identify the rejection point or points for the hypotheses tested in Part A at the 0.05 level of significance.
- D.** Determine whether to reject the null hypothesis at the 0.05 level of significance.

Solution:

- A.** Stating the suspected condition as the alternative hypothesis, we have

$H_0: \mu_A - \mu_B \leq 0$ versus $H_a: \mu_A - \mu_B > 0$,

where

μ_A = the population mean value of Analyst A's forecast errors

μ_B = the population mean value of Analyst B's forecast errors

- B.** We have two normally distributed populations with unknown variances. Based on the samples, it is reasonable to assume that the population variances are equal. The samples are assumed to be independent; this assumption is reasonable because the analysts cover different industries. The appropriate test statistic is t using a pooled estimate of the common variance:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}, \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The number of degrees of freedom is $n_A + n_B - 2 = 10 + 15 - 2 = 23$.

- C. For $df = 23$, according to the table, the rejection point for a one-sided (right side) test at the 0.05 significance level is 1.714.

- D. We first calculate the pooled estimate of variance:

$$s_p^2 = \frac{(10 - 1)0.01 + (15 - 1)0.0081}{10 + 15 - 2} = 0.0088435.$$

We then calculate the t -distributed test statistic:

$$t = \frac{(0.05 - 0.02) - 0}{\sqrt{\frac{0.0088435}{10} + \frac{0.0088435}{15}}} = \frac{0.03}{0.0383916} = 0.78142.$$

Because $0.78142 < 1.714$, we fail to reject the null hypothesis. There is not sufficient evidence to indicate that the mean for Analyst A exceeds that for Analyst B.

An investment consultant gathers two independent random samples of five-year performance data for US and European absolute return hedge funds. Noting a return advantage of 50 bps for US managers, the consultant decides to test whether the two means are different from one another at a 0.05 level of significance. The two populations are assumed to be normally distributed with unknown but equal variances. Results of the hypothesis test are contained in the Exhibit 11.

Exhibit 11: Hypothesis Test Results

	Sample Size	Mean Return (%)	Standard Deviation
US managers	50	4.7	5.4
European managers	50	4.2	4.8
Null and alternative hypotheses		$H_0: \mu_{US} - \mu_E = 0; H_a: \mu_{US} - \mu_E \neq 0$	
Calculated test statistic		0.4893	
Critical value rejection points		± 1.984	

Note: The mean return for US funds is μ_{US} , and μ_E is the mean return for European funds.

3. The consultant should conclude that the:

- A. null hypothesis is not rejected.
- B. alternative hypothesis is statistically confirmed.
- C. difference in mean returns is statistically different from zero.

Solution:

A is correct. The calculated t -statistic value of 0.4893 falls within the bounds of the critical t -values of ± 1.984 . Thus, H_0 cannot be rejected; the result is not statistically significant at the 0.05 level.

4. During a 10-year period, the standard deviation of annual returns on a portfolio you are analyzing was 15 percent a year. You want to see whether this record is sufficient evidence to support the conclusion that the portfolio's

underlying variance of return was less than 400, the return variance of the portfolio's benchmark.

- A. Formulate null and alternative hypotheses consistent with your objective.
- B. Identify the test statistic for conducting a test of the hypotheses in Part A, and calculate the degrees of freedom.
- C. Determine whether the null hypothesis is rejected or not rejected at the 0.05 level of significance using a critical value of 3.325.

Solution:

- A. We have a “less than” alternative hypothesis, where σ^2 is the variance of return on the portfolio. The hypotheses are $H_0: \sigma^2 \geq 400$ versus $H_a: \sigma^2 < 400$, where 400 is the hypothesized value of variance. This means that the rejection region is on the left side of the distribution.
- B. The test statistic is chi-square distributed with $10 - 1 = 9$ degrees of freedom.
- C. The test statistic is calculated as

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{9 \times 15^2}{400} = \frac{2,025}{400} = 5.0625, \text{ or } 5.06.$$

Because 5.06 is not less than 3.325, we do not reject the null hypothesis; the calculated test statistic falls to the right of the critical value, where the critical value separates the left-side rejection region from the region where we fail to reject.

We can determine the critical value for this test using software:

Excel [CHISQ.INV(0.05, 9)]

[qchisq(.05, 9)]

Python

[from scipy.stats import chi2 and chi2.ppf(.05, 9)]

We can determine the *p*-value for the calculated test statistic of 17.0953 using software:

Excel [CHISQ.DIST(5.06, 9, TRUE)]

[pchisq(5.06, 9, lower.tail=TRUE)]

Python

[from scipy.stats import chi2 and chi2.cdf(5.06, 9)]

5. You are investigating whether the population variance of returns on an index changed subsequent to a market disruption. You gather the following

data for 120 months of returns before the disruption and for 120 months of returns after the disruption. You have specified a 0.05 level of significance.

Exhibit 12: Data for 120 Months of Returns

Time Period	N	Mean Monthly Return (%)	Variance of Returns
Before disruption	120	1.416	22.367
After disruption	120	1.436	15.795

- A. Formulate null and alternative hypotheses consistent with the research goal.
- B. Identify the test statistic for conducting a test of the hypotheses in Part A, and calculate the degrees of freedom.
- C. Determine whether to reject the null hypothesis at the 0.05 level of significance if the critical values are 0.6969 and 1.4349.

Solution:

- A. We have a “not equal to” alternative hypothesis:

$$H_0 : \sigma_{\text{Before}}^2 = \sigma_{\text{After}}^2 \text{ versus } H_a : \sigma_{\text{Before}}^2 \neq \sigma_{\text{After}}^2$$

- B. To test a null hypothesis of the equality of two variances, we use an F-test:

$$F = \frac{s_1^2}{s_2^2}$$

$$F = 22.367/15.795 = 1.416, \text{ with } 120 - 1 = 119 \text{ numerator and } 120 - 1 = 119 \text{ denominator degrees of freedom.}$$

- C. Because this is a two-tailed test, we use critical values for the $0.05/2 = 0.025$ level. The calculated test statistic falls within the bounds of the critical values (i.e., between 0.6969 and 1.4349), so we fail to reject the null hypothesis; there is not enough evidence to indicate that the variances are different before and after the disruption. Note that we could also have formed the F-statistic as $15.795/22.367 = 0.706$ and draw the same conclusion.

We could also use software to calculate the critical values:

Excel

```
[F.INV(0.025, 119, 119) and  
F.INV(0.975, 119, 119)]  
[qf(c(.025, .975), 119, 119)]
```

Python

```
from scipy.stats import f and f.ppf  
[(.025, 119, 119) and  
f.ppf(.975, 119, 119)]
```

Additionally, we could use software to calculate the *p*-value of the calculated test statistic, which is 5.896 percent, and then compare it with the level of significance:

Excel

```
[ (1-F.DIST(22.367/15.796,119,119,TRUE)) *2 or  
F.DIST(15.796/22.367,119,119,TRUE) *2]
```

R

```
[ (1-pf(22.367/15.796,119,119)) *2 or  
pf(15.796/22.367,119,119) *2]
```

Python

```
from scipy.stats import f and f.cdf  
[(15.796/22.367,119,119)*2 or  
(1-f.cdf(22.367/15.796,119,119))*2]
```

4**PARAMETRIC VERSUS NONPARAMETRIC TESTS**

compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test

The hypothesis testing procedures we have discussed up to this point have two characteristics in common. First, they are concerned with parameters, and second, their validity depends on a definite set of assumptions. Mean and variance, for example, are two parameters, or defining quantities, of a normal distribution. The tests also make specific assumptions—in particular, assumptions about the distribution of the population producing the sample. Any test or procedure with either of these two characteristics is a **parametric test** or procedure.

In some cases, however, we are concerned about quantities other than parameters of distributions. In other cases, we may believe that the assumptions of parametric tests do not hold. In cases where we are examining quantities other than population parameters or where assumptions of the parameters are not satisfied, a nonparametric test or procedure can be useful.

A **nonparametric test** is a test that is not concerned with a parameter or a test that makes minimal assumptions about the population from which the sample comes. Exhibit 13 presents examples of nonparametric alternatives to the parametric, *t*-distributed tests concerning means.

Exhibit 13: Nonparametric Alternatives to Parametric Tests Concerning Means

	Parametric	Nonparametric
Tests concerning a single mean	<i>t</i> -distributed test <i>z</i> -distributed test	Wilcoxon signed-rank test

Tests concerning differences between means	<i>t</i> -distributed test	Mann–Whitney <i>U</i> test (Wilcoxon rank sum test)
Tests concerning mean differences (paired comparisons tests)	<i>t</i> -distributed test	Wilcoxon signed-rank test Sign test

Uses of Nonparametric Tests

Nonparametric procedures are primarily used in four situations: (1) when the data do not meet distributional assumptions, (2) when there are outliers, (3) when the data are given in ranks or use an ordinal scale, or (4) when the relevant hypotheses do not concern a parameter.

The first situation occurs when the data available for analysis suggest that the distributional assumptions of the parametric test are not satisfied. For example, we may want to test a hypothesis concerning the mean of a population but believe that neither *t*- nor *z*-distributed tests are appropriate because the sample is small and may come from a markedly non-normally distributed population. In that case, we may use a nonparametric test. The nonparametric test will frequently involve the conversion of observations (or a function of observations) into ranks according to magnitude, and sometimes it will involve working with only “greater than” or “less than” relationships (using the + and – signs to denote those relationships). Characteristically, one must refer to specialized statistical tables to determine the rejection points of the test statistic, at least for small samples. Such tests, then, typically interpret the null hypothesis as a hypothesis about ranks or signs.

Second, whereas the underlying distribution of the population may be normal, there may be extreme values or outliers that influence the parametric statistics but not the nonparametric statistics. For example, we may want to use a nonparametric test of the median, in the case of outliers, instead of a test of the mean.

Third, we may have a sample in which observations are ranked. In those cases, we also use nonparametric tests because parametric tests generally require a stronger measurement scale than ranks. For example, if our data were the rankings of investment managers, we would use nonparametric procedures to test the hypotheses concerning those rankings.

A fourth situation in which we use nonparametric procedures occurs when our question does not concern a parameter. For example, if the question concerns whether a sample is random or not, we use the appropriate nonparametric test (a “runs test”). The nonparametric runs test is used to test whether stock price changes can be used to forecast future stock price changes—in other words, a test of the random walk theory. Another type of question that nonparametric methods can address is whether a sample came from a population following a particular probability distribution.

Nonparametric Inference: Summary

Nonparametric statistical procedures extend the reach of inference because they make few assumptions, can be used on ranked data, and may address questions unrelated to parameters. Quite frequently, nonparametric tests are reported alongside parametric tests; the user can then assess how sensitive the statistical conclusion is to the assumptions underlying the parametric test. However, if the assumptions of the parametric test are met, the parametric test (where available) is generally preferred over the nonparametric test because the parametric test may have more power—that is, a greater ability to reject a false null hypothesis.

PRACTICE PROBLEMS

1. An analyst suspects that, in the most recent year, excess returns on stocks have fallen below 5%. She wants to study whether the excess returns are less than 5%. Designating the population mean as μ , which hypotheses are most appropriate for her analysis?
 - A. $H_0: \mu = 5\%$ versus $H_a: \mu \neq 5\%$
 - B. $H_0: \mu \geq 5\%$ versus $H_a: \mu < 5\%$
 - C. $H_0: \mu \leq 5\%$ versus $H_a: \mu > 5\%$
2. Which of the following statements about hypothesis testing is correct?
 - A. The null hypothesis is the condition a researcher hopes to support.
 - B. The alternative hypothesis is the proposition considered true without conclusive evidence to the contrary.
 - C. The alternative hypothesis exhausts all potential parameter values not accounted for by the null hypothesis.
3. Which of the following statements regarding the null hypothesis is correct?
 - A. It can be stated as “not equal to” provided the alternative hypothesis is stated as “equal to.”
 - B. Along with the alternative hypothesis, it considers all possible values of the population parameter.
 - C. In a two-tailed test, it is rejected when evidence supports equality between the hypothesized value and the population parameter.
4. Which of the following statements regarding a one-tailed hypothesis test is correct?
 - A. The rejection region increases in size as the level of significance becomes smaller.
 - B. A one-tailed test more strongly reflects the beliefs of the researcher than a two-tailed test.
 - C. The absolute value of the critical value is larger than that for a two-tailed test at the same level of significance.
5. If a researcher selects a 5 percent level of significance for a hypothesis test, the confidence level is:
 - A. 2.5 percent.
 - B. 5 percent.
 - C. 95 percent.
6. A hypothesis test for a normally distributed population, at a 0.05 significance

level, implies a:

- A. 95 percent probability of rejecting a true null hypothesis.
 - B. 95 percent probability of a Type I error for a two-tailed test.
 - C. 5 percent critical value rejection region for a one-tailed test.
7. The value of a test statistic is *best* described as the basis for deciding whether to:
- A. reject the null hypothesis.
 - B. accept the null hypothesis.
 - C. reject the alternative hypothesis.
8. Which of the following *best* describes a Type I error?
- A. Rejecting a true null hypothesis
 - B. Rejecting a false null hypothesis
 - C. Failing to reject a false null hypothesis
9. A Type II error is *best* described as:
- A. rejecting a true null hypothesis.
 - B. failing to reject a false null hypothesis.
 - C. failing to reject a false alternative hypothesis.
10. The level of significance of a hypothesis test is *best* used to:
- A. calculate the test statistic.
 - B. define the test's rejection points.
 - C. specify the probability of a Type II error.
11. The probability of correctly rejecting the null hypothesis is the:
- A. *p*-value.
 - B. power of a test.
 - C. level of significance.
12. The power of a hypothesis test is:
- A. equivalent to the level of significance.
 - B. the probability of not making a Type II error.
 - C. unchanged by increasing a small sample size.
13. In the step “stating a decision rule” in testing a hypothesis, which of the following elements must be specified?
- A. Critical value
 - B. Power of a test

- C. Value of a test statistic
14. A pooled estimator is used when testing a hypothesis concerning the:
- A. equality of the variances of two normally distributed populations.
 - B. difference between the means of two approximately normally distributed populations with unknown but assumed equal variances.
 - C. difference between the means of two at least approximately normally distributed populations with unknown and assumed unequal variances.
15. When evaluating mean differences between two dependent samples, the *most* appropriate test is a:
- A. *z*-test.
 - B. chi-square test.
 - C. paired comparisons test.
16. A chi-square test is *most* appropriate for tests concerning:
- A. a single variance.
 - B. differences between two population means with variances assumed to be equal.
 - C. differences between two population means with variances not assumed to be equal.
17. Which of the following tests should be used to test the difference between the variances of two normally distributed populations with random independent samples?
- A. *t*-test
 - B. *F*-test
 - C. Paired comparisons test
18. A nonparametric test is most appropriate when the:
- A. data consist of ranked values.
 - B. validity of the test depends on many assumptions.
 - C. sample sizes are large but are drawn from a population that may be non-normal.
19. In which of the following situations would a nonparametric test of a hypothesis *most likely* be used?
- A. The sample data are ranked according to magnitude.
 - B. The sample data come from a normally distributed population.
 - C. The test validity depends on many assumptions about the nature of the population.

20. An analyst is examining the monthly returns for two funds over one year. Both funds' returns are non-normally distributed. To test whether the mean return of one fund is greater than the mean return of the other fund, the analyst can use:

- A. a parametric test only.
- B. a nonparametric test only.
- C. both parametric and nonparametric tests.

SOLUTIONS

1. B is correct. The null hypothesis is what she wants to reject in favor of the alternative, which is that population mean excess return is less than 5%. This is a one-sided (left-side) alternative hypothesis.
2. C is correct. Together, the null and alternative hypotheses account for all possible values of the parameter. Any possible values of the parameter not covered by the null must be covered by the alternative hypothesis (e.g., $H_0: \mu \leq 5$ versus $H_a: \mu > 5$). A is incorrect because the null hypothesis is what the researcher wants to reject; the “hoped for” or “suspected” condition is often set up as the alternative hypothesis. B is incorrect because the null (not the alternative) hypothesis is considered to be true unless the sample used to conduct the hypothesis test gives convincing evidence that the null hypothesis is false.
3. B is correct. The null and alternative hypotheses are complements of one another and must be both mutually and collectively exhaustive. Differently put: all possible values or outcomes need to be contained in either the null or the alternative hypothesis.
A is incorrect because the null hypothesis must always include the equality sign (less than or equal to, equal to, or greater than or equal to). C is incorrect because, in a two-tailed test, the null hypothesis is generally set up as equality between the hypothesized value and the population parameter. If evidence supports equality, then the null hypothesis would not be rejected.
4. B is correct. One-tailed tests in which the alternative is “greater than” or “less than” represent the beliefs of the researcher more firmly than a “not equal to” alternative hypothesis.
A is incorrect because a smaller significance level implies a smaller rejection region. C is incorrect because the absolute value of the critical value for a one-tailed hypothesis test is smaller than that of a two-tailed test.
For example, for a two-tailed t-test with 30 degrees of freedom at the 5% significance level, the corresponding critical value is ± 2.042 (2.5% in each tail) whereas the corresponding critical value for a one-tailed t-test is $+1.697$ or -1.697 (5% in the left or right tail). Thus, the absolute value of the critical value is smaller for the one-tailed test than it is for the two-tailed test for the same level of significance.
5. C is correct. The 5 percent level of significance (i.e., probability of a Type I error) corresponds to $1 - 0.05 = 0.95$, or a 95 percent confidence level (i.e., probability of not rejecting a true null hypothesis). The level of significance is the complement to the confidence level; in other words, they sum to 1.00, or 100 percent.
6. C is correct. For a one-tailed hypothesis test, there is a 5 percent rejection region in one tail of the distribution. A is incorrect because a 5 percent significance level implies a 5 percent probability of rejecting the null hypothesis and a 95 percent confidence interval. B is incorrect because the probability of a Type I error (mistakenly rejecting a true null) is the stated 5 percent significance level.
7. A is correct. In hypothesis testing, a test statistic is a quantity whose value is the basis for deciding whether to reject the null hypothesis.
8. A is correct. The definition of a Type I error is when a true null hypothesis is rejected.

9. B is correct. A Type II error occurs when a false null hypothesis is not rejected.
10. B is correct. The level of significance is used to establish the rejection points of the hypothesis test. A is correct because the significance level is not used to calculate the test statistic; rather, it is used to determine the critical value. C is incorrect because the significance level specifies the probability of making a Type I error.
11. B is correct. The power of a test is the probability of rejecting the null hypothesis when it is false. A is incorrect because the p-value is the smallest level of significance at which the null hypothesis can be rejected. C is incorrect because the level of significance is the probability of mistakenly rejecting the null hypothesis (Type I error).
12. B is correct. The power of a hypothesis test is the probability of correctly rejecting the null when it is false. Failing to reject the null when it is false is a Type II error. Thus, the power of a hypothesis test is the probability of not committing a Type II error.
13. A is correct. The critical value in a decision rule is the rejection point for the test. It is the point with which the test statistic is compared to determine whether to reject the null hypothesis, which is part of the fourth step in hypothesis testing. B is incorrect because the power of a test refers to the probability of rejecting the null hypothesis when it is false. C is incorrect because the value of the test statistic is specified in the ‘Identify the appropriate test statistic and its probability distribution’ step.
14. B is correct. The assumption that the variances are equal allows for the combining of both samples to obtain a pooled estimate of the common variance.
15. C is correct. A paired comparisons test is appropriate to test the mean differences of two samples believed to be dependent. A is incorrect because a z-test is used to determine whether two population means are different when the variances are known and the sample size is large. B is incorrect because a chi-square test is used for tests concerning the variance of a single normally distributed population.
16. A is correct. A chi-square test is used for tests concerning the variance of a single normally distributed population.
17. B is correct. An *F*-test is used to conduct tests concerning the difference between the variances of two normally distributed populations with random independent samples.
18. A is correct. When the samples consist of ranked values, parametric tests are not appropriate. In such cases, nonparametric tests are most appropriate.
19. A is correct. A nonparametric test is used when the data are given in ranks.
20. B is correct. There are only 12 (monthly) observations over the one year of the sample and thus the samples are small. Additionally, the funds’ returns are non-normally distributed. Therefore, the samples do not meet the distributional assumptions for a parametric test. The Mann–Whitney U test (a nonparametric test) could be used to test the differences between population means.

LEARNING MODULE

9

Parametric and Non-Parametric Tests of Independence

by Pamela Peterson Drake, PhD, CFA.

Pamela Peterson Drake, PhD, CFA, is at James Madison University (USA).

LEARNING OUTCOMES

Mastery	The candidate should be able to:
<input type="checkbox"/>	explain parametric and nonparametric tests of the hypothesis that the population correlation coefficient equals zero, and determine whether the hypothesis is rejected at a given level of significance
<input type="checkbox"/>	explain tests of independence based on contingency table data

INTRODUCTION

1

In many contexts in investments, we want to assess the strength of the linear relationship between two variables—that is, we want to evaluate the **correlation** between them. A significance test of a correlation coefficient allows us to assess whether the relationship between two random variables is the result of chance. Lesson 1 covers a parametric and a non-parametric approach to testing the correlation between two variables. If we decide that the relationship does not result from chance, then we can use this information in modeling or forecasting using regression models or machine learning covered in later Learning Modules.

When faced with categorical or discrete data, however, we cannot use the methods discussed in the first lesson to test whether the classifications of such data are independent. If we want to test whether there is a relationship between categorical or discreet data, we can perform a test of independence using a nonparametric test statistic. The second lesson covers the use of contingency tables in implementing this non-parametric test.

LEARNING MODULE OVERVIEW



- There are three ways to formulate hypotheses. Let θ indicate the population parameters:
 1. Two-sided alternative: $H_0: \theta = \theta_0$ versus $H_a: \theta \neq \theta_0$
 2. One-sided (right side): $H_0: \theta \leq \theta_0$ versus $H_a: \theta > \theta_0$

3. One-sided alternative (left side): $H_0: \theta \geq \theta_0$ versus $H_a: \theta < \theta_0$
 where θ_0 is a hypothesized value of the population parameter and θ is the true value of the population parameter.

- A parametric test is a hypothesis test concerning a population parameter or a hypothesis test based on specific distributional assumptions. In contrast, a nonparametric test either is not concerned with a parameter or makes minimal assumptions about the population from which the sample comes.
- A nonparametric test is primarily used when data do not meet distributional assumptions, when there are outliers, when data are given in ranks, or when the hypothesis we are addressing does not concern a parameter.
- In tests concerning correlation, we use a t-statistic to test whether a population correlation coefficient is different from zero. If we have n observations for two variables, this test statistic has a t -distribution with $n - 2$ degrees of freedom.
- The Spearman rank correlation coefficient is calculated on the ranks of two variables within their respective samples.
- A chi-square distributed test statistic is used to test for independence of two categorical variables. This nonparametric test compares actual frequencies with those expected on the basis of independence. This test statistic has degrees of freedom of $(r - 1)(c - 2)$, where r is the number of categories for the first variable and c is the number of categories of the second variable.

2

TESTS CONCERNING CORRELATION



explain parametric and nonparametric tests of the hypothesis that the population correlation coefficient equals zero, and determine whether the hypothesis is rejected at a given level of significance

The most common hypotheses concerning correlation occur when comparing the population correlation coefficient with zero because we often ask whether a relationship exists, which implies a null of the correlation coefficient equal to zero (i.e., no relationship). Hypotheses concerning the population correlation coefficient may be two- or one-sided, as we have seen in other tests. Let ρ represent the population correlation coefficient. The possible hypotheses are as follows:

Two sided: $H_0: \rho = 0$ versus $H_a: \rho \neq 0$

One sided (right side): $H_0: \rho \leq 0$ versus $H_a: \rho > 0$

One sided (left side): $H_0: \rho \geq 0$ versus $H_a: \rho < 0$

We use the sample correlation to test these hypotheses on the population correlation.

Parametric Test of a Correlation

The parametric pairwise correlation coefficient is often referred to as **Pearson correlation**, the **bivariate correlation**, or simply the correlation. Our focus is on the testing of the correlation and not the actual calculation of this statistic, but it helps distinguish this correlation from the nonparametric correlation if we look at the formula for the sample correlation. Consider two variables, X and Y . The sample correlation, r_{XY} , is as follows:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \quad (1)$$

where s_{XY} is the sample covariance between the X and Y variables, s_X is the standard deviation of the X variable, and s_Y is the standard deviation of the Y variable. We often drop the subscript to represent the correlation as simply r .

Therefore, you can see from this formula that each observation is compared with its respective variable mean and that, because of the covariance, it matters how much each observation differs from its respective variable mean. Note that the covariance drives the sign of the correlation.

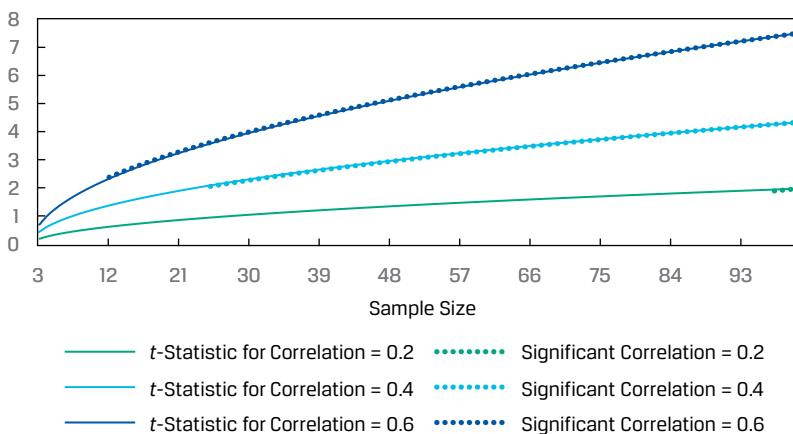
If the two variables are normally distributed, we can test to determine whether the null hypothesis ($H_0: \rho = 0$) should be rejected using the sample correlation, r . The formula for the t -test is as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \quad (2)$$

This test statistic is t -distributed with $n - 2$ degrees of freedom. One practical observation concerning Equation 2 is that the magnitude of r needed to reject the null hypothesis decreases as sample size n increases, for two reasons. First, as n increases, the number of degrees of freedom increases and the absolute value of the critical value of the t -statistic decreases. Second, the absolute value of the numerator increases with larger n , resulting in a larger magnitude of the calculated t -statistic. For example, with sample size $n = 12$, $r = 0.35$ results in a t -statistic of 1.182, which is not different from zero at the 0.05 level ($t_{\alpha/2} = \pm 2.228$). With a sample size of $n = 32$, the same sample correlation, $r = 0.35$, yields a t -statistic of 2.046, which is just significant at the 0.05 level ($t_{\alpha/2} = \pm 2.042$).

Another way to make this point is that when sampling from the same population, a false null hypothesis is more likely to be rejected (i.e., the power of the test increases) as we increase the sample size, all else equal, because a higher number of observations increases the numerator of the test statistic. We show this in Exhibit 1 for three different sample correlation coefficients, with the corresponding calculated t -statistics and significance at the 5 percent level for a two-sided alternative hypothesis. As the sample size increases, significance is more likely to be indicated, but the rate of achieving this significance depends on the sample correlation coefficient; the higher the sample correlation, the faster significance is achieved when increasing the sample size. As the sample sizes increase as ever-larger datasets are examined, the null hypothesis is almost always rejected and other tools of data analysis must be applied.

Exhibit 1: Calculated Test Statistics for Different Sample Sizes and Sample Correlations with a 5 Percent Level of Significance

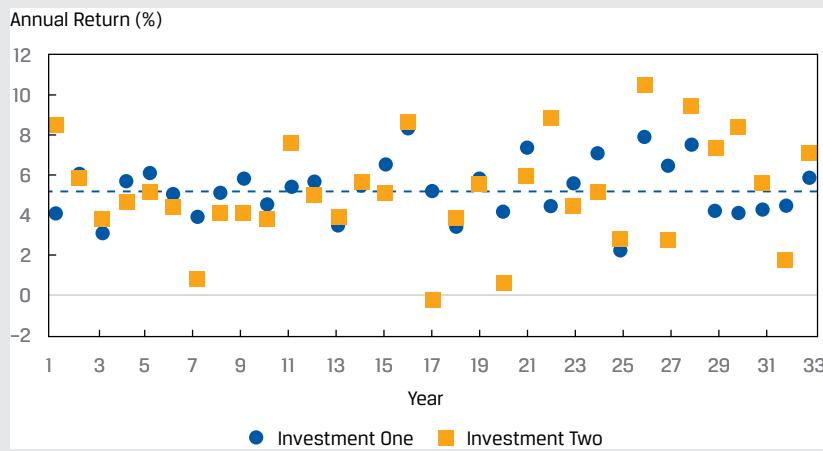
Calculated t -Statistic

EXAMPLE 1

Examining the Relationship between Returns on Investment One and Investment Two

An analyst is examining the annual returns for Investment One and Investment Two over 33 years, as displayed in Exhibit 2.

Exhibit 2: Returns for Investments One and Two over 33 Years



Although this time series plot provides some useful information, the analyst is most interested in quantifying how the returns of these two series are related, so she calculates the correlation coefficient, equal to 0.43051, between these series.

1. Is there a significant positive correlation between these two return series if she uses a 1 percent level of significance?

Solution:

Step 1	State the hypotheses.	$H_0: \rho \leq 0$ versus $H_a: \rho > 0$
Step 2	Identify the appropriate test statistic.	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
Step 3	Specify the level of significance.	1%
Step 4	State the decision rule.	With $33 - 2 = 31$ degrees of freedom and a one-sided test with a 1% level of significance, the critical value is 2.45282. We reject the null hypothesis if the calculated t -statistic is greater than 2.45282.
Step 5	Calculate the test statistic.	$t = \frac{0.43051\sqrt{33-2}}{\sqrt{1-0.18534}} \\ = 2.65568$
Step 6	Make a decision.	Reject the null hypothesis because the calculated t -statistic is greater than 2.45282. Evidence is sufficient to reject the H_0 in favor of H_a , that the correlation between the annual returns of these two investments is positive.

EXAMPLE 2**Correlation with the S&P 500 Returns**

- Exhibit 3 shows the sample correlations between the monthly returns for four different mutual funds and the S&P 500. The correlations are based on 36 monthly observations. The funds are as follows:

Exhibit 3: Sample Correlations between Monthly Returns and the S&P 500

	Fund 1	Fund 2	Fund 3	Fund 4	S&P 500
Fund 1	1				
Fund 2	0.9231	1			
Fund 3	0.4771	0.4156	1		

Fund 4	0.7111	0.7238	0.3102	1
S&P 500	0.8277	0.8223	0.5791	0.7515

Test the null hypothesis that each of these correlations, individually, is equal to zero against the alternative hypothesis that it is not equal to zero. Use a 5 percent significance level and critical t -values of ± 2.032 .

Solution:

The hypotheses are $H_0: \rho = 0$ and $H_a: \rho \neq 0$. The calculated test statistics are based on the formula

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

For example, the calculated t -statistic for the correlation of Fund 3 and Fund 4 is as follows

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.3102\sqrt{36-2}}{\sqrt{1-0.3102^2}} = 1.903.$$

Repeating this calculation for the entire matrix of correlations gives the following:

Calculated t -Statistics for Correlations

	Fund 1	Fund 2	Fund 3	Fund 4	S&P 500
Fund 1					
Fund 2	13.997				
Fund 3	3.165	2.664			
Fund 4	5.897	6.116	1.903		
S&P 500	8.600	8.426	4.142	6.642	

With critical values of ± 2.032 , with the exception of the correlation between Fund 3 and Fund 4 returns, we reject the null hypothesis for these correlations. In other words, evidence is sufficient to indicate that the correlations are different from zero, with the exception of the correlation of returns between Fund 3 and Fund 4.

We could use software to determine the critical values:

Excel

[T.INV(0.025, 34) and T.INV(0.975, 34)]

R

[qt(c(.025, .975), 34)]

Python

[from scipy.stats import t and t.ppf(.025, 34)
and t.ppf(.975, 34)]

We also could use software to determine the p -value for the calculated test statistic to enable a comparison with the level of significance. For example, for $t = 2.664$, the p -value is 0.01172:

Excel

```
[ (1-T.DIST(2.664,34,TRUE))*2]
```

R

```
[ (1-pt(2.664,34))*2]
```

Python

```
[from scipy.stats import t  
(1-t.cdf(2.664,34))*2]
```

Non-Parametric Test of Correlation: The Spearman Rank Correlation Coefficient

When we believe that the population under consideration meaningfully departs from normality, we can use a test based on the **Spearman rank correlation coefficient**, r_S . The Spearman rank correlation coefficient is essentially equivalent to the Pearson correlation coefficient as defined earlier, but it is calculated on the *ranks* of the two variables (e.g., X and Y) within their respective samples. The calculation of r_S requires the following steps:

1. Rank the observations on X from largest to smallest. Assign the number 1 to the observation with the largest value, the number 2 to the observation with second largest value, and so on. In case of ties, assign to each tied observation the average of the ranks that they jointly occupy. For example, if the third and fourth largest values are tied, we assign both observations the rank of 3.5 (the average of 3 and 4). Perform the same procedure for the observations on Y .
2. Calculate the difference, d_i , between the ranks for each pair of observations on X and Y , and then calculate d_i^2 (the squared difference in ranks).
3. With n as the sample size, the Spearman rank correlation is given as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2 - 1)}. \quad (3)$$

Suppose an analyst is examining the relationship between returns for two investment funds, A and B, of similar risk over 35 years. She is concerned that the assumptions for the parametric correlation may not be met, so she decides to test Spearman rank correlations. Her hypotheses are $H_0: r_S = 0$ and $H_a: r_S \neq 0$. She gathers the returns, ranks the returns for each fund, and calculates the difference in ranks and the squared differences. A partial table is provided in Exhibit 4.

Exhibit 4: Differences and Squared Differences in Ranks for Fund A and Fund B over 35 Years

Year	Fund A	Fund B	Rank of A	Rank of B	d	d^2
1	2.453	1.382	27	31	-4	16
2	3.017	3.110	24	24	0	0
3	4.495	6.587	19	7	12	144
4	3.627	3.300	23	23	0	0
:						

Year	Fund A	Fund B	Rank of A	Rank of B	<i>d</i>	<i>d</i> ²
30	2.269	0.025	28	35	-7	49
31	6.354	4.428	10	19	-9	81
32	6.793	4.165	8	20	-12	144
33	7.300	7.623	5	5	0	0
34	6.266	4.527	11	18	-7	49
35	1.257	4.704	34	16	18	324
						Sum = 2,202

The Spearman rank correlation is as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 1 - \frac{6(2,202)}{35(1,225 - 1)} = 0.6916.$$

The test of hypothesis for the Spearman rank correlation depends on whether the sample is small or large ($n > 30$). For small samples, the researcher requires a specialized table of critical values, but for large samples, we can conduct a *t*-test using the test statistic in Equation 2, which is *t*-distributed with $n - 2$ degrees of freedom.

In this example, for a two-tailed test with a 5 percent significance level, the critical values for $n - 2 = 35 - 2 = 33$ degrees of freedom are ± 2.0345 . For the sample information in Exhibit 4, the calculated test statistic is as follows:

$$t = \frac{0.6916 \sqrt{33}}{\sqrt{1 - (0.6916)^2}} = 5.5005$$

Accordingly, we reject the null hypothesis ($H_0: r_S = 0$), concluding that evidence is sufficient to indicate that the correlation between the returns of Fund A and Fund B is different from zero.

EXAMPLE 3

Testing the Exchange Rate Correlation

An analyst gathers exchange rate data for five currencies relative to the US dollar. Upon inspection of the distribution of these exchange rates, she observes a departure from normality, especially with negative skewness for four of the series and positive skewness for the fifth. Therefore, she decides to examine the relationships among these currencies using Spearman rank correlations. She calculates these correlations between the currencies over 180 days, which are shown in the correlogram in Exhibit 5. In this correlogram, the lower triangle reports the pairwise correlations and the upper triangle provides a visualization of the magnitude of the correlations, with larger circles indicating the larger absolute value of the correlations and darker circles indicating correlations that are negative.

Exhibit 5: Spearman Rank Correlations between Exchanges Rates Relative to the US Dollar



- For any of these pairwise Spearman rank correlations, can we reject the null hypothesis of no correlation ($H_0: r_S = 0$ and $H_a: r_S \neq 0$) at the 5 percent level of significance?

Solution:

The critical t -values for 2.5 percent in each tail of the distribution are ± 1.97338 .

There are five exchange rates, so there are $5C2 = \{5! / [2!(5-2)!]\}$, or 10, unique correlation pairs. Therefore, we need to calculate 10 t -statistics. For example, the correlation between EUR/USD and AUD/USD is 0.6079. The calculated t -statistic is

$$\frac{0.6079 \sqrt{180-2}}{\sqrt{1-0.6079^2}} = \frac{8.11040}{0.79401} = 10.2144.$$

Repeating this t -statistic calculation for each pair of exchange rates yields the test statistics shown in Exhibit 6.

Exhibit 6: Calculated Test Statistics for Test of Spearman Rank Correlations

	AUD/USD	CAD/USD	EUR/USD	GBP/USD
CAD/USD	29.7409			
EUR/USD	10.2144	9.1455		
GBP/USD	12.4277	13.2513	7.4773	
JPY/USD	-2.6851	-3.6726	5.2985	-2.7887

The analyst should reject all 10 null hypotheses, because the calculated t -statistics for all exchange rate pairs fall outside the bounds of the two critical values. She should conclude that all the exchange rate pair correlations are different from zero at the 5 percent level.

QUESTION SET

You are interested in whether excess risk-adjusted return (alpha) is correlated with mutual fund expense ratios for US large-cap growth funds. The following table presents the sample.

Mutual Fund	Alpha	Expense Ratio
1	-0.52	1.34
2	-0.13	0.40
3	-0.50	1.90
4	-1.01	1.50
5	-0.26	1.35
6	-0.89	0.50
7	-0.42	1.00
8	-0.23	1.50
9	-0.60	1.45

1. Formulate null and alternative hypotheses consistent with the verbal description of the research goal.

Solution:

We have a “not equal to” alternative hypothesis:

$$H_0: \rho = 0 \text{ versus } H_a: \rho \neq 0$$

2. Identify and justify the test statistic for conducting a test of the hypotheses in Part A.

Solution:

Mutual fund expense ratios are bounded from above and below; in practice, there is at least a lower bound on alpha (as any return cannot be less than -100 percent), and expense ratios cannot be negative. These variables may not be normally distributed, and the assumptions of a parametric test are not likely to be fulfilled. Thus, a nonparametric test appears to be appropriate.

We would use the nonparametric Spearman rank correlation coefficient to conduct the test:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

with the *t*-distributed test statistic of

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}.$$

The calculation of the Spearman rank correlation coefficient is given in the following table.

Mutual Fund	Alpha	Expense Ratio	Rank by Alpha	Rank by Expense Ratio	Difference in Rank	Difference Squared
1	-0.52	1.34	6	6	0	0
2	-0.13	0.40	1	9	-8	64
3	-0.50	1.90	5	1	4	16
4	-1.01	1.50	9	2.5	6.5	42.25
5	-0.26	1.35	3	5	-2	4
6	-0.89	0.50	8	8	0	0
7	-0.42	1.00	4	7	-3	9
8	-0.23	1.50	2	2.5	-0.5	0.25
9	-0.60	1.45	7	4	3	9
						144.5

$$r_s = 1 - \frac{6(144.5)}{9(80)} = -0.20416.$$

3. Determine whether to reject the null hypothesis at the 0.05 level of significance if the critical values are ± 2.306 .

The calculated test statistic, using the *t*-distributed test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \text{ is } t = \frac{-0.2416\sqrt{7}}{\sqrt{1-0.041681}} = \frac{-0.540156}{0.978937} = -0.55177.$$

On the basis of this value falling within the range of ± 2.306 , we fail to reject the null hypothesis that the Spearman rank correlation coefficient is zero.

TESTS OF INDEPENDENCE USING CONTINGENCY TABLE DATA

3



explain tests of independence based on contingency table data

When faced with categorical or discrete data, we cannot use the methods that we have discussed up to this point to test whether the classifications of such data are independent. Suppose we observe the following **frequency table** of 1,594 exchange-traded funds (ETFs) based on two classifications: size (i.e., market capitalization) and investment type (value, growth, or blend), as shown in Exhibit 7. The classification of the investment type is discrete, so we cannot use correlation to assess the relationship between size and investment type.

Exhibit 7: Size and Investment Type Classifications of 1,594 ETFs

Investment Type	Size Based on Market Capitalization			Total
	Small	Medium	Large	
Value	50	110	343	503
Growth	42	122	202	366
Blend	56	149	520	725
Total	148	381	1,065	1,594

Exhibit 7 is called a **contingency table** or a **two-way table** (because there are two classifications, or classes—size and investment type).

If we want to test whether a relationship exists between the size and investment type, we can perform a test of independence using a nonparametric test statistic that is chi-square distributed:

$$\chi^2 = \sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad (4)$$

where:

m = the number of cells in the table, which is the number of groups in the first class multiplied by the number of groups in the second class;

O_{ij} = the number of observations in each cell of row i and column j (i.e., observed frequency); and

E_{ij} = the expected number of observations in each cell of row i and column j , assuming independence (i.e., expected frequency).

This test statistic has $(r - 1)(c - 1)$ degrees of freedom, where r is the number of rows and c is the number of columns.

In Exhibit 7, size class has three groups (small, medium, and large) and investment type class has three groups (value, growth, and blend), so m is 9 ($= 3 \times 3$). The number of ETFs in each cell (O_{ij}), the observed frequency, is given, so to calculate the chi-square test statistic, we need to estimate E_{ij} , the expected frequency, which is the number of ETFs we would expect to be in each cell if size and investment type are completely independent. The expected number of ETFs (E_{ij}) is calculated using the following:

$$E_{ij} = \frac{(\text{Total row } i) \times (\text{Total column } j)}{\text{Overall total}}. \quad (5)$$

Consider one combination of size and investment type, small-cap value:

$$E_{ij} = \frac{503 \times 148}{1,594} = 46.703.$$

We repeat this calculation for each combination of size and investment type (i.e., $m = 9$ pairs) to arrive at the expected frequencies, shown in Panel A of Exhibit 8.

Next, we calculate

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

the squared difference between observed and expected frequencies scaled by expected frequency, for each cell as shown in Panel B of Exhibit 8. Finally, by summing the values of

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

for each of the m cells, we calculate the chi-square statistic as 32.08025.

**Exhibit 8: Inputs to Chi-Square Test Statistic Calculation for 1,594 ETFs
Assuming Independence of Size and Investment Type**

A. Expected Frequency of ETFs by Size and Investment Type

Investment Type	Size Based on Market Capitalization		
	Small	Medium	Large
Value	46.703	120.228	336.070
Growth	33.982	87.482	244.536
Blend	67.315	173.290	484.395
Total	148.000	381.000	1,065.000

B. Scaled Squared Deviation for Each Combination of Size and Investment Type

Investment Type	Size Based on Market Capitalization		
	Small	Medium	Large
Value	0.233	0.870	0.143
Growth	1.892	13.620	7.399
Blend	1.902	3.405	2.617

In our ETF example, we test the null hypothesis of independence between the two classes (i.e., no relationship between size and investment type) versus the alternative hypothesis of dependence (i.e., a relationship between size and investment type) using a 5 percent level of significance, as shown in Exhibit 9. If the observed values are equal to the expected values, the calculated test statistic would be zero. If, however, the observed and expected values are different, these differences are squared, so the calculated chi-square statistic will be positive. Therefore, for the test of independence using a contingency table, there is only one rejection region, on the right side.

Exhibit 9: Test of Independence of Size and Investment Type for 1,594 ETFs

- | | | |
|---------------|--|--|
| Step 1 | State the hypotheses. | H_0 : ETF size and investment type are not related, so these classifications are independent;
H_a : ETF size and investment type are related, so these classifications are not independent. |
| Step 2 | Identify the appropriate test statistic. | $\chi^2 = \sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ |
| Step 3 | Specify the level of significance. 5% | |

Step 4 State the decision rule.

With $(3 - 1) \times (3 - 1) = 4$ degrees of freedom and a one-sided test with a 5% level of significance, the critical value is 9.4877.

We reject the null hypothesis if the calculated χ^2 statistic is greater than 9.4877.

Excel CHISQ.INV(0.95, 4)

R qchisq(.95, 4)

```
Python from scipy.stats import  
chi2  
chi2.ppf(.95, 4)
```

Step 5 Calculate the test statistic.

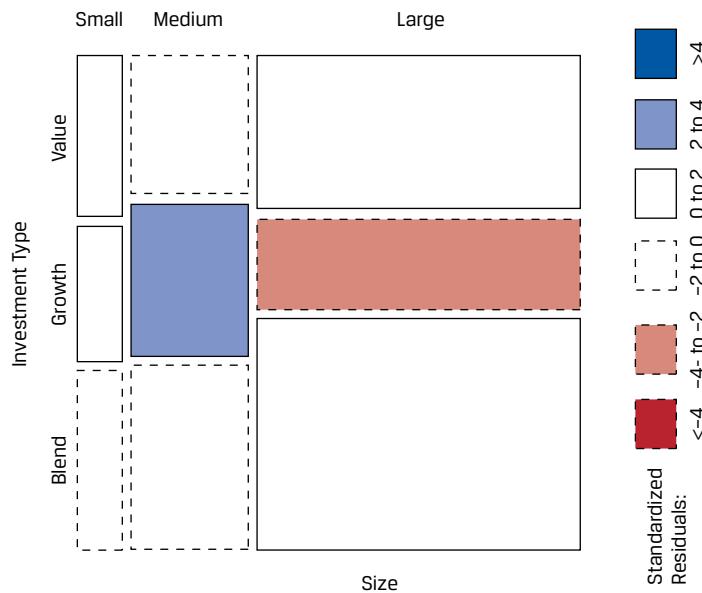
$$\chi^2 = 32.08025$$

Step 6 Make a decision.

Reject the null hypothesis of independence because the calculated χ^2 test statistic is greater than 9.4877. Evidence is sufficient to conclude that ETF size and investment type are related (i.e., not independent).

We can visualize the contingency table in a graphic referred to as a mosaic. In a mosaic, a grid reflects the comparison between the observed and expected frequencies. Consider Exhibit 10, which represents the ETF contingency table.

Exhibit 10: Mosaic of the ETF Contingency Table



The width of the rectangles in Exhibit 10 reflect the proportion of ETFs that are small, medium, and large, whereas the height reflects the proportion that are value, growth, and blend. The darker shading indicates whether the number of observations is more than expected under the null hypothesis of independence, whereas the lighter shading indicates that the number of observations is less than expected, with "more than" and "less than" determined by reference to the standardized residual boxes. The standardized residual, also referred to as a Pearson residual, is as follows:

$$\text{Standardized residual} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} \quad (6)$$

The interpretation for this ETF example is that there are more medium-size growth ETFs (standardized residual of 3.69) and fewer large-size growth ETFs (standardized residual of -2.72) than would be expected if size and investment type were independent.

EXAMPLE 4

Using Contingency Tables to Test for Independence

Consider the contingency table in Exhibit 11, which classifies 500 randomly selected companies on the basis of two environmental, social, and governance (ESG) rating dimensions: environmental rating and governance rating.

Exhibit 11: Classification of 500 Randomly Selected Companies Based on Environmental and Governance Ratings

Environmental Rating	Governance Rating			Total
	Progressive	Average	Poor	
Progressive	35	40	5	80
Average	80	130	50	260
Poor	40	60	60	160
Total	155	230	115	500

- What are the expected frequencies for these two ESG rating dimensions if these categories are independent?

Solution:

The expected frequencies based on independence of the governance rating and the environmental rating are shown in Panel A of Exhibit 12. For example, using Equation 5, the expected frequency for the combination of progressive governance and progressive environmental ratings is

$$E_{ij} = \frac{155 \times 80}{500} = 24.80.$$

The scaled squared deviations for each combination of environmental and governance rating are shown in Panel B of Exhibit 12. For example, using Equation 4, the scaled squared deviation for the combination of progressive governance and progressive environmental ratings is as follows:

$$= \frac{(35 - 24.8)^2}{24.8} = 4.195.$$

Exhibit 12: Inputs to Chi-Square Test Statistic Calculation Assuming Independence of Environmental and Governance Ratings

A. Expected Frequencies of Environmental and Governance Ratings Assuming Independence

Environmental Rating	Governance Rating		
	Progressive	Average	Poor
Progressive	24.8	36.8	18.4
Average	80.6	119.6	59.8
Poor	49.6	73.6	36.8

B. Scaled Squared Deviation for Each Combination of Environmental and Governance Ratings

Environmental Rating	Governance Rating		
	Progressive	Average	Poor
Progressive	4.195	0.278	9.759
Average	0.004	0.904	1.606
Poor	1.858	2.513	14.626

2. Using a 5 percent level of significance, determine whether these two ESG rating dimensions are independent of one another.

Solution:

Step 1 State the hypotheses.
 H_0 : Governance and environmental ratings are not related, so these ratings are independent;
 H_a : Governance and environmental ratings are related, so these ratings are not independent.

Step 2 Identify the appropriate test statistic.
 $\chi^2 = \sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Step 3 Specify the level of significance.
5%

Step 4 State the decision rule.
With $(3 - 1) \times (3 - 1) = 4$ degrees of freedom and a one-sided test with a 5% level of significance, the critical value is 9.487729.

We reject the null hypothesis if the calculated χ^2 statistic is greater than 9.487729.

Excel

```
CHISQ.INV(0.95, 4)
```

R

```
qchisq(.95, 4)
```

Python

```
from scipy.stats import chi2
chi2.ppf(.95, 4)
```

Step 5	Calculate the test statistic.	$\chi^2 = 35.74415$
		To calculate the test statistic, we first calculate the squared difference between observed and expected frequencies scaled by expected frequency for each cell, as shown in Panel B of Exhibit 12. Then, summing the values in each of the m cells (see Equation 4), we calculate the chi-square statistic as 35.74415.
Step 6	Make a decision.	Reject the null hypothesis because the calculated χ^2 test statistic is greater than 9.487729. Evidence is sufficient to indicate that the environmental and governance ratings are related, so they are not independent.

QUESTION SET

An analyst group follows 250 firms and classifies them in two dimensions. First, they use dividend payment history and earnings forecasts to classify firms into one of three groups, with 1 indicating the dividend stars and 3 the dividend laggards. Second, they classify firms on the basis of financial leverage, using debt ratios, debt features, and corporate governance to classify the firms into three groups, with 1 indicating the least risky firms based on financial leverage and 3 indicating the riskiest. The classification of the 250 firms is as follows:

Financial Leverage Group	Dividend Group		
	1	2	3
1	40	40	40
2	30	10	20
3	10	50	10

Using the classification of the 250 firms, answer the following questions:

- What are the null and alternative hypotheses to test whether the dividend and financial leverage groups are independent of one another?

Solution:

The hypotheses are as follows:

H_0 : Dividend and financial leverage ratings are not related, so these groupings are independent.

H_a : Dividend and financial leverage ratings are related, so these groupings are not independent.

- What is the appropriate test statistic to use in this type of test?

Solution:

The appropriate test statistic is

$$\chi^2 = \sum_{i=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where O_{ij} represents the observed frequency for the i and j group and E_{ij} represents the expected frequency for the i and j group if the groupings are independent.

The expected frequencies based on independence are as follows:

Financial Leverage Group	Dividend Group			Sum
	1	2	3	
1	38.4	48	33.6	120
2	19.2	24	16.8	60
3	22.4	28	19.6	70
Sum	80	100	70	250

The scaled squared deviations for each combination of financial leverage and dividend grouping are:

Financial Leverage Group	Dividend Group		
	1	2	3
1	0.06667	1.33333	1.21905
2	6.07500	8.16667	0.60952
3	6.86429	17.28571	4.70204

3. If the critical value for the 0.05 level of significance is 9.4877, what is your conclusion?

Solution:

The sum of these scaled squared deviations is the calculated chi-square statistic of 46.3223. Because this calculated value exceeds the critical value of 9.4877, we reject the null hypothesis that these groupings are independent.

PRACTICE PROBLEMS

1. Jill Batten is analyzing how the returns on the stock of Stellar Energy Corp. are related with the previous month's percentage change in the US Consumer Price Index for Energy (CPIENG). Based on 248 observations, she has computed the sample correlation between the Stellar and CPIENG variables to be -0.1452 . She also wants to determine whether the sample correlation is significantly different from zero. The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. Batten should conclude that the statistical relationship between Stellar and CPIENG is:
 - A. significant, because the calculated test statistic is outside the bounds of the critical values for the test statistic.
 - B. significant, because the calculated test statistic has a lower absolute value than the critical value for the test statistic.
 - C. insignificant, because the calculated test statistic is outside the bounds of the critical values for the test statistic.
2. Which of the following statements is correct regarding the chi-square test of independence?
 - A. The test has a one-sided rejection region.
 - B. The null hypothesis is that the two groups are dependent.
 - C. If there are two categories, each with three levels or groups, there are six degrees of freedom.

SOLUTIONS

1. A is correct. The calculated test statistic is

$$\begin{aligned} t &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\ &= \frac{-0.1452\sqrt{248-2}}{\sqrt{1-(-0.1452)^2}} = -2.30177. \end{aligned}$$

Because the value of $t = -2.30177$ is outside the bounds of ± 1.96 , we reject the null hypothesis of no correlation and conclude that evidence is sufficient to indicate that the correlation is different from zero.

2. A is correct. The test statistic includes squared differences between the observed and expected values, so the test involves only one side, the right side. B is incorrect because the null hypothesis is that the groups are independent, and C is incorrect because with three levels of groups for the two categorical variables, there are four degrees of freedom.

LEARNING MODULE

10

Simple Linear Regression

by Pamela Peterson Drake, PhD, CFA.

Pamela Peterson Drake, PhD, CFA, is at James Madison University (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	describe a simple linear regression model, how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients
<input type="checkbox"/>	explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated
<input type="checkbox"/>	calculate and interpret measures of fit and formulate and evaluate tests of fit and of regression coefficients in a simple linear regression
<input type="checkbox"/>	describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the standard error of estimate in a simple linear regression
<input type="checkbox"/>	calculate and interpret the predicted value for the dependent variable, and a prediction interval for it, given an estimated linear regression model and a value for the independent variable
<input type="checkbox"/>	describe different functional forms of simple linear regressions

INTRODUCTION

1

- describe a simple linear regression model, how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients

Financial analysts often need to examine whether a variable is useful for explaining another variable. For example, the analyst may want to know whether earnings or cash flow growth help explain a company's market value. **Regression analysis** is a tool for examining this type of issue.

Linear regression allows us to test hypotheses about the relationship between two variables by quantifying the strength of the relationship between the two variables, and to use one variable to make predictions about the other variable. Our focus is on linear regression with a single independent variable—that is, simple linear regression.

LEARNING MODULE OVERVIEW



- Simple linear regression is a mathematical process for determining how the variation in one variable can explain the variation in another variable.
- The variable we wish to explain is called the dependent variable, and the variable that we use to explain the dependent variable is called the independent variable.
- Simple linear regression uses the ordinary least squares approach to calculate the slope and intercept parameters that characterize a linear relationship between the two variables.
- Simple linear regression requires that we make four assumptions: linearity, homoskedasticity, independence, and normality.
- Linearity requires that the regression residuals be random and that the independent variable not be random. Homoskedasticity, which refers to variance being constant across observations, cannot be assumed when we see residuals clustering in multiple groups because the clustering indicates multiple regimes with different variances within our time period.
- Independence means that the X-Y pairs are uncorrelated; a pattern in a plot of the residuals (e.g., seasonality) suggests that there is autocorrelation across observations and that we cannot assume independence. Normality means that residuals must be normally distributed and does not require that the data itself be normally distributed. Non-normality is of particular concern for small sample sizes, but for large sample sizes, the central limit theorem tells us that we may be able to relax the normality requirement.
- The total variation in the dependent variable, called the sum of squares total (SST), can be decomposed into two parts: the explained variation, called the sum of squares regression (SSR), and the unexplained variation, called the sum of squares error (SSE).
- There are several ways to measure a regression model's goodness of fit. These include the coefficient of determination, the F -statistic for the test of fit, and the standard error of the regression.
- Hypothesis testing can be used to determine, at a specified confidence level, whether the slope or intercept differs from zero or another specified value, or whether the slope is positive or negative. We can use indicator variables to determine whether our regression parameters differ between data points that either have or do not have a particular characteristic (e.g., monthly price data in cases in which only some months have earnings announcements).
- An analysis of variance (ANOVA) table presents the sums of squares, degrees of freedom, mean squares, and F -statistic for a regression model.

- The standard error of the estimate is a measure of the distance between the observed values of the dependent variable and those predicted from the estimated regression. The smaller this value, the better the fit of the model.
- The standard error of the forecast is used to provide an interval estimate around the estimated regression line. It is necessary because the regression line does not describe the relationship between the dependent and independent variables perfectly.
- The simple linear regression model on non-linear data can be adjusted by using different functional forms that transform the dependent or independent variables.
- Three common functional forms for transforming data include the log-lin model, the lin-log model, and the log-log model.
- The key to fitting the appropriate functional form of a simple linear regression is examining the goodness-of-fit measures—the coefficient of determination (R^2), the F -statistic, and the standard error of the estimate (s_e)—as well as examining whether there are patterns in the residuals.

ESTIMATION OF THE SIMPLE LINEAR REGRESSION MODEL

2

- describe a simple linear regression model, how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients

Introduction to Linear Regression

Suppose an analyst is examining the return on assets (ROA) for an industry and observes the ROA for the six companies shown in Exhibit 1. The average of these ROAs is 12.5 percent, but the range is from 4 percent to 20 percent.

Exhibit 1: Return on Assets of Selected Companies

Company	ROA (%)
A	6.0
B	4.0
C	15.0
D	20.0
E	10.0
F	20.0

In trying to understand why the ROAs differ among these companies, we could look at why the ROA of Company A differs from that of Company B, why the ROA of Company A differs from that of Company D, why the ROA of Company F differs from that of Company C, and so on, comparing each pair of ROAs. We can simplify this exercise by instead comparing each company's ROA to the mean ROA of 12.5 percent. To do this, we look at the sum of the squared deviations of the observations from the mean to capture variations in ROA from their mean. Let Y represent the variable that we would like to explain, which in this case is the ROA. Let Y_i represent an observation of a company's ROA, and let \bar{Y} represent the mean ROA for the sample of size n . We can describe the variation of the ROAs as follows:

$$\text{Variation of } Y = \sum_{i=1}^n (Y_i - \bar{Y})^2. \quad (1)$$

Our goal is to understand what drives these ROAs or, in other words, what explains the variation of Y . The variation of Y is often referred to as the **sum of squares total (SST)**, or the total sum of squares.

We now ask whether it is possible to explain the variation of the ROA using another variable that also varies among the companies; note that if this other variable is constant or random, it would not explain the ROA differences. Suppose the analyst believes that the capital expenditures in the previous period, scaled by the prior period's beginning property, plant, and equipment, are a driver for the ROA variable. Let us represent this scaled capital expenditures variable as CAPEX, as we show in Exhibit 2.

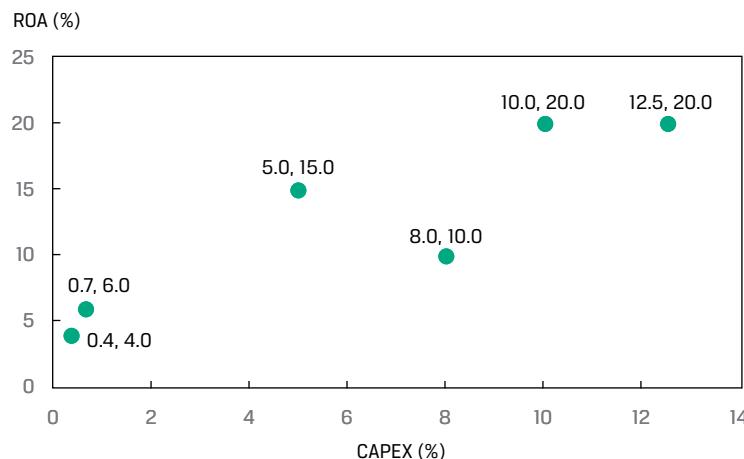
Exhibit 2: Return on Assets and Scaled Capital Expenditures

Company	ROA (%)	CAPEX (%)
A	6.0	0.7
B	4.0	0.4
C	15.0	5.0
D	20.0	10.0
E	10.0	8.0
F	20.0	12.5
Arithmetic mean	12.5	6.1

Let X represent the explanatory variable, in this case, CAPEX. Then X_i will represent an observation of our explanatory variable, and \bar{X} will represent the mean value for the explanatory variable, that is, the mean of all of our CAPEX values. The variation of X is calculated as follows:

$$\text{Variation of } X = \sum_{i=1}^n (X_i - \bar{X})^2. \quad (2)$$

We can see the relation between ROA and CAPEX in the **scatter plot** (or scattergram) in Exhibit 3, which represents the two variables in two dimensions. Typically, we present the variable whose variation we want to explain along the vertical axis and the variable whose variation we want to use to explain that variation along the horizontal axis. Each point in this scatter plot represents a paired observation that consists of CAPEX and ROA. From a casual visual inspection, a positive relation is apparent between ROA and CAPEX: Companies with higher CAPEX tend to have a higher ROA.

Exhibit 3: Scatter Plot of ROA and CAPEX

In the ROA example, we use the capital expenditures to explain the ROAs. We refer to the variable whose variation is being explained as the **dependent variable**, or the explained variable; it is typically denoted by Y . We refer to the variable whose variation is being used to explain the variation of the dependent variable as the **independent variable**, or the explanatory variable; it is typically denoted by X . Therefore, in our example, the ROA is the dependent variable (Y) and CAPEX is the independent variable (X).

A common method used to relate the dependent and independent variables is through the estimation of a linear relationship, which implies describing the relation between the two variables as represented by a straight line. If we have only one independent variable, we refer to the method as **simple linear regression (SLR)**, or linear regression; if we have more than one independent variable, we refer to the method as multiple regression.

Linear regression allows us to test hypotheses about the relationship between two variables by quantifying the strength of the relationship between the two variables and to use one variable to make predictions about the other variable.

IDENTIFYING THE DEPENDENT AND INDEPENDENT VARIABLES IN A REGRESSION



An analyst is researching the relationship between corporate earnings growth and stock returns. Specifically, she is interested in whether earnings revisions are correlated with stock price returns in the same period. She collects five years of monthly data on earnings per share (EPS) revisions and stock price returns for a sample of 100 companies.

1. What are the dependent and independent variables in her model?

Solution:

The dependent variable is monthly stock price returns, and the independent variable is EPS revisions. In the analyst's model, the variation in monthly stock price returns is being explained by the variation in EPS revisions.

Estimating the Parameters of a Simple Linear Regression

The Basics of Simple Linear Regression

Regression analysis begins with the dependent variable, the variable whose variation you are seeking to explain. The independent variable is the variable whose variation you are using to explain changes in the dependent variable. For example, you might try to explain small-stock returns (the dependent variable) using returns to the S&P 500 Index (the independent variable). Or you might try to explain a country's inflation rate (the dependent variable) as a function of growth in its money supply (the independent variable).

As the name implies, linear regression assumes a linear relationship between the dependent and the independent variables. The goal is to fit a line to the observations on Y and X to minimize the squared deviations from the line; this is the least squares criterion—hence, the name least squares regression. Because of its common use, linear regression is often referred to as ordinary least squares (OLS) regression.

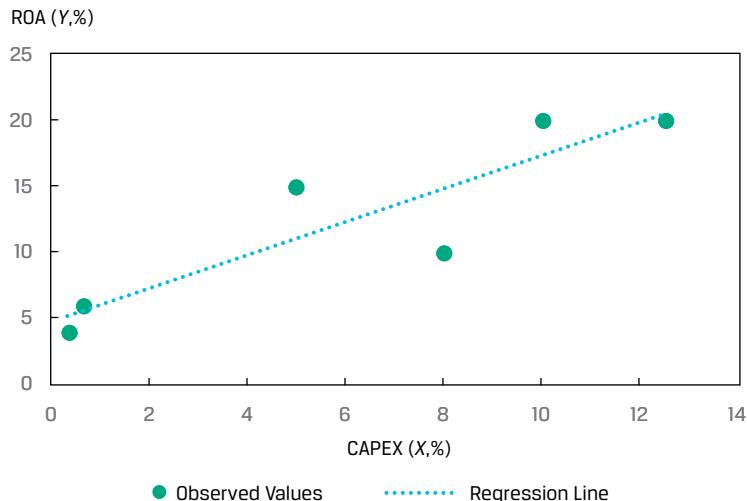
Using notation, the linear relation between the dependent and independent variables is described as follows:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, i = 1, \dots, n. \quad (3)$$

Equation 3 is a model that does not require that every (X, Y) pair for an observation fall on the regression line. This equation states that the dependent variable, Y , is equal to the **intercept**, b_0 , plus a **slope coefficient**, b_1 , multiplied by the independent variable, X , plus an **error term**, ε . The error term, or simply the error, represents the difference between the observed value of Y and that expected from the true underlying population relation between Y and X . We refer to the intercept, b_0 , and the slope coefficient, b_1 , as the **regression coefficients**? A way that we often describe this simple linear regression relation is that Y is regressed on X .

Consider the ROA and CAPEX scatter diagram from Exhibit 3, which we elaborate on in Exhibit 4 by including the fitted regression line. This line represents the average relationship between ROA and CAPEX; not every observation falls on the line, but the line describes the mean relation between ROA and CAPEX.

Exhibit 4: Fitted Regression Line of ROA and CAPEX



Estimating the Regression Line

We cannot observe the population parameter values b_0 and b_1 in a regression model. Instead, we observe only \hat{b}_0 and \hat{b}_1 , which are estimates (as indicated by the “hats” above the coefficients) of the population parameters based on the sample. Thus, predictions must be based on the parameters’ estimated values, and testing is based on estimated values in relation to the hypothesized population values.

We estimate the regression line as the line that best fits the observations. In simple linear regression, the estimated intercept, \hat{b}_0 , and slope, \hat{b}_1 , are such that the sum of the squared vertical distances from the observations to the fitted line is minimized. The focus is on the sum of the squared differences between the observations on Y_i and the corresponding estimated value, \hat{Y}_i , on the regression line.

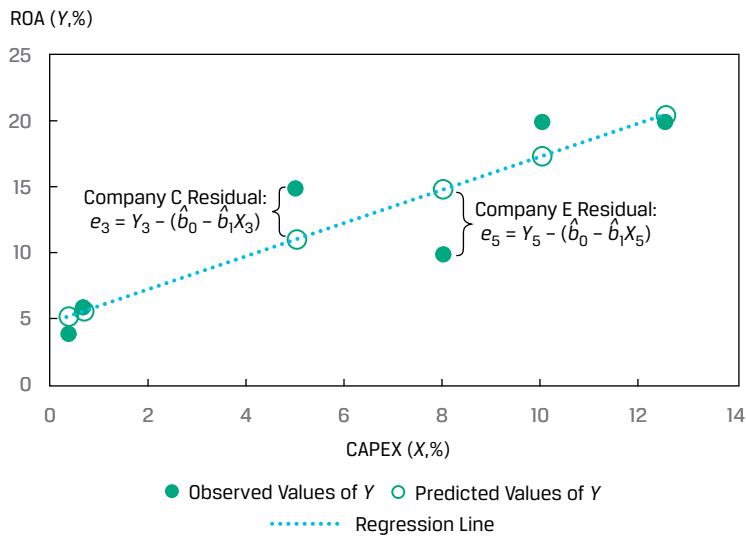
We represent the value of the dependent variable for the i th observation that falls on the line as \hat{Y}_i , which is equal to $\hat{b}_0 + \hat{b}_1 X_i$. \hat{Y}_i is what the estimated value of the Y variable would be for the i th observation based on the mean relationship between Y and X . The **residual** for the i th observation, e_i , is how much the observed value of Y_i differs from the \hat{Y}_i estimated using the regression line: $e_i = Y_i - \hat{Y}_i$. Note the subtle difference between the error term and the residual: The error term refers to the true underlying population relationship, whereas the residual refers to the fitted linear relation based on the sample.

Fitting the line requires minimizing the sum of the squared residuals, the **sum of squares error (SSE)**, also known as the residual sum of squares:

$$\begin{aligned} \text{Sum of squares error} &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n [Y_i - (\hat{b}_0 + \hat{b}_1 X_i)]^2 \\ &= \sum_{i=1}^n e_i^2. \end{aligned} \tag{4}$$

Using least squares regression to estimate the values of the population parameters of b_0 and b_1 , we can fit a line through the observations of X and Y that explains the value that Y takes for any particular value of X .

As seen in Exhibit 5, the residuals are represented by the vertical distances from the fitted line (see the third and fifth observations, Companies C and E, respectively) and are, therefore, in the units of measurement represented by the dependent variable. The residual is in the same unit of measurement as the dependent variable: If the dependent variable is in euros, the error term is in euros, and if the dependent variable is in growth rates, the error term is in growth rates.

Exhibit 5: Residuals of the Linear Regression

How do we calculate the intercept (\hat{b}_0) and the slope (\hat{b}_1) for a given sample of (Y , X) pairs of observations? The slope is the ratio of the covariance between Y and X to the variance of X , where \bar{Y} is the mean of the Y variable and \bar{X} is the mean of X variable:

$$\hat{b}_1 = \frac{\text{Covariance of } Y \text{ and } X}{\text{Variance of } X} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\frac{n-1}{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

Simplifying,

$$\hat{b}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \quad (5)$$

Once we estimate the slope, we can then estimate the intercept using the mean of Y and the mean of X :

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}. \quad (6)$$

Incremental values to calculate the slope and the intercept are in Exhibit 6.

Exhibit 6: Estimating Slope and Intercept for the ROA Model

Company	ROA (Y _i)	CAPEX (X _i)	(Y _i - \bar{Y}) ²	(X _i - \bar{X}) ²	(Y _i - \bar{Y})(X _i - \bar{X})
A	6.0	0.7	42.25	29.16	35.10
B	4.0	0.4	72.25	32.49	48.45
C	15.0	5.0	6.25	1.21	-2.75
D	20.0	10.0	56.25	15.21	29.25
E	10.0	8.0	6.25	3.61	-4.75
F	20.0	12.5	56.25	40.96	48.00

Company	ROA (Y _i)	CAPEX (X _i)	(Y _i - \bar{Y}) ²	(X _i - \bar{X}) ²	(Y _i - \bar{Y})(X _i - \bar{X})
Sum	75.0	36.6	239.50	122.64	153.30
Arithmetic mean	12.5	6.1			

Slope coefficient: $\hat{b}_1 = \frac{153.30}{122.64} = 1.25$.

Intercept: $\hat{b}_0 = 12.5 - (1.25 \times 6.10) = 4.875$.

ROW regression model: $\hat{Y}_i = 4.875 + 1.25 X_i + \varepsilon_i$.

Notice the similarity of the formula for the slope coefficient and that of the pairwise correlation. The sample correlation, r , is the ratio of the covariance to the product of the standard deviations:

$$r = \frac{\text{Covariance of } Y \text{ and } X}{(\text{Standard deviation of } Y)(\text{Standard deviation of } X)} \quad (7)$$

The subtle difference between the slope and the correlation formulas is in the denominator: For the slope, this is the variance of the independent variable, but for the correlation, the denominator is the product of the standard deviations. For our ROA and CAPEX analysis,

$$\text{Covariance of } Y \text{ and } X: \text{cov}_{XY} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{n - 1} = \frac{153.30}{5} = 30.6600.$$

Standard deviation of Y and X :

$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}} = \sqrt{\frac{239.50}{5}} = 6.9210 ;$$

$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} = \sqrt{\frac{122.64}{5}} = 4.9526.$$

$$r = \frac{30.66}{(6.9210)(4.9526)} = 0.8945.$$

Because the denominators of both the slope and the correlation are positive, the sign of the slope and the correlation are driven by the numerator: If the covariance is positive, both the slope and the correlation are positive, and if the covariance is negative, both the slope and the correlation are negative.

EXAMPLE 1

How Do Analysts Perform Simple Linear Regression?

Typically, an analyst will use the data analysis functions on a spreadsheet, such as Microsoft Excel, or a statistical package in the R or Python programming languages to perform linear regression analysis. The following are some of the more common choices in practice.

Simple Linear Regression: Intercept and Slope

- *Excel*: Use the INTERCEPT, SLOPE functions.
- *R*: Use the lm function.
- *Python*: Use the sm.OLS function in the statsmodels package.

Correlations

- *Excel*: Use the CORREL function.

- *R*: Use the `cor` function in the `stats` library.
- *Python*: Use the `corrcoef` function in the `numpy` library.

Note that in R and Python, there are many choices for regression and correlation analysis.

Interpreting the Regression Coefficients

What is the meaning of the regression coefficients? The intercept is the value of the dependent variable if the value of the independent variable is zero. Importantly, this does not make sense in some contexts, especially if it is unrealistic that the independent variable would be zero. For example, if we have a model in which money supply explains GDP growth, the intercept has no meaning because, practically speaking, zero money supply is not possible. If the independent variable were money supply growth, however, the intercept is meaningful. The slope is the change in the dependent variable for a one-unit change in the independent variable. If the slope is positive, then the change in the independent variable and that of the dependent variable will be in the same direction; if the slope is negative, the change in the independent variable and that of the dependent variable will be in opposite directions.

EXAMPLE 2

Interpreting Positive and Negative Slopes

Suppose the dependent variable (Y) is in millions of euros and the independent variable (X) is in millions of US dollars.

If the slope is positive 1.2, then

↑ USD1 million → ↑ EUR1.2 million

↓ USD1 million → ↓ EUR1.2 million

If the slope is negative 1.2, then

↑ USD1 million → ↓ EUR1.2 million

↓ USD1 million → ↑ EUR1.2 million

Using the ROA regression model from Exhibit 6, we would interpret the estimated coefficients as follows:

- The return on assets for a company is 4.875% if the company makes no capital expenditures.
- If CAPEX increases by one unit—say, from 4% to 5%—ROA increases by 1.25%.

Using the estimated regression coefficients, we can determine the values of the dependent variable if they follow the average relationship between the dependent and independent variables. A result of the mathematics of the least squares fitting of the regression line is that the expected value of the residual term is zero: $E(\epsilon) = 0$.

We show the calculation of the predicted dependent variable and residual term for each observation in the ROA example in Exhibit 7. Note that the sum and average of Y_i and \hat{Y}_i are the same, and the sum of the residuals is zero.

Exhibit 7: Calculation of the Dependent Variable and Residuals for the ROA and CAPEX Model

Company	(1)	(2)	(3)	(4)
	ROA (Y_i)	CAPEX(X_i)	Predicted ROA (\hat{Y}_i)	(1) – (3) (2) Residual (e_i)
A	6.0	0.7	5.750	0.250
B	4.0	0.4	5.375	-1.375
C	15.0	5.0	11.125	3.875
D	20.0	10.0	17.375	2.625
E	10.0	8.0	14.875	-4.875
F	20.0	12.5	20.500	-0.500
Sum	75.0	36.6	75.000	0.000
Average	12.5	6.1	12.5	0.000

For Company C ($i = 3$),

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i + \varepsilon_i = 4.875 + 1.25 X_i + \varepsilon_i$$

$$\hat{Y}_i = 4.875 + (1.25 \times 5.0) = 4.875 + 6.25 = 11.125$$

$$Y_i - \hat{Y}_i = e_i = 15.0 - 11.125 = 3.875, \text{ the vertical distance in Exhibit 5.}$$

Whereas the sum of the residuals must equal zero by design, the focus of fitting the regression line in a simple linear regression is minimizing the sum of the squared residual terms.

Cross-Sectional versus Time-Series Regressions

Regression analysis uses two principal types of data: cross sectional and time series. A cross-sectional regression involves many observations of X and Y for the same time period. These observations could come from different companies, asset classes, investment funds, countries, or other entities, depending on the regression model. For example, a cross-sectional model might use data from many companies to test whether predicted EPS growth explains differences in price-to-earnings ratios during a specific time period. Note that if we use cross-sectional observations in a regression, we usually denote the observations as $i = 1, 2, \dots, n$.

Time-series data use many observations from different time periods for the same company, asset class, investment fund, country, or other entity, depending on the regression model. For example, a time-series model might use monthly data from many years to test whether a country's inflation rate determines its short-term interest rates. If we use time-series data in a regression, we usually denote the observations as $t = 1, 2, \dots, T$. Note that in the sections that follow, we primarily use the notation $i = 1, 2, \dots, n$, even for time series.

QUESTION SET



An analyst is exploring the relationship between a company's net profit margin and research and development expenditures. He collects data for an industry and calculates the ratio of research and development expenditures to revenues (RDR) and the net profit margin (NPM) for eight companies. Specifically, he wants to explain the net profit margin variation by using the variation observed in the companies' research and development spending. He reports the data in Exhibit 8.

Exhibit 8: Observations on NPM and RDR for Eight Companies

Company	NPM (%)	RDR (%)
1	4	8
2	5	10
3	10	6
4	9	5
5	5	7
6	6	9
7	12	5
8	3	10

1. What is the slope coefficient for this simple linear regression model?

Solution:

The slope coefficient for the regression model is -1.3 , and the details for the inputs to this calculation are in Exhibit 9.

Exhibit 9: Details of Calculation of Slope of NPM Regressed on RDR

Company	NPM (%) (Y _i)	RDR (%) (X _i)	$Y_i - \bar{Y}$	$X_i - \bar{X}$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})(X_i - \bar{X})$
1	4	8	-2.8	0.5	7.5625	0.25	-1.375
2	5	10	-1.8	2.5	3.0625	6.25	-4.375
3	10	6	3.3	-1.5	10.5625	2.25	-4.875
4	9	5	2.3	-2.5	5.0625	6.25	-5.625
5	5	7	-1.8	-0.5	3.0625	0.25	0.875
6	6	9	-0.8	1.5	0.5625	2.25	-1.125
7	12	5	5.3	-2.5	27.5625	6.25	-13.125
8	3	10	-3.8	2.5	14.0625	6.25	-9.375
Sum	54	60	0.0	0.0	71.5000	30.00	-39.000
Average	6.75	7.5					

$$\text{Slope coefficient: } \hat{b}_1 = \frac{-39}{30} = -1.3.$$

2. What is the intercept for this regression model?

Solution:

The intercept of the regression model is 16.5 :

$$\text{Intercept: } \hat{b}_0 = 6.75 - (-1.3 \times 7.5) = 6.75 + 9.75 = 16.5$$

3. How is this estimated linear regression model represented?

Solution:

The regression model is represented by $\hat{Y}_i = 16.5 - 1.3X_i + \varepsilon_i$.

4. What is the pairwise correlation between NPM and RDR?

Solution:

The pairwise correlation is -0.8421 :

$$r = \frac{-39.7}{\sqrt{71.57} \sqrt{30.7}} = \frac{-5.5714}{(3.1960)(2.0702)} = -0.8421.$$

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL

3



explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated

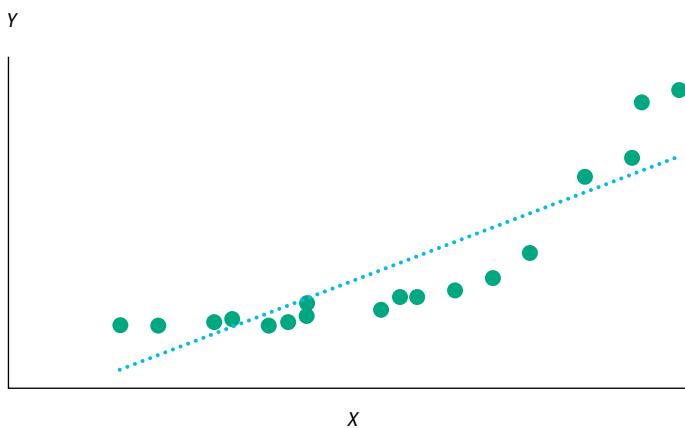
We have discussed how to interpret the coefficients in a simple linear regression model. Now we turn to the statistical assumptions underlying this model. Suppose that we have n observations of both the dependent variable, Y , and the independent variable, X , and we want to estimate the simple linear regression of Y regressed on X . We need to make the following four key assumptions to be able to draw valid conclusions from a simple linear regression model:

1. Linearity: The relationship between the dependent variable, Y , and the independent variable, X , is linear.
2. Homoskedasticity: The variance of the regression residuals is the same for all observations.
3. Independence: The observations, pairs of Y s and X s, are independent of one another. This implies the regression residuals are uncorrelated across observations.
4. Normality: The regression residuals are normally distributed.

Now we take a closer look at each of these assumptions and introduce the “best practice” of examining residual plots of regression results to identify potential violations of these key assumptions.

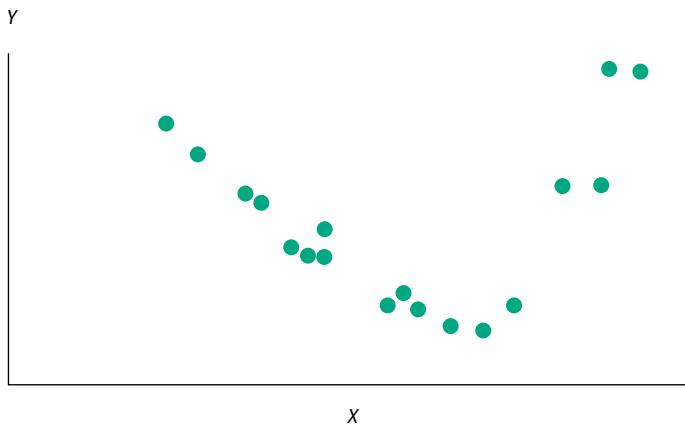
Assumption 1: Linearity

We are fitting a linear model, so we must assume that the true underlying relationship between the dependent and independent variables is linear. If the relationship between the independent and dependent variables is nonlinear in the parameters, estimating that relation with a simple linear regression model will produce invalid results: The model will be biased, because it will under- and overestimate the dependent variable at certain points. For example, $Y_i = b_0 e^{b_1 X_i} + \varepsilon_i$ is nonlinear in b_1 , so we should not apply the linear regression model to it. Exhibit 10 shows an example of this exponential model, with a regression line indicated. You can see that this line does not fit this relationship well: For lower and higher values of X , the linear model underestimates the Y , whereas for the middle values, the linear model overestimates Y .

Exhibit 10: Illustration of Nonlinear Relationship Estimated as a Linear Relationship

Another implication of this assumption is that the independent variable, X , must not be random; that is, it is non-stochastic. If the independent variable is random, there would be no linear relation between the dependent and independent variables. Although we may initially assume that the independent variable in the regression model is not random, that assumption may not always be true.

When we look at the residuals of a model, what we would like to see is that the residuals are random. The residuals should not exhibit a pattern when plotted against the independent variable. As we show in Exhibit 11, the residuals from the Exhibit 10 linear regression do not appear to be random but, rather, they exhibit a relationship with the independent variable, X , falling for some range of X and rising in another.

Exhibit 11: Illustration of Residuals in a Nonlinear Relationship Estimated as a Linear Relationship

Assumption 2: Homoskedasticity

Assumption 2, that the variance of the residuals is the same for all observations, is known as the **homoskedasticity** assumption. In terms of notation, this assumption relates to the squared residuals:

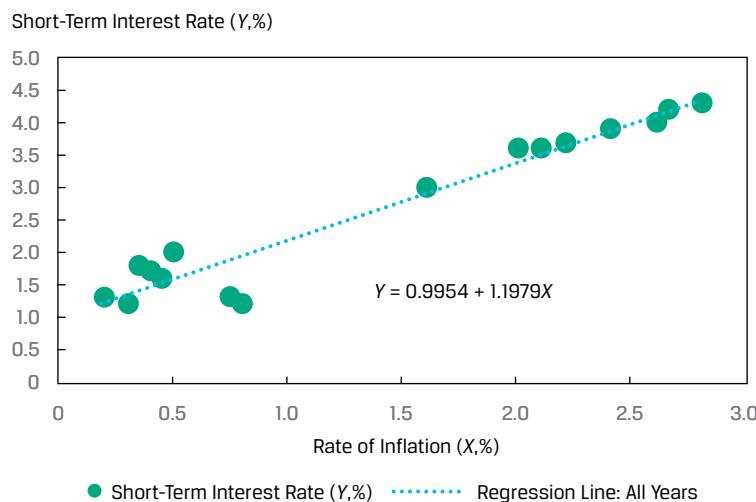
$$E(\varepsilon_i^2) = \sigma_\varepsilon^2, i = 1, \dots, n. \quad (8)$$

If the residuals are not homoskedastic, that is, if the variance of residuals differs across observations, then we refer to this as **heteroskedasticity**.

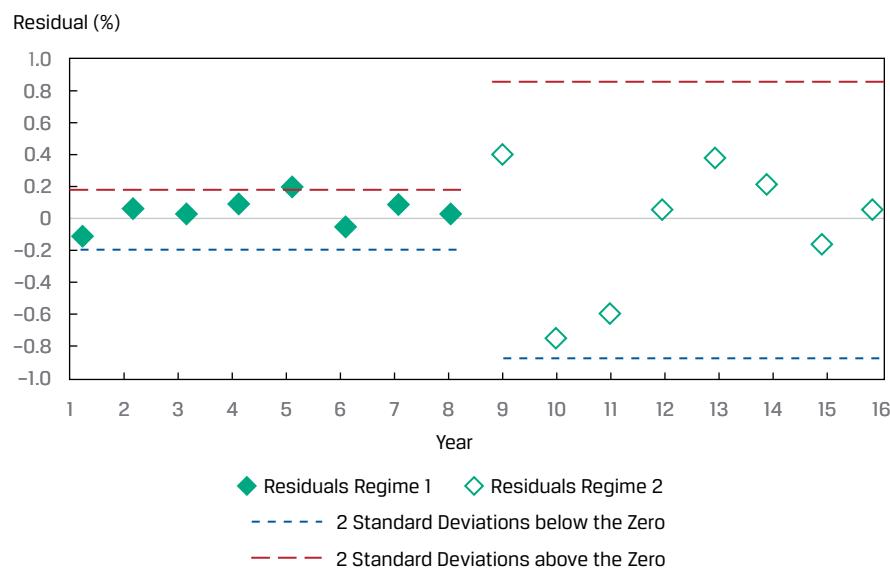
Suppose you are examining a time series of short-term interest rates as the dependent variable and inflation rates as the independent variable over 16 years. We may believe that short-term interest rates (Y) and inflation rates (X) should be related (i.e., interest rates are higher with higher rates of inflation). If this time series spans many years, with different central bank actions that force short-term interest rates to be (artificially) low for the last eight years of the series, then it is likely that the residuals in this estimated model will appear to come from two different models. We will refer to the first eight years as Regime 1 (normal rates) and the second eight years as Regime 2 (low rates). If the model fits differently in the two regimes, the residuals and their variances will be different.

You can see this situation in Exhibit 12, which shows a scatter plot with an estimated regression line. The slope of the regression line over all 16 years is 1.1979.

Exhibit 12: Scatter Plot of Interest Rates (Y) and Inflation Rates (X)

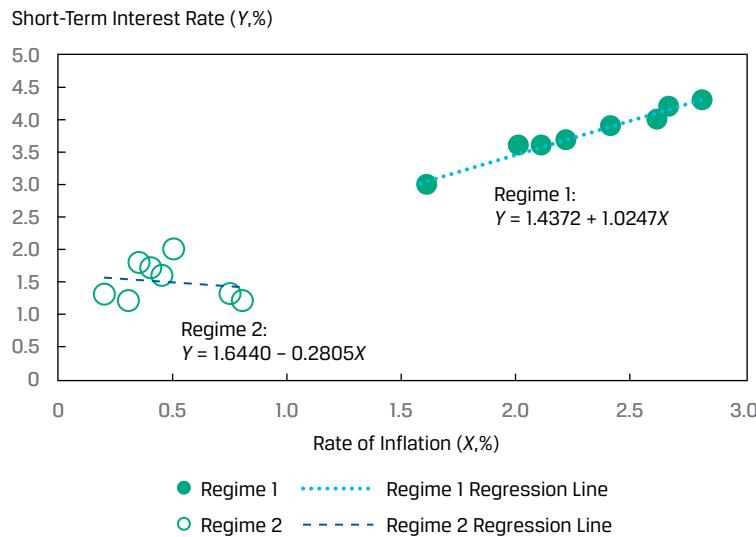


We plot the residuals of this model in Exhibit 13 against the years. In this plot, we indicate the distance that is two standard deviations from zero (the mean of the residuals) for the first eight years' residuals and then do the same for the second eight years. As you can see, the residuals appear different for the two regimes: the variation in the residuals for the first eight years is much smaller than the variation for the second eight years.

Exhibit 13: Residual Plot for Interest Rates (Y) vs. Inflation Rates (X) Model

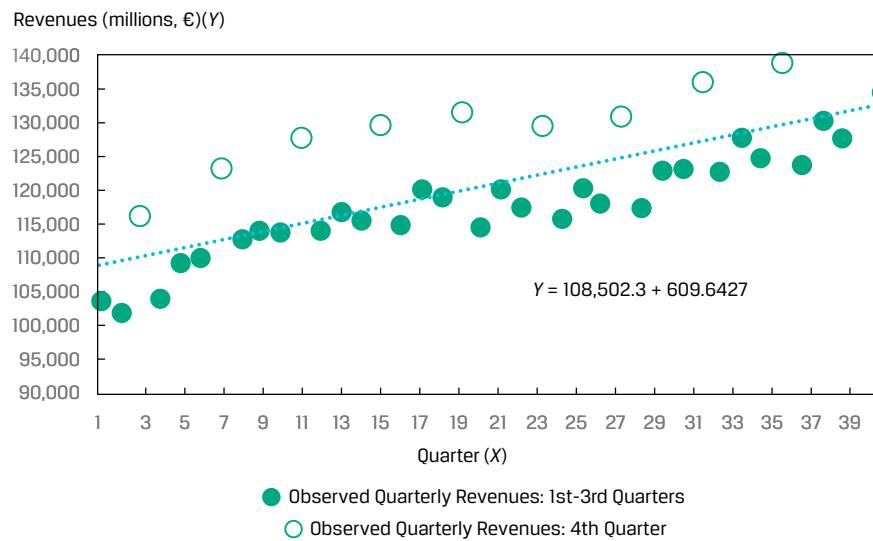
Why does this happen? The model seems appropriate, but when we examine the residuals (Exhibit 13), an important step in assessing the model fit, we see that the model fits better in some years compared with others. The difference in variance of residuals between the two regimes is apparent from the much wider band around residuals for Regime 2 (the low-rate period). This indicates a clear violation of the homoskedasticity assumption.

If we estimate a regression line for each regime, we can see that the model for the two regimes is quite different, as we show in Exhibit 14. In the case of Regime 1 (normal rates), the slope is 1.0247, whereas in Regime 2 (low rates) the slope is -0.2805. In sum, the clustering of residuals in two groups with much different variances clearly indicates the existence of distinct regimes for the relationship between short-term interest rates and the inflation rate.

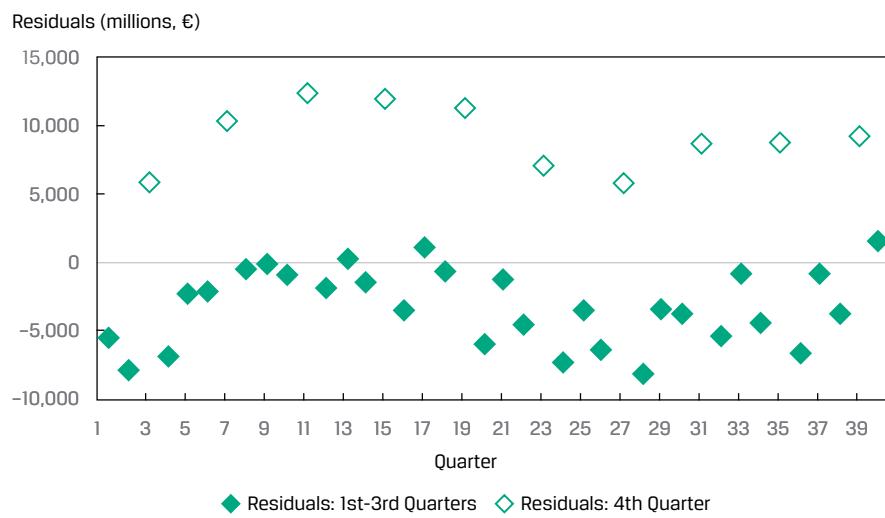
Exhibit 14: Fitted Regression Lines for the Two Regimes**Assumption 3: Independence**

We assume that the observations (Y and X pairs) are uncorrelated with one another, meaning they are independent. If there is correlation between observations (i.e., autocorrelation), they are not independent and the residuals will be correlated. The assumption that the residuals are uncorrelated across observations is also necessary for correctly estimating the variances of the **estimated parameters** of b_0 and b_1 (i.e., \hat{b}_0 and \hat{b}_1) that we use in hypothesis tests of the intercept and slope, respectively. It is important to examine whether the residuals exhibit a pattern, suggesting a violation of this assumption. Therefore, we need to visually and statistically examine the residuals for a regression model.

Consider the quarterly revenues of a company regressed over 40 quarters, as shown in Exhibit 15, with the regression line included. It is clear that these revenues display a seasonal pattern, an indicator of autocorrelation.

Exhibit 15: Regression of Quarterly Revenues vs. Time (40 Quarters)

In Exhibit 16, we plot the residuals from this model and see that there is a pattern. These residuals are correlated, specifically jumping up in Quarter 4 and then falling back the subsequent quarter. In sum, the patterns in both Exhibit 15 and Exhibit 16 indicate a violation of the assumption of independence.

Exhibit 16: Residual Plot for Quarterly Revenues vs. Time Model**Assumption 4: Normality**

The assumption of normality requires that the residuals be normally distributed. This does not mean that the dependent and independent variables must be normally distributed; it only means that the residuals from the model are normally distributed. However, in estimating any model, it is good practice to understand the distribution of

the dependent and independent variables to explore for outliers. An outlier in either or both variables can substantially influence the fitted line such that the estimated model will not fit well for most of the other observations.

With normally distributed residuals, we can test a particular hypothesis about a linear regression model. For large sample sizes, we may be able to drop the assumption of normality by appealing to the central limit theorem; asymptotic theory (which deals with large samples) shows that in many cases, the test statistics produced by standard regression programs are valid even if the model's residuals are not normally distributed.

QUESTION SET



An analyst is investigating a company's revenues and estimates a simple linear time-series model by regressing revenues against time, where time—1, 2, . . . , 15—is measured in years. She plots the company's observed revenues and the estimated regression line, as shown in Exhibit 17. She also plots the residuals from this regression model, as shown in Exhibit 18.

Exhibit 17: Revenues vs. Time Using Simple Linear Regression

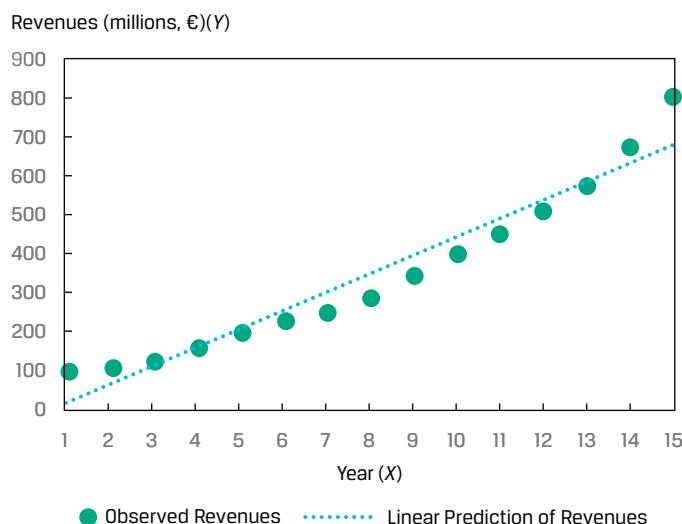
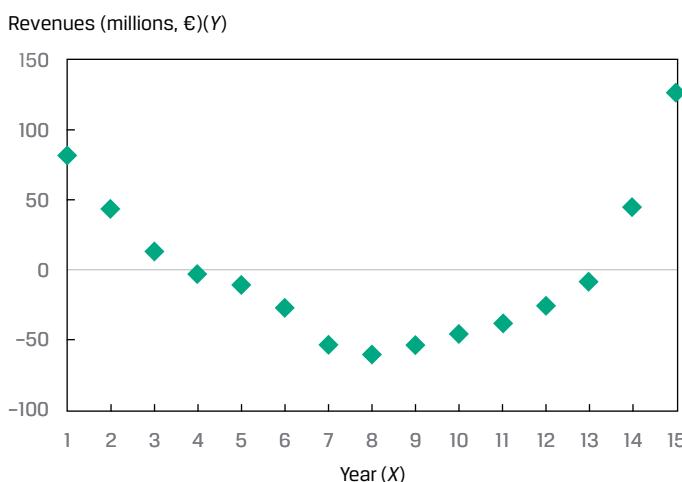


Exhibit 18: Residual Plot for Revenues vs. Time



- Based on Exhibit 17 and Exhibit 18, describe which assumption(s) of simple linear regression the analyst's model may be violating.

Solution:

The correct model is not linear, as evident from the pattern of the revenues in Exhibit 17. In the earlier years (i.e., 1 and 2) and later years (i.e., 14 and 15), the linear model underestimates revenues, whereas for the middle years (i.e., 7–11), the linear model overestimates revenues. Moreover, the curved pattern of residuals in Exhibit 18 indicates potential heteroskedasticity (residuals have unequal variances), lack of independence of observations, and non-normality (a concern given the small sample size of $n = 15$). In sum, the analyst should be concerned that her model violates all the assumptions governing simple linear regression (linearity, homoskedasticity, independence, and normality).

4**HYPOTHESIS TESTS IN THE SIMPLE LINEAR REGRESSION MODEL**

calculate and interpret measures of fit and formulate and evaluate tests of fit and of regression coefficients in a simple linear regression

Analysis of Variance

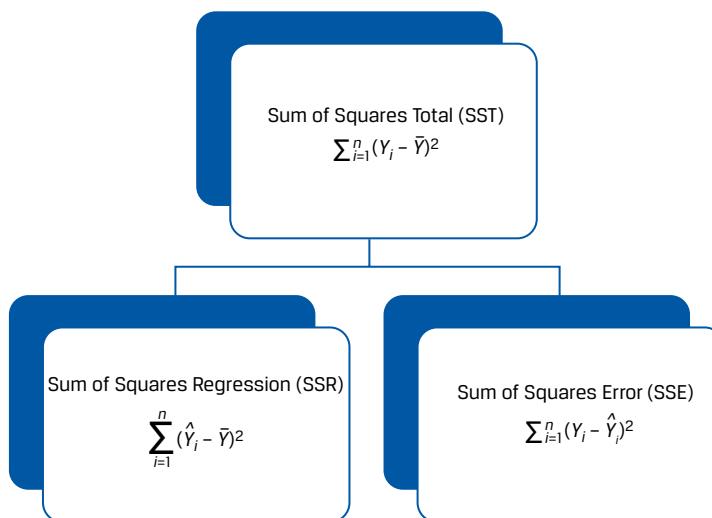
The simple linear regression model sometimes describes the relationship between two variables quite well, but sometimes it does not. We must be able to distinguish between these two cases to use regression analysis effectively. Remember our goal is to explain the variation of the dependent variable. So, how well has this goal been achieved, given our choice of independent variable?

Breaking Down the Sum of Squares Total into Its Components

We begin with the sum of squares total and then break it down into two parts: the sum of squares error and the **sum of squares regression (SSR)**. The sum of squares regression is the sum of the squared differences between the predicted value of the dependent variable, \hat{Y}_i , based on the estimated regression line, and the mean of the dependent variable, \bar{Y} :

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2. \quad (9)$$

We have already defined the sum of squares total, which is the total variation in Y , and the sum of squares error, the unexplained variation in Y . Note that the sum of squares regression is the explained variation in Y . So, as illustrated in Exhibit 19, $SST = SSR + SSE$, meaning total variation in Y equals explained variation in Y plus unexplained variation in Y .

Exhibit 19: Breakdown of Variation of Dependent Variable

We show the breakdown of the sum of squares total formula for our ROA regression example in Exhibit 20. The total variation of ROA that we want to explain (SST) is 239.50. This number includes the variation unexplained (SSE), 47.88, and the variation explained (SSR), 191.63. These sum of squares values are important inputs into measures of the fit of the regression line.

Exhibit 20: Breakdown of the Sum of Squares Total for ROA Model

Company	ROA (Y _i)	CAPEX (X _i)	Predicted ROA (\hat{Y})	Variation to Be Explained $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
A	6.0	0.7	5.750	42.25	0.063	45.563
B	4.0	0.4	5.375	72.25	1.891	50.766
C	15.0	5.0	11.125	6.25	15.016	1.891
D	20.0	10.0	17.375	56.25	6.891	23.766
E	10.0	8.0	14.875	6.25	23.766	5.641
F	20.0	12.5	20.500	56.25	0.250	64.000
				239.50	47.88	191.625
Mean	12.50					

Sum of squares total = 239.50.

Sum of squares error = 47.88.

Sum of squares regression = 191.63.

Measures of Goodness of Fit

We can use several measures to evaluate goodness of fit—that is, how well the regression model fits the data. These include the coefficient of determination, the *F*-statistic for the test of fit, and the standard error of the regression.

The **coefficient of determination (R^2)**, also referred to as the R -squared or R^2 , is the percentage of the variation of the dependent variable that is explained by the independent variable:

$$\text{Coefficient of determination} = \frac{\text{Sum of squares regression}}{\text{Sum of squares total}} \quad (10)$$

$$\text{Coefficient of determination} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

By construction, the coefficient of determination ranges from 0 percent to 100 percent. In our ROA example, the coefficient of determination is $191.625 \div 239.50$, or 0.8001, so 80.01 percent of the variation in ROA is explained by CAPEX. In a simple linear regression, the square of the pairwise correlation is equal to the coefficient of determination:

$$r^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = R^2. \quad (11)$$

In our earlier ROA regression analysis, $r = 0.8945$, so we now see that r^2 is indeed equal to the coefficient of determination (R^2), since $(0.8945)^2 = 0.8001$.

Whereas the coefficient of determination—the portion of the variation of the dependent variable explained by the independent variable—is descriptive, it is not a statistical test. To see if our regression model is likely to be statistically meaningful, we will need to construct an F -distributed test statistic.

In general, we use an F -distributed test statistic to compare two variances. In regression analysis, we can use an F -distributed test statistic to test whether the slopes in a regression are equal to zero, with the slopes designated as b_i , against the alternative hypothesis that at least one slope is not equal to zero:

$$H_0: b_1 = b_2 = b_3 = \dots = b_k = 0.$$

$$H_a: \text{At least one } b_k \text{ is not equal to zero.}$$

For simple linear regression, these hypotheses simplify to

$$H_0: b_1 = 0.$$

$$H_a: b_1 \neq 0.$$

The F -distributed test statistic is constructed by using the sum of squares regression and the sum of squares error, each adjusted for degrees of freedom; in other words, it is the ratio of two variances. We divide the sum of squares regression by the number of independent variables, represented by k . In the case of a simple linear regression, $k = 1$, so we arrive at the **mean square regression (MSR)**, which is the same as the sum of squares regression:

$$\text{MSR} = \frac{\text{Sum of squares regression}}{k} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1}. \quad (12)$$

So, for simple linear regression,

$$\text{MSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2. \quad (13)$$

Next, we calculate the **mean square error (MSE)**, which is the sum of squares error divided by the degrees of freedom, which are $n - k - 1$. In simple linear regression, $n - k - 1$ becomes $n - 2$:

$$\text{MSE} = \frac{\text{Sum of squares error}}{n - k - 1},$$

$$\text{MSE} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}. \quad (14)$$

Therefore, the F -distributed test statistic (MSR/MSE) is

$$F = \frac{\frac{\text{Sum of squares regression}}{k}}{\frac{\text{Sum of squares error}}{n - k - 1}} = \frac{\text{MSR}}{\text{MSE}},$$

$$F = \frac{\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1}}{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}}, \quad (15)$$

which is distributed with 1 and $n - 2$ degrees of freedom in simple linear regression. The F -statistic in regression analysis is one sided, with the rejection region on the right side, because we are interested in whether the variation in Y explained (the numerator) is larger than the variation in Y unexplained (the denominator).

Hypothesis Testing of Individual Regression Coefficients

Hypothesis Tests of the Slope Coefficient

We can use the F -statistic to test for the significance of the slope coefficient (i.e., whether it is significantly different from zero), but we also may want to perform other hypothesis tests for the slope coefficient—for example, testing whether the population slope is different from a specific value or whether the slope is positive. We can use a t -distributed test statistic to test such hypotheses about a regression coefficient.

Suppose we want to check a stock's valuation using the market model; we hypothesize that the stock has an average systematic risk (i.e., risk similar to that of the market), as represented by the coefficient on the market returns variable. Or we may want to test the hypothesis that economists' forecasts of the inflation rate are unbiased (i.e., on average, not overestimating or underestimating actual inflation rates). In each case, does the evidence support the hypothesis? Such questions as these can be addressed with hypothesis tests on the regression slope. To test a hypothesis about a slope, we calculate the test statistic by subtracting the hypothesized population slope (B_1) from the estimated slope coefficient (\hat{b}_1) and then dividing this difference by the standard error of the slope coefficient, $s_{\hat{b}_1}$:

$$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}. \quad (16)$$

This test statistic is t -distributed with $n - k - 1$ or $n - 2$ degrees of freedom because two parameters (an intercept and a slope) were estimated in the regression.

The **standard error of the slope coefficient** ($s_{\hat{b}_1}$) for a simple linear regression is the ratio of the model's standard error of the estimate (s_e), introduced later, to the square root of the variation of the independent variable:

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}. \quad (17)$$

We compare the calculated t -statistic with the critical values to test hypotheses. Note that the greater the variability of the independent variable, the lower the standard error of the slope (Equation 17) and hence the greater the calculated t -statistic (Equation 16). If the calculated t -statistic is outside the bounds of the critical t -values, we reject the null hypothesis, but if the calculated t -statistic is within the bounds of the critical values, we fail to reject the null hypothesis. Similar to tests of the mean, the alternative hypothesis can be two sided or one sided.

Consider our previous simple linear regression example with ROA as the dependent variable and CAPEX as the independent variable. Suppose we want to test whether the slope coefficient of CAPEX is different from zero to confirm our intuition of a

significant relationship between ROA and CAPEX. We can test the hypothesis concerning the slope using the six-step process, as we show in Exhibit 21. As a result of this test, we conclude that the slope is different from zero; that is, CAPEX is a significant explanatory variable of ROA.

Exhibit 21: Test of the Slope for the Regression of ROA on CAPEX

Step 1	State the hypotheses.	$H_0: b_1 = 0$ versus $H_a: b_1 \neq 0$
Step 2	Identify the appropriate test statistic.	$t = \frac{\hat{b}_1 - B_1}{s\hat{b}_1}$ with $6 - 2 = 4$ degrees of freedom.
Step 3	Specify the level of significance.	$\alpha = 5\%$.
Step 4	State the decision rule.	Critical t -values = ± 2.776 . We can determine this from Excel <i>Lower: T.INV(0.025, 4)</i> <i>Upper: T.INV(0.975, 4)</i> R: <code>qt(c(.025, .975), 4)</code> Python: <code>from scipy.stats import t</code> <i>Lower: t.ppf(.025, 4)</i> <i>Upper: t.ppf(.975, 4)</i> We reject the null hypothesis if the calculated t -statistic is less than -2.776 or greater than $+2.776$. The slope coefficient is 1.25 (Exhibit 6). The mean square error is 11.96875 (Exhibit 39). The variation of CAPEX is 122.640 (Exhibit 6). $s_e = \sqrt{11.96875} = 3.459588$.
Step 5	Calculate the test statistic.	Reject the null hypothesis of a zero slope. There is sufficient evidence to indicate that the slope is different from zero.
Step 6	Make a decision.	

A feature of simple linear regression is that the t -statistic used to test whether the slope coefficient is equal to zero and the t -statistic to test whether the pairwise correlation is zero (i.e., $H_0: \rho = 0$ versus $H_a: \rho \neq 0$) are the same value. Just as with a test of a slope, both two-sided and one-sided alternatives are possible for a test of a correlation—for example, $H_0: \rho \leq 0$ versus $H_a: \rho > 0$. The test-statistic to test whether the correlation is equal to zero is as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \quad (18)$$

In our example of ROA regressed on CAPEX, the correlation (r) is 0.8945. To test whether this correlation is different from zero, we perform a test of hypothesis, shown in Exhibit 22. As you can see, we draw a conclusion similar to that for our test of the slope, but it is phrased in terms of the correlation between ROA and CAPEX: There is a significant correlation between ROA and CAPEX.

Exhibit 22: Test of the Correlation between ROA and CAPEX

Step 1	State the hypotheses.	$H_0: \rho = 0$ versus $H_a: \rho \neq 0$
Step 2	Identify the appropriate test statistic.	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with $6 - 2 = 4$ degrees of freedom.
Step 3	Specify the level of significance.	$\alpha = 5\%$.
Step 4	State the decision rule.	Critical t -values = ± 2.776 . Reject the null if the calculated t -statistic is less than -2.776 or greater than $+2.776$.
Step 5	Calculate the test statistic.	$t = \frac{0.8945\sqrt{4}}{\sqrt{1-0.8001}} = 4.00131$.
Step 6	Make a decision.	Reject the null hypothesis of no correlation. There is sufficient evidence to indicate that the correlation between ROA and CAPEX is different from zero.

Another interesting feature of simple linear regression is that the test-statistic used to test the fit of the model (i.e., the F -distributed test statistic) is related to the calculated t -statistic used to test whether the slope coefficient is equal to zero: $t^2 = F$; therefore, $4.00131^2 = 16.0104$.

What if instead we want to test whether there is a one-to-one relationship between ROA and CAPEX, implying a slope coefficient of 1.0. The hypotheses become $H_0: b_1 = 1$ and $H_a: b_1 \neq 1$. The calculated t -statistic is as follows:

$$t = \frac{1.25 - 1}{0.312398} = 0.80026.$$

This calculated test statistic falls within the bounds of the critical values, ± 2.776 , so we fail to reject the null hypothesis: There is not sufficient evidence to indicate that the slope is different from 1.0.

What if instead we want to test whether there is a positive slope or positive correlation, as our intuition suggests? In this case, all the steps are the same as in Exhibit 21 and Exhibit 22 except the critical values because the tests are one sided. For a test of a positive slope or positive correlation, the critical value for a 5 percent level of significance is $+2.132$. We show the test of hypotheses for a positive slope and a positive correlation in Exhibit 23. Our conclusion is that evidence is sufficient to support both a positive slope and a positive correlation.

Exhibit 23: One-Sided Tests for the Slope and Correlation

		Test of the Slope	Test of the Correlation
Step 1	State the hypotheses.	$H_0: b_1 \leq 0$ versus $H_a: b_1 > 0$	$H_0: \rho \leq 0$ versus $H_a: \rho > 0$
Step 2	Identify the appropriate test statistic.	$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}$ with $6 - 2 = 4$ degrees of freedom.	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ with $6 - 2 = 4$ degrees of freedom.
Step 3	Specify the level of significance.	$\alpha = 5\%$.	$\alpha = 5\%$.
Step 4	State the decision rule.	Critical t -value = 2.132. Reject the null if the calculated t -statistic is greater than 2.132.	Critical t -value = 2.132. Reject the null if the calculated t -statistic is greater than 2.132.

Step 5	Calculate the test statistic.	$t = \frac{1.25 - 0}{0.312398} = 4.00131$	$t = \frac{0.8945\sqrt{4}}{\sqrt{1 - 0.8001}} = 4.00131$
Step 6	Make a decision.	Reject the null hypothesis. Evidence is sufficient to indicate that the slope is greater than zero.	Reject the null hypothesis. Evidence is sufficient to indicate that the correlation is greater than zero.

Hypothesis Tests of the Intercept

We may want to test whether the population intercept is a specific value. As a reminder on how to interpret the intercept, consider the simple linear regression with a company's revenue growth rate as the dependent variable (Y), and the GDP growth rate of its home country as the independent variable (X). The intercept is the company's revenue growth rate if the GDP growth rate is 0 percent.

The equation for the standard error of the intercept, $s_{\hat{b}_0}$, is as follows:

$$s_{\hat{b}_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}. \quad (19)$$

We can test whether the intercept is different from the hypothesized value, B_0 , by comparing the estimated intercept (\hat{b}_0) with the hypothesized intercept and then dividing the difference by the standard error of the intercept:

$$t_{intercept} = \frac{\hat{b}_0 - B_0}{s_{\hat{b}_0}} = \frac{\hat{b}_0 - B_0}{\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}}. \quad (20)$$

In the ROA regression example, the intercept is 4.875 percent. Suppose we want to test whether the intercept is greater than 3 percent. The one-sided hypothesis test is shown in Exhibit 24. As you can see, we reject the null hypothesis. In other words, evidence is sufficient that if there are no capital expenditures (CAPEX = 0), ROA is greater than 3 percent.

Exhibit 24: Test of Hypothesis for Intercept for Regression of ROA on CAPEX

Step 1	State the hypotheses.	$H_0: b_0 \leq 3\%$ versus $H_a: b_0 > 3\%$
Step 2	Identify the appropriate test statistic.	$t_{intercept} = \frac{\hat{b}_0 - B_0}{s_{\hat{b}_0}}$ with $6 - 2 = 4$ degrees of freedom.
Step 3	Specify the level of significance.	$\alpha = 5\%$.
Step 4	State the decision rule.	Critical t -value = 2.132. Reject the null if the calculated t -statistic is greater than 2.132.
Step 5	Calculate the test statistic.	$t_{intercept} = \frac{4.875 - 3.0}{\sqrt{\frac{1}{6} + \frac{6.1^2}{122.64}}} = \frac{1.875}{0.68562} = 2.73475$
Step 6	Make a decision.	Reject the null hypothesis. There is sufficient evidence to indicate that the intercept is greater than 3%.

Hypothesis Tests of Slope When the Independent Variable Is an Indicator Variable

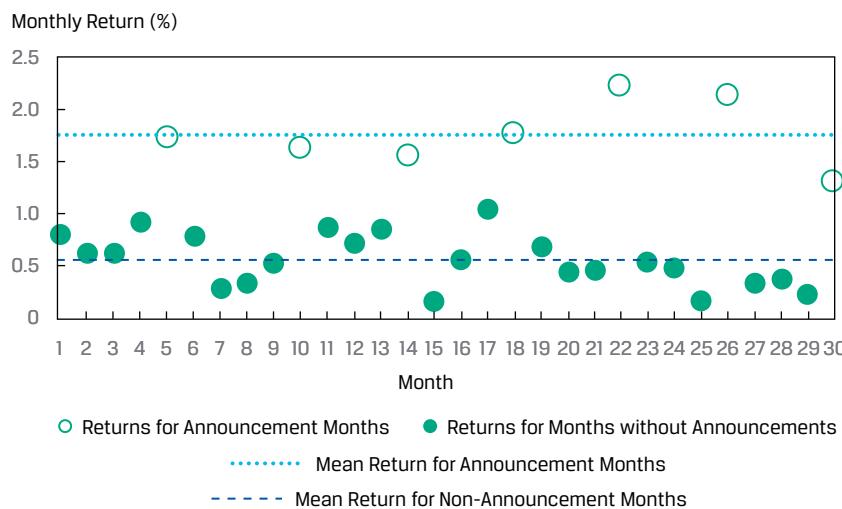
Suppose we want to examine whether a company's quarterly earnings announcements influence its monthly stock returns. In this case, we could use an **indicator variable**, or dummy variable, that takes on only the values 0 or 1 as the independent variable.

Consider the case of a company's monthly stock returns over a 30-month period. A simple linear regression model for investigating this question would be monthly returns, RET, regressed on the indicator variable, EARN, that takes on a value of 0 if there is no earnings announcement that month and 1 if there is an earnings announcement:

$$\text{RET}_i = b_0 + b_1 \text{EARN}_i + \varepsilon_i \quad (21)$$

This regression setup allows us to test whether there are different returns for earnings-announcement months versus non-earnings-announcement months. The observations and regression results are shown graphically in Exhibit 25.

Exhibit 25: Earnings Announcements, Dummy Variable, and Stock Returns



Clearly there are some months in which the returns are different from other months, and these correspond to months in which there was an earnings announcement. We estimate the simple linear regression model and perform hypothesis testing in the same manner as if the independent variable were a continuous variable. In a simple linear regression, the interpretation of the intercept is the predicted value of the dependent variable if the indicator variable is zero. Moreover, the slope when the indicator variable is 1 is the difference in the means if we grouped the observations by the indicator variable. The results of the regression are given in Panel A of Exhibit 26.

Exhibit 26: Regression and Test of Differences Using an Indicator Variable

Regression Estimation Results

Estimated Coefficients	Standard Error of Coefficients	Calculated Test Statistic
Intercept	0.5629	0.0560
EARN	1.2098	0.1158

Regression Estimation Results

Degrees of freedom = 28.

Critical t-values = +2.0484 (5% significance).

Test of Differences in Means

	RET for Earnings-Announcement Months	RET for Non-Earnings-Announcement Months	Difference in Means
Mean	1.7727	0.5629	1.2098
Variance	0.1052	0.0630	
Observations	7	23	
Pooled variance			0.07202
Calculated test statistic			10.4435

Test of Differences in Means

Degrees of freedom = 28.

Critical t-values = +2.0484 (5% significance).

We can see the following from Panel A of Exhibit 26:

- The intercept (0.5629) is the mean of the returns for non-earnings-announcement months.
- The slope coefficient (1.2098) is the difference in means of returns between earnings-announcement and non-earnings-announcement months.
- We reject the null hypothesis that the slope coefficient on EARN is equal to zero. We also reject the null hypothesis that the intercept is zero. The reason is that in both cases, the calculated test statistic exceeds the critical *t*-value.

We could also test whether the mean monthly return is the same for both the non-earnings-announcement months and the earnings-announcement months by testing the following:

$$H_0: \mu_{RETearnings} = \mu_{RETNon-earnings} \text{ and } H_a: \mu_{RETearnings} \neq \mu_{RETNon-earnings}$$

The results of this hypothesis test are gleaned from Panel B of Exhibit 26. As you can see, we reject the null hypothesis that there is no difference in the mean RET for the earnings-announcement and non-earnings-announcements months at the 5 percent level of significance, because the calculated test statistic (10.4435) exceeds the critical value (2.0484).

Test of Hypotheses: Level of Significance and *p*-Values

The choice of significance level in hypothesis testing is always a matter of judgment. Analysts often choose the 0.05 level of significance, which indicates a 5 percent chance of rejecting the null hypothesis when, in fact, it is true (a Type I error, or false positive). Of course, decreasing the level of significance from 0.05 to 0.01 decreases the probability of a Type I error, but it also increases the probability of a Type II error—failing to reject the null hypothesis when, in fact, it is false (i.e., a false negative).

The *p*-value is the smallest level of significance at which the null hypothesis can be rejected. The smaller the *p*-value, the smaller the chance of making a Type I error (i.e., rejecting a true null hypothesis), so the greater the likelihood the regression model is valid. For example, if the *p*-value is 0.005, we reject the null hypothesis that the true parameter is equal to zero at the 0.5 percent significance level (99.5 percent confidence). In most software packages, the *p*-values provided for regression coefficients are for a test of null hypothesis that the true parameter is equal to zero against the alternative that the parameter is not equal to zero.

In our ROA regression example, the calculated t -statistic for the test of whether the slope coefficient is zero is 4.00131. The p -value corresponding to this test statistic is 0.016, which means there is just a 0.16 percent chance of rejecting the null hypotheses when it is true. Comparing this p -value with the level of significance of 5 percent (and critical values of ± 2.776) leads us to easily reject the null hypothesis of $H_0: b_1 = 0$.

How do we determine the p -values? Because this is the area in the distribution outside the calculated test statistic, we need to resort to software tools. For the p -value corresponding to the $t = 4.00131$ from the ROA regression example, we could use the following:

- **Excel** $(1-T.DIST(4.00131,4,TRUE))^*2$
- **R** $(1-pt(4.00131,4))^*2$
- **Python** from `scipy.stats import t` and $(1 - t.cdf(4.00131,4))^*2$

QUESTION SET



The following applies to questions 1–3:

Julie Moon is an energy analyst examining electricity, oil, and natural gas consumption in different regions over different seasons. She ran a simple regression explaining the variation in energy consumption as a function of temperature. The total variation of the dependent variable was 140.58, and the explained variation was 60.16. She had 60 monthly observations.

1. Calculate the coefficient of determination.

Solution:

The coefficient of determination is 0.4279:

$$\frac{\text{Explained variation}}{\text{Total variation}} = \frac{60.16}{140.58} = 0.4279.$$

2. Calculate the F -statistic to test the fit of the model.

Solution:

$$F = \frac{60.16/1}{(140.58 - 60.16)/(60 - 2)} = \frac{60.16}{1.3866} = 43.3882.$$

3. Calculate the sample standard deviation of monthly energy consumption.

Solution:

The sample variance of the dependent variable uses the total variation of the dependent variable and divides it by the number of observations less one:

$$\sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{n-1} = \frac{\text{Total variation}}{n-1} = \frac{140.58}{60-1} = 2.3827.$$

The sample standard deviation of the dependent variable is the square root of the variance, or $\sqrt{2.3827} = 1.544$.

The following applies to questions 4–6:

An analyst is interested in interpreting the results of and performing tests of hypotheses for the market model estimation that regresses the daily return on ABC stock on the daily return on the fictitious Europe–Asia–Africa (EAA) Equity Index, his proxy for the stock market. He has generated the regression results presented in Exhibit 27.

**Exhibit 27: Hypothesis Testing of Simple Linear Regression
Results Selected Results of Estimation of Market Model for
ABC Stock**

Standard error of the estimate (s_e)	1.26
Standard deviation of ABC stock returns	0.80
Standard deviation of EAA Equity Index returns	0.70
Number of observations	1,200

<i>Coefficients</i>	
Intercept	0.010
Slope of EAA Equity Index returns	0.982

4. If the critical t -values are ± 1.96 (at the 5 percent significance level), is the slope coefficient different from zero?

Solution:

First, we calculate the variation of the independent variable using the standard deviation of the independent variable:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \times (n-1).$$

So,

$$\sum_{i=1}^n (X_i - \bar{X})^2 = 0.70^2 \times 1,199 = 587.51.$$

Next, the standard error of the estimated slope coefficient is

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{1.26}{\sqrt{587.51}} = 0.051983,$$

and the test statistic is

$$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}} = \frac{0.982 - 0}{0.051983} = 18.8907.$$

The calculated test statistic is outside the bounds of ± 1.96 , so we reject the null hypothesis of a slope coefficient equal to zero.

5. If the critical t -values are ± 1.96 (at the 5 percent significance level), is the slope coefficient different from 1.0?

Solution:

The calculated test statistic for the test of whether the slope coefficient is equal to 1.0 is

$$t = \frac{0.982 - 1}{0.051983} = -0.3463.$$

The calculated test statistic is within the bounds of ± 1.96 , so we fail to reject the null hypothesis of a slope coefficient equal to 1.0, which is evidence that the true population slope may be 1.0.

6. An economist collected the monthly returns for KDL's portfolio and a diversified stock index. The data collected are shown in Exhibit 28:

Exhibit 28: Monthly Returns for KDL

Month	Portfolio Return (%)	Index Return (%)
1	1.11	-0.59
2	72.10	64.90
3	5.12	4.81
4	1.01	1.68
5	-1.72	-4.97
6	4.06	-2.06

The economist calculated the correlation between the two returns and found it to be 0.996. The regression results with the portfolio return as the dependent variable and the index return as the independent variable are given in Exhibit 29:

Exhibit 29: Regression Results**Regression Statistics**

R^2	0.9921
Standard error	2.8619
Observations	6

Source	df	Sum of Squares	Mean Square	F	p-Value
Regression	1	4,101.6205	4,101.6205	500.7921	0.0000
Residual	4	32.7611	8.1903		
Total	5	4,134.3815			

	Coefficients	Standard Error	t-Statistic	p-Value
Intercept	2.2521	1.2739	1.7679	0.1518
Index return (%)	1.0690	0.0478	22.3784	0.0000

When reviewing the results, Andrea Fusilier suspected that they were unreliable. She found that the returns for Month 2 should have been 7.21 percent and 6.49 percent, instead of the large values shown in the first table. Correcting these values resulted in a revised correlation of 0.824 and the revised regression results in Exhibit 30:

Exhibit 30: Revised Regression Results**Regression Statistics**

R^2	0.6784
Standard error	2.0624
Observations	6

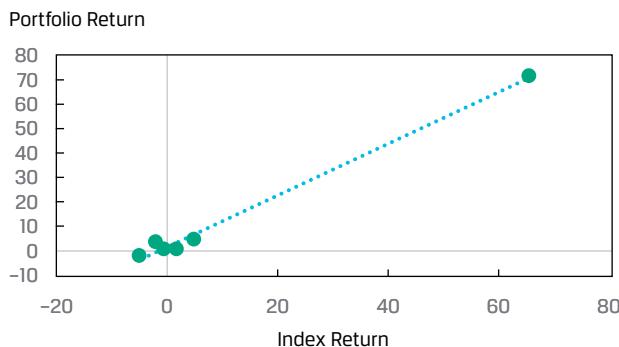
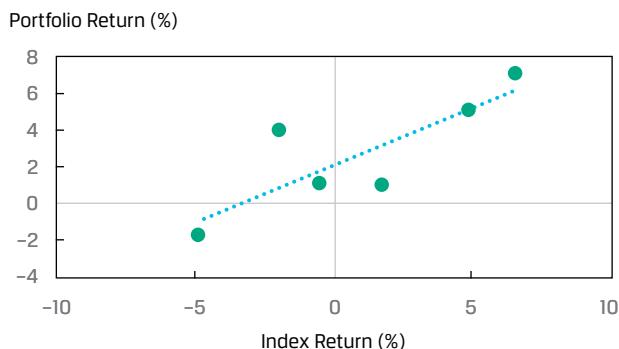
Source	df	Sum of Squares	Mean Square	F	p-Value
Regression	1	35.8950	35.8950	8.4391	0.044
Residual	4	17.0137	4.2534		
Total	5	52.91			

	Coefficients	Standard Error	t-Statistic	p-Value
Intercept	2.2421	0.8635	2.5966	0.060
Slope	0.6217	0.2143	2.9050	0.044

Explain how the bad data affected the results.

Solution:

The Month 2 data point is an outlier, lying far away from the other data values. Because this outlier was caused by a data entry error, correcting the outlier improves the validity and reliability of the regression. In this case, revised R^2 is lower (from 0.9921 to 0.6784). The outliers created the illusion of a better fit from the higher R^2 ; the outliers altered the estimate of the slope. The standard error of the estimate is lower when the data error is corrected (from 2.8619 to 2.0624), as a result of the lower mean square error. At a 0.05 level of significance, both models fit well. The difference in the fit is illustrated in Exhibit 31.

Exhibit 31: Fit of the Model with and without Data Errors**A. Before the Data Errors Are Corrected****B. After the Data Errors Are Corrected**

PREDICTION IN THE SIMPLE LINEAR REGRESSION MODEL

5

- describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the standard error of estimate in a simple linear regression
- calculate and interpret the predicted value for the dependent variable, and a prediction interval for it, given an estimated linear regression model and a value for the independent variable

ANOVA and Standard Error of Estimate in Simple Linear Regression

We often represent the sums of squares from a regression model in an **analysis of variance (ANOVA)** table, as shown in Exhibit 32, which presents the sums of squares, the degrees of freedom, the mean squares, and the F -statistic. Notice that the variance of the dependent variable is the ratio of the sum of squares total to $n - 1$.

Exhibit 32: Analysis of Variance Table for Simple Linear Regression

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Regression	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1}$	$F = \frac{MSR}{MSE} = \frac{\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{1}}{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$
Error	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}$	
Total	$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$		

From the ANOVA table, we can also calculate the **standard error of the estimate** (s_e), which is also known as the standard error of the regression or the root mean square error. The s_e is a measure of the distance between the observed values of the dependent variable and those predicted from the estimated regression—the smaller the s_e , the better the fit of the model. The s_e , along with the coefficient of determination and the F-statistic, is a measure of the goodness of the fit of the estimated regression line. Unlike the coefficient of determination and the F-statistic, which are relative measures of fit, the standard error of the estimate is an absolute measure of the distance of the observed dependent variable from the regression line. Thus, the s_e is an important statistic used to evaluate a regression model and is used to calculate prediction intervals and to perform tests on the coefficients. The calculation of s_e is straightforward once we have the ANOVA table because it is the square root of the MSE:

$$\text{Standard error of the estimate } (s_e) = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}}. \quad (22)$$

We show the ANOVA table for our ROA regression example in Exhibit 32, using the information from Exhibit 33. For a 5 percent level of significance, the critical F-value for the test of whether the model is a good fit (i.e., whether the slope coefficient is different from zero) is 7.71. We can get this critical value in the following ways:

- Excel [F.INV(0.95,1,4)]
- R [qf(.95,1,4)]
- Python [from scipy.stats import f and f.ppf(.95,1,4)]

With a calculated F-statistic of 16.0104 and a critical F-value of 7.71, we reject the null hypothesis and conclude that the slope of our simple linear regression model for ROA is different from zero.

Exhibit 33: ANOVA Table for ROA Regression Model

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Regression	191.625	1	191.625	16.0104
Error	47.875	4	11.96875	
Total	239.50	5		

The calculations to derive the ANOVA table and ultimately to test the goodness of fit of the regression model can be time consuming, especially for samples with many observations. However, statistical packages, such as SAS, SPSS Statistics, and Stata, as well as software, such as Excel, R, and Python, produce the ANOVA table as part of the output for regression analysis.

Prediction Using Simple Linear Regression and Prediction Intervals

Financial analysts often want to use regression results to make predictions about a dependent variable. For example, we might ask, "How fast will the sales of XYZ Corporation grow this year if real GDP grows by 4 percent?" But we are not merely interested in making these forecasts; we also want to know how certain we can be about the forecasts' results. A forecasted value of the dependent variable, \hat{Y}_f , is determined using the estimated intercept and slope, as well as the expected or forecasted independent variable, X_f :

$$\hat{Y}_f = \hat{b}_0 + \hat{b}_1 X_f \quad (23)$$

In our ROA regression model, if we forecast a company's CAPEX to be 6 percent, the forecasted ROA based on our estimated equation is 12.375 percent:

$$\hat{Y}_f = 4.875 + (1.25 \times 6) = 12.375.$$

However, we need to consider that the estimated regression line does not describe the relation between the dependent and independent variables perfectly; it is an average of the relation between the two variables. This is evident because the residuals are not all zero.

Therefore, an interval estimate of the forecast is needed to reflect this uncertainty. The estimated variance of the prediction error, s_f^2 , of Y , given X , is

$$s_f^2 = s_e^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_X^2} \right] = s_e^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right],$$

and the **standard error of the forecast** is

$$s_f = s_e \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}. \quad (24)$$

The standard error of the forecast depends on the following:

- the standard error of the estimate, s_e ;
- the number of observations, n ;
- the forecasted value of the independent variable, X_f , used to predict the dependent variable and its deviation from the estimated mean, \bar{X} ; and
- the variation of the independent variable.

We can see the following from the equation for the standard error of the forecast:

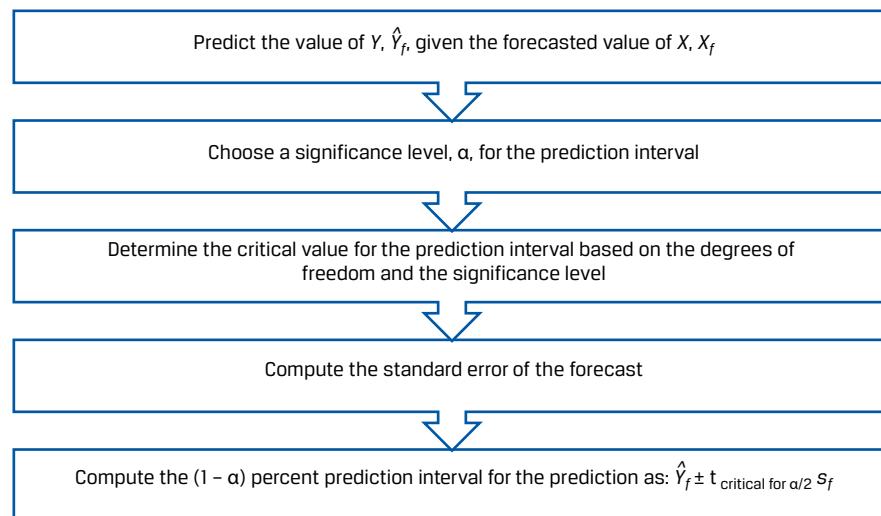
1. The better the fit of the regression model, the smaller the standard error of the estimate (s_e) and, therefore, the smaller standard error of the forecast.
2. The larger the sample size (n) in the regression estimation, the smaller the standard error of the forecast.
3. The closer the forecasted independent variable (X_f) is to the mean of the independent variable (\bar{X}) used in the regression estimation, the smaller the standard error of the forecast.

Once we have this estimate of the standard error of the forecast, determining a prediction interval around the predicted value of the dependent variable (\hat{Y}_f) is very similar to estimating a confidence interval around an estimated parameter. The prediction interval is

$$\hat{Y}_f \pm t_{critical\ for\ \alpha/2} s_f$$

We outline the steps for developing the prediction interval in Exhibit 34.

Exhibit 34: Creating a Prediction Interval around the Predicted Dependent Variable



For our ROA regression model, given that the forecasted value of CAPEX is 6.0, the predicted value of Y is 12.375:

$$\hat{Y}_f = 4.875 + 1.25 X_f = 4.875 + (1.25 \times 6.0) = 12.375.$$

Assuming a 5 percent significance level (α), two sided, with $n - 2$ degrees of freedom (so, $df = 4$), the critical values for the prediction interval are ± 2.776 .

The standard error of the forecast is

$$s_f = 3.459588 \sqrt{1 + \frac{1}{6} + \frac{(6 - 6.1)^2}{122.640}} = 3.459588 \sqrt{1.166748} = 3.736912.$$

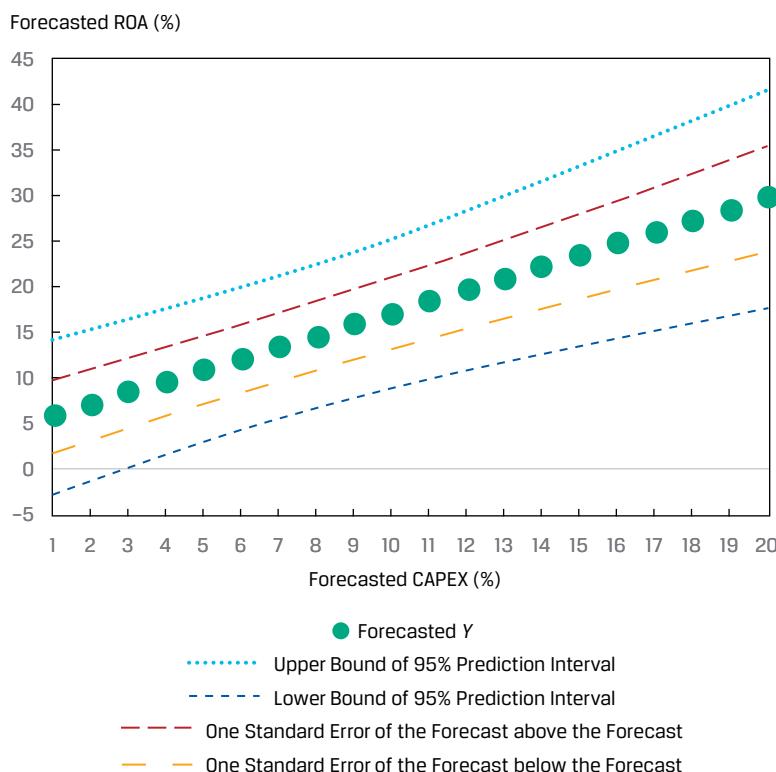
The 95 percent prediction interval then becomes

$$12.375 \pm 2.776 (3.736912)$$

$$12.375 \pm 10.3737$$

$$\{2.0013 < \hat{Y}_f < 22.7487\}$$

For our ROA regression example, we can see how the standard error of the forecast (s_f) changes as our forecasted value of the independent variable gets farther from the mean of the independent variable ($X_f - \bar{X}$) in Exhibit 35. The mean of CAPEX is 6.1 percent, and the band that represents one standard error of the forecast, above and below the forecast, is minimized at that point and increases as the independent variable gets farther from \bar{X} .

Exhibit 35: ROA Forecasts and Standard Error of the Forecast**QUESTION SET**

Suppose you run a cross-sectional regression for 100 companies, where the dependent variable is the annual return on stock and the independent variable is the lagged percentage of institutional ownership (INST). The results of this simple linear regression estimation are shown in Exhibit 36. Evaluate the model by answering questions 1–4.

Exhibit 36: ANOVA Table for Annual Stock Return Regressed on Institutional Ownership

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression	576.1485	1	576.1485
Error	1,873.5615	98	19.1180
Total	2,449.7100		

- What is the coefficient of determination for this regression model?

Solution:

The coefficient of determination is sum of squares regression/sum of squares total: $576.148 \div 2,449.71 = 0.2352$, or 23.52 percent.

2. What is the standard error of the estimate for this regression model?

Solution:

The standard error of the estimate is the square root of the mean square error: $\sqrt{19.1180} = 4.3724$.

3. At a 5 percent level of significance, do we reject the null hypothesis of the slope coefficient equal to zero if the critical F -value is 3.938?

Solution:

Using a six-step process for testing hypotheses, we get the following:

Step 1	State the hypotheses.	$H_0: b_1 = 0$ versus $H_a: b_1 \neq 0$
Step 2	Identify the appropriate test statistic.	$F = \frac{MSR}{MSE}$ with 1 and 98 degrees of freedom.
Step 3	Specify the level of significance.	$\alpha = 5\%$ (one tail, right side).
Step 4	State the decision rule.	Critical F -value = 3.938. Reject the null hypothesis if the calculated F -statistic is greater than 3.938.
Step 5	Calculate the test statistic.	$F = \frac{576.1485}{19.1180} = 30.1364$
Step 6	Make a decision.	Reject the null hypothesis because the calculated F -statistic is greater than the critical F -value. Evidence is sufficient to indicate that the slope coefficient is different from 0.0.

4. Based on your answers to the preceding questions, evaluate this simple linear regression model.

Solution:

The coefficient of determination indicates that variation in the independent variable explains 23.52 percent of the variation in the dependent variable. Also, the F -statistic test confirms that the model's slope coefficient is different from 0 at the 5 percent level of significance. In sum, the model seems to fit the data reasonably well.

The following applies to questions 5-11.

Suppose we want to forecast a company's net profit margin (NPM) based on its research and development expenditures scaled by revenues (RDR), using the model estimated in Example 2 and the details provided in Exhibit 8. The regression model was estimated using data on eight companies as

$$\hat{Y}_f = 16.5 - 1.3X_f$$

with a standard error of the estimate (s_e) of 1.8618987 and variance of RDR, $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n - 1)}$, of 4.285714, as given.

5. What is the predicted value of NPM if the forecasted value of RDR is 5?

Solution:

The predicted value of NPM is 10: $16.5 - (1.3 \times 5) = 10$.

6. What is the standard error of the forecast (s_f) if the forecasted value of RDR is 5?

Solution:

To derive the standard error of the forecast (s_f), we first have to calculate the variation of RDR. Then, we have all the pieces to calculate s_f :

$$\sum_{i=1}^n (X_i - \bar{X})^2 = 4.285714 \times 7 = 30.$$

$$s_f = 1.8618987 \sqrt{1 + \frac{1}{8} + \frac{(5 - 7.5)^2}{30}} = 2.1499.$$

7. What is the 95 percent prediction interval for the predicted value of NPM using critical t -values ($df = 6$) of ± 2.447 ?

Solution:

The 95 percent prediction interval for the predicted value of NPM is

$$\{10 \pm 2.447(2.1499)\},$$

$$\{4.7392 < \hat{Y}_f < 15.2608\}.$$

8. What is the predicted value of NPM if the forecasted value of RDR is 15?

Solution:

The predicted value of NPM is -3 : $16.5 - (1.3 \times 15) = -3$.

9. Referring to exhibit 9, what is the standard error of the forecast if the forecasted value of RDR is 15?

Solution:

To derive the standard error of the forecast, we first must calculate the variation of RDR. Then, we can calculate s_f :

$$\sum_{i=1}^n (X_i - \bar{X})^2 = 4.285714 \times 7 = 30.$$

$$s_f = 1.8618987 \sqrt{1 + \frac{1}{8} + \frac{(15 - 7.5)^2}{30}} = 3.2249.$$

10. What is the 95 percent prediction interval for the predicted value of NPM using critical t -values ($df = 6$) of ± 2.447 ?

Solution:

The 95 percent prediction interval for the predicted value of NPM is

$$\{-3 \pm 2.447(3.2249)\},$$

$$\{-10.8913 < \hat{Y}_f < 4.8913\}.$$

The following applies to questions 12-17.

You are examining the results of a regression estimation that attempts to explain the unit sales growth of a business you are researching. The analysis of variance output for the regression is given in the Exhibit 37. The regression was based on five observations ($n = 5$).

Exhibit 37: ANOVA Output

Source	df	Sum of Squares	Mean Square	F	p-Value
Regression	1	88.0	88.0	36.667	0.00904
Residual	3	7.2	2.4		
Total	4	95.2			

11. Calculate the sample variance of the dependent variable using information in the table.

Solution:

The sample variance of the dependent variable is the sum of squares total divided by its degrees of freedom ($n - 1 = 5 - 1 = 4$, as given). Thus, the sample variance of the dependent variable is $95.2 \div 4 = 23.8$.

12. Calculate the coefficient of determination for this estimated model.

Solution:

The coefficient of determination = $88.0 \div 95.2 = 0.92437$.

13. What hypothesis does the *F*-statistic test?

Solution:

The *F*-statistic tests whether all the slope coefficients in a linear regression are equal to zero.

14. Is the *F*-test significant at the 0.05 significance level?

Solution:

The calculated value of the *F*-statistic is 36.667, as shown in Exhibit 32. The corresponding *p*-value is less than 0.05, so you reject the null hypothesis of a slope equal to zero.

15. Calculate the standard error of the estimate.

Solution:

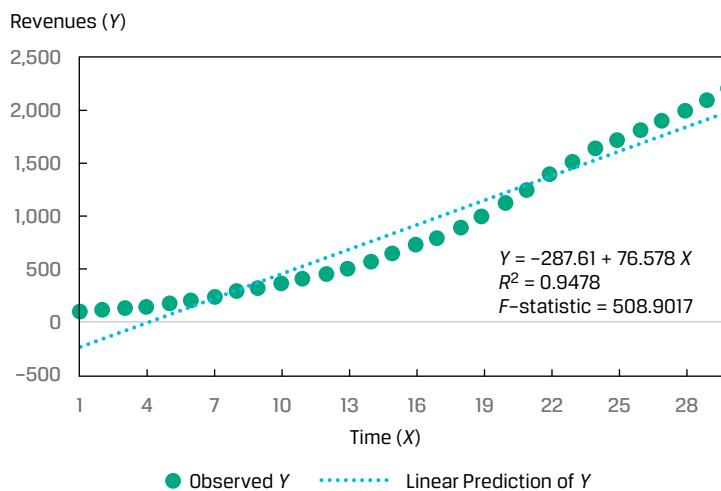
The standard error of the estimate is the square root of the mean square error: $s_e = \sqrt{2.4} = 1.54919$.

6**FUNCTIONAL FORMS FOR SIMPLE LINEAR REGRESSION**

describe different functional forms of simple linear regressions

Not every set of independent and dependent variables has a linear relation. In fact, we often see non-linear relationships in economic and financial data. Consider the revenues of a company over time illustrated in Exhibit 38, with revenues as the dependent

(Y) variable and time as the independent (X) variable. Revenues grow at a rate of 15 percent per year for several years, but then the growth rate eventually declines to just 5 percent per year. Estimating this relationship as a simple linear model would understate the dependent variable, revenues, and, for some ranges of the independent variable, time, and would overstate it for other ranges of the independent variable.

Exhibit 38: Company Revenues over Time


We can still use the simple linear regression model, but we need to modify either the dependent or the independent variables to make it work well. This is the case with many different financial or economic data that you might use as dependent and independent variables in your regression analysis.

Several different functional forms can be used to potentially transform the data to enable their use in linear regression. These transformations include using the log (i.e., natural logarithm) of the dependent variable, the log of the independent variable, the reciprocal of the independent variable, the square of the independent variable, or the differencing of the independent variable. We illustrate and discuss three often-used functional forms, each of which involves log transformation:

1. the **log-lin model**, in which the dependent variable is logarithmic but the independent variable is linear;
2. the **lin-log model**, in which the dependent variable is linear but the independent variable is logarithmic; and
3. the **log-log model**, in which both the dependent and independent variables are in logarithmic form.

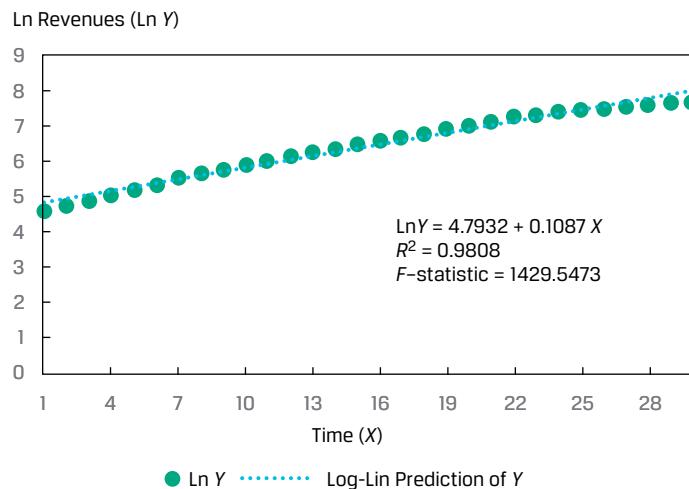
The Log-Lin Model

In the log-lin model, the dependent variable is in logarithmic form and the independent variable is not, as follows:

$$\ln Y_i = b_0 + b_1 X_i \quad (25)$$

The slope coefficient in this model is the relative change in the dependent variable for an absolute change in the independent variable. We can transform the Y variable (revenues) in Exhibit 38 into its natural log (\ln) and then fit the regression line, as shown in Exhibit 39. From this chart, we see that the log-lin model is a better-fitting model than the simple linear regression model.

Exhibit 39: Log-Lin Model Applied to Company Revenues over Time



It is important to note that in working with a log-lin model, you must take care when making a forecast. For example, suppose the estimated regression model is $\ln Y = -7 + 2X$. If X is 2.5 percent, then the forecasted value of $\ln Y$ is -2 . In this case, the predicted value of Y is the antilog of -2 , or $e^{-2} = 0.135335$. Another caution is that you cannot directly compare a log-lin model with a lin-lin model (i.e., the regression of Y on X without any transformation) because the dependent variables are not in the same form—we would have to transform the R^2 and F -statistic to enable a comparison. Looking at the residuals, however, is helpful:

The Lin-Log Model

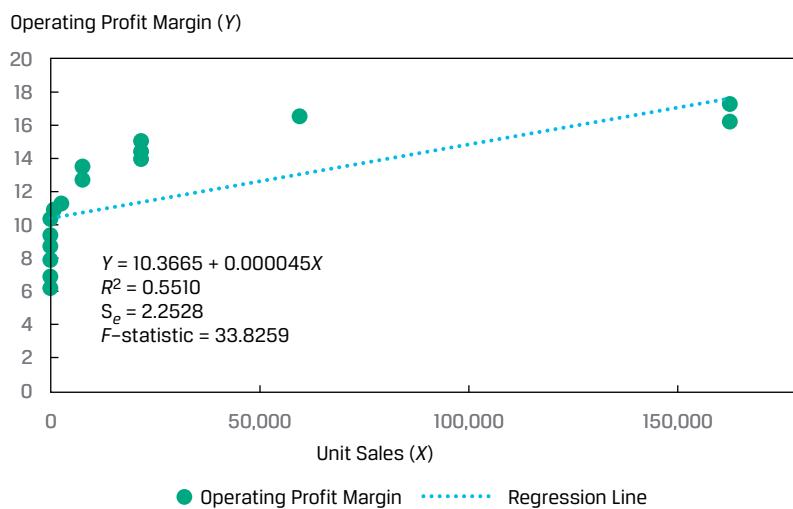
The lin-log model is similar to the log-lin model, but only the independent variable is in logarithmic form:

$$Y_i = b_0 + b_1 \ln X_i \quad (26)$$

The slope coefficient in this regression model provides the absolute change in the dependent variable for a relative change in the independent variable.

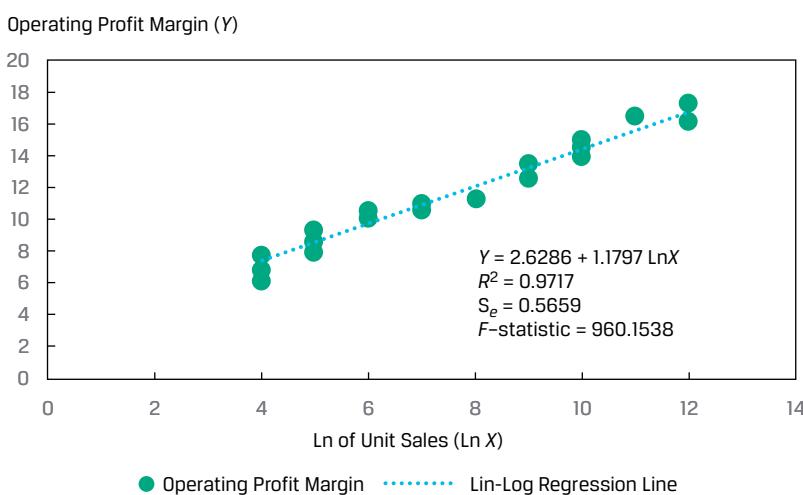
Suppose an analyst is examining the cross-sectional relationship between operating profit margin, the dependent variable (Y), and unit sales, the independent variable (X), and gathers data on a sample of 30 companies. The scatter plot and regression line for these observations are shown in Exhibit 40. Although the slope is different from zero at the 5 percent level (the calculated t -statistic on the slope is 5.8616, compared with critical t -values of ± 2.048), given the R^2 of 55.10 percent, the issue is whether we can get a better fit by using a different functional form.

Exhibit 40: Relationship between Operating Profit Margin and Unit Sales



If instead we use the natural log of the unit sales as the independent variable in our model, we get a very different picture, as shown in Exhibit 41. The R^2 for the model of operating profit margin regressed on the natural log of unit sales jumps to 97.17 percent. Because the dependent variable is the same in both the original and transformed models, we can compare the standard error of the estimate: 2.2528 with the original independent variable and a much lower 0.5629 with the transformed independent variable. Clearly the log-transformed explanatory variable has resulted in a better-fitting model.

Exhibit 41: Relationship Between Operating Profit Margin and Natural Logarithm of Unit Sales



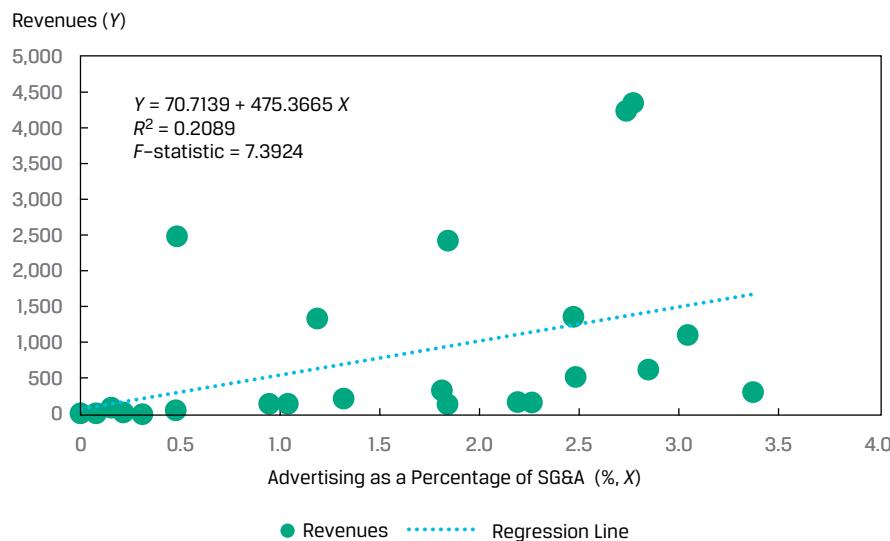
The Log-Log Model

The log-log model, in which both the dependent variable and the independent variable are linear in their logarithmic forms, is also referred to as the double-log model:

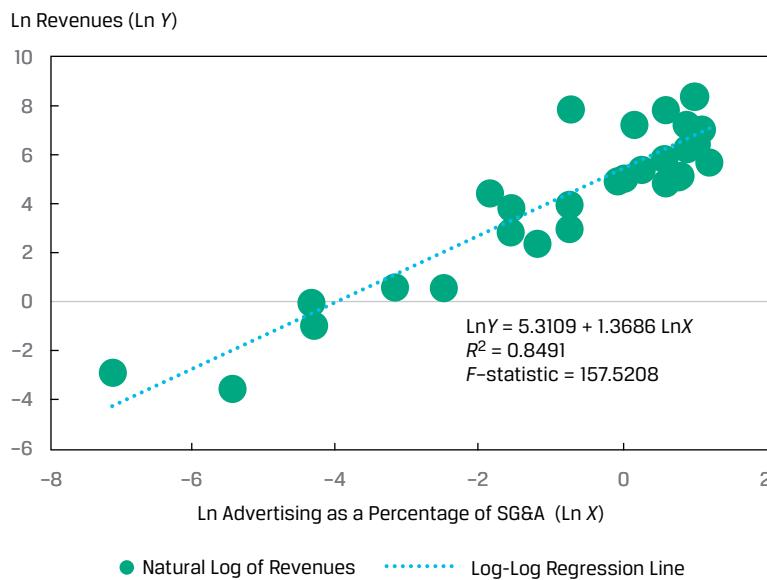
$$\ln Y_i = b_0 + b_1 \ln X_i \quad (27)$$

This model is useful in calculating elasticities because the slope coefficient is the relative change in the dependent variable for a relative change in the independent variable. Consider a cross-sectional model of company revenues (the Y variable) regressed on advertising spending as a percentage of selling, general, and administrative expenses, ADVERT (the X variable). As shown in Exhibit 42, a simple linear regression model results in a shallow regression line, with a coefficient of determination of just 20.89 percent.

Exhibit 42: Fitting a Linear Relation Between Revenues and Advertising Spending



If instead we use the natural logarithms of both the revenues and ADVERT, we get a much different picture of this relationship. As shown in Exhibit 43, the estimated regression line has a significant positive slope; the log-log model's R^2 increases by more than four times, from 20.89 percent to 84.91 percent; and the F -statistic jumps from 7.39 to 157.52. So, using the log-log transformation dramatically improves the regression model fit relative to our data.

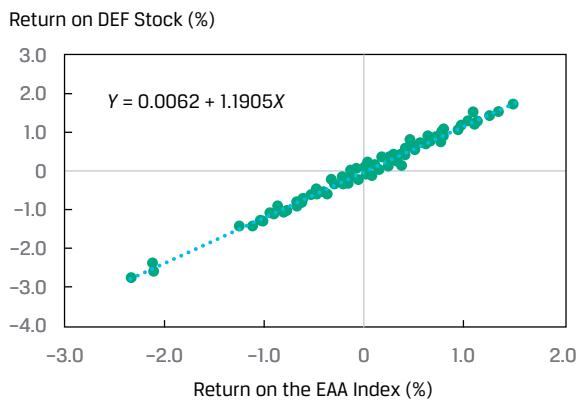
Exhibit 43: Fitting a Log-Log Model of Revenues and Advertising Spending**Selecting the Correct Functional Form**

The key to fitting the appropriate functional form of a simple linear regression is examining the goodness-of-fit measures—the coefficient of determination (R^2), the F -statistic, and the standard error of the estimate (s_e)—as well as examining whether there are patterns in the residuals. In addition to fit statistics, most statistical packages provide plots of residuals as part of the regression output, which enables you to visually inspect the residuals. To reiterate an important point, what you want to see in these plots is random residuals.

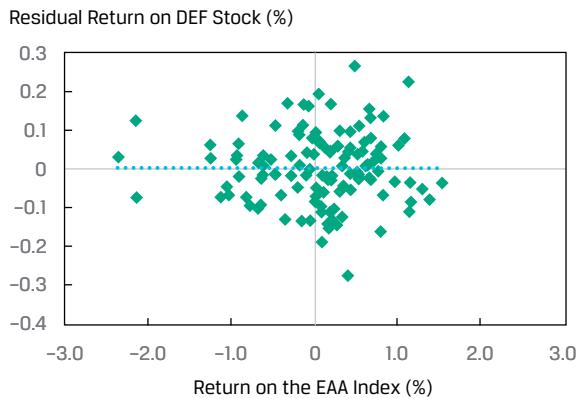
As an example, consider the relationship between the monthly returns on DEF stock and the monthly returns of the EAA Equity Index, as depicted in Panel A of Exhibit 44, with the regression line indicated. Using the equation for this regression line, we calculate the residuals and plot them against the EAA Equity Index, as shown in Panel B of Exhibit 44. The residuals appear to be random, bearing no relation to the independent variable. The distribution of the residuals, shown in Panel C of Exhibit 44, shows that the residuals are approximately normal. Using statistical software, we can investigate further by examining the distribution of the residuals, including using a normal probability plot or statistics to test for normality of the residuals.

Exhibit 44: Monthly Returns on DEF Stock Regressed on Returns on the EAA Index

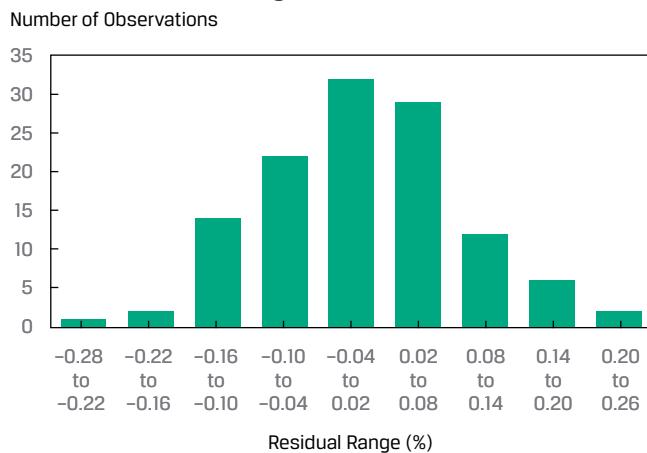
A. Scatterplot of Returns on DEF Stock and Return on the EAA Index



B. Scatterplot of Residuals and the Returns on the EAA Index



C. Histogram of Residuals



QUESTION SET



An analyst is investigating the relationship between the annual growth in consumer spending (CONS) in a country and the annual growth in the country's GDP (GGDP). The analyst estimates the two models in Exhibit 45.

Exhibit 45: Model Estimates

	Model 1	Model 2
	$GGDP_i = b_0 + b_1 \text{CONS}_i + \varepsilon_i$	$GGDP_i = b_0 + b_1 \ln(\text{CONS}_i) + \varepsilon_i$
Intercept	1.040	1.006
Slope	0.669	1.994
R^2	0.788	0.867
Standard error of the estimate	0.404	0.320
<i>F</i> -statistic	141.558	247.040

1. Identify the functional form used in these models.

Solution:

Model 1 is a simple linear regression model with no variable transformation, whereas Model 2 is a lin-log model with the natural log of the variable CONS as the independent variable.

2. Explain which model has better goodness of fit with the sample data.

Solution:

The lin-log model, Model 2, fits the data better. Since the dependent variable is the same for the two models, we can compare the fit of the models using either the relative measures (R^2 or *F*-statistic) or the absolute measure of fit, the standard error of the estimate. The standard error of the estimate is lower for Model 2, whereas the R^2 and *F*-statistic are higher for Model 2 compared with Model 1.

PRACTICE PROBLEMS

The following information relates to questions 1-3

An analyst has estimated a model that regresses a company's return on equity (ROE) against its growth opportunities (GO), defined as the company's three-year compounded annual growth rate in sales, over 20 years, and produces the following estimated simple linear regression:

$$\text{ROE}_i = 4 + 1.8 \text{ GO}_i + \varepsilon_i$$

Both variables are stated in percentages, so a GO observation of 5 percent is included as 5.

1. The predicted value of the company's ROE if its GO is 10 percent is closest to:
 - A. 1.8 percent.
 - B. 15.8 percent.
 - C. 22.0 percent.
2. The change in ROE for a change in GO from 5 percent to 6 percent is closest to:
 - A. 1.8 percent.
 - B. 4.0 percent.
 - C. 5.8 percent.
3. The residual in the case of a GO of 8 percent and an observed ROE of 21 percent is closest to:
 - A. -1.8 percent.
 - B. 2.6 percent.
 - C. 12.0 percent.

4. Homoskedasticity is best described as the situation in which the variance of the residuals of a regression is:
 - A. zero.
 - B. normally distributed.
 - C. constant across observations.

The following information relates to questions 5-8

An analyst is examining the annual growth of the money supply for a country over the past 30 years. This country experienced a central bank policy shift 15 years ago, which altered the approach to the management of the money supply. The analyst estimated a model using the annual growth rate in the money supply regressed on the variable (SHIFT) that takes on a value of 0 before the policy shift and 1 after. She estimated the values in Exhibit 1:

Exhibit 1: SHIFT Estimates

	Coefficients	Standard Error	t-Stat.
Intercept	5.767264	0.445229	12.95348
SHIFT	-5.13912	0.629649	-8.16188

*Critical t-values, level of significance of 0.05:
eOne-sided, left side: -1.701
fOne-sided, right side: +1.701
gTwo-sided: ±2.048*

5. The variable SHIFT is best described as:
 - A. an indicator variable.
 - B. a dependent variable.
 - C. a continuous variable.
6. The interpretation of the intercept is the mean of the annual growth rate of the money supply:
 - A. before the shift in policy.
 - B. over the entire period.
 - C. after the shift in policy.
7. The interpretation of the slope is the:
 - A. change in the annual growth rate of the money supply per year.
 - B. average annual growth rate of the money supply after the shift in policy.
 - C. difference in the average annual growth rate of the money supply from before to after the shift in policy.
8. Testing whether there is a change in the money supply growth after the shift in policy, using a 0.05 level of significance, we conclude that there is:
 - A. sufficient evidence that the money supply growth changed.
 - B. not enough evidence that the money supply growth is different from zero.
 - C. not enough evidence to indicate that the money supply growth changed.

The following information relates to questions 9-12

Kenneth McCoin, CFA, is a challenging interviewer. Last year, he handed each job applicant a sheet of paper with the information in Exhibit 1, and he then asked several questions about regression analysis. Some of McCoin's questions, along with a sample of the answers he received to each, are given below. McCoin told the applicants that the independent variable is the ratio of net income to sales for restaurants with a market cap of more than \$100 million and the dependent variable is the ratio of cash flow from operations to sales for those restaurants. Which of the choices provided is the best answer to each of McCoin's questions?

Exhibit 1: Regression Analysis

Regression Statistics					
R^2		0.7436			
Standard error		0.0213			
Observations		24			
Analysis of Variance					
Source	df	Sum of Squares	Mean Square	F	p-Value
Regression	1	0.029	0.029000	63.81	0
Residual	22	0.010	0.000455		
Total	23	0.040			
Coefficients					
	Coefficients	Standard Error	t-Statistic	p-Value	
Intercept	0.077	0.007	11.328	0	
Net income to sales (%)	0.826	0.103	7.988	0	

9. The coefficient of determination is *closest* to:
- 0.7436.
 - 0.8261.
 - 0.8623.
10. The correlation between X and Y is *closest* to:
- 0.7436.
 - 0.7436.
 - 0.8623.
11. If the ratio of net income to sales for a restaurant is 5 percent, the predicted ratio

of cash flow from operations (CFO) to sales is *closest* to:

- A. -4.054.
- B. 0.524.
- C. 4.207.

12. Is the relationship between the ratio of cash flow to operations and the ratio of net income to sales significant at the 0.05 level?

- A. No, because the R^2 is greater than 0.05
 - B. No, because the p -values of the intercept and slope are less than 0.05
 - C. Yes, because the p -values for F and t for the slope coefficient are less than 0.05
-

The following information relates to questions 13-17

Howard Golub, CFA, is preparing to write a research report on Stellar Energy Corp. common stock. One of the world's largest companies, Stellar is in the business of refining and marketing oil. As part of his analysis, Golub wants to evaluate the sensitivity of the stock's returns to various economic factors. For example, a client recently asked Golub whether the price of Stellar Energy Corp. stock has tended to rise following increases in retail energy prices. Golub believes the association between the two variables is negative, but he does not know the strength of the association.

Golub directs his assistant, Jill Batten, to study the relationships between (1) Stellar monthly common stock returns and the previous month's percentage change in the US Consumer Price Index for Energy (CPIENG) and (2) Stellar monthly common stock returns and the previous month's percentage change in the US Producer Price Index for Crude Energy Materials (PPICEM). Golub wants Batten to run both a correlation and a linear regression analysis. In response, Batten compiles the summary statistics shown in Exhibit 1 for 248 months. All the data are in decimal form, where 0.01 indicates a 1 percent return. Batten also runs a regression analysis using Stellar monthly returns as the dependent variable and the monthly change in CPIENG as the independent variable. Exhibit 2 displays the results of this regression model.

Exhibit 1: Descriptive Statistics

	Stellar Common Stock Monthly Return	Lagged Monthly Change	
		CPIENG	PPICEM
Mean	0.0123	0.0023	0.0042
Standard deviation	0.0717	0.0160	0.0534
Covariance, Stellar vs. CPIENG		-0.00017	

	Stellar Common Stock Monthly Return	Lagged Monthly Change	
		CPIENG	PPICEM
Covariance, Stellar vs. PPICEM	-0.00048		
Covariance, CPIENG vs. PPICEM	0.00044		
Correlation, Stellar vs. CPIENG	-0.1452		

Exhibit 2: Regression Analysis with CPIENG

Regression Statistics			
	Coefficients	Standard Error	t-Statistic
R^2		0.0211	
Standard error of the estimate		0.0710	
Observations		248	

	Coefficients	Standard Error	t-Statistic
Intercept	0.0138	0.0046	3.0275
CPIENG (%)	-0.6486	0.2818	-2.3014

Critical t-values

One-sided, left side: -1.651

One-sided, right side: +1.651

Two-sided: ±1.967

13. Which of the following best describes Batten's regression?
- Time-series regression
 - Cross-sectional regression
 - Time-series and cross-sectional regression
14. Based on the regression, if the CPIENG *decreases* by 1.0 percent, the expected return on Stellar common stock during the next period is *closest* to:
- 0.0073 (0.73 percent).
 - 0.0138 (1.38 percent).
 - 0.0203 (2.03 percent).
15. Based on Batten's regression model, the coefficient of determination indicates that:
- Stellar's returns explain 2.11 percent of the variability in CPIENG.
 - Stellar's returns explain 14.52 percent of the variability in CPIENG.
 - changes in CPIENG explain 2.11 percent of the variability in Stellar's returns.
16. For Batten's regression model, 0.0710 is the standard deviation of:
- the dependent variable.

- B. the residuals from the regression.
- C. the predicted dependent variable from the regression.
17. For the analysis run by Batten, which of the following is an *incorrect* conclusion from the regression output?
- A. The estimated intercept from Batten's regression is statistically different from zero at the 0.05 level of significance.
- B. In the month after the CPIENG declines, Stellar's common stock is expected to exhibit a positive return.
- C. Viewed in combination, the slope and intercept coefficients from Batten's regression are not statistically different from zero at the 0.05 level of significance.
-

The following information relates to questions 18-26

Anh Liu is an analyst researching whether a company's debt burden affects investors' decision to short the company's stock. She calculates the short interest ratio (the ratio of short interest to average daily share volume, expressed in days) for 50 companies as of the end of the year and compares this ratio with the companies' debt ratio (the ratio of total liabilities to total assets, expressed in decimal form).

Liu provides a number of statistics in Exhibit 1. She also estimates a simple regression to investigate the effect of the debt ratio on a company's short interest ratio. The results of this simple regression, including the analysis of variance (ANOVA), are shown in Exhibit 2.

In addition to estimating a regression equation, Liu graphs the 50 observations using a scatter plot, with the short interest ratio on the vertical axis and the debt ratio on the horizontal axis.

Exhibit 1: Summary Statistics

Statistic	Debt Ratio (X_i)	Short Interest Ratio (Y_i)
Sum	19.8550	192.3000
Sum of squared deviations from the mean	$\sum_{i=1}^n (X_i - \bar{X})^2 = 2.2225.$	$\sum_{i=1}^n (Y_i - \bar{Y})^2 = 412.2042.$
Sum of cross-products of deviations from the mean	$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = -9.2430.$	

Exhibit 2: Regression of the Short Interest Ratio on the Debt Ratio

ANOVA	Degrees of Freedom (df)	Sum of Squares	Mean Square
Regression	1	38.4404	38.4404
Residual	48	373.7638	7.7867
Total	49	412.2042	
Regression Statistics			
R^2	0.0933		
Standard error of estimate	2.7905		
Observations	50		

	Coefficients	Standard Error	t-Statistic
Intercept	5.4975	0.8416	6.5322
Debt ratio (%)	-4.1589	1.8718	-2.2219

Critical t-values for a 0.05 level of significance:

One-sided, left side: -1.677

One-sided, right side: +1.677

Two-sided: ±2.011

Liu is considering three interpretations of these results for her report on the relationship between debt ratios and short interest ratios:

Interpretation 1 Companies' higher debt ratios cause lower short interest ratios.

Interpretation 2 Companies' higher short interest ratios cause higher debt ratios.

Interpretation 3 Companies with higher debt ratios tend to have lower short interest ratios.

She is especially interested in using her estimation results to predict the short interest ratio for MQD Corporation, which has a debt ratio of 0.40.

18. Based on Exhibit 1 and Exhibit 2, if Liu were to graph the 50 observations, the scatter plot summarizing this relation would be *best* described as:

A. horizontal.

B. upward sloping.

C. downward sloping.

19. Based on Exhibit 1, the sample covariance is *closest to*:

A. -9.2430.

B. -0.1886.

C. 8.4123.

20. Based on Exhibit 1 and Exhibit 2, the correlation between the debt ratio and the

short interest ratio is *closest to*:

- A. -0.3054.
- B. 0.0933.
- C. 0.3054.

21. Which of the interpretations *best* describes Liu's findings?

- A. Interpretation 1
- B. Interpretation 2
- C. Interpretation 3

22. The dependent variable in Liu's regression analysis is the:

- A. intercept.
- B. debt ratio.
- C. short interest ratio.

23. Based on Exhibit 2, the number of degrees of freedom for the *t*-test of the slope coefficient in this regression is:

- A. 48.
- B. 49.
- C. 50.

24. Which of the following should Liu conclude from the results shown in Exhibit 2?

- A. The average short interest ratio is 5.4975.
- B. The estimated slope coefficient is different from zero at the 0.05 level of significance.
- C. The debt ratio explains 30.54 percent of the variation in the short interest ratio.

25. Based on Exhibit 2, the short interest ratio expected for MQD Corporation is *closest to*:

- A. 3.8339.
- B. 5.4975.
- C. 6.2462.

26. Based on Liu's regression results in Exhibit 2, the *F*-statistic for testing whether the slope coefficient is equal to zero is *closest to*:

- A. -2.2219.
 - B. 3.5036.
 - C. 4.9367.
-

The following information relates to questions 27-29

Doug Abitbol is a portfolio manager for Polyi Investments, a hedge fund that trades in the United States. Abitbol manages the hedge fund with the help of Robert Olabudo, a junior portfolio manager.

Abitbol looks at economists' inflation forecasts and would like to examine the relationship between the US Consumer Price Index (US CPI) consensus forecast and the actual US CPI using regression analysis. Olabudo estimates regression coefficients to test whether the consensus forecast is unbiased. If the consensus forecasts are unbiased, the intercept should be 0.0 and the slope will be equal to 1.0. Regression results are presented in Exhibit 1. Additionally, Olabudo calculates the 95 percent prediction interval of the actual CPI using a US CPI consensus forecast of 2.8.

Exhibit 1: Regression Output: Estimating US CPI

Regression Statistics

R^2	0.9859
Standard error of estimate	0.0009
Observations	60

	Coefficients	Standard Error	t-Statistic
Intercept	0.0001	0.0002	0.5000
US CPI consensus forecast	0.9830	0.0155	63.4194

Notes:

1. The absolute value of the critical value for the t -statistic is 2.002 at the 5 percent level of significance.
2. The standard deviation of the US CPI consensus forecast is $s_x = 0.7539$.
3. The mean of the US CPI consensus forecast is $\bar{X} = 1.3350$.

Finally, Abitbol and Olabudo discuss the forecast and forecast interval:

- Observation 1. For a given confidence level, the forecast interval is the same no matter the US CPI consensus forecast.
- Observation 2. A larger standard error of the estimate will result in a wider confidence interval.

27. Based on Exhibit 1, Olabudo should:

- A. conclude that the inflation predictions are unbiased.
- B. reject the null hypothesis that the slope coefficient equals one.
- C. reject the null hypothesis that the intercept coefficient equals zero.

28. Based on Exhibit 1, Olabudo should calculate a prediction interval for the actual US CPI closest to:

- A. 2.7506 to 2.7544.

- B. 2.7521 to 2.7529.
 C. 2.7981 to 2.8019.
29. Which of Olabudo's observations of forecasting is correct?
- A. Only Observation 1
 B. Only Observation 2
 C. Both Observation 1 and Observations 2
-

The following information relates to questions 30-34

Elena Vasileva recently joined EnergyInvest as a junior portfolio analyst. Vasileva's supervisor asks her to evaluate a potential investment opportunity in Amtex, a multinational oil and gas corporation based in the United States. Vasileva's supervisor suggests using regression analysis to examine the relation between Amtex shares and returns on crude oil.

Vasileva notes the following assumptions of regression analysis:

- Assumption 1. The error term is uncorrelated across observations.
- Assumption 2. The variance of the error term is the same for all observations.
- Assumption 3. The dependent variable is normally distributed.

Vasileva runs a regression of Amtex share returns on crude oil returns using the monthly data she collected. Selected data used in the regression are presented in Exhibit 1, and selected regression output is presented in Exhibit 2. She uses a 1 percent level of significance in all her tests.

Exhibit 1: Selected Data for Crude Oil Returns and Amtex Share Returns

	Oil Return (X_i)	Amtex Return (Y_i)	Cross-Product ($X_i - \bar{X}$) ($Y_i - \bar{Y}$)	Predicted Amtex Return (\hat{Y}_i)	Regression Residual ($Y_i - \hat{Y}_i$)	Squared Residual ($Y_i - \hat{Y}_i$) ²
Month 1	-0.032000	0.033145	-0.000388	0.002011	-0.031134	0.000969
:	:	:	:	:	:	:
Month 36	0.028636	0.062334	0.002663	0.016282	-0.046053	0.002121
Sum			0.085598			0.071475
Average	-0.018056	0.005293				

Exhibit 2: Selected Regression Output, Dependent Variable: Amtex Share Return

	Coefficient	Standard Error
Intercept	0.0095	0.0078
Oil return	0.2354	0.0760

Critical t -values for a 1 percent level of significance:

One-sided, left side: -2.441

One-sided, right side: +2.441

Two-sided: ± 2.728

Vasileva expects the crude oil return next month, Month 37, to be -0.01. She computes the standard error of the forecast to be 0.0469.

30. Which of Vasileva's assumptions regarding regression analysis is *incorrect*?

- A. Assumption 1
- B. Assumption 2
- C. Assumption 3

31. Based on Exhibit 1, the standard error of the estimate is *closest* to:

- A. 0.04456.
- B. 0.04585.
- C. 0.05018.

32. Based on Exhibit 2, Vasileva should reject the null hypothesis that:

- A. the slope is less than or equal to 0.15.
- B. the intercept is less than or equal to zero.
- C. crude oil returns do not explain Amtex share returns.

33. Based on Exhibit 2 and Vasileva's prediction of the crude oil return for Month 37, the estimate of Amtex share return for Month 37 is *closest* to:

- A. -0.0024.
- B. 0.0071.
- C. 0.0119.

34. Using information from Exhibit 2, the 99 percent prediction interval for Amtex share return for Month 37 is *best* described as:

- A. $\hat{Y}_f \pm 0.0053$.
- B. $\hat{Y}_f \pm 0.0469$.
- C. $\hat{Y}_f \pm 0.1279$.

The following information relates to questions

35-38

Espey Jones is examining the relation between the net profit margin (NPM) of companies, in percent, and their fixed asset turnover (FATO). He collected a sample of 35 companies for the most recent fiscal year and fit several different functional forms, settling on the following model:

$$\ln(\text{NPM}_i) = b_0 + b_1 \text{FATO}_i$$

The results of this estimation are provided in Exhibit 1.

Exhibit 1: Results of Regressing NPM on FATO

Source	df	Sum of Squares	Mean Square	F	p-Value
Regression	1	102.9152	102.9152	1,486.7079	0.0000
Residual	32	2.2152	0.0692		
Total	33	105.1303			
<hr/>					
	Coefficients	Standard Error	t- Statistic	p-Value	
Intercept	0.5987	0.0561	10.6749	0.0000	
FATO	0.2951	0.0077	38.5579	0.0000	

35. The coefficient of determination is *closest* to:

- A. 0.0211.
- B. 0.9789.
- C. 0.9894.

36. The standard error of the estimate is *closest* to:

- A. 0.2631.
- B. 1.7849.
- C. 38.5579.

37. At a 0.01 level of significance, Jones should conclude that:

- A. the mean net profit margin is 0.5987 percent.
- B. the variation of the fixed asset turnover explains the variation of the natural log of the net profit margin.
- C. a change in the fixed asset turnover from three to four times is likely to result in a change in the net profit margin of 0.5987 percent.

38. The predicted net profit margin for a company with a fixed asset turnover of 2

times is *closest* to:

- A. 1.1889 percent.
 - B. 1.8043 percent.
 - C. 3.2835 percent
-

SOLUTIONS

1. C is correct. The predicted value of ROE = $4 + (1.8 \times 10) = 22$.
2. A is correct. The slope coefficient of 1.8 is the expected change in the dependent variable (*ROE*) for a one-unit change in the independent variable (*GO*).
3. B is correct. The predicted value is $ROE = 4 + (1.8 \times 8) = 18.4$. The observed value of *ROE* is 21, so the residual is $2.6 = 21.0 - 18.4$.
4. C is correct. Homoskedasticity is the situation in which the variance of the residuals is constant across the observations.
5. A is correct. SHIFT is an indicator or dummy variable because it takes on only the values 0 and 1.
6. A is correct. In a simple regression with a single indicator variable, the intercept is the mean of the dependent variable when the indicator variable takes on a value of zero, which is before the shift in policy in this case.
7. C is correct. Whereas the intercept is the average of the dependent variable when the indicator variable is zero (i.e., before the shift in policy), the slope is the difference in the mean of the dependent variable from before to after the change in policy.
8. A is correct. The null hypothesis of no difference in the annual growth rate is rejected at the 0.05 level: The calculated test statistic of -8.16188 is outside the bounds of ± 2.048 .
9. A is correct. The coefficient of determination is the same as R^2 , which is 0.7436 in the table.
10. C is correct. Because the slope is positive, the correlation between *X* and *Y* is simply the square root of the coefficient of determination: $\sqrt{0.7436} = 0.8623$.
11. C is correct. To make a prediction using the regression model, multiply the slope coefficient by the forecast of the independent variable and add the result to the intercept. Expected value of CFO to sales = $0.077 + (0.826 \times 5) = 4.207$.
12. C is correct. The *p*-value is the smallest level of significance at which the null hypotheses concerning the slope coefficient can be rejected. In this case, the *p*-value is less than 0.05, and thus the regression of the ratio of cash flow from operations to sales on the ratio of net income to sales is significant at the 5 percent level.
13. A is correct. The data are observations over time.
14. C is correct. From the regression equation, Expected return = $0.0138 + (-0.6486 \times -0.01) = 0.0138 + 0.006486 = 0.0203$, or 2.03 percent.
15. C is correct. R^2 is the coefficient of determination. In this case, it shows that 2.11% of the variability in Stellar's returns is explained by changes in CPIENG.
16. B is correct. The standard error of the estimate is the standard deviation of the regression residuals.
17. C is the correct response because it is a false statement. The slope and intercept are both statistically different from zero at the 0.05 level of significance.

18. C is correct. The slope coefficient (shown in Exhibit 2) is negative.
19. B is correct. The sample covariance is calculated as follows:
- $$\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} = -9.2430 \div 49 = -0.1886.$$
20. A is correct. In simple regression, the R^2 is the square of the pairwise correlation. Because the slope coefficient is negative, the correlation is the negative of the square root of 0.0933, or -0.3055.
21. C is correct. Conclusions cannot be drawn regarding causation; they can be drawn only about association; therefore, Interpretations 1 and 2 are incorrect.
22. C is correct. Liu explains the variation of the short interest ratio using the variation of the debt ratio.
23. A is correct. The degrees of freedom are the number of observations minus the number of parameters estimated, which equals 2 in this case (the intercept and the slope coefficient). The number of degrees of freedom is $50 - 2 = 48$.
24. B is correct. The t -statistic is -2.2219, which is outside the bounds created by the critical t -values of ± 2.011 for a two-tailed test with a 5 percent significance level. The value of 2.011 is the critical t -value for the 5 percent level of significance (2.5 percent in one tail) for 48 degrees of freedom. A is incorrect because the mean of the short interest ratio is $192.3 \div 50 = 3.846$. C is incorrect because the debt ratio explains 9.33 percent of the variation of the short interest ratio.
25. A is correct. The predicted value of the short interest ratio = $5.4975 + (-4.1589 \times 0.40) = 5.4975 - 1.6636 = 3.8339$.

26. C is correct. The calculation is $F = \frac{\text{Mean square regression}}{\text{Mean square error}} = \frac{38.4404}{7.7867} = 4.9367$.

27. A is correct. We fail to reject the null hypothesis of a slope equal to one, and we fail to reject the null hypothesis of an intercept equal to zero. The test of the slope equal to 1.0 is

$$t = \frac{0.9830 - 1.000}{0.0155} = -1.09677.$$

The test of the intercept equal to 0.0 is

$$t = \frac{0.0001 - 0.0000}{.00002} = 0.5000.$$

Therefore, we conclude that the forecasts are unbiased.

28. A is correct. The forecast interval for inflation is calculated in three steps:
Step 1. Make the prediction given the US CPI forecast of 2.8:

$$\begin{aligned}\widehat{Y} &= \widehat{b}_0 + \widehat{b}_1 X \\ &= 0.0001 + (0.9830 \times 2.8) \\ &= 2.7525.\end{aligned}$$

Step 2. Compute the variance of the prediction error:

$$\begin{aligned}s_f^2 &= s_e^2 \left\{ 1 + (1/n) + [(X_f - \bar{X})^2] / [(n-1) \times s_x^2] \right\}. \\ s_f^2 &= 0.0009^2 \left\{ 1 + (1/60) + [(2.8 - 1.3350)^2] / [(60-1) \times 0.7539^2] \right\}. \\ s_f^2 &= 0.00000088. \\ s_f &= 0.0009.\end{aligned}$$

Step 3. Compute the prediction interval:

$$\hat{Y} \pm t_c \times s_f$$

$$2.7525 \pm (2.0 \times 0.0009)$$

$$\text{Lower bound: } 2.7525 - (2.0 \times 0.0009) = 2.7506.$$

$$\text{Upper bound: } 2.7525 + (2.0 \times 0.0009) = 2.7544.$$

Given the US CPI forecast of 2.8, the 95 percent prediction interval is 2.7506 to 2.7544.

29. B is correct. The confidence level influences the width of the forecast interval through the critical t -value that is used to calculate the distance from the forecasted value: The larger the confidence level, the wider the interval. Therefore, Observation 1 is not correct.

Observation 2 is correct. The greater the standard error of the estimate, the greater the standard error of the forecast.

30. C is correct. The assumptions of the linear regression model are that (1) the relationship between the dependent variable and the independent variable is linear in the parameters b_0 and b_1 , (2) the residuals are independent of one another, (3) the variance of the error term is the same for all observations, and (4) the error term is normally distributed. Assumption 3 is incorrect because the dependent variable need not be normally distributed.

31. B is correct. The standard error of the estimate for a linear regression model with one independent variable is calculated as the square root of the mean square error:

$$s_e = \sqrt{\frac{0.071475}{34}} = 0.04585.$$

32. C is correct. Crude oil returns explain the Amtex share returns if the slope coefficient is statistically different from zero. The slope coefficient is 0.2354, and the calculated t -statistic is

$$t = \frac{0.2354 - 0.0000}{0.0760} = 3.0974,$$

which is outside the bounds of the critical values of ± 2.728 .

Therefore, Vasileva should reject the null hypothesis that crude oil returns do not explain Amtex share returns, because the slope coefficient is statistically different from zero.

A is incorrect because the calculated t -statistic for testing the slope against 0.15 is $t = \frac{0.2354 - 0.1500}{0.0760} = 1.1237$, which is less than the critical value of +2.441.

B is incorrect because the calculated t -statistic is $t = \frac{0.0095 - 0.0000}{0.0078} = 1.2179$, which is less than the critical value of +2.441.

33. B is correct. The predicted value of the dependent variable, Amtex share return, given the value of the independent variable, crude oil return, -0.01, is calculated as $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_i = 0.0095 + [0.2354 \times (-0.01)] = 0.0071$.

34. C is correct. The predicted share return is $0.0095 + [0.2354 \times (-0.01)] = 0.0071$. The lower limit for the prediction interval is $0.0071 - (2.728 \times 0.0469) = -0.1208$, and the upper limit for the prediction interval is $0.0071 + (2.728 \times 0.0469) = 0.1350$.
- A is incorrect because the bounds of the interval should be based on the standard error of the forecast and the critical t -value, not on the mean of the dependent variable.
- B is incorrect because bounds of the interval are based on the product of the standard error of the forecast *and* the critical t -value, not simply the standard error of the forecast.
35. B is correct. The coefficient of determination is $102.9152 \div 105.1303 = 0.9789$.
36. A is correct. The standard error is the square root of the mean square error, or $\sqrt{0.0692} = 0.2631$.
37. B is correct. The p -value corresponding to the slope is less than 0.01, so we reject the null hypothesis of a zero slope, concluding that the fixed asset turnover explains the natural log of the net profit margin.
38. C is correct. The predicted natural log of the net profit margin is $0.5987 + (2 \times 0.2951) = 1.1889$. The predicted net profit margin is $e^{1.1889} = 3.2835\%$.

LEARNING MODULE

11

Introduction to Big Data Techniques

by Robert Kissell, PhD, and Barbara J. Mack.

Robert Kissell, PhD, is at Molloy College and Kissell Research Group (USA). Barbara J. Mack is at Pingry Hill Enterprises, Inc. (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	describe aspects of “fintech” that are directly relevant for the gathering and analyzing of financial data.
<input type="checkbox"/>	describe Big Data, artificial intelligence, and machine learning
<input type="checkbox"/>	describe applications of Big Data and Data Science to investment management

INTRODUCTION

1

The meeting of finance and technology, commonly known as *fintech*, is changing the landscape of investment management. Advancements include the use of Big Data, artificial intelligence, and machine learning to evaluate investment opportunities, optimize portfolios, and mitigate risks. These developments are affecting not only quantitative asset managers but also fundamental asset managers who make use of these tools and technologies to engage in hybrid forms of investment decision making.

LEARNING MODULE OVERVIEW



- Big Data is characterized by the three Vs—volume, velocity, and variety—and includes both traditional and non-traditional (or alternative) datasets. When Big Data is used for inference or prediction, it is important to consider a fourth V: veracity.
- Among the main sources of alternative data are data generated by individuals, business processes, and sensors.
- Artificial intelligence (AI) computer systems are capable of performing tasks that traditionally required human intelligence at levels comparable (or superior) to those of human beings.

- Machine learning (ML) seeks to extract knowledge from large amounts of data by “learning” from known examples and then generating structure or predictions. Simply put, ML algorithms aim to “find the pattern, apply the pattern.” Main types of ML include supervised learning, unsupervised learning, and deep learning.
- Natural language processing (NLP) is an application of text analytics that uses insight into the structure of human language to analyze and interpret text- and voice-based data.

2

HOW IS FINTECH USED IN QUANTITATIVE INVESTMENT ANALYSIS?

- describe aspects of “fintech” that are directly relevant for the gathering and analyzing of financial data.
- describe Big Data, artificial intelligence, and machine learning

In its broadest sense, the term **fintech** generally refers to technology-driven innovation occurring in the financial services industry. For our purposes, fintech refers to technological innovation in the design and delivery of financial services and products. In common usage, fintech can also refer to companies involved in developing the new technologies and their applications, as well as the business sector that includes such companies. Many of these innovations are challenging the traditional business models of incumbent financial services providers.

Early forms of fintech included data processing and the automation of routine tasks. Systems that provided execution of decisions according to specified rules and instructions followed. Fintech has advanced into decision-making applications based on complex machine-learning logic, in which computer programs are able to “learn” how to complete tasks over time. In some applications, advanced computer systems are performing tasks at levels that far surpass human capabilities. Fintech has changed the financial services industry in many ways, giving rise to new systems for investment advice, financial planning, business lending, and payments.

Whereas fintech covers a broad range of services and applications, areas of development that are more directly relevant to quantitative analysis in the investment industry include the following:

- **Analysis of large datasets.** In addition to growing amounts of traditional data, such as security prices, corporate financial statements, and economic indicators, massive amounts of **alternative data** generated from non-traditional data sources, such as social media and sensor networks, can now be integrated into a portfolio manager’s investment decision-making process and used to help generate alpha and reduce losses.
- **Analytical tools.** For extremely large datasets, techniques involving **artificial intelligence (AI)**—computer systems capable of performing tasks that previously required human intelligence—might be better suited to identify complex, non-linear relationships than traditional quantitative methods and statistical analysis. Advances in AI-based techniques are enabling different data analysis approaches. For example, analysts are turning to AI to sort through the enormous amounts of data from company filings, annual

reports, and earnings calls to determine which data are most important and to help uncover trends and generate insights relating to human sentiment and behavior.

Big Data

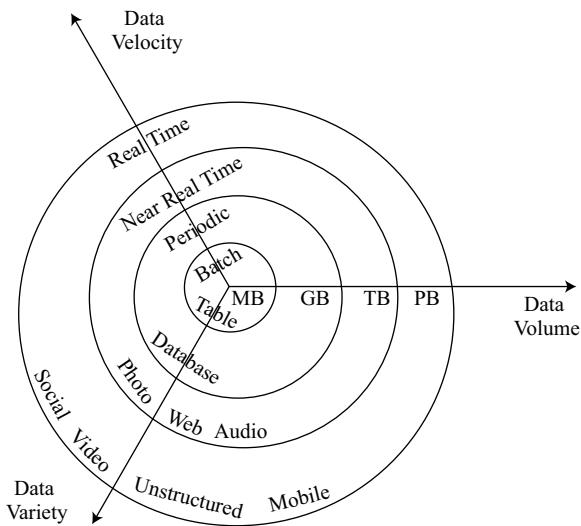
As noted, datasets are growing rapidly in terms of the size and diversity of data types that are available for analysis. The term **Big Data** has been in use since the late 1990s and refers to the vast amount of information being generated by industry, governments, individuals, and electronic devices. Big Data includes data generated from traditional sources—such as stock exchanges, companies, and governments—as well as non-traditional data types, also known as alternative data, arising from the use of electronic devices, social media, sensor networks, and company exhaust (information generated in the normal course of doing business).

Traditional data sources include corporate data in the form of annual reports, regulatory filings, sales and earnings figures, and conference calls with analysts. Traditional data also include data that are generated in the financial markets, including trade prices and volumes. Because the world has become increasingly connected, we can now obtain data from a wide range of devices, including smart phones, cameras, microphones, radio-frequency identification (RFID) readers, wireless sensors, and satellites that are now in use all over the world. As the internet and the presence of such networked devices have grown, the use of non-traditional data sources, or alternative data sources—including social media (posts, tweets, and blogs), email and text communications, web traffic, online news sites, and other electronic information sources—has risen.

The term *Big Data* typically refers to datasets that have the following characteristics:

- **Volume:** The amount of data collected in files, records, and tables is very large, representing many millions, or even billions, of data points.
- **Velocity:** The speed and frequency with which the data are recorded and transmitted has accelerated. Real-time or near-real-time data have become the norm in many areas.
- **Variety:** The data are collected from many different sources and in a variety of formats, including structured data (e.g., SQL tables), semistructured data (e.g., HTML code), and unstructured data (e.g., video messages).

Features relating to big data's volume, velocity, and variety are shown in Exhibit 1.

Exhibit 1: Big Data Characteristics: Volume, Velocity, and Variety

Data Volume Key Bytes of Information

MB	Megabyte	One Million
GB	Gigabyte	One Billion
TB	Terabyte	One Trillion
PB	Petabyte	One Quadrillion

Source: Ivy Wigmore, "Definition: 3Vs (Volume, Variety and Velocity)," WhatIs.com, last updated December 2020, <http://whatis.techtarget.com/definition/3Vs>.

Exhibit 1 shows that data volumes are growing from megabytes and gigabytes to far larger sizes, such as terabytes and petabytes, as more data are being generated, captured, and stored. At the same time, more data, traditional and non-traditional, are available on a real-time or near-real-time basis with far greater variety in data types than ever before.

When Big Data is used for inference or prediction, a “fourth V” comes into play—veracity—which relates to the credibility and reliability of different data sources. Determining the credibility and reliability of data sources is an important part of any empirical investigation. The issue of veracity becomes critically important for Big Data, however, because of the varied sources of these large datasets. Big Data amplifies the age-old challenge of disentangling quality from quantity.

Big Data can be structured, semi-structured, or unstructured data. Structured data items can be organized in tables and are commonly stored in a database where each field represents the same type of information. Unstructured data can be disparate, unorganized data that cannot be represented in tabular form. Unstructured data, such as those generated by social media, email, text messages, voice recordings, pictures, blogs, scanners, and sensors, often require different, specialized applications or custom programs before they can be useful to investment professionals. For example, to analyze data contained in emails or texts, specially developed or customized computer code might be required to first process these files. Semistructured data can have attributes of both structured and unstructured data.

Sources of Big Data

Big Data, therefore, encompasses data generated by the following:

- financial markets (e.g., equity, fixed income, futures, options, and other derivatives),
- businesses (e.g., corporate financials, commercial transactions, and credit card purchases),
- governments (e.g., trade, economic, employment, and payroll data),
- individuals (e.g., credit card purchases, product reviews, internet search logs, and social media posts),
- sensors (e.g., satellite imagery, shipping cargo information, and traffic patterns), and, in particular,
- the Internet of Things, or IoT (e.g., data generated by “smart” buildings, where the building is providing a steady stream of information about climate control, energy consumption, security, and other operational details).

In gathering business intelligence, historically, analysts have tended to draw on traditional data sources, using statistical methods to measure performance, predict future growth, and analyze sector and market trends. In contrast, the analysis of Big Data incorporates the use of alternative data sources.

From retail sales data to social media sentiment to satellite imagery that might reveal information about agriculture, shipping, and oil rigs, alternative datasets can provide additional insights about consumer behavior, firm performance, trends, and other factors important for investment-related activities. Such information is having a significant effect on the way that professional investors, particularly quantitative investors, approach financial analysis and decision-making processes.

The three main sources of alternative data are

- data generated by individuals,
- data generated by business processes, and
- data generated by sensors.

Data generated by individuals are often produced in text, video, photo, and audio formats and also can be generated through such means as website clicks or time spent on a webpage. This type of data tends to be unstructured. The volume of this type of data is growing dramatically as people participate in greater numbers and more frequently in online activities, such as social media and e-commerce, including online reviews of products, services, and entire companies, and as they make personal data available through web searches, email, and other electronic trails.

Business process data include information flows from corporations and other public entities. These data tend to be structured data and include direct sales information, such as credit card data, as well as corporate exhaust. Corporate exhaust includes corporate supply chain information, banking records, and retail point-of-sale scanner data. Business process data can be leading or real-time indicators of business performance, whereas traditional corporate metrics might be reported only on a quarterly or even yearly basis and typically are lagging indicators of performance.

Sensor data are collected from such devices as smart phones, cameras, RFID chips, and satellites that usually are connected to computers through wireless networks. Sensor data can be unstructured, and the volume of data is many orders of magnitude greater than that of individual or business process datastreams. This form of data is growing exponentially because microprocessors and networking technology are increasingly present in a wide array of personal and commercial electronic devices. Extended to office buildings, homes, vehicles, and many other physical forms, this culminates in a network arrangement, known as the **Internet of Things**, which is formed by the vast

array of physical devices, home appliances, smart buildings, vehicles, and other items that are embedded with electronics, sensors, software, and network connections that enable the objects in the system to interact and share information.

Exhibit 2 shows a classification of alternative data sources and includes examples for each.

Exhibit 2: Classification of Alternative Data Sources

Individuals	Business Processes	Sensors
Social media	Transaction data	Satellites
News, reviews	Corporate data	Geolocation
Web searches, personal data		Internet of Things Other sensors

In the search to identify new factors that could affect security prices, enhance asset selection, improve trade execution, and uncover trends, alternative data are being used to support data-driven investment models and decisions. As interest in alternative data has risen, the number of specialized firms that collect, aggregate, and sell alternative datasets has grown.

Although the market for alternative data is expanding, investment professionals should understand the potential legal and ethical issues related to information that is not in the public domain. For example, the **scraping** of web data potentially could capture personal information that is protected by regulations or that might have been published or provided without the explicit knowledge and consent of the individuals involved. Best practices are still in development in many jurisdictions, and because of varying approaches taken by national regulators, the different forms of guidance could conflict.

Big Data Challenges

Big Data poses several challenges when it is used in investment analysis, including the quality, volume, and appropriateness of the data. Key issues revolve around the following questions, among others: Does the dataset have selection bias, missing data, or data outliers? Is the volume of collected data sufficient? Is the dataset well suited for the type of analysis? In most instances, the data must be sourced, cleansed, and organized before analysis can occur. This process can be extremely difficult with alternative data because of the unstructured characteristics of the data involved, which more often are qualitative (e.g., texts, photos, and videos) than quantitative in nature.

Given the size and complexity of alternative datasets, traditional analytical methods cannot always be used to interpret and evaluate these datasets. To address this challenge, AI and machine learning techniques have emerged that support work on such large and complex sources of information.

3

ADVANCED ANALYTICAL TOOLS: ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING



describe Big Data, artificial intelligence, and machine learning

Artificial intelligence (AI) computer systems are capable of performing tasks that traditionally have required human intelligence. AI technology has enabled the development of computer systems that exhibit cognitive and decision-making ability comparable or superior to that of human beings.

An early example of AI was the **expert system**, a type of computer programming that attempted to simulate the knowledge base and analytical abilities of human experts in specific problem-solving contexts. This was often accomplished through the use of “if–then” rules. By the late 1990s, faster networks and more powerful processors enabled AI to be deployed in logistics, data mining, financial analysis, medical diagnosis, and other areas. Since the 1980s, financial institutions have made use of AI—particularly, **neural networks**, programming based on how our brain learns and processes information—to detect abnormal charges or claims in credit card fraud detection systems.

Machine learning (ML) involves computer-based techniques that seek to extract knowledge from large amounts of data without making any assumptions on the data’s underlying probability distribution. The goal of ML algorithms is to automate decision-making processes by generalizing, or “learning,” from known examples to determine an underlying structure in the data. The emphasis is on the ability of the algorithm to generate structure or predictions without any help from a human. Simply put, ML algorithms aim to “find the pattern, apply the pattern.”

As it is currently used in the investing context, ML requires massive amounts of data for “training,” so although some ML techniques have existed for years, insufficient data have historically limited broader application. Previously, these algorithms lacked access to the large amounts of data needed to model relationships successfully. The growth in Big Data has provided ML algorithms, including neural networks, with sufficient data to improve modeling and predictive accuracy, and greater use of ML techniques is now possible.

In ML, the computer algorithm is given “inputs” (a set of variables or datasets) and might be given “outputs” (the target data). The algorithm “learns” from the data provided how best to model inputs to outputs (if provided) or how to identify or describe underlying data structure if no outputs are given. Training occurs as the algorithm identifies relationships in the data and uses that information to refine its learning process.

ML involves splitting the dataset into three distinct subsets: a training dataset, a validation dataset, and a test dataset. The training dataset allows the algorithm to identify relationships between inputs and outputs based on historical patterns in the data. These relationships are then validated, and the model tuned, using the validation dataset. The test dataset is used to test the model’s ability to predict well on new data. Once an algorithm has been trained, validated, and tested, the ML model can be used to predict outcomes based on other datasets.

ML still requires human judgment in understanding the underlying data and selecting the appropriate techniques for data analysis. Before they can be used, the data must be clean and free of biases and spurious data. As noted, ML models also require sufficiently large amounts of data and might not perform well when not enough available data are available to train and validate the model.

Analysts must be cognizant of errors that could arise from **overfitting** the data, because models that overfit the data might discover “false” relationships or “unsubstantiated” patterns that will lead to prediction errors and incorrect output forecasts. Overfitting occurs when the ML model learns the input and target dataset too precisely. In such cases, the model has been “overtrained” on the data and treats noise in the data as true parameters. An ML model that has been overfitted is not able to accurately predict outcomes using a different dataset and might be too complex. When a model has been **underfitted**, the ML model treats true parameters as if they

are noise and is not able to recognize relationships within the training data. In such cases, the model could be too simplistic. Underfitted models typically will fail to fully discover patterns that underlie the data.

In addition, because they are not explicitly programmed, ML techniques can appear to be opaque or “black box” approaches, which arrive at outcomes that might not be entirely understood or explainable.

ML approaches can help identify relationships between variables, detect patterns or trends, and create structure from data, including data classification. ML can be divided broadly into three distinct classes of techniques: supervised learning, unsupervised learning, and deep learning.

In **supervised learning**, computers learn to model relationships based on labeled training data. In supervised learning, inputs and outputs are labeled, or identified, for the algorithm. After learning how best to model relationships for the labeled data, the trained algorithms are used to model or predict outcomes for new datasets. Trying to identify the best signal, or variable; to forecast future returns on a stock; or to predict whether local stock market performance will be up, down, or flat during the next business day are problems that could be approached using supervised learning techniques.

In **unsupervised learning**, computers are not given labeled data but instead are given only data from which the algorithm seeks to describe the data and their structure. For example, grouping companies into peer groups based on their characteristics rather than using standard sector or country groupings is an application of unsupervised learning techniques.

Underlying AI advances have been key developments relating to neural networks. In **deep learning** (or **deep learning nets**), computers use neural networks, often with many hidden layers, to perform multistage, non-linear data processing to identify patterns. Deep learning can use supervised or unsupervised ML approaches. By taking a layered or multistage approach to data analysis, deep learning develops an understanding of simple concepts that informs analysis of more complex concepts.

Neural networks have existed since 1958 and have been used for many applications, such as forecasting and pattern recognition. Improvements in the algorithms underlying neural networks are providing more accurate models that better incorporate and learn from data. As a result, these algorithms are now far better at such activities as image, pattern, and speech recognition. In many cases, the advanced algorithms require less computing power than the earlier neural networks, and their improved solution enables analysts to discover insights and identify relationships that were previously too difficult or too time consuming to uncover.

ADVANCES IN AI OUTSIDE FINANCE

Non-finance-related AI breakthroughs include victories in the general knowledge gameshow *Jeopardy* (by IBM’s Watson in 2011) and in the ancient Chinese board game Go (by Google’s DeepMind in 2016). Not only is AI providing solutions where perfect information exists (all players have equal access to the same information), such as checkers, chess, and Go, but AI is also providing insight in cases in which information might be imperfect and players have hidden information; AI successes at the game of poker (by DeepStack) are an example. AI has also been behind the rise of virtual assistants, such as Siri (from Apple), Google’s Translate app, and Amazon’s product recommendation engine.

The ability to analyze Big Data using ML techniques, alongside more traditional statistical methods, represents a significant development in investment research, supported by the presence of greater data availability and advances in the algorithms. Improvements in computing power and software processing speeds and falling storage costs have further supported this evolution.

ML techniques are being used for Big Data analysis to help predict trends or market events, such as the likelihood of a successful merger or an outcome to a political election. Image recognition algorithms can now analyze data from satellite-imaging systems to provide intelligence on the number of consumers in retail store parking lots, shipping activity and manufacturing facilities, and yields on agricultural crops, to name just a few examples.

Such information could provide insight into individual firms or at national or global levels and might be used as inputs into valuation or economic models.

TACKLING BIG DATA WITH DATA SCIENCE

4

- describe applications of Big Data and Data Science to investment management

Data science can be defined as an interdisciplinary field that harnesses advances in computer science (including ML), statistics, and other disciplines for the purpose of extracting information from Big Data (or data in general). Companies rely on the expertise of data scientists/analysts to extract information and insights from Big Data for a wide variety of business and investment purposes.

An important consideration for the data scientist is the structure of the data. As noted in the discussion on Big Data, because of their unstructured nature, alternative data often require specialized treatment before they can be used for analysis.

Data Processing Methods

To help determine the best data management technique needed for Big Data analysis, data scientists use various data processing methods, including capture, curation, storage, search, and transfer.

- Capture—Data capture refers to how the data are collected and transformed into a format that can be used by the analytical process. Low-latency systems—systems that operate on networks that communicate high volumes of data with minimal delay (latency)—are essential for automated trading applications that make decisions based on real-time prices and market events. In contrast, high-latency systems do not require access to real-time data and calculations.
- Curation—Data curation refers to the process of ensuring data quality and accuracy through a data cleaning exercise. This process consists of reviewing all data to detect and uncover data errors—bad or inaccurate data—and making adjustments for missing data when appropriate.
- Storage—Data storage refers to how the data will be recorded, archived, and accessed and the underlying database design. An important consideration for data storage is whether the data are structured or unstructured and whether analytical needs require low-latency solutions.
- Search—Search refers to how to query data. Big Data has created the need for advanced applications capable of examining and reviewing large quantities of data to locate requested data content.
- Transfer—Transfer refers to how the data will move from the underlying data source or storage location to the underlying analytical tool. This could be through a direct data feed, such as a stock exchange's price feed.

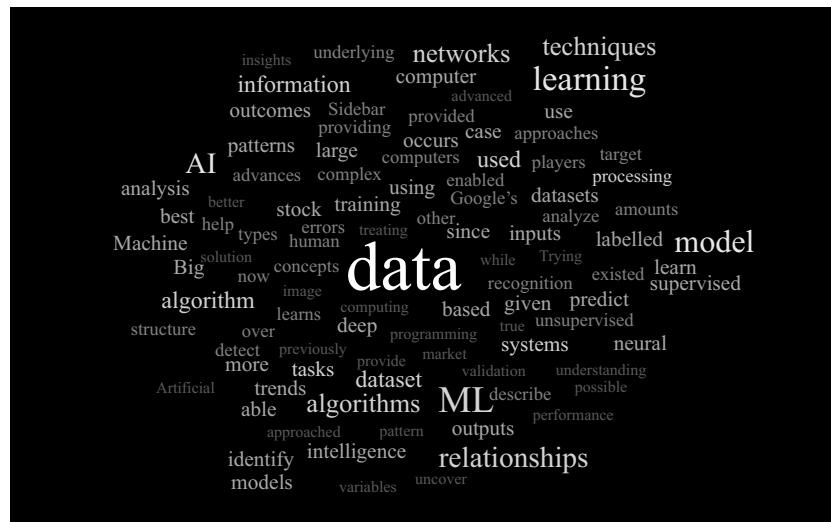
Data Visualization

Data visualization is an important tool for understanding Big Data. Visualization refers to how the data will be formatted, displayed, and summarized in graphical form. Traditional structured data can be visualized using tables, charts, and trends, whereas non-traditional unstructured data require new techniques of data visualization. These visualization tools include, for example, interactive three-dimensional (3D) graphics, in which users can focus in on specified data ranges and rotate the data across 3D axes to help identify trends and uncover relationships. Multidimensional data analysis consisting of more than three variables requires additional data visualization techniques—for example, adding color, shapes, and sizes to the 3D charts. Furthermore, a wide variety of solutions exists to reflect the structure of the data through the geometry of the visualization, with interactive graphics allowing for especially rich possibilities. Examples include heat maps, tree diagrams, and network graphs.

Another valuable Big Data visualization technique that is applicable to textual data is a “tag cloud,” in which words are sized and displayed on the basis of the frequency of the word in the data file. For example, words that appear more often are shown with a larger font, and words that appear less often are shown with a smaller font. A “mind map” is another data visualization technique; it is a variation of the tag cloud, but rather than displaying the frequency of words, a mind map shows how different concepts are related to each other.

Exhibit 3 shows an example of a “tag cloud” based on a section of this reading. The more frequently a word is found within the text, the larger it becomes in the tag cloud. As shown in the tag cloud, the words appearing most frequently in the section include “data,” “ML,” “learning,” “AI,” “techniques,” “model,” and “relationships.”

Exhibit 3: Data Visualization Tag Cloud



Source: "About Word Clouds," WordItOut, <https://worditout.com/word-cloud/create>.

Fintech is being used in numerous areas of investment management. Applications for investment management include text analytics and natural language processing, risk analysis, and algorithmic trading.

Text Analytics and Natural Language Processing

Text analytics involves the use of computer programs to analyze and derive meaning typically from large, unstructured text- or voice-based datasets, such as company filings, written reports, quarterly earnings calls, social media, email, internet postings, and surveys. Text analytics includes using computer programs to perform automated information retrieval from different, unrelated sources to aid the decision-making process. More analytical usage includes lexical analysis, or the analysis of word frequency in a document and pattern recognition based on key words and phrases. Text analytics could be used in predictive analysis to help identify indicators of future performance, such as consumer sentiment.

Natural language processing (NLP) is a field of research at the intersection of computer science, AI, and linguistics that focuses on developing computer programs to analyze and interpret human language. Within the larger field of text analytics, NLP is an important application. Automated tasks using NLP include translation, speech recognition, text mining, sentiment analysis, and topic analysis. NLP also might be used in compliance functions to review employee voice and electronic communications for adherence to company or regulatory policy, for detecting fraud or inappropriate conduct, or for ensuring private or customer information is kept confidential.

Consider that all the public corporations worldwide generate millions of pages of annual reports and tens of thousands of hours of earnings calls each year. This is more information than any individual analyst or team of researchers can assess. NLP, especially when aided by ML algorithms, can analyze annual reports, call transcripts, news articles, social media posts, and other text- and audio-based data to identify trends in shorter time spans and with greater scale and accuracy than is humanly possible.

For example, NLP can be used to monitor analyst commentary to aid investment decision making. Financial analysts might generate earnings-per-share (EPS) forecasts reflecting their views on a company's near-term prospects. Focusing on forecasted EPS numbers could mean investors miss subtleties contained in an analyst's written research report. Because analysts tend not to change their buy, hold, and sell recommendations for a company frequently, they might instead offer nuanced commentary without making a change in their investment recommendation. After analyzing analyst commentary, NLP can assign sentiment ratings ranging from very negative to very positive for each. NLP, therefore, can be used to detect, monitor, and tag shifts in sentiment, potentially ahead of an analyst's recommendation change. Machine capabilities enable this analysis to scale across thousands of companies worldwide, performing work previously done by humans.

Similarly, communications and transcripts from policy makers, such as the European Central Bank or the US Federal Reserve, offer an opportunity for NLP-based analysis because officials at these institutions might send subtle messages through their choice of topics, words, and inferred tone. NLP can analyze nuances within text to provide insights around trending or waning topics of interest, such as interest rate policy, aggregate output, or inflation expectations.

Models using NLP analysis might incorporate non-traditional information to evaluate what people are saying—through their preferences, opinions, likes, or dislikes—in an attempt to identify trends and short-term indicators—for example about a company, a stock, or an economic event—to forecast coming trends that may affect investment performance in the future. For example, past research has evaluated the predictive power of Twitter sentiment regarding initial public offering (IPO) performance as well as the effect of positive and negative news sentiment on stock returns.

PROGRAMMING LANGUAGES AND DATABASES

Some of the more common programming languages used in data science include the following:

- **Python:** Python is an open-source, free programming language that does not require an in-depth understanding of computer programming. Python allows individuals with little or no programming experience to develop computer applications for advanced analytical use and is the basis for many fintech applications.
- **R:** R is an open-source, free programming language traditionally used for statistical analysis. R has mathematical packages for statistical analysis, ML, optimization, econometrics, and financial analysis.
- **Java:** Java is a programming language that can run on different computers, servers, and operating systems. Java is the underlying program language used in many internet applications.
- **C and C++:** Both C and C++ are specialized programming languages that provide the ability to optimize source code to achieve superior calculation speed and processing performance. C and C++ is used in applications for algorithmic and high-frequency trading.
- **Excel VBA:** Excel VBA helps bridge the gap between programming and manual data processing by allowing users to run macros to automate tasks, such as updating data tables and formulas, running data queries and collecting data from different web locations, and performing calculations. Excel VBA allows users to develop customized reports and analyses that rely on data that are updated from different applications and databases.

Some of the more common types of databases in use include the following:

- **SQL:** SQL is a database query language for structured data where the data can be stored in tables with rows and columns. SQL-based databases need to be run on a server that is accessed by users using SQL queries.
- **SQLite:** SQLite is a database for structured data. SQLite databases are embedded into the program and do not need to be run on a server. It is the most common database for mobile apps that require access to data.
- **NoSQL:** NoSQL is a database used for unstructured data where the data cannot be summarized in traditional tables with rows and columns.

PRACTICE PROBLEMS

1. A characteristic of Big Data is that:
 - A. it involves formats with diverse structures.
 - B. one of its traditional sources is business processes.
 - C. real-time communication of it is uncommon due to vast content.
2. Which of the following statements is true in the use of ML:
 - A. some techniques are termed “black box” due to data biases.
 - B. human judgment is not needed because algorithms continuously learn from data.
 - C. training data can be learned too precisely, resulting in inaccurate predictions when used with different datasets.
3. Text analytics is appropriate for application to:
 - A. large, structured datasets.
 - B. public but not private information.
 - C. identifying possible short-term indicators of coming trends.

SOLUTIONS

1. A is correct. Big Data is collected from many different sources and is in a variety of formats, including structured data (e.g., SQL tables), semistructured data (e.g., HTML code), and unstructured data (e.g., video messages).
2. C is correct. Overfitting occurs when the ML model learns the input and target dataset too precisely. In this case, the model has been “overtrained” on the data and is treating noise in the data as true parameters. An ML model that has been overfitted is not able to accurately predict outcomes using a different dataset and might be too complex.
3. C is correct. Through the text analytics application of NLP, models using NLP analysis might incorporate non-traditional information to evaluate what people are saying—through their preferences, opinions, likes, or dislikes—in an attempt to identify trends and short-term indicators—for example, about a company, a stock, or an economic event—to forecast coming trends that may affect investment performance in the future.

LEARNING MODULE

12

Appendices A–E

APPENDICES A–E

1

Appendix A	Cumulative Probabilities for a Standard Normal Distribution
Appendix B	Table of the Student's <i>t</i> -Distribution (One-Tailed Probabilities)
Appendix C	Values of X^2 (Degrees of Freedom, Level of Significance)
Appendix D	Table of the <i>F</i> -Distribution
Appendix E	Critical Values for the Durbin-Watson Statistic ($\alpha = .05$)

Appendix A**Cumulative Probabilities for a Standard Normal Distribution** $P(Z \leq x) = N(x)$ for $x \geq 0$ or $P(Z \leq z) = N(z)$ for $z \geq 0$

<i>x or z</i>	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.00	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.10	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.20	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.30	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.40	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.50	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.60	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.70	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.80	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.90	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.00	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

For example, to find the *z*-value leaving 2.5 percent of the area/probability in the upper tail, find the element 0.9750 in the body of the table. Read 1.90 at the left end of the element's row and 0.06 at the top of the element's column, to give $1.90 + 0.06 = 1.96$. *Table generated with Excel.*

Quantitative Methods for Investment Analysis, Second Edition, by Richard A. DeFusco, CFA, Dennis W. McLeavey, CFA, Jerald E. Pinto, CFA, and David E. Runkle, CFA. Copyright © 2004 by CFA Institute.

Appendix A (continued)**Cumulative Probabilities for a Standard Normal Distribution** $P(Z \leq x) = N(x)$ for $x \leq 0$ or $P(Z \leq z) = N(z)$ for $z \leq 0$

<i>x or z</i>	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
-0.10	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.20	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.30	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.40	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.50	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.60	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.70	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.80	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.90	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.00	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-1.10	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.20	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.30	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.40	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.50	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.60	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.70	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.80	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.90	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-2.00	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-2.10	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.20	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.30	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.40	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.50	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.60	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.70	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.80	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.90	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-3.00	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-3.10	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.20	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.30	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.40	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.50	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.60	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.70	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.80	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
-4.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

For example, to find the *z*-value leaving 2.5 percent of the area/probability in the lower tail, find the element 0.0250 in the body of the table. Read -1.90 at the left end of the element's row and 0.06 at the top of the element's column, to give $-1.90 - 0.06 = -1.96$. *Table generated with Excel.*

Appendix B**Table of the Student's t-Distribution (One-Tailed Probabilities)**

df	p = 0.10	p = 0.05	p = 0.025	p = 0.01	p = 0.005	df	p = 0.10	p = 0.05	p = 0.025	p = 0.01	p = 0.005
1	3.078	6.314	12.706	31.821	63.657	31	1.309	1.696	2.040	2.453	2.744
2	1.886	2.920	4.303	6.965	9.925	32	1.309	1.694	2.037	2.449	2.738
3	1.638	2.353	3.182	4.541	5.841	33	1.308	1.692	2.035	2.445	2.733
4	1.533	2.132	2.776	3.747	4.604	34	1.307	1.691	2.032	2.441	2.728
5	1.476	2.015	2.571	3.365	4.032	35	1.306	1.690	2.030	2.438	2.724
6	1.440	1.943	2.447	3.143	3.707	36	1.306	1.688	2.028	2.434	2.719
7	1.415	1.895	2.365	2.998	3.499	37	1.305	1.687	2.026	2.431	2.715
8	1.397	1.860	2.306	2.896	3.355	38	1.304	1.686	2.024	2.429	2.712
9	1.383	1.833	2.262	2.821	3.250	39	1.304	1.685	2.023	2.426	2.708
10	1.372	1.812	2.228	2.764	3.169	40	1.303	1.684	2.021	2.423	2.704
11	1.363	1.796	2.201	2.718	3.106	41	1.303	1.683	2.020	2.421	2.701
12	1.356	1.782	2.179	2.681	3.055	42	1.302	1.682	2.018	2.418	2.698
13	1.350	1.771	2.160	2.650	3.012	43	1.302	1.681	2.017	2.416	2.695
14	1.345	1.761	2.145	2.624	2.977	44	1.301	1.680	2.015	2.414	2.692
15	1.341	1.753	2.131	2.602	2.947	45	1.301	1.679	2.014	2.412	2.690
16	1.337	1.746	2.120	2.583	2.921	46	1.300	1.679	2.013	2.410	2.687
17	1.333	1.740	2.110	2.567	2.898	47	1.300	1.678	2.012	2.408	2.685
18	1.330	1.734	2.101	2.552	2.878	48	1.299	1.677	2.011	2.407	2.682
19	1.328	1.729	2.093	2.539	2.861	49	1.299	1.677	2.010	2.405	2.680
20	1.325	1.725	2.086	2.528	2.845	50	1.299	1.676	2.009	2.403	2.678
21	1.323	1.721	2.080	2.518	2.831	60	1.296	1.671	2.000	2.390	2.660
22	1.321	1.717	2.074	2.508	2.819	70	1.294	1.667	1.994	2.381	2.648
23	1.319	1.714	2.069	2.500	2.807	80	1.292	1.664	1.990	2.374	2.639
24	1.318	1.711	2.064	2.492	2.797	90	1.291	1.662	1.987	2.368	2.632
25	1.316	1.708	2.060	2.485	2.787	100	1.290	1.660	1.984	2.364	2.626
26	1.315	1.706	2.056	2.479	2.779	110	1.289	1.659	1.982	2.361	2.621
27	1.314	1.703	2.052	2.473	2.771	120	1.289	1.658	1.980	2.358	2.617
28	1.313	1.701	2.048	2.467	2.763	200	1.286	1.653	1.972	2.345	2.601
29	1.311	1.699	2.045	2.462	2.756	∞	1.282	1.645	1.960	2.326	2.576
30	1.310	1.697	2.042	2.457	2.750						

To find a critical *t*-value, enter the table with *df* and a specified value for α , the significance level. For example, with 5 df, $\alpha = 0.05$ and a one-tailed test, the desired probability in the tail would be $p = 0.05$ and the critical *t*-value would be $t(5, 0.05) = 2.015$. With $\alpha = 0.05$ and a two-tailed test, the desired probability in each tail would be $p = 0.025 = \alpha/2$, giving $t(0.025) = 2.571$. *Table generated using Excel.*

Quantitative Methods for Investment Analysis, Second Edition, by Richard A. DeFusco, CFA, Dennis W. McLeavy, CFA, Jerald E. Pinto, CFA, and David E. Runkle, CFA. Copyright © 2004 by CFA Institute.

Appendix C
Values of χ^2 (Degrees of Freedom, Level of Significance)

Degrees of Freedom	Probability in Right Tail								
	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000157	0.000982	0.003932	0.0158	2.706	3.841	5.024	6.635	7.879
2	0.020100	0.050636	0.102586	0.2107	4.605	5.991	7.378	9.210	10.597
3	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.345	12.838
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.994
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.335
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
100	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.170

To have a probability of 0.05 in the right tail when $df = 5$, the tabled value is $\chi^2(5, 0.05) = 11.070$.

Quantitative Methods for Investment Analysis, Second Edition, by Richard A. DeFusco, CFA, Dennis W. McLeavey, CFA, Jerald E. Pinto, CFA, and David E. Runkle, CFA. Copyright © 2004 by CFA Institute.

Appendix D**Table of the F-Distribution****Panel A. Critical values for right-hand tail area equal to 0.05**

	df₁: 1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	25	30	40	60	120	∞
df₂: 1	161	200	216	225	230	234	237	239	241	242	243	244	246	248	248	249	249	249	249	250	251	252	253	254
2	18.5	19.0	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.70	8.66	8.65	8.64	8.64	8.63	8.62	8.59	8.57	8.55	8.53	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.86	5.80	5.79	5.79	5.78	5.77	5.75	5.72	5.69	5.66	5.63	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.62	4.56	4.55	4.54	4.53	4.52	4.50	4.46	4.43	4.40	4.37	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.94	3.87	3.86	3.86	3.85	3.84	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.51	3.44	3.43	3.43	3.42	3.41	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.22	3.15	3.14	3.13	3.12	3.11	3.08	3.04	3.01	2.97	2.93	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.01	2.94	2.93	2.92	2.91	2.90	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.85	2.77	2.76	2.75	2.75	2.74	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.72	2.65	2.64	2.63	2.62	2.61	2.60	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.62	2.54	2.53	2.52	2.51	2.51	2.50	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.53	2.46	2.45	2.44	2.43	2.42	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.46	2.39	2.38	2.37	2.36	2.35	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.40	2.33	2.32	2.31	2.30	2.29	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.35	2.28	2.26	2.25	2.24	2.24	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.31	2.23	2.22	2.21	2.20	2.19	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.27	2.19	2.18	2.17	2.16	2.15	2.14	2.11	2.07	2.03	1.98	1.93
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.23	2.16	2.14	2.13	2.12	2.11	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.20	2.12	2.11	2.10	2.09	2.08	2.07	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.18	2.10	2.08	2.07	2.06	2.05	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.15	2.07	2.06	2.05	2.04	2.03	2.02	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.13	2.05	2.04	2.02	2.01	2.01	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.11	2.03	2.01	2.00	1.99	1.98	1.97	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.09	2.01	2.00	1.98	1.97	1.96	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.01	1.93	1.92	1.91	1.90	1.89	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.92	1.84	1.83	1.81	1.80	1.79	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.84	1.75	1.73	1.72	1.71	1.70	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83	1.75	1.66	1.64	1.63	1.62	1.61	1.60	1.55	1.50	1.43	1.35	1.25
Infinity	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75	1.67	1.57	1.56	1.54	1.53	1.52	1.51	1.46	1.39	1.32	1.22	1.00

Appendix D (continued)
Table of the F-Distribution

		Panel B. Critical values for right-hand tail area equal to 0.025												Numerator: df_1 and Denominator: df_2										
		df ₁ : 1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	25	30	40	60	∞
	df ₂ : 1	648	799	864	900	922	937	948	957	963	969	973	977	985	993	994	995	996	997	998	1001	1006	1014	1018
2	38.51	39.00	391.17	392.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.47	39.45	39.45	39.45	39.46	39.46	39.47	39.48	39.49	39.49	39.50	
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.37	14.34	14.25	14.17	14.16	14.14	14.13	14.12	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.79	8.75	8.66	8.56	8.53	8.52	8.51	8.50	8.46	8.41	8.36	8.31	8.26	
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.57	6.52	6.43	6.33	6.31	6.30	6.29	6.28	6.27	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.41	5.37	5.27	5.17	5.14	5.13	5.12	5.11	5.07	5.01	4.96	4.90	4.85	
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.71	4.67	4.57	4.47	4.45	4.44	4.43	4.41	4.40	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.24	4.20	4.10	4.00	3.98	3.97	3.96	3.95	3.94	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.91	3.87	3.77	3.67	3.65	3.64	3.63	3.61	3.60	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.66	3.62	3.52	3.42	3.40	3.39	3.38	3.37	3.35	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.47	3.43	3.33	3.23	3.21	3.20	3.18	3.17	3.16	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.32	3.28	3.18	3.07	3.06	3.04	3.03	3.02	3.01	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.20	3.15	3.05	2.95	2.93	2.92	2.91	2.89	2.88	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.09	3.05	2.95	2.84	2.83	2.81	2.80	2.79	2.78	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	3.01	2.96	2.86	2.76	2.74	2.73	2.71	2.70	2.69	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.93	2.89	2.79	2.68	2.67	2.65	2.64	2.63	2.61	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.87	2.82	2.72	2.62	2.60	2.59	2.57	2.56	2.55	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.81	2.77	2.67	2.56	2.54	2.53	2.52	2.50	2.49	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.76	2.72	2.62	2.51	2.49	2.48	2.46	2.45	2.44	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.72	2.68	2.57	2.46	2.45	2.43	2.42	2.41	2.40	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.68	2.64	2.53	2.42	2.41	2.39	2.38	2.37	2.36	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.65	2.60	2.50	2.39	2.37	2.36	2.34	2.33	2.32	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.62	2.57	2.47	2.36	2.34	2.33	2.31	2.30	2.29	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.59	2.54	2.44	2.33	2.31	2.30	2.28	2.27	2.26	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.56	2.51	2.41	2.30	2.28	2.27	2.26	2.24	2.23	2.18	2.12	2.05	1.98	1.91
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.46	2.41	2.31	2.20	2.18	2.16	2.15	2.14	2.12	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.33	2.29	2.18	2.07	2.05	2.03	2.02	2.01	1.99	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.22	2.17	2.06	1.94	1.93	1.91	1.90	1.88	1.87	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.10	2.05	1.94	1.82	1.81	1.79	1.77	1.76	1.75	1.69	1.61	1.53	1.43	1.31
Infinity	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.99	1.94	1.83	1.71	1.69	1.67	1.64	1.63	1.57	1.48	1.39	1.27	1.00	

Appendix D (continued)
Table of the F-Distribution

		Numerator: df_1 and Denominator: df_2																						
		1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	30	40	60	120	∞
$df_2: 1$	4052	5000	5403	5625	5764	5859	5928	6023	6056	6083	6106	6157	6209	6216	6223	6229	6235	6240	6261	6287	6313	6339	6366	
2	98.5	99.0	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.1	27.1	26.9	26.7	26.6	26.6	26.6	26.6	26.6	26.6	26.4	26.3	26.2	26.1	
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	14.0	14.0	13.9	13.9	13.8	13.7	13.7	13.6	13.5	13.5	
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	10.0	9.89	9.72	9.55	9.53	9.51	9.49	9.45	9.38	9.29	9.20	9.11	9.02	
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.56	7.40	7.37	7.35	7.33	7.31	7.30	7.23	7.14	7.06	6.97	
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.31	6.16	6.13	6.11	6.09	6.07	6.06	5.99	5.91	5.82	5.74	
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.52	5.36	5.34	5.32	5.30	5.28	5.26	5.20	5.12	5.03	4.95	
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	4.96	4.81	4.79	4.77	4.75	4.73	4.71	4.65	4.57	4.48	4.40	
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.56	4.41	4.38	4.36	4.34	4.33	4.31	4.25	4.17	4.08	4.00	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.25	4.10	4.08	4.06	4.04	4.02	4.01	3.94	3.86	3.78	3.69	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.01	3.86	3.84	3.82	3.80	3.78	3.76	3.70	3.62	3.54	3.45	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.82	3.66	3.64	3.62	3.60	3.59	3.57	3.51	3.43	3.34	3.25	
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.66	3.51	3.48	3.46	3.44	3.43	3.41	3.35	3.27	3.18	3.09	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.52	3.37	3.35	3.33	3.31	3.29	3.28	3.21	3.13	3.05	2.96	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55	3.41	3.26	3.24	3.22	3.20	3.18	3.16	3.10	3.02	2.93	2.75	
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46	3.31	3.16	3.14	3.12	3.10	3.08	3.07	3.00	2.92	2.83	2.75	
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37	3.23	3.08	3.05	3.03	3.02	3.00	2.98	2.92	2.84	2.75	2.66	
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.15	3.00	2.98	2.96	2.94	2.92	2.91	2.84	2.76	2.67	2.49	
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23	3.09	2.94	2.92	2.90	2.88	2.86	2.84	2.78	2.69	2.61	2.52	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.24	3.17	3.03	2.88	2.86	2.84	2.82	2.80	2.79	2.72	2.64	2.55	2.46	
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	2.98	2.83	2.81	2.78	2.77	2.75	2.73	2.67	2.58	2.50	2.40	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.93	2.78	2.76	2.74	2.72	2.70	2.69	2.62	2.54	2.45	2.26	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03	2.89	2.74	2.72	2.70	2.68	2.66	2.64	2.58	2.49	2.40	2.31	
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	3.06	2.99	2.85	2.70	2.68	2.66	2.64	2.62	2.60	2.53	2.45	2.36	2.27	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84	2.70	2.55	2.51	2.49	2.47	2.45	2.39	2.30	2.21	2.11	2.01	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66	2.52	2.37	2.35	2.31	2.29	2.27	2.20	2.11	2.02	1.92	1.80	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.35	2.20	2.17	2.15	2.13	2.12	2.10	2.03	1.94	1.84	1.73	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.19	2.03	2.01	1.99	1.97	1.95	1.93	1.86	1.76	1.66	1.53	
Infinity	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18	2.04	1.88	1.85	1.81	1.79	1.77	1.70	1.59	1.47	1.32	1.00	

Appendix D (continued)
Table of the F-Distribution

Panel D. Critical values for right-hand tail area equal to 0.005											Numerator: df ₁ and Denominator: df ₂													
df ₁ : 1	2	3	4	5	6	7	8	9	10	11	12	15	20	21	22	23	24	25	30	40	60	120	∞	
df₂: 1	16211	20000	21615	22500	23457	23715	23925	24091	24222	24334	24426	24630	24836	24863	24892	24915	24940	24959	25044	25146	25253	25359	25464	
2	198.5	199.0	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	200		
3	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13	43.88	43.68	43.52	43.39	43.08	42.78	42.73	42.69	42.66	42.62	42.59	42.47	42.31	42.15	41.99	
4	31.33	26.28	24.26	23.15	22.46	21.98	21.62	21.35	21.14	20.97	20.82	20.70	20.44	20.17	20.13	20.09	20.06	20.03	20.00	19.89	19.75	19.61	19.47	
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.49	13.38	13.15	12.90	12.87	12.84	12.81	12.78	12.76	12.66	12.53	12.40	12.27	
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.13	10.03	9.81	9.59	9.56	9.53	9.50	9.47	9.45	9.36	9.24	9.12	9.00	8.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.27	8.18	7.97	7.75	7.72	7.69	7.67	7.64	7.62	7.53	7.42	7.31	7.19	7.08
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.10	7.01	6.81	6.61	6.58	6.55	6.53	6.50	6.48	6.40	6.29	6.18	6.06	5.95
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.31	6.23	6.03	5.83	5.80	5.78	5.75	5.73	5.71	5.62	5.52	5.41	5.30	5.19
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.75	5.66	5.47	5.27	5.25	5.22	5.20	5.17	5.15	5.07	4.97	4.86	4.75	4.64
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.32	5.24	5.05	4.86	4.83	4.80	4.78	4.76	4.74	4.65	4.55	4.45	4.34	4.23
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.99	4.91	4.72	4.53	4.50	4.48	4.45	4.43	4.41	4.33	4.23	4.12	4.01	3.90
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.72	4.64	4.46	4.27	4.24	4.22	4.19	4.17	4.15	4.07	3.97	3.87	3.76	3.65
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.51	4.43	4.25	4.06	4.03	4.01	3.98	3.96	3.94	3.86	3.76	3.66	3.55	3.44
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.33	4.25	4.07	3.88	3.86	3.83	3.81	3.79	3.77	3.69	3.59	3.48	3.37	3.26
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.18	4.10	3.92	3.73	3.71	3.68	3.66	3.64	3.62	3.54	3.44	3.33	3.22	3.11
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	4.05	3.97	3.79	3.61	3.58	3.56	3.53	3.51	3.49	3.41	3.31	3.21	3.10	2.98
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.94	3.86	3.68	3.50	3.47	3.45	3.42	3.40	3.38	3.30	3.20	3.10	2.99	2.87
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.84	3.76	3.59	3.40	3.37	3.35	3.33	3.31	3.29	3.21	3.11	3.00	2.89	2.78
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.76	3.68	3.50	3.32	3.29	3.27	3.24	3.22	3.20	3.12	3.02	2.92	2.81	2.69
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.68	3.60	3.43	3.24	3.22	3.19	3.17	3.15	3.13	3.05	2.95	2.84	2.73	2.61
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.61	3.54	3.36	3.18	3.15	3.12	3.10	3.08	3.06	2.98	2.88	2.77	2.66	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.55	3.47	3.30	3.12	3.09	3.06	3.04	3.02	3.00	2.92	2.82	2.71	2.60	2.48
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.50	3.42	3.25	3.06	3.04	3.01	2.99	2.97	2.95	2.87	2.77	2.66	2.55	2.43
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.45	3.37	3.20	3.01	2.99	2.96	2.94	2.92	2.90	2.82	2.72	2.61	2.50	2.38
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.25	3.18	3.01	2.82	2.80	2.77	2.75	2.73	2.71	2.63	2.52	2.42	2.30	2.18
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	3.03	2.95	2.78	2.60	2.57	2.55	2.52	2.50	2.48	2.40	2.30	2.18	2.06	1.93
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.82	2.74	2.57	2.39	2.36	2.33	2.31	2.29	2.27	2.19	2.08	1.96	1.83	1.69
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.62	2.54	2.37	2.19	2.16	2.13	2.11	2.09	2.07	1.98	1.87	1.75	1.61	1.43
Infinity	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.43	2.36	2.19	2.00	1.97	1.95	1.92	1.90	1.88	1.79	1.67	1.53	1.36	1.00

With 1 degree of freedom (df) in the numerator and 3 df in the denominator, the critical *F*-value is 10.1 for a right-hand tail area equal to 0.05.*Quantitative Methods for Investment Analysis*, Second Edition, by Richard A. DeFusco, CFA, Dennis W. McLeavey, CFA, Jerald E. Pinto, CFA, and David E. Runkle, CFA. Copyright © 2004 by CFA Institute.

Appendix E**Critical Values for the Durbin-Watson Statistic ($\alpha = .05$)**

<i>n</i>	<i>K</i> = 1		<i>K</i> = 2		<i>K</i> = 3		<i>K</i> = 4		<i>K</i> = 5	
	<i>d_l</i>	<i>d_u</i>								
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Note: *K* = the number of slope parameters in the model.

Source: From J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression, II." *Biometrika* 38 (1951): 159–178.

Economics

LEARNING MODULE

1

Firms and Market Structures

by Gary L. Arbogast, PhD, CFA, Richard V. Eastin, PhD, Fritz Richard, PhD, and Gambera Michele, PhD, CFA.

Gary L. Arbogast, PhD, CFA (USA). Richard V. Eastin, PhD, is at the University of Southern California (USA). Richard Fritz, PhD, is at the School of Economics at Georgia Institute of Technology (USA). Michele Gambera, PhD, CFA, is at UBS Asset Management and the University of Illinois at Urbana-Champaign (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	determine and interpret breakeven and shutdown points of production, as well as how economies and diseconomies of scale affect costs under perfect and imperfect competition
<input type="checkbox"/>	describe characteristics of perfect competition, monopolistic competition, oligopoly, and pure monopoly
<input type="checkbox"/>	explain supply and demand relationships under monopolistic competition, including the optimal price and output for firms as well as pricing strategy
<input type="checkbox"/>	explain supply and demand relationships under oligopoly, including the optimal price and output for firms as well as pricing strategy
<input type="checkbox"/>	identify the type of market structure within which a firm operates and describe the use and limitations of concentration measures

INTRODUCTION

1

This learning module addresses several important concepts that extend the basic market model of demand and supply to the assessment of a firm's breakeven and shutdown points of production. Demand concepts covered include own-price elasticity of demand, cross-price elasticity of demand, and income elasticity of demand. Supply concepts covered include total, average, and marginal product of labor; total, variable, and marginal cost of labor; and total and marginal revenue. These concepts are used to calculate the breakeven and shutdown points of production.

This learning module surveys how economists classify market structures. We analyze distinctions between the different structures that are important for understanding demand and supply relations, optimal price and output, and the factors affecting long-run profitability. We also provide guidelines for identifying market structure in practice.

LEARNING MODULE OVERVIEW



- Firms under conditions of perfect competition have no pricing power and, therefore, face a perfectly horizontal demand curve at the market price. For firms under conditions of perfect competition, price is identical to marginal revenue (MR).
- Firms under conditions of imperfect competition face a negatively sloped demand curve and have pricing power. For firms under conditions of imperfect competition, MR is less than price.
- Economic profit equals total revenue (TR) minus total economic cost, whereas accounting profit equals TR minus total accounting cost.
- Economic cost considers the total opportunity cost of all factors of production.
- Opportunity cost is the next best alternative use of a resource forgone in making a decision.
- Maximum economic profit requires that (1) MR equals marginal cost (MC) and (2) MC not be falling with output.
- The breakeven point occurs when TR equals total cost (TC), otherwise stated as the output quantity at which average total cost (ATC) equals price.
- Shutdown occurs when a firm is better off not operating than continuing to operate.
- If all fixed costs are sunk costs, then shutdown occurs when the market price falls below the minimum average variable cost. After shutdown, the firm incurs only fixed costs and loses less money than it would operating at a price that does not cover variable costs.
- In the short run, it may be rational for a firm to continue to operate while earning negative economic profit if some unavoidable fixed costs are covered.
- Economies of scale is defined as decreasing long-run cost per unit as output increases. Diseconomies of scale is defined as increasing long-run cost per unit as output increases.
- Long-run ATC is the cost of production per unit of output under conditions in which all inputs are variable.
- Specialization efficiencies and bargaining power in input price can lead to economies of scale.
- Bureaucratic and communication breakdowns and bottlenecks that raise input prices can lead to diseconomies of scale.
- The minimum point on the long-run ATC curve defines the minimum efficient scale for the firm.
- Economic market structures can be grouped into four categories: perfect competition, monopolistic competition, oligopoly, and monopoly.

- The categories of economic market structures differ because of the following characteristics: The number of producers is many in perfect and monopolistic competition, few in oligopoly, and one in monopoly. The degree of product differentiation, the pricing power of the producer, the barriers to entry of new producers, and the level of non-price competition (e.g., advertising) are all low in perfect competition, moderate in monopolistic competition, high in oligopoly, and generally highest in monopoly.
- A financial analyst must understand the characteristics of market structures to better forecast a firm's future profit stream.
- The optimal MR equals MC. Only in perfect competition, however, does the MR equal price. In the remaining structures, price generally exceeds MR because a firm can sell more units only by reducing the per unit price.
- The quantity sold is highest in perfect competition. The price in perfect competition is usually lowest, but this depends on factors such as demand elasticity and increasing returns to scale (which may reduce the producer's MC). Monopolists, oligopolists, and producers in monopolistic competition attempt to differentiate their products so that they can charge higher prices.
- Typically, monopolists sell a smaller quantity at a higher price. Investors may benefit from being shareholders of monopolistic firms that have large margins and substantial positive cash flows.
- In perfect competition, firms do not earn economic profit. The market will compensate for the rental of capital and of management services, but the lack of pricing power implies that there will be no extra margins.
- In the short run, firms in any market structure can have economic profits, the more competitive a market is and the lower the barriers to entry, the faster the extra profits will fade. In the long run, new entrants shrink margins and push the least efficient firms out of the market.
- Oligopoly is characterized by the importance of strategic behavior. Firms can change the price, quantity, quality, and advertisement of the product to gain an advantage over their competitors. Several types of equilibrium (e.g., Nash, Cournot, kinked demand curve) may occur that affect the likelihood of each of the incumbents (and potential entrants in the long run) having economic profits. Price wars may be started to force weaker competitors to abandon the market.
- Measuring market power is complicated. Ideally, econometric estimates of the elasticity of demand and supply should be computed. However, because of the lack of reliable data and the fact that elasticity changes over time (so that past data may not apply to the current situation), regulators and economists often use simpler measures. The concentration ratio is simple, but the Herfindahl-Hirschman index (HHI), with a little more computation required, often produces a better figure for decision making.

2

PROFIT MAXIMIZATION: PRODUCTION BREAKEVEN, SHUTDOWN AND ECONOMIES OF SCALE



determine and interpret breakeven and shutdown points of production, as well as how economies and diseconomies of scale affect costs under perfect and imperfect competition

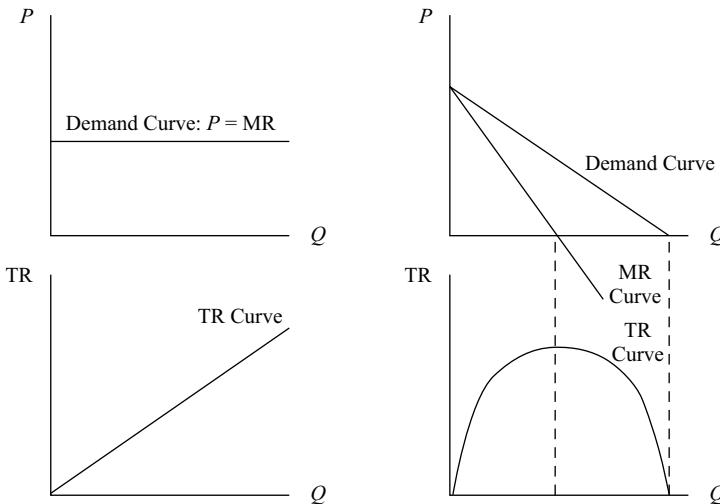
Firms generally can be classified as operating in either a perfectly competitive or an imperfectly competitive environment. The difference between the two manifests in the slope of the demand curve facing the firm. If the environment of the firm is perfectly competitive, it must take the market price of its output as given, so it faces a perfectly elastic, horizontal demand curve. In this case, the firm's marginal revenue (MR) and the price of its product are identical. Additionally, the firm's **average revenue** (AR), or revenue per unit, is also equal to price per unit. A firm that faces a negatively sloped demand curve, however, must lower its price to sell an additional unit, so its MR is less than price (P).

These characteristics of MR are also applicable to the total revenue (TR) functions. Under conditions of perfect competition, TR (as always) is equal to price times quantity: $TR = (P)(Q)$. But under conditions of perfect competition, price is dictated by the market; the firm has no control over price. As the firm sells one more unit, its TR rises by the exact amount of price per unit.

Under conditions of imperfect competition, price is a variable under the firm's control, and therefore price is a function of quantity: $P = f(Q)$, and $TR = f(Q) \times Q$. For simplicity, suppose the firm is monopolistic and faces the market demand curve, which we will assume is linear and negatively sloped. Because the monopolist is the only seller, its TR is identical to the total expenditure of all buyers in the market. When price is reduced and quantity sold increases in this environment, a decrease in price initially increases total expenditure by buyers and TR to the firm because the decrease in price is outweighed by the increase in units sold. But as price continues to fall, the decrease in price overshadows the increase in quantity, and total expenditure (revenue) falls. We can now depict the demand and TR functions for firms under conditions of perfect and imperfect competition, as shown in Exhibit 1.

Exhibit 1: Demand and Total Revenue Functions for Firms under Conditions of Perfect and Imperfect Competition

A. Perfectly Competitive Firm B. Imperfectly Competitive Firm

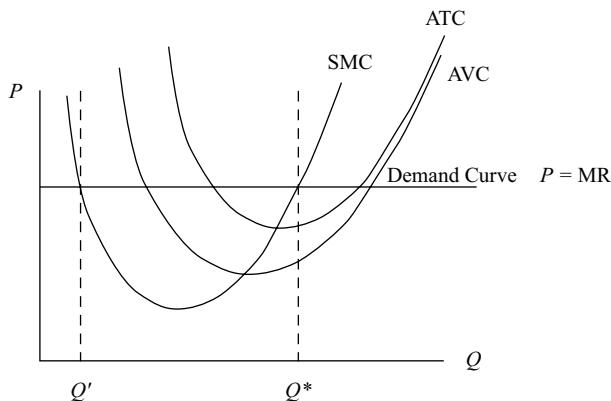


Panel A of Exhibit 1 depicts the demand curve (upper graph) and total revenue curve (lower graph) for the firm under conditions of perfect competition. Notice that the vertical axis in the upper graph is price per unit (e.g., GBP/bushel), whereas TR is measured on the vertical axis in the lower graph (e.g., GBP/week). The same is true for the respective axes in Panel B, which depicts the demand and total revenue curves for the monopolist. The TR curve for the firm under conditions of perfect competition is linear, with a slope equal to price per unit. The TR curve for the monopolist first rises (in the range where MR is positive and demand is elastic) and then falls (in the range where MR is negative and demand is inelastic) with output.

Profit-Maximization, Breakeven, and Shutdown Points of Production

We can now combine the firm's short-run TC curves with its TR curves to represent profit maximization in the cases of perfect competition and imperfect competition. Exhibit 2 shows both the AR and average cost curves in one graph for the firm under conditions of perfect competition.

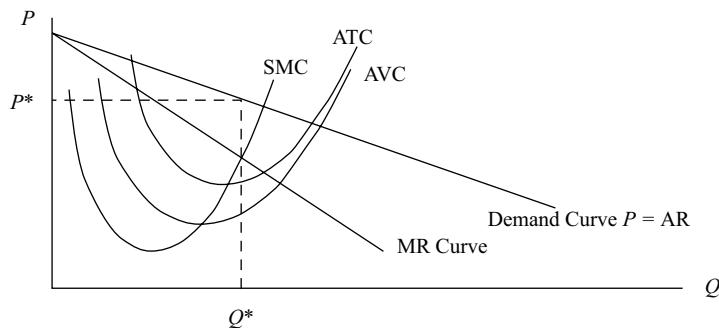
Exhibit 2: Demand and Average and Marginal Cost Curves for the Firm under Conditions of Perfect Competition



The firm is maximizing profit by producing Q^* , where price is equal to short-run marginal cost (SMC) and SMC is rising. Note at another output level, Q' , where $P = SMC$, SMC is still falling, so this cannot be a profit-maximizing solution. If market price were to rise, the firm's demand and MR curve would simply shift upward, and the firm would reach a new profit-maximizing output level to the right of Q^* . If, however, market price were to fall, the firm's demand and MR curve would shift downward, resulting in a new and lower level of profit-maximizing output. As depicted, this firm is currently earning a positive economic profit because market price exceeds average total cost (ATC), at output level Q^* . This profit is possible in the short run, but in the long run, competitors would enter the market to capture some of those profits and would drive the market price down to a level equal to each firm's ATC.

Exhibit 3 depicts the cost and revenue curves for the monopolist that is facing a negatively sloped market demand curve. The MR and demand curves are not identical for this firm. But the profit-maximizing rule is still the same: Find the level of Q that equates SMC to MR—in this case, Q^* . Once that level of output is determined, the optimal price to charge is given by the firm's demand curve at P^* . This monopolist is earning positive economic profit because its price exceeds its ATC. The barriers to entry that give this firm its monopolistic power mean that outside competitors would not be able to compete away this firm's profits.

Exhibit 3: Demand and Average and Marginal Cost Curves for the Monopolistic Firm



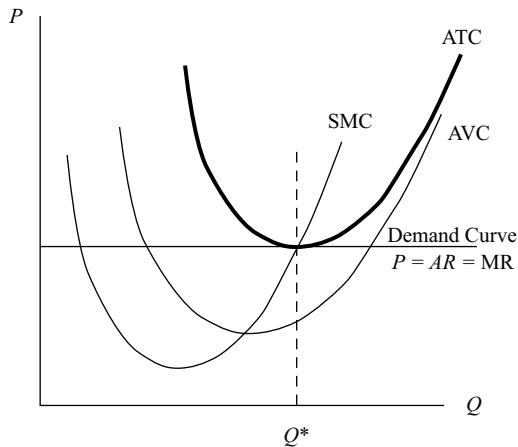
Breakeven Analysis and Shutdown Decision

A firm is said to break even if its TR is equal to its TC. It also can be said that a firm breaks even if its price (AR) is exactly equal to its ATC, which is true under conditions of perfect and imperfect competition. Of course, the goal of management is not just to breakeven but to maximize profit. However, perhaps the best the firm can do is cover all of its economic costs. Economic costs are the sum of total accounting costs and implicit opportunity costs. A firm whose revenue is equal to its economic costs is covering the opportunity cost of all of its factors of production, including capital. Economists would say that such a firm is earning normal profit, but not positive economic profit. It is earning a rate of return on capital just equal to the rate of return that an investor could expect to earn in an equivalently risky alternative investment (opportunity cost). Firms that are operating in a competitive environment with no barriers to entry from other competitors can expect, in the long run, to be unable to earn a positive economic profit; the excess rate of return would attract entrants who would produce more output and ultimately drive the market price down to the level at which each firm is, at best, just earning a normal profit. This situation, of course, does not imply that the firm is earning zero accounting profit.

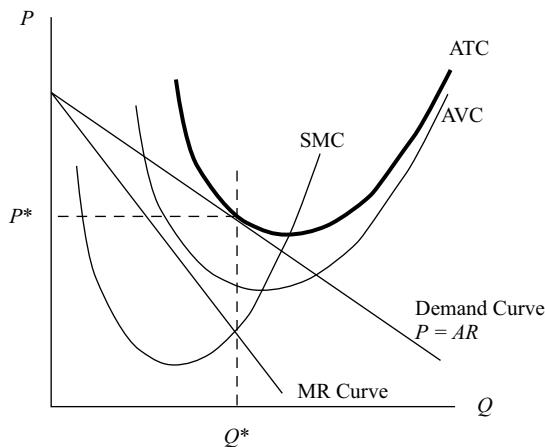
Exhibit 4 depicts the condition for both a firm under conditions of perfect competition (Panel A) and a monopolist (Panel B) in which the best each firm can do is to break even. Note that at the level of output at which SMC is equal to MR, price is equal to ATC. Hence, economic profit is zero, and the firms are breaking even.

Exhibit 4: Examples of Firms under Perfect Competition and Monopolistic Firms That Can, at Best, Break Even

A. Perfect Competition



B. Monopolist



The Shutdown Decision

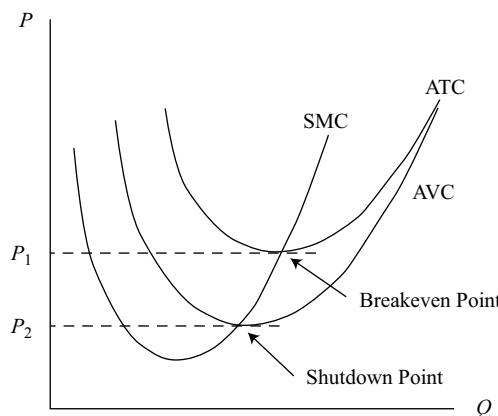
In the long run, if a firm cannot earn at least a zero economic profit, it will not operate because it is not covering the opportunity cost of all of its factors of production, labor, and capital. In the short run, however, a firm might find it advantageous to continue to operate even if it is not earning at least a zero economic profit. The discussion that follows addresses the decision to continue to operate and earn negative profit or shut down operations.

Recall that typically some or all of a firm's fixed costs are incurred regardless of whether the firm operates. The firm might have a lease on its building that it cannot avoid paying until the lease expires. In that case, the lease payment is a sunk cost: It cannot be avoided, no matter what the firm does. Sunk costs must be ignored in the decision to continue to operate in the short run. As long as the firm's revenues cover

at least its variable cost, the firm is better off continuing to operate. If price is greater than average variable cost (AVC), the firm is covering not only all of its variable cost but also a portion of fixed cost.

In the long run, unless market price increases, this firm would exit the industry. But in the short run, it will continue to operate at a loss. Exhibit 5 depicts a firm under conditions of perfect competition facing three alternative market price ranges for its output. At any price above P_1 , the firm can earn a positive profit and clearly should continue to operate. At a price below P_2 , the minimum AVC, the firm could not even cover its variable cost and should shut down. At prices between P_2 and P_1 , the firm should continue to operate in the short run because it is able to cover all of its variable cost and contribute something toward its unavoidable fixed costs. Economists refer to the minimum AVC point as the **shutdown point** and the minimum ATC point as the **breakeven point**.

Exhibit 5: A Firm under Conditions of Perfect Competition Will Choose to Shut Down If Market Price Is Less Than Minimum AVC



EXAMPLE 1

Breakeven Analysis and Profit Maximization When the Firm Faces a Negatively Sloped Demand Curve under Imperfect Competition

Revenue and cost information for a future period including all opportunity costs is presented in Exhibit 6 for WR International, a newly formed corporation that engages in the manufacturing of low-cost, prefabricated dwelling units for urban housing markets in emerging economies. (Note that quantity increments are in blocks of 10 for a 250 change in price.) The firm has few competitors in a market setting of imperfect competition.

Exhibit 6: Revenue and Cost Information for WR International

Quantity (Q)	Price (P)	Total Revenue (TR)	Total Cost (TC) ^a	Profit
0	10,000	0	100,000	-100,000
10	9,750	97,500	170,000	-72,500

Quantity (Q)	Price (P)	Total Revenue (TR)	Total Cost (TC) ^a	Profit
20	9,500	190,000	240,000	-50,000
30	9,250	277,500	300,000	-22,500
40	9,000	360,000	360,000	0
50	8,750	437,500	420,000	17,500
60	8,500	510,000	480,000	30,000
70	8,250	577,500	550,000	27,500
80	8,000	640,000	640,000	0
90	7,750	697,500	710,000	-12,500
100	7,500	750,000	800,000	-50,000

^a Includes all opportunity costs

1. How many units must WR International sell to initially break even?

Solution:

WR International will initially break even at 40 units of production, where TR and TC equal 360,000.

2. Where is the region of profitability?

Solution:

The region of profitability will range from greater than 40 units to less than 80 units. Any production quantity of less than 40 units and any quantity greater than 80 units will result in an economic loss.

3. At what point will the firm maximize profit? At what points are there economic losses?

Solution:

Maximum profit of 30,000 will occur at 60 units. Lower profit will occur at any output level that is higher or lower than 60 units. From 0 units to less than 40 units and for quantities greater than 80 units, economic losses occur.

Given the relationships between TR, total variable costs (TVC), and total fixed costs (TFC), Exhibit 7 summarizes the decisions to operate, shut down production, or exit the market in both the short run and the long run. The firm must cover its variable cost to remain in business in the short run; if TR cannot cover TVC, the firm shuts down production to minimize loss. The loss would be equal to the amount of fixed cost. If TVC exceeds TR in the long run, the firm will exit the market to avoid the loss associated with fixed cost at zero production. By exiting the market, the firm's investors do not suffer the erosion of their equity capital from economic losses. When TR is enough to cover TVC but not all of TFC, the firm can continue to produce in the short run but will not be able to maintain financial solvency in the long run.

Exhibit 7: Short-Run and Long-Run Decisions to Operate or Not

Revenue–Cost Relationship	Short-Run Decision	Long-Term Decision
$TR = TC$	Stay in market	Stay in market
$TR = TVC$ but $< TC$	Stay in market	Exit market
$TR < TVC$	Shut down production	Exit market

EXAMPLE 2**Shutdown Analysis**

For the most recent financial reporting period, a London-based business has revenue of GBP2 million and TC of GBP2.5 million, which are or can be broken down into TFC of GBP1 million and TVC of GBP1.5 million. The net loss on the firm's income statement is reported as GBP500,000 (ignoring tax implications). In prior periods, the firm had reported profits on its operations.

1. What decision should the firm make regarding operations over the short term?

Solution:

In the short run, the firm is able to cover all of its TVC but only half of its GBP1 million in TFC. If the business ceases to operate, its loss would be GBP1 million, the amount of TFC, whereas the net loss by operating would be minimized at GBP500,000. The firm should attempt to operate by negotiating special arrangements with creditors to buy time to return operations back to profitability.

2. What decision should the firm make regarding operations over the long term?

Solution:

If the revenue shortfall is expected to persist over time, the firm should cease operations, liquidate assets, and pay debts to the extent possible. Any residual for shareholders would decrease the longer the firm is allowed to operate unprofitably.

3. Assume the same business scenario except that revenue is now GBP1.3 million, which creates a net loss of GBP1.2 million. What decision should the firm make regarding operations in this case?

Solution:

The firm would minimize loss at GBP1 million of TFC by shutting down. If the firm decided to continue to do business, the loss would increase to GBP1.2 million. Shareholders would save GBP200,000 in equity value by pursuing this option. Unquestionably, the business would have a rather short life expectancy if this loss situation were to continue.

When evaluating profitability, particularly of start-up firms and businesses using turnaround strategies, analysts should consider highlighting breakeven and shutdown points in their financial research. Identifying the unit sales levels at which the firm

enters or leaves the production range for profitability and at which the firm can no longer function as a viable business entity provides invaluable insight when making investment decisions.

Economies and Diseconomies of Scale with Short-Run and Long-Run Cost Analysis

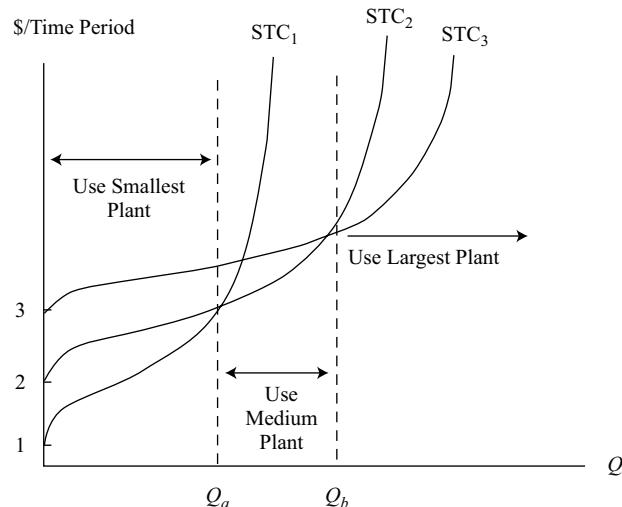
Rational behavior dictates that the firm select an operating size or scale that maximizes profit over any time frame. The time frame that defines the short run and long run for any firm is based on the ability of the firm to adjust the quantities of the fixed resources it uses. The short run is the time period during which at least one of the factors of production, such as technology, physical capital, and plant size, is fixed. The long run is defined as the time period during which all factors of production are variable. Additionally, in the long run, firms can enter or exit the market based on decisions regarding profitability. The long run is often referred to as the “planning horizon” in which the firm can choose the short-run position or optimal operating size that maximizes profit over time. The firm is always operating in the short run but planning in the long run.

The time required for long-run adjustments varies by industry. For example, the long run for a small business using very little technology and physical capital may be less than a year, whereas for a capital-intensive firm, the long run may be more than a decade. Given enough time, however, all production factors are variable, which allows the firm to choose an operating size or plant capacity based on different technologies and physical capital. In this regard, costs and profits will differ between the short run and the long run.

Short- and Long-Run Cost Curves

Recall that when we addressed the short-run cost curves of the firm, we assumed that the capital input was held constant. That meant that the only way to vary output in the short run was to change the level of the variable input—in our case, labor. If the capital input—namely, plant and equipment—were to change, however, we would have an entirely new set of short-run cost curves, one for each level of capital input.

The short-run total cost includes all the inputs—labor and capital—the firm is using to produce output. For reasons discussed earlier, the typical short-run total cost (STC) curve might rise with output, first at a decreasing rate because of specialization economies and then at an increasing rate, reflecting the law of diminishing marginal returns to labor. TFC, the quantity of capital input multiplied by the rental rate on capital, determines the vertical intercept of the STC curve. At higher levels of fixed input, TFC is greater, but the production capacity of the firm is also greater. Exhibit 8 shows three different STC curves for the same technology but using three distinct levels of capital input—points 1, 2, and 3 on the vertical axis.

Exhibit 8: Short-Run Total Cost Curves for Various Plant Sizes

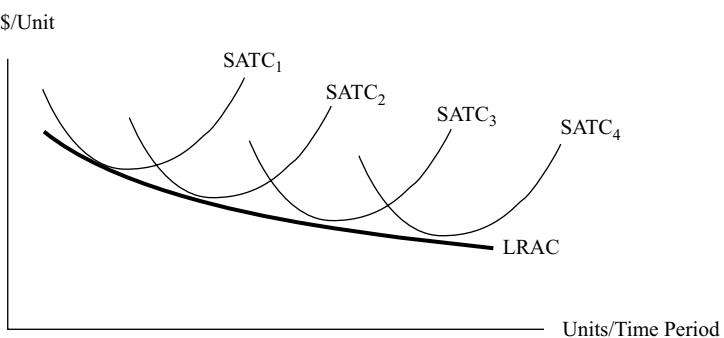
Plant Size 1 is the smallest and, of course, has the lowest fixed cost; hence, its STC_1 curve has the lowest vertical intercept. Note that STC_1 begins to rise more steeply with output, reflecting the lower plant capacity. Plant Size 3 is the largest of the three and reflects that size with both a higher fixed cost and a lower slope at any level of output. If a firm decided to produce an output between zero and Q_a , it would plan on building Plant Size 1 because for any output level in that range, its cost would be less than it would be for Plant Size 2 or 3. Accordingly, if the firm were planning to produce output greater than Q_b , it would choose Plant Size 3 because its cost for any of those levels of output would be lower than it would be for Plant Size 1 or 2. Of course, Plant Size 2 would be chosen for output levels between Q_a and Q_b . The long-run total cost curve is derived from the lowest level of STC for each level of output because in the long run, the firm is free to choose which plant size it will operate. This curve is called an “envelope curve.” In essence, this curve envelopes—encompasses—all possible combinations of technology, plant size, and physical capital.

For each STC curve, there is also a corresponding **short-run average total cost** (SATC) curve and a corresponding **long-run average total cost** (LRAC) curve, the envelope curve of all possible short-run average total cost curves. The shape of the LRAC curve reflects an important concept called **economies of scale** and **diseconomies of scale**.

Defining Economies of Scale and Diseconomies of Scale

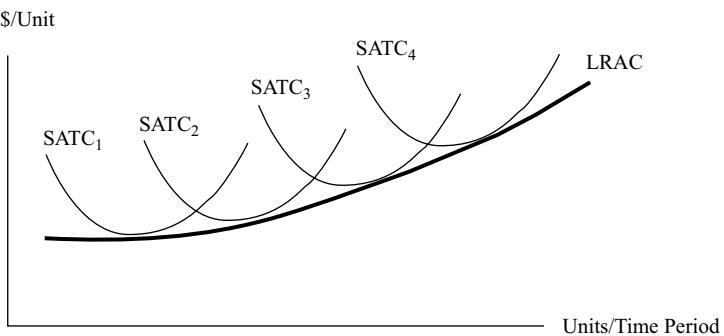
When a firm increases all of its inputs to increase its level of output (obviously, a long-run concept), it is said to *scale up* its production. *Scaling down* is the reverse—decreasing all of its inputs to produce less in the long run. Economies of scale occur if, as the firm increases its output, cost per unit of production falls. Graphically, this definition translates into an LRAC curve with a negative slope. Exhibit 9 depicts several SATC curves, one for each plant size, and the LRAC curve representing economies of scale.

Exhibit 9: Short-Run Average Total Cost Curves for Various Plant Sizes and Their Envelope Curve, LRAC: Economies of Scale



Diseconomies of scale occur if cost per unit rises as output increases. Graphically, diseconomies of scale translate into an LRAC curve with a positive slope. Exhibit 10 depicts several SATC curves, one for each plant size, and their envelope curve, the LRAC curve, representing diseconomies of scale.

Exhibit 10: Short-Run Average Total Cost Curves for Various Plant Sizes and Their Envelope Curve, LRAC: Diseconomies of Scale



As the firm grows in size, economies of scale and a lower ATC can result from the following factors:

- Achieving **increasing returns to scale** when a production process allows for increases in output that are proportionately larger than the increase in inputs.
- Having a division of labor and management in a large firm with numerous workers, which allows each worker to specialize in one task rather than perform many duties, as in the case of a small business (as such, workers in a large firm become more proficient at their jobs).
- Being able to afford more expensive, yet more efficient equipment and to adapt the latest in technology that increases productivity.
- Effectively reducing waste and lowering costs through marketable by-products, less energy consumption, and enhanced quality control.
- Making better use of market information and knowledge for more effective managerial decision making.

- Obtaining discounted prices on resources when buying in larger quantities.

A classic example of a business that realizes economies of scale through greater physical capital investment is an electric utility. By expanding output capacity to accommodate a larger customer base, the utility company's per-unit cost will decline. Economies of scale help explain why electric utilities have naturally evolved from localized entities to regional and multiregional enterprises. Walmart is an example of a business that has used its bulk purchasing power to obtain deep discounts from suppliers to keep costs and prices low. Walmart also uses the latest technology to monitor point-of-sale transactions to gather timely market information to respond to changes in customer buying behavior, which leads to economies of scale through lower distribution and inventory costs.

Factors that can lead to diseconomies of scale, inefficiencies, and rising costs when a firm increases in size include the following:

- Realizing **decreasing returns to scale** when a production process leads to increases in output that are proportionately smaller than the increase in inputs.
- Being so large that it cannot be properly managed.
- Overlapping and duplicating business functions and product lines.[
- Experiencing higher resource prices because of supply constraints when buying inputs in large quantities.

Before its restructuring, General Motors (GM) was an example of a business that had realized diseconomies of scale by becoming too large. Scale diseconomies occurred through product overlap and duplication (i.e., similar or identical automobile models), and the fixed cost for these models was not spread over a large volume of output. In 2009, GM decided to discontinue three brands (Saturn, Pontiac, and Hummer) and also to drop various low-volume product models that overlapped with others. GM had numerous manufacturing plants around the world and sold vehicles in more than a hundred countries. Given this geographic dispersion in production and sales, the company had communication and management coordination problems, which resulted in higher costs. In 2017, GM sold its European arm, Opel, to Groupe PSA, the maker of Peugeot and Citroën. GM also had significantly higher labor costs than its competitors. As the largest producer in the market, it had been a target of labor unions for higher compensation and benefits packages relative to other firms.

Economies and diseconomies of scale can occur at the same time; the impact on LRAC depends on which dominates. If economies of scale dominate, LRAC decreases with increases in output. The reverse holds true when diseconomies of scale prevail. LRAC may fall (economies of scale) over a range of output and then LRAC might remain constant over another range, which could be followed by a range over which diseconomies of scale prevail, as depicted in Exhibit 11.

The minimum point on the LRAC curve is referred to as the **minimum efficient scale**. The minimum efficient scale is the optimal firm size under perfect competition over the long run. Theoretically, perfect competition forces the firm to operate at the minimum point on the LRAC curve because the market price will be established at this level over the long run. If the firm is not operating at this least-cost point, its long-term viability will be threatened.

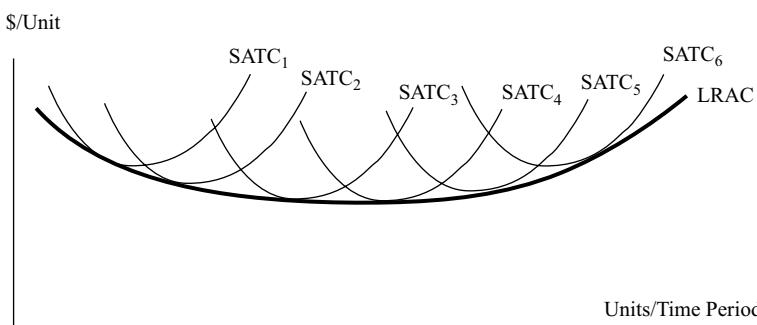
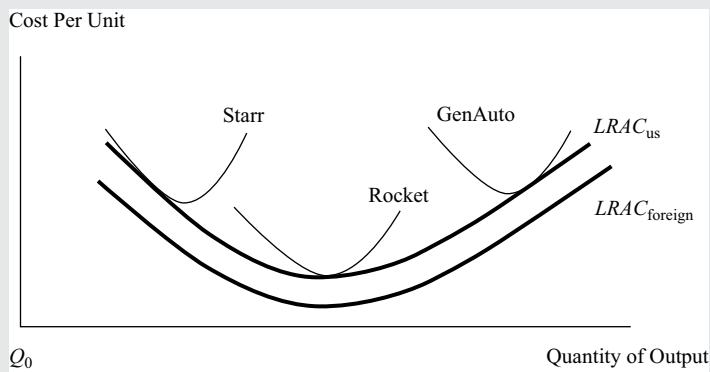
Exhibit 11: LRAC Can Exhibit Economies and Diseconomies of Scale**EXAMPLE 3****Long-Run Average Total Cost Curve**

Exhibit 12 displays the long-run average total cost curve ($LRAC_{US}$) and the short-run average total cost curves for three hypothetical US-based automobile manufacturers—Starr Vehicles (Starr), Rocket Sports Cars (Rocket), and General Auto (GenAuto). The $LRAC$ curve for foreign-owned automobile companies that compete in the US auto market ($LRAC_{foreign}$) is also indicated in the graph. (The market structure implicit in the exhibit is imperfect competition.)

1. To what extent are the cost relationships depicted in Exhibit 12 useful for an economic and financial analysis of the three US-based auto firms?

Exhibit 12: Long-Run Average Total Cost Curves for Three Auto Manufacturers**Solution:**

First, it is observable that the foreign auto companies have a lower $LRAC$ compared with that of the US automobile manufacturers. This competitive position places the US firms at a cost—and possibly, pricing—disadvantage in the market, with the potential to lose market share to the lower-cost foreign competitors. Second, only Rocket operates at the minimum point of the $LRAC_{US}$, whereas GenAuto is situated in the region of diseconomies of scale and Starr is positioned in the economies of scale portion of the curve. To become more efficient and competitive, GenAuto needs to downsize and restructure, which means moving down the $LRAC_{US}$ curve to a smaller

yet lower-cost production volume. In contrast, Starr has to grow in size to become more efficient and competitive by lowering per-unit costs. From a long-term investment prospective and given its cost advantage, Rocket has the potential to create more investment value relative to GenAuto and Starr. Over the long run, if GenAuto and Starr can lower their ATC, they will become more attractive to investors. But if any of the three US auto companies cannot match the cost competitiveness of the foreign firms, they may be driven from the market. In the long run, the lower-cost foreign automakers pose a severe competitive challenge to the survival of the US manufacturers and their ability to maintain and grow shareholders' wealth.

QUESTION SET

1. An agricultural firm operating in a perfectly competitive market supplies wheat to manufacturers of consumer food products and animal feeds. If the firm were able to expand its production and unit sales by 10%, the *most likely* result would be:
 - A. a 10% increase in total revenue.
 - B. a 10% increase in average revenue.
 - C. a less than 10% increase in total revenue.

Solution:

A is correct. In a perfectly competitive market, an increase in supply by a single firm will not affect price. Therefore, an increase in units sold by the firm will be matched proportionately by an increase in revenue.

2. The marginal revenue per unit sold for a firm doing business under conditions of perfect competition will *most likely* be:
 - A. equal to average revenue.
 - B. less than average revenue.
 - C. greater than average revenue.

Solution:

A is correct. Under perfect competition, a firm is a price taker at any quantity supplied to the market, and $AR = MR = \text{Price}$.

3. A profit maximum is *least likely* to occur when:
 - A. average total cost is minimized.
 - B. marginal revenue is equal to marginal cost.
 - C. the difference between total revenue and total cost is maximized.

Solution:

A is correct. The quantity at which average total cost is minimized does not necessarily correspond to a profit maximum.

4. The short-term breakeven point of production for a firm operating under perfect competition will *most likely* occur when:
 - A. price is equal to average total cost.
 - B. marginal revenue is equal to marginal cost.

- C. marginal revenue is equal to average variable costs.

Solution:

A is correct. Under perfect competition, price is equal to marginal revenue. A firm breaks even when marginal revenue equals average total cost.

3

INTRODUCTION TO MARKET STRUCTURES



describe characteristics of perfect competition, monopolistic competition, oligopoly, and pure monopoly

Different market structures result in different sets of choices facing a firm's decision makers. Thus, an understanding of market structure is a powerful tool in analyzing issues, such as a firm's pricing of its products and, more broadly, its potential to increase profitability. In the long run, a firm's profitability will be determined by the forces associated with the market structure within which it operates. In a highly competitive market, long-run profits will be driven down by the forces of competition. In less competitive markets, large profits are possible even in the long run; in the short run, any outcome is possible. Therefore, understanding the forces behind the market structure will aid the financial analyst in determining firms' short- and long-term prospects.

Market structures address questions such as the following: What determines the degree of competition associated with each market structure? Given the degree of competition associated with each market structure, what decisions are left to the management team developing corporate strategy? How does a chosen pricing and output strategy evolve into specific decisions that affect the profitability of the firm? The answers to these questions are related to the forces of the market structure within which the firm operates.

Analysis of Market Structures

Traditionally, economists classify a market into one of four structures: perfect competition, monopolistic competition, oligopoly, and monopoly.

Economists define a market as a group of buyers and sellers that are aware of each other and can agree on a price for the exchange of goods and services. Although internet access has extended a number of markets worldwide, certain markets remain limited by geographic boundaries. For example, the internet search engine Google operates in a worldwide market. In contrast, the market for premixed cement is limited to the area within which a truck can deliver the mushy mix from the plant to a construction site before the compound becomes useless. Thomas L. Friedman's international best seller *The World Is Flat* challenges the concept of the geographic limitations of the market. If the service being provided by the seller can be digitized, its market expands worldwide. For example, a technician can scan your injury in a clinic in Switzerland. That radiographic image can be digitized and sent to a radiologist in India to be read. As a customer (i.e., patient), you may never know that part of the medical service provided to you was the result of a worldwide market.

Some markets are highly concentrated, with the majority of total sales coming from a small number of firms. For example, in the market for internet search, three firms controlled 98.9 percent of the US market (Google 63.5 percent, Microsoft 24 percent, and Oath (formerly Yahoo!) 11.4 percent) as of January 2018. Other markets

are fragmented, such as automobile repairs, in which small independent shops often dominate and large chains may or may not exist. New products can lead to market concentration. For example, Apple introduced its first digital audio player (iPod) in 2001 and despite the entry of competitors had a world market share of more than 70 percent among digital audio players in 2009.

THE IMPORTANCE OF MARKET STRUCTURE

Consider the evolution of television broadcasting. As the market environment for television broadcasting evolved, the market structure changed, resulting in a new set of challenges and choices. In the early days, viewers had only one choice: the “free” analog channels that were broadcast over the airwaves. Most countries had one channel, owned and run by the government. In the United States, some of the more populated markets were able to receive more channels because local channels were set up to cover a market with more potential viewers. By the 1970s, new technologies made it possible to broadcast by way of cable connectivity and the choices offered to consumers began to expand rapidly. Cable television challenged the “free” broadcast channels by offering more choice and a better-quality picture. The innovation was expensive for consumers and profitable for the cable companies. By the 1990s, a new alternative began to challenge the existing broadcast and cable systems: satellite television. Satellite providers offered a further expanded set of choices, albeit at a higher price, than the free broadcast and cable alternatives. In the early 2000s, satellite television providers lowered their pricing to compete directly with the cable providers.

Today, cable program providers, satellite television providers, and terrestrial digital broadcasters that offer premium and pay-per-view channels compete for customers who are increasingly finding content on the internet and on their mobile devices. Companies like Netflix, Apple, and Amazon offered alternative ways for consumers to access content. Over time, these companies had moved beyond the repackaging of existing shows to developing their own content, mirroring the evolution of cable channels, such as HBO and ESPN a decade earlier.

This is a simple illustration of the importance of market structure. As the market for television broadcasting became increasingly competitive, managers have had to make decisions regarding product packaging, pricing, advertising, and marketing to survive in the changing environment. In addition, mergers and acquisitions as a response to these competitive pressures have changed the essential structure of the industry.

Market structure can be broken down into four distinct categories: perfect competition, monopolistic competition, oligopoly, and monopoly.

We start with the most competitive environment, **perfect competition**. Unlike some economic concepts, perfect competition is not merely an ideal based on assumptions. Perfect competition is a reality—for example, in several commodities markets, in which sellers and buyers have a strictly homogeneous product and no single producer is large enough to influence market prices. Perfect competition’s characteristics are well recognized and its long-run outcome is unavoidable. Profits under the conditions of perfect competition are driven to the required rate of return paid by the entrepreneur to borrow capital from investors (so-called normal profit or rental cost of capital). This does not mean that all perfectly competitive industries are doomed to extinction by a lack of profits. On the contrary, millions of businesses that do very well are living under the pressures of perfect competition.

Monopolistic competition is also highly competitive; however, it is considered a form of imperfect competition. Two economists, Edward H. Chamberlin (United States) and Joan Robinson (United Kingdom), identified this hybrid market and came

up with the term because this market structure not only has strong elements of competition but also some monopoly-like conditions. The competitive characteristic is a notably large number of firms, while the monopoly aspect is the result of product differentiation. That is, if the seller can convince consumers that its product is uniquely different from other, similar products, then the seller can exercise some degree of pricing power over the market. A good example is the brand loyalty associated with soft drinks such as Coca-Cola. Many of Coca-Cola's customers believe that their beverages are truly different from and better than all other soft drinks. The same is true for fashion creations and cosmetics.

The **oligopoly** market structure is based on a relatively small number of firms supplying the market. The small number of firms in the market means that each firm must consider what retaliatory strategies the other firms will pursue when prices and production levels change. Consider the pricing behavior of commercial airline companies. Pricing strategies and route scheduling are based on the expected reaction of the other carriers in similar markets. For any given route—say, from Paris, France, to Chennai, India—only a few carriers are in competition. If one of the carriers changes its pricing package, others likely will retaliate. Understanding the market structure of oligopoly markets can help identify a logical pattern of strategic price changes for the competing firms.

Finally, the least competitive market structure is the **monopoly**. In pure monopoly markets, no other good substitutes exist for the given product or service. A single seller, which, if allowed to operate without constraint, exercises considerable power over pricing and output decisions. In most market-based economies around the globe, pure monopolies are regulated by a governmental authority. The most common example of a regulated monopoly is the local electrical power provider. In most cases, the monopoly power provider is allowed to earn a normal return on its investment and prices are set by the regulatory authority to allow that return.

Factors That Determine Market Structure

The following five factors determine market structure:

1. The number and relative size of firms supplying the product;
2. The degree of product differentiation;
3. The power of the seller over pricing decisions;
4. The relative strength of the barriers to market entry and exit; and
5. The degree of non-price competition.

The number and relative size of firms in a market influence market structure. When many firms exist, the degree of competition increases. With fewer firms supplying a good or service, consumers are limited in their market choices. One extreme case is the monopoly market structure, with only one firm supplying a unique good or service. Another extreme is perfect competition, with many firms supplying a similar product. Finally, an example of relative size is the automobile industry, in which a small number of large international producers (e.g., Volkswagen and Toyota) are the leaders in the global market, and a number of small companies either have market power because they are niche players (e.g., Ferrari or McLaren) or have limited market power because of their narrow range of models or limited geographical presence (e.g., Mazda or Stellantis).

In the case of monopolistic competition, many firms are providing products to the market, as with perfect competition. However, one firm's product is differentiated in some way that makes it appear to be better than similar products from other firms. If a firm is successful in differentiating its product, this differentiation will provide pricing leverage. The more dissimilar the product appears, the more the market will

resemble the monopoly market structure. A firm can differentiate its product through aggressive advertising campaigns; frequent styling changes; the linking of its product with other complementary products; or a host of other methods.

When the market dictates the price based on aggregate supply and demand conditions, the individual firm has no control over pricing. The typical hog farmer in Nebraska and the milk producer in Bavaria are **price takers**. That is, they must accept whatever price the market dictates. This is the case under the market structure of perfect competition. In the case of monopolistic competition, the success of product differentiation determines the degree with which the firm can influence price. In the case of oligopoly, there are so few firms in the market that price control becomes possible. However, the small number of firms in an oligopoly market invites complex pricing strategies. Collusion, price leadership by dominant firms, and other pricing strategies can result.

The degree to which one market structure can evolve into another and the difference between potential short-run outcomes and long-run equilibrium conditions depend on the strength of the barriers to entry and the possibility that firms fail to recoup their original costs or lose money for an extended period of time and therefore are forced to exit the market. Barriers to entry can result from large capital investment requirements, as in the case of petroleum refining. Barriers may also result from patents, as in the case of some electronic products and drug formulas. Another entry consideration is the possibility of high exit costs. For example, plants that are specific to a special line of products, such as aluminum smelting plants, are non-redeployable, and exit costs would be high without a liquid market for the firm's assets. High exit costs deter entry and therefore also are considered barriers to entry. In the case of farming, the barriers to entry are low. Production of corn, soybeans, wheat, tomatoes, and other produce is an easy process to replicate; therefore, those are highly competitive markets.

Non-price competition dominates those market structures in which product differentiation is critical. Therefore, monopolistic competition relies on competitive strategies that may not include pricing changes. An example of non-price competition is product differentiation through marketing. In other circumstances, non-price competition may occur because the few firms in the market feel dependent on each other. Each firm fears retaliatory price changes that would reduce total revenue for all of the firms in the market. Because oligopoly industries have so few firms, each firm feels dependent on the pricing strategies of the others. Therefore, non-price competition becomes a dominant strategy.

Characteristics of Market Structure

Exhibit 13: Characteristics of Market Structure

Market Structure	Number of Sellers	Degree of Product Differentiation	Barriers to Entry	Pricing Power of Firm	Non-Price Competition
Perfect competition	Many	Homogeneous/ Standardized	Very Low	None	None
Monopolistic competition	Many	Differentiated	Low	Some	Advertising and Product Differentiation
Oligopoly	Few	Homogeneous/ Standardized	High	Some or Considerable	Advertising and Product Differentiation
Monopoly	One	Unique Product	Very High	Considerable	Advertising

From the perspective of the owners of the firm, the most desirable market structure is that with the most control over price, because this control can lead to large profits (Exhibit 13). Monopoly and oligopoly markets offer the greatest potential control over price; monopolistic competition offers less control. Firms operating under perfectly competitive market conditions have no control over price. From the consumers' perspective, the most desirable market structure is that with the greatest degree of competition because prices are generally lower. Thus, consumers would prefer as many goods and services as possible to be offered in competitive markets.

As often happens in economics, there is a trade-off. While perfect competition gives the largest quantity of a good at the lowest price, other market forms may spur more innovation. Specifically, firms may incur high costs in researching a new product, and they will incur such costs only if they expect to earn an attractive return on their research investment. This is the case often made for medical innovations, for example—the cost of clinical trials and experiments to create new medicines would bankrupt perfectly competitive firms but may be acceptable in an oligopoly market structure. Therefore, consumers can benefit from less-than-perfectly-competitive markets.

PORTER'S FIVE FORCES AND MARKET STRUCTURE

A financial analyst aiming to establish market conditions and consequent profitability of incumbent firms should start with the questions posed earlier: How many sellers are there? Is the product differentiated? Moreover, in the case of monopolies and quasi-monopolies, the analyst should evaluate the legislative and regulatory framework: Can the company set prices freely, or are there governmental controls? Finally, the analyst should consider the threat of competition from potential entrants.

This analysis is often summarized by students of corporate strategy as “Porter’s five forces,” named after Harvard Business School professor Michael E. Porter. His book, *Competitive Strategy*, presented a systematic analysis of the practice of market strategy. Porter identified the five forces as follows:

- Threat of entry;
- Power of suppliers;
- Power of buyers (customers);
- Threat of substitutes; and
- Rivalry among existing competitors.

It is easy to note the parallels between four of these five forces and the questions posed earlier. The only “orphan” is the power of suppliers, which is not at the core of the theoretical economic analysis of competition, but which has substantial weight in the practical analysis of competition and profitability.

Some stock analysts use the term “economic moat” to suggest that some of the factors protecting the profitability of a firm are similar to the moats (ditches full of water) that were used to protect some medieval castles. A deep moat means that there is little or no threat of entry by invaders (i.e., competitors). It also means that customers are locked in because of high switching costs.

QUESTION SET

1. A market structure characterized by many sellers with each having some pricing power and product differentiation is *best* described as:

- A. oligopoly.
- B. perfect competition.
- C. monopolistic competition.

Solution:

C is correct. Monopolistic competition is characterized by many sellers, differentiated products, and some pricing power.

2. A market structure with relatively few sellers of a homogeneous or standardized product is *best* described as:

- A. oligopoly.
- B. monopoly.
- C. perfect competition.

Solution:

A is correct. Few sellers of a homogeneous or standardized product characterizes an oligopoly.

MONOPOLISTIC COMPETITION**4**

explain supply and demand relationships under monopolistic competition, including the optimal price and output for firms as well as pricing strategy

Many market structures exhibit characteristics of strong competitive forces; however, other distinct non-competitive factors can also play important roles in the market. As the name implies, monopolistic competition is a hybrid market. *The most distinctive factor in monopolistic competition is product differentiation.* Recall the characteristics outlined earlier:

1. The market has a large number of potential buyers and sellers.
2. The products offered by each seller are close substitutes for the products offered by other firms, and each firm tries to make its product look different.
3. Entry into and exit from the market are possible with fairly low costs.
4. Firms have some pricing power.
5. Suppliers differentiate their products through advertising and other non-price strategies.

While the market is made up of many firms that compose the product group, each producer attempts to distinguish its product from that of the others. Product differentiation is accomplished in a variety of ways. For example, consider the wide variety of communication devices available today. Decades ago, when each communication

market was controlled by a regulated single seller (the telephone company), all telephones were alike. In today's deregulated market, the variety of physical styles and colors is extensive. All versions accomplish many of the same tasks.

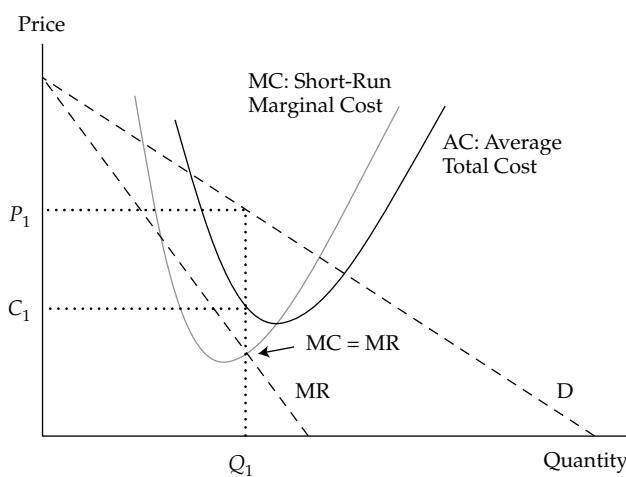
The communication device manufacturers and providers differentiate their products with different colors, styles, networks, bundled applications, conditional contracts, functionality, and more. Advertising is usually the avenue pursued to convince consumers that the goods in the product group are different. Successful advertising and trademark branding result in customer loyalty. A good example is the brand loyalty associated with Harley-Davidson motorcycles. Harley-Davidson's customers believe that their motorcycles are truly different from and better than all other motorcycles.

The extent to which the producer is successful in product differentiation determines pricing power in the market. Very successful differentiation results in a market structure that resembles the single-seller market (monopoly). Because of relatively low entry and exit costs, competition will, in the long run, drive prices and revenues down toward an equilibrium similar to perfect competition. Thus, the hybrid market displays characteristics found in both perfectly competitive and monopoly markets.

Demand Analysis in Monopolistically Competitive Markets

Because each good sold in the product group is somewhat different from the others, the demand curve for each firm in the monopolistic competition market structure is downward sloping to the right. Price and the quantity demanded are negatively related. Lowering the price will increase the quantity demanded and raising the price will decrease the quantity demanded. There will be ranges of prices within which demand is elastic and (lower) prices at which demand is inelastic. Exhibit 14 illustrates the demand, marginal revenue, and cost structures facing a monopolistically competitive firm in the short run.

Exhibit 14: Short-Run Equilibrium in Monopolistic Competition



In the short run, the profit-maximizing choice is the level of output at which $MR = MC$. Because the product is somewhat different from that of the competitors, the firm can charge the price determined by the demand curve. Therefore, in Exhibit 14, Q_1 is the ideal level of output and P_1 is the price consumers are willing to pay to acquire that quantity. Total revenue is the area of the rectangle $P_1 \times Q_1$.

Supply Analysis in Monopolistically Competitive Markets

In perfect competition, the firm's supply schedule is represented by the marginal cost schedule. In monopolistic competition, there is no well-defined supply function. The information used to determine the appropriate level of output is based on the intersection of MC and MR. However, the price that will be charged is based on the market demand schedule. The firm's supply curve should measure the quantity the firm is willing to supply at various prices. That information is not represented by either marginal cost or average cost.

Optimal Price and Output in Monopolistically Competitive Markets

In the short run, the profit-maximizing choice is the level of output at which $MR = MC$ and total revenue is the area of the rectangle $P_1 \times Q_1$ shown in Exhibit 14.

The average cost of producing Q_1 units of the product is C_1 , and the total cost is the area of the rectangle $C_1 \times Q_1$. The difference between TR and TC is economic profit. The profit relationship is described as follows:

$$\pi = TR - TC,$$

where π is total profit, TR is total revenue, and TC is total cost.

THE BENEFITS OF IMPERFECT COMPETITION

Is monopolistic competition indeed imperfect—that is, is it a bad thing? At first, one would say that it is an inefficient market structure because prices are higher and the quantity supplied is less than in perfect competition. At the same time, in the real world, we see more markets characterized by monopolistic competition than markets meeting the strict conditions of perfect competition. If monopolistic competition were that inefficient, one wonders, why would it be so common?

A part of the explanation goes back to Schumpeter. Firms try to differentiate their products to meet the needs of customers. Differentiation provides a profit incentive to innovate, experiment with new products and services, and potentially improve the standard of living.

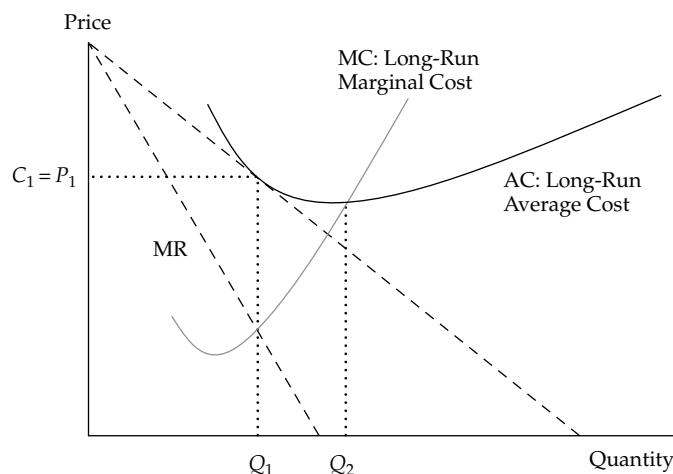
Moreover, because each customer has differing tastes and preferences, slight variations of each good or service are likely to capture the niche of the market that prefers them. An example is the market for candy, where one can find chocolate, licorice, mint, fruit, and many other flavors.

Another reason why monopolistic competition may be good is that people like variety. Traditional economic theories of international trade suggested that countries should buy products from other countries that they cannot produce domestically. Therefore, Norway should buy bananas from a tropical country and sell crude oil in exchange. But this is not the only kind of exchange that happens in reality: For example, Germany imports Honda, Subaru, and Toyota cars from Japan and sells Volkswagen, Porsche, Mercedes, and BMW cars to Japan. In theory, this should not occur because each of the countries produces good cars domestically and does not need to import them. The truth, however, is that consumers in both countries enjoy variety. Some Japanese drivers prefer to be at the steering wheel of a BMW; others like Hondas, and the same happens in Germany. Variety and product differentiation, therefore, are not necessarily bad things.

Long-Run Equilibrium in Monopolistic Competition

Because total cost includes all costs associated with production, including opportunity cost, economic profit is a signal to the market, and that signal will attract more competition. Just as with the perfectly competitive market structure, with relatively low entry costs, more firms will enter the market and lure some customers away from the firm making an economic profit. The loss of customers to new entrant firms will drive down the demand for all firms producing similar products. In the long run for the monopolistically competitive firm, economic profit will fall to zero. Exhibit 15 illustrates the condition of long-run equilibrium for monopolistic competition.

Exhibit 15: Long-Run Equilibrium in Monopolistic Competition



In long-run equilibrium, output is still optimal at the level at which $MR = MC$, which is Q_1 in Exhibit 15. Again, the price consumers are willing to pay for any amount of the product is determined from the demand curve. That price is P_1 for the quantity Q_1 in Exhibit 15, and total revenue is the area of the rectangle $P_1 \times Q_1$. Notice that unlike long-run equilibrium in perfect competition, in the market of monopolistic competition, the equilibrium position is at a higher level of average cost than the level of output that minimizes average cost. Average cost does not reach its minimum until output level Q_2 is achieved. Total cost in this long-run equilibrium position is the area of the rectangle $C_1 \times Q_1$. Economic profit is total revenue minus total cost. In Exhibit 15, economic profit is zero because total revenue equals total cost: $P_1 \times Q_1 = C_1 \times Q_1$.

In the hybrid market of monopolistic competition, zero economic profit in long-run equilibrium resembles perfect competition. However, the long-run level of output, Q_1 , is less than Q_2 , which corresponds to the minimum average cost of production and would be the long-run level of output in a perfectly competitive market. In addition, the economic cost in monopolistic competition includes some cost associated with product differentiation, such as advertising. In perfect competition, no costs are associated with advertising or marketing because all products are homogeneous. Prices are lower, but consumers may have little variety.

QUESTION SET

1. A company doing business in a monopolistically competitive market will *most likely* maximize profits when its output quantity is set such that:
- average cost is minimized.
 - marginal revenue is equal to average cost.
 - marginal revenue is equal to marginal cost.

Solution:

C is correct. The profit maximizing choice is the level of output at which marginal revenue equals marginal cost.

OLIGOPOLY**5**

explain supply and demand relationships under oligopoly, including the optimal price and output for firms as well as pricing strategy

Oligopoly and Pricing Strategies

An oligopoly market structure is characterized by only a few firms doing business in a relevant market. The products must all be similar and generally are substitutes for one another. In some oligopoly markets, the goods or services may be differentiated by marketing and strong brand recognition, as in the markets for breakfast cereals and for bottled or canned beverages. Other examples of oligopoly markets are made up of homogeneous products with little or no attempt at product differentiation, such as petroleum and cement. *The most distinctive characteristic of oligopoly markets is the small number of firms that dominate the market. There are so few firms in the relevant market that their pricing decisions are interdependent.* That is, each firm's pricing decision is based on the expected retaliation by the other firms. Recall the characteristics of oligopoly markets:

- There are a small number of potential sellers.
- The products offered by each seller are close substitutes for the products offered by other firms and may be differentiated by brand or homogeneous and unbranded.
- Entry into the market is difficult, with fairly high costs and significant barriers to competition.
- Firms typically have substantial pricing power.
- Products are often highly differentiated through marketing, features, and other non-price strategies.

Because there are so few firms, each firm can have some degree of pricing power, which can result in substantial profits. Another by-product of the oligopoly market structure is the attractiveness of price collusion. Even without price collusion, a dominant firm may easily become the price maker in the market. Oligopoly markets

without collusion typically have the most sophisticated pricing strategies. Perhaps the most well-known oligopoly market with collusion is the Organization of the Petroleum Exporting Countries (OPEC) cartel, which seeks to control prices in the petroleum market by fostering agreements among oil-producing countries.

Demand Analysis and Pricing Strategies in Oligopoly Markets

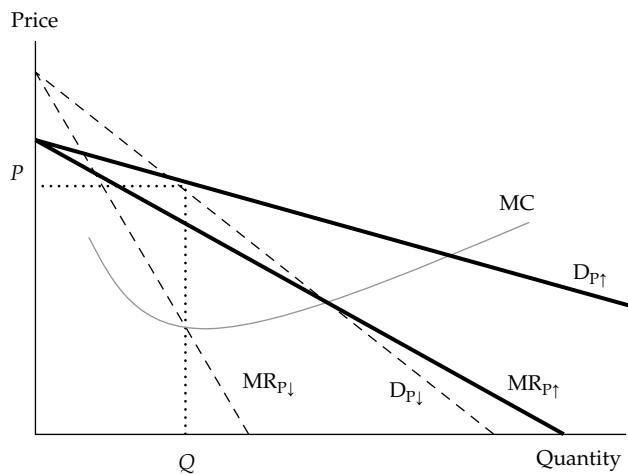
Oligopoly markets' demand curves depend on the degree of pricing interdependence. In a market in which collusion is present, the aggregate market demand curve is divided by the individual production participants. Under non-colluding market conditions, each firm faces an individual demand curve. Furthermore, non-colluding oligopoly market demand characteristics depend on the pricing strategies adopted by the participating firms. The three basic pricing strategies are pricing interdependence, the Cournot assumption, and the Nash equilibrium.

The first pricing strategy assumes pricing interdependence among the firms in the oligopoly. A good example of this situation is any market in which there are "price wars," such as the commercial airline industry. For example, flying out of their hubs in Atlanta, both Delta Air Lines and AirTran Airways jointly serve several cities. AirTran is a low-cost carrier and typically offers lower fares to destinations out of Atlanta. Delta tends to match the lower fares for those cities also served by AirTran when the departure and arrival times are similar to its own. When Delta offers service to the same cities at different time slots, however, Delta's ticket prices are higher.

The most common pricing strategy assumption in these price war markets is that competitors will match a price reduction and ignore a price increase. The logic is that by lowering its price to match a competitor's price reduction, the firm will not experience a reduction in customer demand. Conversely, by not matching the price increase, the firm stands to attract customers away from the firm that raised its prices. The oligopolist's demand relationship must represent the potential increase in market share when rivals' price increases are not matched and no significant change in market share when rivals' price decreases are matched.

Given a prevailing price, the price elasticity of demand will be much greater if the price is increased and less if the price is decreased. The firm's customers are more responsive to price increases because its rivals have lower prices. Alternatively, the firm's customers are less responsive to price decreases because its rivals will match its price change.

This implies that the oligopolistic firm faces two different demand structures: one associated with price increases and another relating to price reductions. Each demand function will have its own marginal revenue structure as well. Consider the demand and marginal revenue functions in Exhibit 16(A). The functions $D_{P\uparrow}$ and $MR_{P\uparrow}$ represent the demand and marginal revenue schedules associated with higher prices, whereas the functions $D_{P\downarrow}$ and $MR_{P\downarrow}$ represent the lower prices' demand and marginal revenue schedules. The two demand schedules intersect at the prevailing price (i.e., the price at which the price increase and price decrease are both equal to zero).

Exhibit 16: Kinked Demand Curve in Oligopoly Market

This oligopolistic pricing strategy results in a kinked demand curve, with the two segments representing the different competitor reactions to price changes. The kink in the demand curve also yields a discontinuous marginal revenue structure, with one part associated with the price increase segment of demand and the other relating to the price decrease segment. Therefore, the firm's overall demand equals the relevant portion of $D_{P\uparrow}$ and the relevant portion of $D_{P\downarrow}$. Exhibit 16(B) represents the firm's new demand and marginal revenue schedules. The firm's demand schedule shown in Exhibit 16(B) is segment $D_{P\uparrow}$ and $D_{P\downarrow}$, where overall demand $D = D_{P\uparrow} + D_{P\downarrow}$.

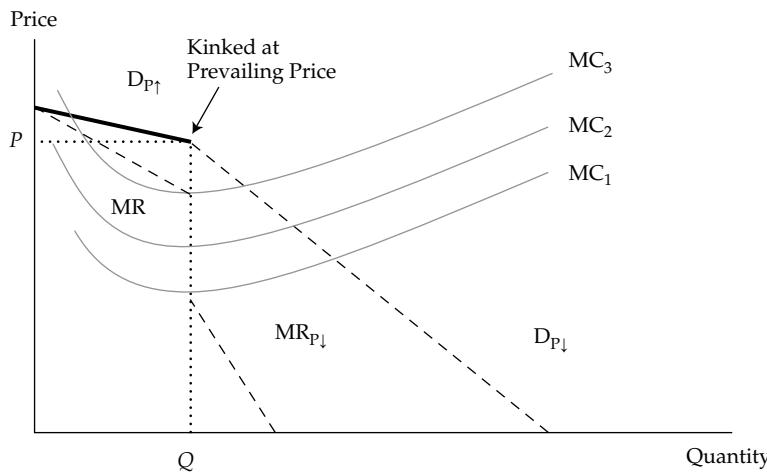
Exhibit 17: Kinked Demand Curve in Oligopoly Market

Exhibit 16(B) shows that a wide variety of cost structures are consistent with the prevailing price. If the firm has relatively low marginal costs, MC_1 , the profit-maximizing pricing rule established earlier, $MR = MC$, still holds for the oligopoly firm. Marginal cost can rise to MC_2 and MC_3 before the firm's profitability is challenged. If the

marginal cost curve MC_2 passes through the gap in marginal revenue, the most profitable price and output combination remains unchanged at the prevailing price and original level of output.

Criticism of the kinked demand curve analysis focuses on its inability to determine what the prevailing price is from the outset. The kinked demand curve analysis helps explain why stable prices have been observed in oligopoly markets and therefore is a useful tool for analyzing such markets. However, because it cannot determine the original prevailing price, it is considered to be an incomplete pricing analysis.

The Cournot Assumption

The second pricing strategy was first developed by French economist Augustin Cournot in 1838. In the **Cournot assumption**, each firm determines its profit-maximizing production level by assuming that the other firms' output will not change. This assumption simplifies pricing strategy because there is no need to guess what the other firm will do to retaliate. It also provides a useful approach to analyzing real-world behavior in oligopoly markets. Take the most basic oligopoly market situation, a two-firm duopoly market. In equilibrium, neither firm has an incentive to change output, given the other firm's production level. Each firm attempts to maximize its own profits under the assumption that the other firm will continue producing the same level of output in the future. The Cournot strategy assumes that this pattern continues until each firm reaches its long-run equilibrium position. In long-run equilibrium, output and price are stable: No change in price or output will increase profits for either firm.

Consider this example of a duopoly market. Assume that the aggregate market demand has been estimated to as follows:

$$Q_D = 450 - P.$$

The supply function is represented by constant marginal cost $MC = 30$.

The Cournot strategy's solution can be found by setting $Q_D = q_1 + q_2$, where q_1 and q_2 represent the output levels of the two firms. Each firm seeks to maximize profit, and each firm believes the other firm will not change output as it changes its own output (i.e., Cournot's assumption). The firm will maximize profit where $MR = MC$. Rearranging the aggregate demand function in terms of price, we get:

$$P = 450 - Q_D = 450 - (q_1 + q_2), \text{ and } MC = 30.$$

Total revenue for each of the two firms is found by multiplying price and quantity:

$$TR_1 = P q_1 = (450 - q_1 - q_2) q_1 = 450 q_1 - q_1^2 - q_1 q_2,$$

and

$$TR_2 = P q_2 = (450 - q_1 - q_2) q_2 = 450 q_2 - q_2 q_1 - q_2^2.$$

Marginal revenue is defined as the change in total revenue, given a change in sales (q_1 or q_2). For the profit-maximizing output, set $MR = MC$, or

$$450 - 2q_1 - q_2 = 30,$$

and

$$450 - q_1 - 2q_2 = 30.$$

Then find the simultaneous equilibrium for the two firms by solving the two equations with two unknowns:

$$450 - 2q_1 - q_2 = 450 - q_1 - 2q_2.$$

Because $q_2 = q_1$ under Cournot's assumption, insert this solution into the demand function and solve as follows:

$$450 - 2q_1 - q_1 = 450 - 3q_1 = 30.$$

Therefore, $q_1 = 140$, $q_2 = 140$, and $Q = 280$.

The price is $P = 450 - 280 = 170$.

In the Cournot strategic pricing solution, the market equilibrium price will be 170, and the aggregate output will be 280 units. This result, known as the Cournot equilibrium, differs from the perfectly competitive market equilibrium because the perfectly competitive price will be lower and the perfectly competitive output will be higher. In general, non-competitive markets have higher prices and lower levels of output in equilibrium when compared with perfect competition. In competition, the equilibrium is reached when price equals marginal cost:

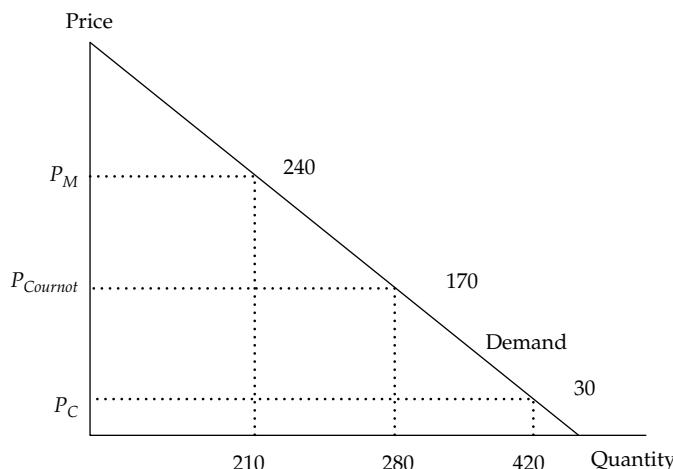
$$P_C = MR_C = MC, \text{ so } 450 - Q = 30,$$

where P_C is the competitive firm's equilibrium price.

$$Q = 420, \text{ and } P_C = 30.$$

Exhibit 17 describes the oligopoly, competitive, and monopoly market equilibrium positions, where P_M is the monopoly optimum price, P_C is the competitive price, and $P_{Cournot}$ is the oligopoly price under the Cournot assumption.

Exhibit 18: Cournot Equilibrium in Duopoly Market



In the later discussion regarding monopoly market structure, equilibrium will be established where $MR = MC$. That solution is also shown in Exhibit 17. The monopoly firm's demand schedule is the aggregate market demand schedule. Therefore, the solution is

$$MR = MC.$$

For the market demand function, total revenue is $P \times Q = 450Q - Q^2$ and $MR = 450 - 2Q$; therefore,

$$450 - 2Q = 30 \quad \text{and} \quad Q = 210.$$

From the aggregate demand function, solve for price:

$$P_M = 450 - 210 = 240.$$

Note that the Cournot solution falls between the competitive equilibrium and the monopoly solution.

As the number of firms increases from two to three, from three to four, and so on, the output and price equilibrium positions move toward the competitive equilibrium solution. Historically, this result has been the theoretical basis for the antitrust policies established in the United States.

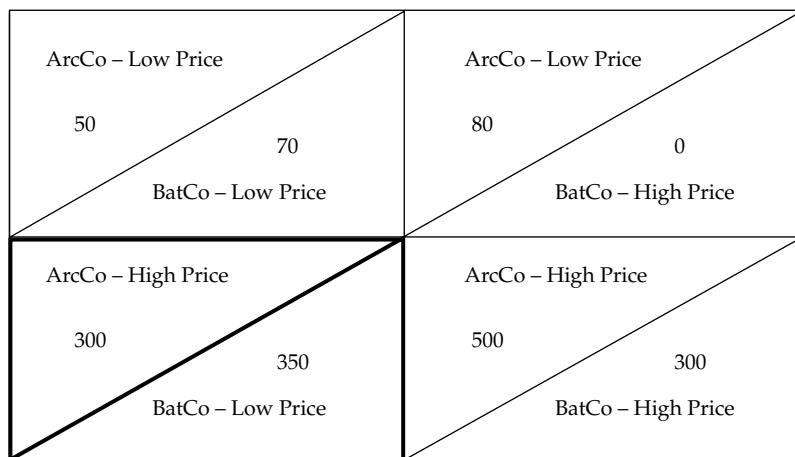
The Nash Equilibrium

The third pricing strategy is attributed to one of the 1994 Nobel Prize winners, John Nash, who first developed the general concepts. In the previous analysis, the concept of market equilibrium occurs when firms are achieving their optimum remuneration under the circumstances they face. In this optimum environment, the firm has no motive to change price or output level. Existing firms are earning a normal return (zero economic profit), leaving no motive for entry to or exit from the market. All of the firms in the market are producing at the output level at which price equals the average cost of production.

In **game theory** (the set of tools that decision makers use to consider responses by rival decision makers), the **Nash equilibrium** is present when two or more participants in a non-cooperative game have no incentive to deviate from their respective equilibrium strategies after they have considered and anticipated their opponent's rational choices or strategies. In the context of oligopoly markets, the Nash equilibrium is an equilibrium defined by the characteristic that none of the oligopolists can increase its profits by unilaterally changing its pricing strategy. The assumption is made that each participating firm does the best it can, given the reactions of its rivals. Each firm anticipates that the other firms will react to any change made by competitors by doing the best they can under the altered circumstances. The firms in the oligopoly market have interdependent actions. Their actions are non-cooperative, with each firm making decisions that maximize its own profits. The firms do not collude in an effort to maximize joint profits. The equilibrium is reached when all firms are doing the best they can, given the actions of their rivals.

Exhibit 18 illustrates the duopoly result from the Nash equilibrium. Assume there are two firms in the market, ArcCo and BatCo. ArcCo and BatCo can charge high prices or low prices for the product. The market outcomes are shown in Exhibit 18.

Exhibit 19: Nash Equilibrium in Duopoly Market



For example, the top left solution indicates that when both ArcCo and BatCo offer the product at low prices, ArcCo earns a profit of 50 and BatCo earns 70. The top right solution shows that if ArcCo offers the product at a low price, BatCo earns zero profits. The solution with the maximum joint profits is the lower right equilibrium, where both firms charge high prices for the product. Joint profits are 800 in this solution.

The Nash equilibrium, however, requires that each firm behaves in its own best interest. BatCo can improve its position by offering the product at low prices when ArcCo is charging high prices. In the lower left solution, BatCo maximizes its profits at 350. Although ArcCo can earn 500 in its best solution, it can do so only if BatCo also agrees to charge high prices. This option is clearly not in BatCo's best interest because it can increase its return from 300 to 350 by charging lower prices.

This scenario brings up the possibility of collusion. If ArcCo agrees to share at least 51 of its 500 when both companies are charging high prices, BatCo should also be willing to charge high prices. In general, such collusion is unlawful in most countries, but it remains a tempting alternative. Clearly, conditions in oligopolistic industries encourage collusion, with a small number of competitors and interdependent pricing behavior. Collusion is motivated by several factors: increased profits, reduced cash flow uncertainty, and improved opportunities to construct barriers to entry.

When collusive agreements are made openly and formally, the firms involved are called a **cartel**. In some cases, collusion is successful; other times, the forces of competition overpower collusive behavior. The following six major factors affect the chances of successful collusion:

1. *The number and size distribution of sellers.* Successful collusion is more likely if the number of firms is small or if one firm is dominant. Collusion becomes more difficult as the number of firms increases or if the few firms have similar market shares. When the firms have similar market shares, the competitive forces tend to overshadow the benefits of collusion.
2. *The similarity of the products.* When the products are homogeneous, collusion is more successful. The more differentiated the products, the less likely it is that collusion will succeed.
3. *Cost structure.* The more similar the firms' cost structures, the more likely it is that collusion will succeed.
4. *Order size and frequency.* Successful collusion is more likely when orders are frequent, received on a regular basis, and relatively small. Frequent small orders, received regularly, diminish the opportunities and rewards for cheating on the collusive agreement.
5. *The strength and severity of retaliation.* Oligopolists will be less likely to break the collusive agreement if the threat of retaliation by the other firms in the market is severe.
6. *The degree of external competition.* The main reason to enter into the formal collusion is to increase overall profitability of the market, and rising profits attract competition. For example, in 2016 the average extraction cost of a barrel of crude oil from Saudi Arabia was approximately \$9, while the average cost from United States shale oil fields was roughly \$23.50. The cost of extracting oil from the Canadian tar sands in 2016 was roughly \$27 per barrel. It is more likely that crude oil producers in the gulf countries will successfully collude because of the similarity in their cost structures (roughly \$9–\$10 per barrel). If OPEC had held crude oil prices down below \$30 per barrel, there would not have been a viable economic argument to develop US shale oil fields through fracking or expand extraction from Canada's tar

sands. OPEC's successful cartel raised crude oil prices to the point at which outside sources became economically possible and, in doing so, increased the competition the cartel faces.

Other possible oligopoly strategies are associated with decision making based on game theory. The Cournot equilibrium and the Nash equilibrium are examples of specific strategic games. A strategic game is any interdependent behavioral choice employed by individuals or groups that share a common goal (e.g., military units, sports teams, or business decision makers). Another prominent decision-making strategy in oligopolistic markets is the first-mover advantage in the **Stackelberg model**, named after the economist who first conceptualized the strategy. The important difference between the Cournot model and the Stackelberg model is that Cournot assumes that in a duopoly market, decision making is simultaneous, whereas Stackelberg assumes that decisions are made sequentially. In the Stackelberg model, the leader firm chooses its output first and then the follower firm chooses after observing the leader's output. It can be shown that the leader firm has a distinct advantage—that is, being a first mover. In the Stackelberg game, the leader can aggressively overproduce to force the follower to scale back its production or even punish or eliminate the weaker opponent. This approach is sometimes referred to as a "top dog" strategy. The leader earns more than in Cournot's simultaneous game, while the follower earns less. Many other strategic games are possible in oligopoly markets. The important conclusion is that the optimal strategy of the firm depends on what its adversary does. The price and marginal revenue the firm receives for its product depend on both its decisions and its adversary's decisions.

Oligopoly Markets: Optimal Price, Output, and Long-Run Equilibrium

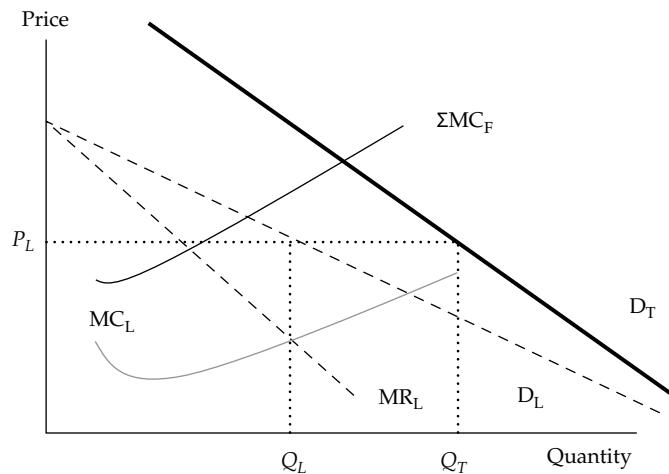
As in monopolistic competition, the oligopolist does not have a well-defined supply function. That is, there is no way to determine the oligopolist's optimal levels of output and price independent of demand conditions and competitor's strategies. However, the oligopolist still has a cost function that determines the optimal level of supply. Therefore, the profit-maximizing rule established earlier is still valid: The level of output that maximizes profit is where $MR = MC$. The price to charge is determined by what price consumers are willing to pay for that quantity of the product. Therefore, the equilibrium price comes from the demand curve, whereas the output level comes from the relationship between marginal revenue and marginal cost.

Consider an oligopoly market in which one of the firms is dominant and thus able to be the price leader. Dominant firms generally have 40 percent or greater market share. When one firm dominates an oligopoly market, it does so because it has greater capacity, has a lower cost structure, was first to market, or has greater customer loyalty than other firms in the market.

Assuming there is no collusion, the dominant firm becomes the price maker, and therefore its actions are similar to monopoly behavior in its segment of the market. The other firms in the market follow the pricing patterns of the dominant firm. Why wouldn't the price followers attempt to gain market share by undercutting the dominant firm's price? The most common explanation is that the dominant firm's supremacy often stems from a lower cost of production. Usually, the price followers would rather charge a price that is even higher than the dominant firm's price choice. If they attempt to undercut the dominant firm, the followers risk a price war with a lower-cost producer that can threaten their survival. Some believe that one explanation for the price leadership position of the dominant firm is simply convenience. Only one firm has to make the pricing decisions, and the others can simply follow its lead.

Exhibit 19 establishes the dominant firm's pricing decision. The dominant firm's demand schedule, D_L , is a substantial share of the total market demand, D_T . The low-cost position of the dominant firm is represented by its marginal cost, MC_L . The sum of the marginal costs of the price followers is established as ΣMC_F and represents a higher cost of production than that of the price leader.

Exhibit 20: Dominant Oligopolist's Price Leadership



An important reason why the total demand curve and the leader demand curve are not parallel is illustrated in Exhibit 19: The leader is the low-cost producer. Therefore, as price decreases, fewer of the smaller suppliers will be able to profitably remain in the market, and several will exit because they do not want to sell below cost. Therefore, the leader will have a larger market share as P decreases, which implies that Q_L increases at a low price, exactly as shown by a steeper D_T in the diagram.

The price leader identifies its profit-maximizing output where $MR_L = MC_L$, at output Q_L . This is the quantity it wants to supply; however, the price it will charge is determined by its segment of the total demand function, D_L . At price P_L , the dominant firm will supply quantity Q_L of total demand, D_T . The price followers will supply the difference to the market, $(Q_T - Q_L) = Q_F$. Therefore, neither the dominant firm nor the follower firms have a single functional relationship that determines the quantity supplied at various prices.

Optimal Price and Output in Oligopoly Markets

As this discussion shows, clearly no single optimum price and output analysis fits all oligopoly market situations. The interdependence among the few firms that make up the oligopoly market provides a complex set of pricing alternatives, depending on the circumstances in each market. In the case of the kinked demand curve, the optimum price is the prevailing price at the kink in the demand function. As noted, however, the kinked demand curve analysis does not provide insight into what established the prevailing price in the first place.

Perhaps the case of the dominant firm, with the other firms following the price leader, is the most obvious. In that case, the optimal price is determined at the output level where $MR = MC$. The profit-maximizing price is then determined by the output position of the segment of the demand function faced by the dominant firm. The price

followers have little incentive to change the leader's price. In the case of the Cournot assumption, each firm assumes that the other firms will not alter their output following the dominant firm's selection of its price and output level.

Therefore, again, the optimum price is determined by the output level where $MR = MC$. In the case of the Nash equilibrium, each firm will react to the circumstances it faces, maximizing its own profit. These adjustments continue until prices and levels of output are stable. Because of the interdependence, the individual firm's price and output level remain uncertain.

Factors Affecting Long-Run Equilibrium in Oligopoly Markets

Long-run economic profits are possible for firms operating in oligopoly markets. History has shown that, however, the market share of the dominant firm declines. Profits attract entry by other firms into the oligopoly market. Over time, the marginal costs of the entrant firms decrease because they adopt more efficient production techniques, the dominant firm's demand and marginal revenue shrink, and the profitability of the dominant firm declines. In the early 1900s, J.P. Morgan, Elbert Gary, Andrew Carnegie, and Charles M. Schwab created the United States Steel Corporation (US Steel). When it was first formed in 1901, US Steel controlled 66 percent of the market. By 1920, US Steel's market share had declined to 46 percent, and by 1925 its market share was 42 percent.

In the long run, optimal pricing strategy must include the reactions of rival firms. History has proven that pricing wars should be avoided because any gains in market share are temporary. Decreasing prices to drive away competitors lowers total revenue to all participants in the oligopoly market. Innovation may be a way—though sometimes an uneconomical one—to maintain market leadership.

OLIGOPOLIES: APPEARANCE VERSUS BEHAVIOR

When is an oligopoly not an oligopoly? There are two extreme cases of this situation. A normal oligopoly has a few firms producing a differentiated good, and this differentiation gives them pricing power.

At one end of the spectrum, we have the oligopoly with a credible threat of entry. In practice, if the oligopolists are producing a good or service that can be easily replicated, has limited economies of scale, and is not protected by brand recognition or patents, they will not be able to charge high prices. The easier it is for a new supplier to enter the market, the lower the margins. In practice, this oligopoly will behave very much like a perfectly competitive market.

At the opposite end of the spectrum, we have the case of the cartel. Here, the oligopolists collude and act as if they were a single firm. In practice, a very effective cartel enacts a cooperative strategy. Instead of going to a Nash equilibrium, the cartel participants go to the more lucrative (for them) cooperative equilibrium.

A cartel may be explicit (i.e., based on a contract) or implicit (based on signals). An example of signals in a duopoly would be that one of the firms reduces its prices and the other does not. Because the firm not cutting prices refuses to start a price war, the firm that cut prices may interpret this signal as a "suggestion" to raise prices to a higher level than before, so that profits may increase for both.

QUESTION SET

1. Oligopolistic pricing strategy *most likely* results in a demand curve that is:

- A. kinked.
- B. vertical.
- C. horizontal.

Solution:

A is correct. The oligopolist faces two different demand structures, one for price increases and another for price decreases. Competitors will lower prices to match a price reduction, but will not match a price increase. The result is a kinked demand curve.

2. Collusion is *less likely* in a market when:

- A. the product is homogeneous.
- B. companies have similar market shares.
- C. the cost structures of companies are similar.

Solution:

B is correct. When companies have similar market shares, competitive forces tend to outweigh the benefits of collusion.

3. In an industry composed of three companies, which are small-scale manufacturers of an easily replicable product unprotected by brand recognition or patents, the *most* representative model of company behavior is:

- A. oligopoly.
- B. perfect competition.
- C. monopolistic competition.

Solution:

B is correct. The credible threat of entry holds down prices and multiple incumbents are offering undifferentiated products.

4. Deep River Manufacturing is one of many companies in an industry that makes a food product. Deep River units are identical up to the point they are labeled. Deep River produces its labeled brand, which sells for \$2.20 per unit, and "house brands" for seven different grocery chains, which sell for \$2.00 per unit. Each grocery chain sells both the Deep River brand and its house brand. The *best* characterization of Deep River's market is:

- A. oligopoly.
- B. perfect competition.
- C. monopolistic competition.

Solution:

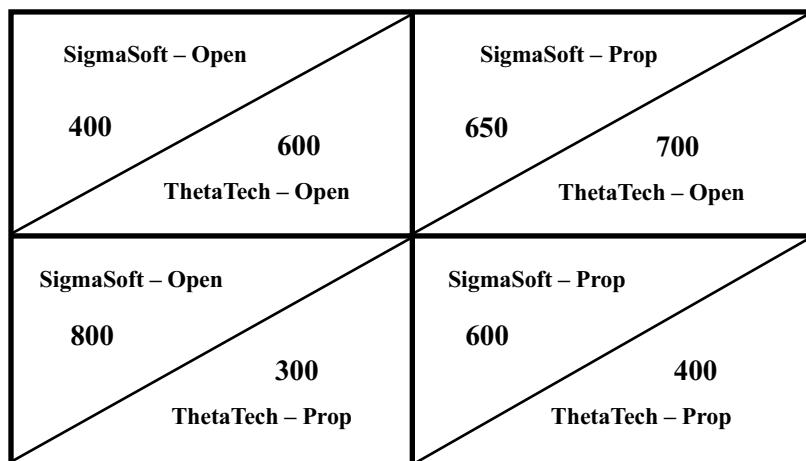
C is correct. There are many competitors in the market, but some product differentiation exists, as the price differential between Deep River's brand and the house brands indicates.

5. SigmaSoft and ThetaTech are the dominant makers of computer system software. The market has two components: a large mass-market component in which demand is price sensitive, and a smaller performance-oriented component in which demand is much less price sensitive. SigmaSoft's product is considered to be technically superior. Each company can choose one of two strategies:

Open architecture (Open): Mass market focus allowing other software vendors to develop products for its platform.

Proprietary (Prop): Allow only its own software applications to run on its platform.

Depending upon the strategy each company selects, their profits would be:



6. The Nash equilibrium for these companies is:

- A. proprietary for SigmaSoft and proprietary for ThetaTech.
- B. open architecture for SigmaSoft and proprietary for ThetaTech.
- C. proprietary for SigmaSoft and open architecture for ThetaTech.

Solution:

C is correct. In the Nash model, each company considers the other's reaction in selecting its strategy. In equilibrium, neither company has an incentive to change its strategy. ThetaTech is better off with open architecture regardless of what SigmaSoft decides. Given this choice, SigmaSoft is better off with a proprietary platform. Neither company will change its decision unilaterally.

6

DETERMINING MARKET STRUCTURE



identify the type of market structure within which a firm operates and describe the use and limitations of concentration measures

Monopoly markets and other situations in which companies have pricing power can be inefficient because producers constrain output to cause an increase in prices. Therefore, less of the good will be consumed, and it will be sold at a higher price, which is generally inefficient for the overall market. As a result, many countries have introduced competition law to regulate the degree of competition in many industries.

Market power in the real world is not always as clear as it is in textbook examples. Governments and regulators often have the difficult task of measuring market power and establishing whether a firm has a dominant position that may resemble a monopoly. A few historical examples of this are as follows:

1. In the 1990s, US regulators prosecuted agricultural corporation Archer Daniels Midland for conspiring with Japanese competitors to fix the price of lysine, an amino acid used as an animal feed additive. The antitrust action resulted in a settlement that involved more than US\$100 million in fines paid by the cartel members.
2. In the 1970s, US antitrust authorities broke up the local telephone monopoly, leaving AT&T the long-distance business (and opening that business to competitors), and required AT&T to divest itself of the local telephone companies it owned. This antitrust decision brought competition, innovation, and lower prices to the US telephone market.
3. European regulators (specifically, the European Commission) have affected the mergers and monopoly positions of European corporations (as in the case of the companies Roche, Rhone-Poulenc, and BASF, which were at the center of a vitamin price-fixing case) as well as non-European companies (such as Intel) that do business in Europe. Moreover, the merger between the US company General Electric and the European company Honeywell was denied by the European Commission on grounds of excessive market concentration.

Quantifying excessive market concentration is difficult. Sometimes, regulators need to measure whether something that has not yet occurred might generate excessive market power. For example, a merger between two companies might allow the combined company to be a monopolist or quasi-monopolist in a certain market.

A financial analyst hearing news about a possible merger should always consider the impact of competition law (sometimes called antitrust law)—that is, whether a proposed merger may be blocked by regulators in the interest of preserving a competitive market.

Econometric Approaches

How should one measure market power? The theoretical answer is to estimate the elasticity of demand and supply in a market. If demand is very elastic, the market must be very close to perfect competition. If demand is rigid (inelastic), companies *may* have market power.

From the econometric point of view, this estimation requires some attention. The problem is that observed price and quantity are the equilibrium values of price and quantity and do not represent the value of either supply or demand. Technically, this is called the problem of endogeneity, in the sense that the equilibrium price and quantity are jointly determined by the interaction of demand and supply. Therefore, to have an appropriate estimation of demand and supply, we need to use a model with two equations: an equation of demanded quantity (as a function of price, income of the buyers, and other variables) and an equation of supplied quantity (as a function of price, production costs, and other variables). The estimated parameters will then allow us to compute elasticity.

Regression analysis is useful in computing elasticity but requires a large number of observations. Therefore, one may use a time-series approach and, for example, look at 20 years of quarterly sales data for a market. The market structure may have changed radically over those 20 years, however, and the estimated elasticity may not apply to the current situation. Moreover, the supply curve may change because of a merger among large competitors, and the estimation based on past data may not be informative regarding the future state of the market postmerger.

An alternative approach is a cross-sectional regression analysis. Instead of looking at total sales and average prices in a market over time (the time-series approach mentioned earlier), we can look at sales from different companies in the market during the same year, or even at single transactions from many buyers and companies. Clearly, this approach requires a substantial data-gathering effort, and therefore, this estimation method can be complicated. Moreover, different specifications of the explanatory variables (e.g., using total GDP rather than median household income or per-capita GDP to represent income) may lead to dramatically different estimates.

Simpler Measures

Trying to avoid these drawbacks, analysts often use simpler measures to estimate elasticity. The simplest measure is the concentration ratio, which is the sum of the market shares of the largest N firms. To compute this ratio, one would, for example, add the sales values of the largest 10 firms and divide this figure by total market sales. This number is always between 0 (perfect competition) and 100 percent (monopoly).

The main advantage of the concentration ratio is that it is simple to compute, as shown previously. The disadvantage is that it does not directly quantify market power. In other words, is a high concentration ratio a clear signal of monopoly power? A company may be the only incumbent in a market, but if the barriers to entry are low, the simple presence of a *potential* entrant may be sufficient to convince the incumbent to behave like a firm in perfect competition. For example, a sugar wholesaler may be the only one in a country, but the knowledge that other large wholesalers in the food industry might easily add imported sugar to their range of products should convince the sugar wholesaler to price its product as if it were in perfect competition.

Another disadvantage of the concentration ratio is that it tends to be unaffected by mergers among the top market incumbents. For example, if the largest and second-largest incumbents merge, the pricing power of the combined entity is likely to be larger than that of the two preexisting companies. But the concentration ratio may not change much.

CALCULATING THE CONCENTRATION RATIO

Suppose there are eight producers of a certain good in a market. The largest producer has 35 percent of the market, the second largest has 25 percent, the third has 20 percent, the fourth has 10 percent, and the remaining four have 2.5 percent each. If we computed the concentration ratio of the top three producers, it would be $35 + 25 + 20 = 80$ percent, while the concentration ratio of the top four producers would be $35 + 25 + 20 + 10 = 90$ percent.

If the two largest companies merged, the new concentration ratio for the top three producers would be 60 (the sum of the market shares of the merged companies) $+ 20 + 10 = 90$ percent, and the concentration ratio for the four top producers would be 92.5 percent. Therefore, this merger affects the concentration ratio very mildly, even though it creates a substantial entity that controls 60 percent of the market.

For example, the effect of consolidation in the US retail gasoline market has resulted in increasing degrees of concentration. In 1992, the top four companies in the US retail gasoline market shared 33 percent of the market. By 2001, the top four companies controlled 78 percent of the market (Exxon Mobil 24 percent, Shell 20 percent, BP/Amoco/Arco 18 percent, and Chevron/Texaco 16 percent).

To avoid the known issues with concentration ratios, economists O. C. Herfindahl and A. O. Hirschman suggested an index in which the market shares of the top N companies are first squared and then added. If one firm controls the whole market (a monopoly), the Herfindahl–Hirschman index (HHI) equals 1. If there are M firms in the industry with equal market shares, then the HHI equals $(1/M)$. This provides a useful gauge for interpreting an HHI. For example, an HHI of 0.20 would be analogous to having the market shared equally by five firms.

The HHI for the top three companies in the example in the box above would be $0.35^2 + 0.25^2 + 0.20^2 = 0.225$ before the merger, whereas after the merger, it would be $0.60^2 + 0.20^2 + 0.10^2 = 0.410$, which is substantially higher than the initial 0.225. The HHI is widely used by competition regulators; however, just like the concentration ratio, the HHI does not take the possibility of entry into account, nor does it consider the elasticity of demand. Therefore, the HHI has limited use for a financial analyst trying to estimate the potential profitability of a company or group of companies.

EXAMPLE 4

The Herfindahl–Hirschman Index

- Suppose a market has 10 suppliers, each of them with 10 percent of the market. What are the concentration ratio and the HHI of the top four firms?
 - Concentration ratio 4 percent and HHI 40
 - Concentration ratio 40 percent and HHI 0.4
 - Concentration ratio 40 percent and HHI 0.04

Solution:

C is correct. The concentration ratio for the top four firms is $10 + 10 + 10 + 10 = 40$ percent, and the HHI is $0.10^2 \times 4 = 0.01 \times 4 = 0.04$.

QUESTION SET



- An analyst gathers the following market share data for an industry:

Company	Sales (in millions of euros)
ABC	300
Brown	250
Coral	200
Delta	150
Erie	100
All others	50

- The industry's four-company concentration ratio is closest to:

- 71%.

- B. 86%.

- C. 95%.

Solution:

B is correct. The top four companies in the industry account for 86 percent of industry sales: $(300 + 250 + 200 + 150)/(300 + 250 + 200 + 150 + 100 + 50) = 900/1050 = 86\%$.

3. An analyst gathered the following market share data for an industry composed of five companies:

Company	Market Share (%)
Zeta	35
Yusef	25
Xenon	20
Waters	10
Vlastos	10

4. The industry's three-firm Herfindahl–Hirschman index is *closest* to:

- A. 0.185.

- B. 0.225.

- C. 0.235.

Solution:

B is correct. The three-firm Herfindahl–Hirschman index is $0.35^2 + 0.25^2 + 0.20^2 = 0.225$.

5. One disadvantage of the Herfindahl–Hirschman index is that the index:

- A. is difficult to compute.

- B. fails to reflect low barriers to entry.

- C. fails to reflect the effect of mergers in the industry.

Solution:

B is correct. The Herfindahl–Hirschman index does not reflect low barriers to entry that may restrict the market power of companies currently in the market.

PRACTICE PROBLEMS

1. The short-term shutdown point of production for a firm operating under perfect competition will *most likely* occur when:
 - A. price is equal to average total cost.
 - B. marginal revenue is equal to marginal cost.
 - C. marginal revenue is equal to average variable costs.
2. Under conditions of perfect competition, a company will break even when market price is equal to the minimum point of the:
 - A. average total cost curve.
 - B. average variable cost curve.
 - C. short-run marginal cost curve.
3. A company will shut down production in the short run if total revenue is less than total:
 - A. fixed costs.
 - B. variable costs.
 - C. opportunity costs.
4. A company has total variable costs of \$4 million and fixed costs of \$3 million. Based on this information, the company will stay in the market in the long term if total revenue is at least:
 - A. \$3.0 million.
 - B. \$4.5 million.
 - C. \$7.0 million.
5. When total revenue is greater than total variable costs but less than total costs, in the short term, a firm will *most likely*:
 - A. exit the market.
 - B. stay in the market.
 - C. shut down production.
6. Under conditions of perfect competition, in the long run, firms will *most likely* earn:
 - A. normal profits.
 - B. positive economic profits.
 - C. negative economic profits.
7. A firm that increases its quantity produced without any change in per-unit cost is

- experiencing:
- A. economies of scale.
 - B. diseconomies of scale.
 - C. constant returns to scale.
8. A company is experiencing economies of scale when:
- A. cost per unit increases as output increases.
 - B. it is operating at a point on the LRAC curve at which the slope is negative.
 - C. it is operating beyond the minimum point on the long-run average total cost curve.
9. Diseconomies of scale *most likely* result from:
- A. specialization in the labor force.
 - B. overlap of business functions and product lines.
 - C. discounted prices on resources when buying in larger quantities.
10. A firm is operating beyond minimum efficient scale in a perfectly competitive industry. To maintain long-term viability, the *most likely* course of action for the firm is to:
- A. operate at the current level of production.
 - B. increase its level of production to gain economies of scale.
 - C. decrease its level of production to the minimum point on the long-run average total cost curve.
11. Companies *most likely* have a well-defined supply function when the market structure is:
- A. oligopoly.
 - B. perfect competition.
 - C. monopolistic competition.
12. Aquarius, Inc. is the dominant company and the price leader in its market. One of the other companies in the market attempts to gain market share by undercutting the price set by Aquarius. The market share of Aquarius will *most likely*:
- A. increase.
 - B. decrease.
 - C. stay the same.
13. Over time, the market share of the dominant company in an oligopolistic market will *most likely*:
- A. increase.
 - B. decrease.

- C. remain the same.

SOLUTIONS

1. C is correct. The firm should shut down production when marginal revenue is less than or equal to average variable cost.
2. A is correct. A company is said to break even if its total revenue is equal to its total cost. Under conditions of perfect competition, a company will break even when market price is equal to the minimum point of the average total cost curve.
3. B is correct. A company will shut down production in the short run when total revenue is below total variable costs.
4. C is correct. A company will stay in the market in the long term if total revenue is equal to or greater than total cost. Because total costs are \$7 million (\$4 million variable costs and \$3 million fixed costs), the company will stay in the market in the long term if total revenue equals at least \$7 million.
5. B is correct. When total revenue is enough to cover variable costs but not total fixed costs in full, the firm can survive in the short run but would not be able to maintain financial solvency in the long run.
6. A is correct. Competition should drive prices down to long-run marginal cost, resulting in only normal profits being earned.
7. C is correct. Output increases in the same proportion as input increases occur at constant returns to scale.
8. B is correct. Economies of scale occur if, as the firm increases output, cost per unit of production falls. Graphically, this definition translates into an LRAC with a negative slope.
9. B is correct. As the firm increases output, diseconomies of scale and higher average total costs can result when business functions and product lines overlap or are duplicated.
10. C is correct. The firm operating at greater than long-run efficient scale is subject to diseconomies of scale. It should plan to decrease its level of production.
11. B is correct. A company in a perfectly competitive market must accept whatever price the market dictates. The marginal cost schedule of a company in a perfectly competitive market determines its supply function.
12. A is correct. As prices decrease, smaller companies will leave the market rather than sell below cost. The market share of Aquarius, the price leader, will increase.
13. B is correct. The dominant company's market share tends to decrease as profits attract entry by other companies.

LEARNING MODULE

2

Understanding Business Cycles

by Gambera Michele, PhD, CFA, Ezrati Milton, and Cao Bolong, PhD, CFA.

Michele Gambera, PhD, CFA, is with UBS Asset Management and the University of Illinois at Urbana-Champaign (USA). Milton Ezrati (USA). Bolong Cao, PhD, CFA, is at Ohio University (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	describe the business cycle and its phases
<input type="checkbox"/>	describe credit cycles
<input type="checkbox"/>	describe how resource use, consumer and business activity, housing sector activity, and external trade sector activity vary over the business cycle and describe their measurement using economic indicators

INTRODUCTION

1

A typical economy's output of goods and services fluctuates around its long-term path. We now turn our attention to those recurring, cyclical fluctuations in economic output. Some of the factors that influence short-term changes in the economy—such as changes in population, technology, and capital—are the same as those that affect long-term sustainable economic growth. But forces that cause shifts in aggregate demand and aggregate supply curves—such as expectations, political developments, natural disasters, and fiscal and monetary policy decisions—also influence economies, particularly in the short run.

We first describe a typical business cycle and its phases. While each cycle is different, analysts and investors need to be familiar with the typical cycle phases and what they mean for the expectations and decisions of businesses and households that influence the performance of sectors and companies. These behaviors also affect financial conditions and risk appetite, thus affecting the setting of expectations and choices of portfolio exposures to different investment sectors or styles.

In the lessons that follow, we describe credit cycles, introduce several theories of business cycles, and explain how different economic schools of thought interpret the business cycle and their recommendations with respect to it. We also discuss

variables that demonstrate predictable relationships with the economy, focusing on those whose movements have value in predicting the future course of the economy. We then proceed to explain measures and features of unemployment and inflation.

LEARNING MODULE OVERVIEW



- Business cycles are recurrent expansions and contractions in economic activity affecting broad segments of the economy.
- Classical cycle refers to fluctuations in the level of economic activity (e.g., measured by GDP in volume terms).
- Growth cycle refers to fluctuations in economic activity around the long-term potential or trend growth level.
- Growth rate cycle refers to fluctuations in the growth rate of economic activity (e.g., GDP growth rate).
- The overall business cycle can be split into four phases: recovery, expansion, slowdown, and contraction.
- In the recovery phase of the business cycle, the economy is going through the “trough” of the cycle, where actual output is at its lowest level relative to potential output.
- In the expansion phase of the business cycle, output increases, and the rate of growth is above average. Actual output rises above potential output, and the economy enters the so-called boom phase.
- In the slowdown phase of the business cycle, output reaches its highest level relative to potential output (i.e., the largest positive output gap). The growth rate begins to slow relative to potential output growth, and the positive output gap begins to narrow.
- In the contraction phase of the business cycle, actual economic output falls below potential economic output.
- Credit cycles describe the changing availability—and pricing—of credit.
- Strong peaks in credit cycles are closely associated with subsequent systemic banking crises.
- Economic indicators are variables that provide information on the state of the overall economy.
 - Leading economic indicators have turning points that usually precede those of the overall economy.
 - Coincident economic indicators have turning points that usually are close to those of the overall economy.
 - Lagging economic indicators have turning points that take place later than those of the overall economy.
- A diffusion index reflects the proportion of a composite index of leading, lagging and coincident indicators that are moving in a pattern consistent with the overall index. Analysts often rely on these diffusion indexes to provide a measure of the breadth of the change in a composite index.

OVERVIEW OF THE BUSINESS CYCLE

2

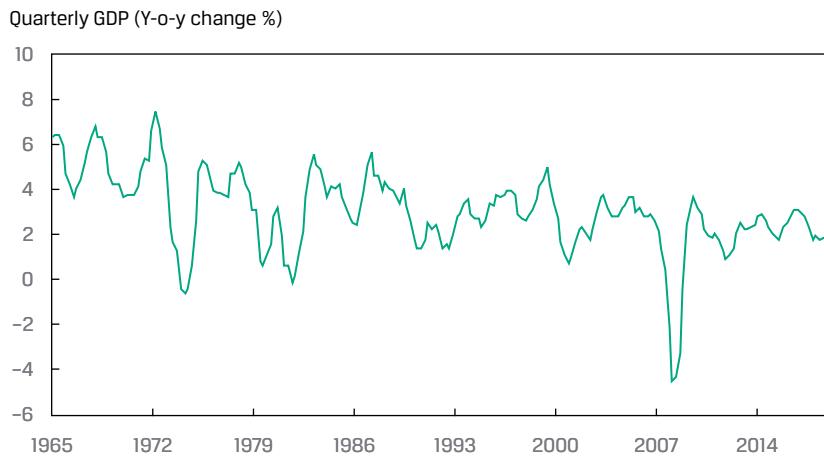
- describe the business cycle and its phases

Business cycles are recurrent expansions and contractions in economic activity affecting broad segments of the economy. In their 1946 book “Measuring Business Cycles”, Burns and Mitchell define the business cycle as follows:

Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of events is recurrent but not periodic; in duration, business cycles vary from more than one year to 10 or 12 years.

This definition is rich with important insight. First, business cycles are typical of economies that rely mainly on business enterprises—therefore, not agrarian societies or centrally planned economies. Second, a cycle has an expected sequence of phases, alternating between expansion and contraction, or upswings and downturns. Third, such phases occur at about the same time throughout the economy. Finally, cycles are recurrent; they happen again and again over time but not in a periodic way; they do not all have the exact same intensity and duration. Exhibit 1 provides an illustration of the pattern of economic growth rate in developed markets.

Exhibit 1: Fluctuations of Growth in OECD Countries over Time



Note: The Organisation for Economic Co-operation and Development (OECD) includes more than 30 large member countries.

Source: OECD.Stat (<https://stats.oecd.org>), year-over-year change in quarterly GDP in OECD countries.

Burns and Mitchell's definition remains helpful. History never repeats itself in quite the same way, but it certainly does offer patterns that can be used when analyzing the present and forecasting the future. Business cycle analysis is a wide-ranging topic with conflicting perspectives held by industry participants.

Phases of the Business Cycle

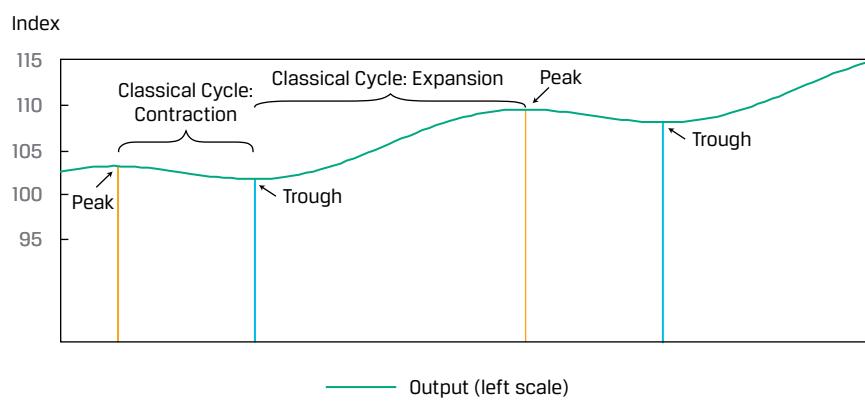
Business cycles are recurring sequences of alternating upswings and downturns. The business cycle can be broken into phases in various ways. The most obvious way is to divide it into two primary segments: the expansion, or the upswing, and the contraction, or the downturn, with two key turning points, or peaks and troughs (see Exhibits 2 and 3). These two periods are fairly easy to identify in retrospect. Subdividing the cycle more finely, however, becomes ambiguous, even in retrospect, because it requires identifying more nuanced changes, such as acceleration or deceleration of growth without a change in its direction. It thus is useful to divide the cycle into several phases distinguished through both economic and financial market characteristics. Our focus is on economic characteristics of the different phases, but we also will highlight their implication for the behavior of different segments of the financial markets.

The timing of these periods will depend on the type of cycle. Before moving on to the description of the four distinct phases to which we will refer in the subsequent sections, we first explain the different cycle concepts that analysts should be aware of given the range of different opinions, interpretations, and descriptions that practitioners use.

Types of Cycles

- **Classical cycle** refers to fluctuations in the level of economic activity (e.g., measured by GDP in volume terms). The contraction phases between peaks and troughs are often short, whereas expansion phases are much longer. Exhibit 2 shows the classical cycle of economic activity. In practice, the classical cycle is not used extensively by academics and practitioners because it does not easily allow the breakdown of movements in GDP between short-term fluctuations and long-run trends. In addition, an absolute decline in activity between peaks and troughs does not occur frequently.

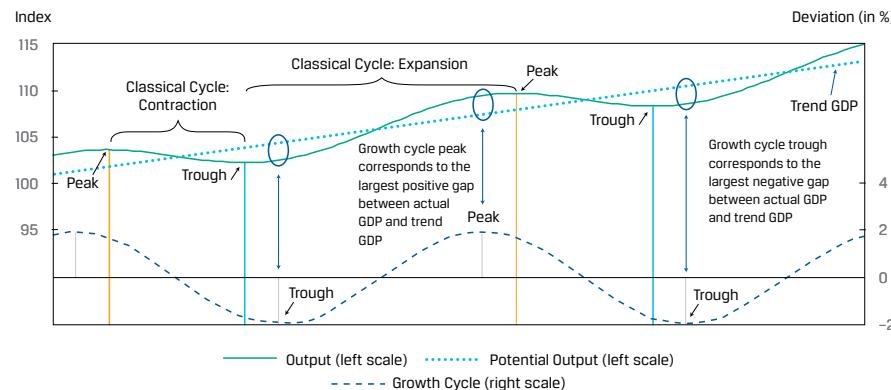
Exhibit 2: Classical Cycle



-
- **Growth cycle** refers to fluctuations in economic activity around the long-term potential or trend growth level. The focus is on how much actual economic activity is below or above trend growth in economic activity. The dashed "wave" in the lower part of Exhibit 3 captures the fluctuation of actual activity from trend growth activity. Exhibit 3 shows "gaps" between actual and trend output. The growth cycle definition comes closest to how mainstream economists think: It dissects overall economic activity into a

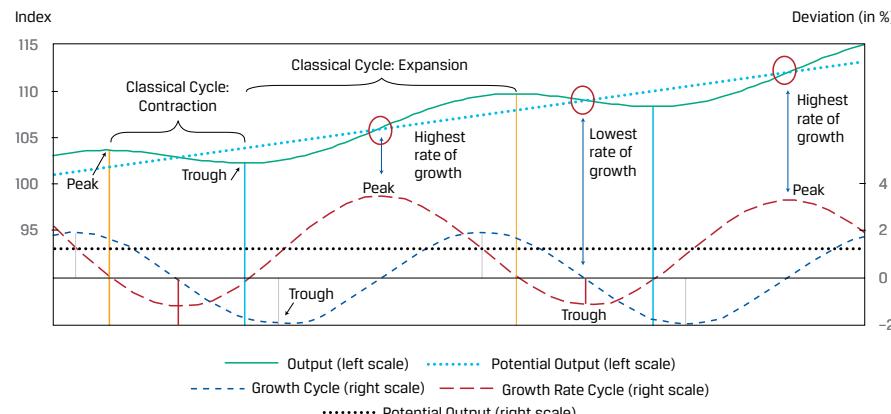
part driven by long-run trends and a part reflecting short-run fluctuations. Compared with the classical view of business cycles, peaks generally are reached earlier and troughs later in time. The time periods below and above trend growth are of similar length.

Exhibit 3: Classical and Growth Cycles



- **Growth rate cycle** refers to fluctuations in the growth rate of economic activity (e.g., GDP growth rate). Peaks and troughs are mostly recognized earlier than when using the other two definitions (see Exhibit 4). One advantage of this approach is that it is not necessary to first estimate a long-run growth path. Nevertheless, economists often refer to economic growth being above or below potential growth rate, reflecting upswings or downturns.

Exhibit 4: Classical, Growth, and Growth Rate Cycles



Notes: The vertical lines indicate troughs and peaks when using either the classical, growth, or growth rate cycle definition of a business cycle. The growth cycle reflects the percentage deviation of output relative to its trend. The growth rates in the growth rate cycle are calculated as annualized month-over-month growth rates.

Practical Issues

In practice, the definitions of a business cycle are used interchangeably, which often causes confusion regarding how one labels the phases and their timing. The classical cycle definition is rarely used. In line with how most economists and practitioners view the cycle, we will generally be using the growth cycle concept in which business cycles can be thought of as fluctuations around potential output (the trend in potential output is shown as the upward sloping dotted line in Exhibit 3).

Four Phases of the Cycle

The overall business cycle can be split into four phases:

Recovery: The economy is going through the “trough” of the cycle, where actual output is at its lowest level relative to potential output. Economic activity, including consumer and business spending, is below potential but is starting to increase, closing the negative output gap.

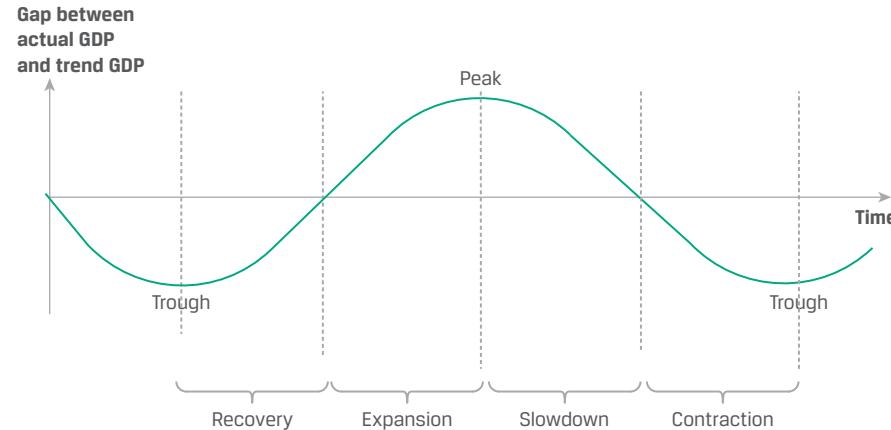
Expansion: The recovery gathers pace, output increases, and the rate of growth is above average. Actual output rises above potential output, and the economy enters the so-called boom phase. Consumers increase spending, and companies increase production, employment, and investment. Prices and interest rates may start increasing. As the expansion continues, the economy may start to experience shortages in factors of production. Overinvestment in productive capacity may lead companies to reduce further investment spending.

Slowdown: Output of the economy reaches its highest level relative to potential output (largest positive output gap). The growth rate begins to slow relative to potential output growth, and the positive output gap begins to narrow. Consumers remain optimistic about the economy, but companies may rely on overtime rather than using new hires to meet demand. Inflation slows at some point, and price levels may decrease.

Contraction: Actual economic output falls below potential economic output. Consumer and business confidence declines. Companies reduce costs by eliminating overtime and reducing employment. The economy may experience declines in absolute economic activity; a recession; or if the fall in activity is particularly large, a depression. If the decline is moderate, this phase tends to be shorter than the expansion phase.

Exhibit 5 provides a summary of the key characteristics of each phase and describes how several important economic variables evolve through the course of a business cycle.

Exhibit 5: Business Cycle Phase Characteristics



Phase	Recovery	Expansion	Slowdown	Contraction
Description	Economy going through a trough. Negative output gap starts to narrow.	Economy enjoying an upswing. Positive output gap opens.	Economy going through a peak. Positive output gap starts to narrow.	Economy weakens and may go into a recession. Negative output gap opens.
Activity levels: consumers and businesses	Activity levels are below potential but start to increase.	Activity measures show above-average growth rates.	Activity measures are above average but decelerating. Moving to below-average rates of growth.	Activity measures are below potential. Growth is lower than normal.
Employment	Layoffs slow. Businesses rely on overtime before moving to hiring. Unemployment remains higher than average.	Businesses move from using overtime and temporary employees to hiring. Unemployment rate stabilizes and starts falling.	Business continue hiring but at a slower pace. Unemployment rate continues to fall but at decreasing rates.	Businesses first cut hours, eliminate overtime, and freeze hiring, followed by outright layoffs. Unemployment rate starts to rise.
Inflation	Inflation remains moderate.	Inflation picks up modestly.	Inflation further accelerates.	Inflation decelerates but with a lag.

Leads and Lags in Business and Consumer Decision Making

The behavior of businesses and households is key to the cycle and frequently incorporates leads and lags relative to what are established as turning points. For example, at the beginning of an expansion phase, companies may want to fully use their existing workforce and wait to hire new employees until they are sure that the economy is indeed growing. However, gradually all economic variables are going to revert toward their normal range of values (e.g., GDP growth will be close to potential, or average, growth).

Market Conditions and Investor Behavior

Many economic variables and sectors of the economy have distinctive cyclical patterns. Knowledge of these patterns can offer insight into likely cyclical directions overall, or it can be particularly applicable to an investment strategy that requires more specific rather than general cyclical insights for investment success.

Recovery Phase

When asset markets expect the end of a recession and the beginning of an expansion phase, risky assets will be repriced upward. When an expansion is expected, the markets will start incorporating higher profit expectations into the prices of corporate bonds and stocks. Typically, equity markets will hit a trough about three to six months before the economy bottoms out and well before the economic indicators turn up. Indeed, as we will see later, the equity market is classified as a leading indicator of the economy.

Expansion Phase

When an economy's expansion is well-established, a later part of an expansion, referred to as a “**boom**,” often follows. The boom is an expansionary phase characterized by economic growth “testing the limits” of the economy, strong confidence, profit, and credit growth. For example, companies may expand so much that they have difficulty finding qualified workers and will compete with other prospective employers by

raising wages and continuing to expand capacity, relying on strong cash flows and borrowing. The government or central bank may step in if it is concerned about the economy overheating.

Slowdown Phase

During the boom, the riskiest assets will often have substantial price increases. Safe assets, such as government bonds that were more highly prized during recession, may have lower prices and thus higher yields. In addition, investors may fear higher inflation, which also contributes to higher nominal yields.

Contraction Phase

During contraction, investors place relatively high values on such safer assets as government securities and shares of companies with steady (or growing) positive cash flows, such as utilities and producers of staple goods. Such preferences reflect the fact that the marginal utility of a safe income stream increases in periods when employment is insecure or declining.

WHEN DO RECESSIONS BEGIN AND END?

A simple and commonly referred to rule is the following: A recession has started when a country or region experiences two consecutive quarters of negative real GDP growth. Real GDP growth is a measure of the “real” or “inflation-adjusted” growth of the overall economy. This rule can be misleading because it does not indicate a recession if real GDP growth is negative in one quarter, slightly positive the next quarter, and again negative in the next quarter. Many analysts question this result. This issue is why some countries have statistical and economic committees that apply the principles stated by Burns and Mitchell to several macroeconomic variables—not just real GDP growth—as a basis to identify business cycle peaks and troughs. The National Bureau of Economic Research (NBER) is an organization that dates business cycles in the United States. Interestingly, the economists and statisticians on NBER’s Business Cycle Dating Committee analyze numerous time series of data focusing on employment, industrial production, and sales. Because the data are available with a delay (preliminary data releases can be revised even several years after the period they refer to), it also means that the Committee’s determinations may take place well after the business cycle turning points have occurred. As we will see later in the reading, practical indicators may help economists understand in advance if a cyclical turning point is about to happen.

1. Which of the following rules is *most likely* to be used to determine whether the economy is in a recession?
 - A. The central bank has run out of foreign reserves.
 - B. Real GDP has two consecutive quarters of negative growth.
 - C. Economic activity experiences a significant decline in two business sectors.

Solution:

B is correct. GDP is a measure of economic activity for the whole economy. Changes in foreign reserves or a limited number of sectors may not have a material impact on the whole economy.

2. Suppose you are interested in forecasting earnings growth for a company active in a country where no official business cycle dating committee (such

as the NBER) exists. The variables you are *most likely* to consider to identify peaks and troughs of a country's business cycle are:

- A. inflation, interest rates, and unemployment.
- B. stock market values and money supply.
- C. unemployment, GDP growth, industrial production, and inflation.

Solution:

C is correct. Unemployment, GDP growth, industrial production, and inflation are measures of economic activity. The discount rate, the monetary base, and stock market indexes are not direct measures of economic activities. The first two are determined by monetary policy, which react to economic activities, whereas the stock market indexes tend to be forward looking or leading indicators of the economy.

QUESTION SET



1. The characteristic business cycle patterns of trough, expansion, peak, and contraction are:

- A. periodic.
- B. recurrent.
- C. of similar duration.

Solution:

B is correct. The stages of the business cycle occur repeatedly over time.

2. During the contraction phase of a business cycle, it is *most likely* that:

- A. inflation indicators are stable.
- B. aggregate economic activity relative to potential output is decreasing.
- C. investor preference for government securities declines.

Solution:

B is correct. The net trend during contraction is negative.

3. An economic peak is *most* closely associated with:

- A. accelerating inflation.
- B. stable unemployment.
- C. declining capital spending.

Solution:

A is correct. Inflation is rising at peaks.

Whereas business cycles mostly use GDP as a measure of economic activity, a body of literature has emerged in which cyclical developments of financial variables are analyzed separately. This is most commonly done in terms of credit and property prices. Credit cycles describe the changing availability—and pricing—of credit. They describe growth in private sector credit (availability and usage of loans), which is essential for business investments and household purchases of real estate. Therefore, they are connected to real economic activity captured by business cycles that describe fluctuations in real GDP.

When the economy is strong or improving, the willingness of lenders to extend credit, and on favorable terms, is high. Conversely, when the economy is weak or weakening, lenders pull back, or “tighten” credit, by making it less available and more expensive. This frequently contributes to the decline of such asset values as real estate, causing further economic weakness and higher defaults. This is because of the importance of credit in the financing of construction and the purchase of property. Credit cycles are a subset of a wider family of so-called financial cycles, a topic that goes beyond the scope of our coverage.

Applications of Credit Cycles

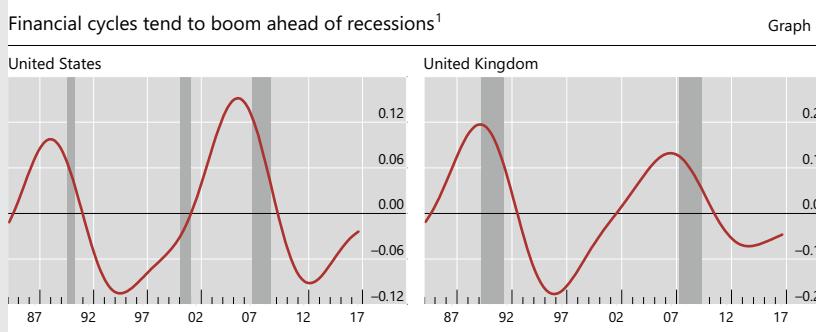
Financial factors were for a long time not prominent on the radar screens of macroeconomists. Monetary and financial phenomena were largely seen as a veil that could be ignored when trying to understand the economy. But loose private sector credit is considered to have contributed to several financial crises, such as the Latam crisis of the 1980s; the Mexican, Brazilian, and Russian crises of the 1990s; the Asian crisis of 1997–1998; and the Global Financial Crisis of 2008–2009. Expansive credit conditions often lead to asset price and real estate bubbles that burst when capital market outflows and drawdowns occur mostly because of weaker fundamentals.

It is recognized that in a world with financial frictions, business cycles can be amplified, with deeper recessions and more extensive expansions because of changes in access to external financing. In line with this belief, it is found that the duration and magnitude of recessions and recoveries are often shaped by linkages between business and credit cycles. In particular, recessions accompanied by financial disruption episodes (notably, house and equity price busts), tend to be longer and deeper. Recoveries combined with rapid growth in credit, risk-taking, and house prices tend to be stronger.

Financial variables tend to co-vary closely with each other and can often help explain the size of an economic expansion or contraction, but they are not always synchronized with the traditional business cycle. Credit cycles tend to be longer, deeper, and sharper than business cycles. Although the length of a business cycle varies from peak to trough, the average length of a credit cycle is mostly found to be longer than that of the business cycle. Exhibit 6 illustrates how credit cycles can be visualized.

VISUALIZING FINANCIAL CYCLES

In an October 2019 working paper titled “Predicting Recessions: Financial Cycle versus Term Spread,” the Bank for International Settlements (BIS) provided a visual presentation of credit cycles, which is reproduced in Exhibit 6. It shows that such cycles tend to boom before recessions.

Exhibit 6: BIS Visualization of Financial Cycles

Notes: Credit cycles are measured by a (composite) proxy calculated from variables that include credit-to-GDP ratio and real house prices. The axis on the right shows the year-on-year change in the proxy.

Source: Bank for International Settlement (BIS) Material (available on the BIS website: www.bis.org).

Consequences for Policy

Investors pay attention to the stage in the credit cycle because (1) it helps them understand developments in the housing and construction markets; (2) it helps them assess the extent of business cycle expansions as well as contractions, particularly the severity of a recession if it coincides with the contraction phase of the credit cycle; and (3) it helps them better anticipate policy makers' actions. Whereas monetary and fiscal policy traditionally concentrate on reducing the volatility of business cycles, macroprudential stabilization policies that aim to dampen financial booms have gained importance. This is further stressed by findings that strong peaks in credit cycles are closely associated with subsequent systemic banking crises.

QUESTION SET



1. A senior portfolio manager at Carnara Asset Management explains her analysis of business cycles to a junior analyst. She makes two statements:

Statement 1 Business cycles measure activity by GDP, whereas credit cycles combine a range of financial variables, such as the amount of and pricing of credit.

Statement 2 Credit cycles and business cycles are unrelated and serve different purposes.

- A. Only Statement 1 is true.
- B. Only Statement 2 is true.
- C. Both statements are true.

Solution:

A is correct. Only Statement 1 is true. Statement 2 is not true because researchers have found linkages between financial and business cycles that help explain the magnitude of business cycle expansions and contractions depending on the state of the credit cycle.

2. With which sector of the economy would analysts most commonly associate credit cycles?

- A. Exports
- B. Construction and purchases of property
- C. Food retail

Solution:

B is correct. Credit cycles are associated with availability of credit, which is important in the financing of construction and the purchase of property.

3. The reason analysts follow developments in the availability of credit is that:

- A. loose private sector credit may contribute to the extent of asset price and real estate bubbles and subsequent crises.
- B. loose credit helps reduce the extent of asset price and real estate bubbles.
- C. credit cycles are of same length and depth as business cycles.

Solution:

A is correct. Studies have shown that loose credit conditions contribute to the extent of asset price and real estate bubbles that tend to be followed by crises.

4

ECONOMIC INDICATORS OVER THE BUSINESS CYCLE



describe how resource use, consumer and business activity, housing sector activity, and external trade sector activity vary over the business cycle and describe their measurement using economic indicators

This lesson provides a broad overview of how the use of resources needed to produce goods and services typically evolves during a business cycle. We start by focusing on circumstances of firms and explore some of the links between fluctuations in inventory, employment, and investment in physical capital with economic fluctuations.

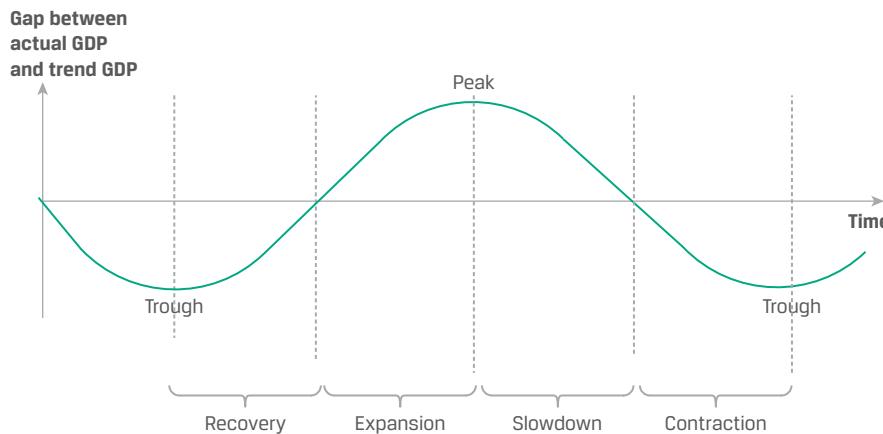
The Workforce and Company Costs

The pattern of hiring and employment is shown in Exhibit 7. When the economy enters contraction, companies reduce costs and eliminate overtime. They may try to retain workers rather than reduce employment only to increase it later. Finding and training new workers is costly, and it may be more cost efficient to keep workers on the payroll even if they are not fully utilized. Companies may also benefit from an implicit bond of loyalty between a company and its workers, boosting productivity in the process. In prolonged contractions, companies will start reducing costs more aggressively—terminating consultants, advertising campaigns, and workers beyond the strict minimum. Capacity utilization will be low, and few companies will invest in new equipment. Companies will try to liquidate their inventories of unsold products. In addition, banks will be reluctant to lend because bankruptcy risks are perceived to be higher, adding to the weakness in the economy.

Decreases in aggregate demand are likely to depress wages or wage growth as well as prices of inputs and capital goods. After a while, all of these input prices will be relatively low. In addition, interest rates may be cut to try to revive the economy.

As prices and interest rates decrease, consumers and companies may begin to purchase more and aggregate demand may begin to rise. This stage is the turning point of the business cycle: Aggregate demand starts to increase and economic activity increases.

Exhibit 7: Business Cycle Phases—Levels of Employment



Phase	Recovery	Expansion	Slowdown	Contraction
Description of activity levels	Economy starts at trough and output below potential. Activity picks up, and gap starts to close.	Economy enjoys an upswing, with activity measures showing above-average growth rates.	Economy at peak. Activity above average but decelerating. The economy may experience shortages of factors of production as demand may exceed supply.	Economy goes into a contraction, (recession, if severe). Activity measures are below potential. Growth is lower than normal.
Employment	Layoffs slow. Businesses rely on overtime before moving to hiring. Unemployment remains higher than average.	Businesses move from using overtime and temporary employees to hiring. Unemployment rate stabilizes and starts falling.	Businesses continue hiring but at a slower pace. Unemployment rate continues to fall but at slowly decreasing rates.	Businesses first cut hours, eliminate overtime, and freeze hiring, followed by outright layoffs. Unemployment rate starts to rise.
Levels of employment lag the cycle				

Fluctuations in Capital Spending

Capital spending—spending on tangible goods, such as property, plant, and equipment—typically fluctuates with the business cycle. Because business profits and cash flows are sensitive to changes in economic activity, capital spending is also highly sensitive to changes in economic activity. In fact, investment is one of the most procyclical and volatile components of GDP. Company spending decisions are driven by business conditions, expectations, and levels of capacity utilization, all of which fluctuate over the cycle. With regard to efficiency, firms will run “lean production” to generate maximum output with the fewest number of workers at the end of contractions. Exhibit 8 provides a description of capital spending over the cycle. Note that new orders statistics include orders that will be delivered over several years. For example, it is common for airlines to order 40 airplanes to be delivered over five years. Where relevant, analysts use “core” orders that exclude defense and aircrafts for a better understanding of the economy’s trend.

Exhibit 8: Capital Spending during the Economic Cycle

Phase of the Cycle	Business conditions and expectations	Capital spending	Examples
Recovery	Excess capacity during trough, low utilization, little need for capacity expansion. Interest rates tend to be low—supporting investment.	Low but increasing as companies start to enjoy better conditions. Capex focus on efficiency rather than capacity. Upturn most pronounced in orders for light producer equipment. Typically, the orders initially reinstated are for equipment with a high rate of obsolescence, such as software, systems, and technological hardware.	Software, systems, and hardware (high rates of obsolescence) orders placed or re-instated first.
Expansion	Companies enjoy favorable conditions. Capacity utilization increases from low levels. Over time, productive capacity may begin to limit ability to respond to demand. Growth in earnings and cash flow gives businesses the financial ability to increase investment spending.	Customer orders and capacity utilization increase. Companies start to focus on capacity expansion. The composition of the economy's capacity may not be optimal for the current structure of demand, necessitating spending on new types of equipment. Orders precede actual shipments, so orders for capital equipment are a widely watched indicator of the future direction of capital spending.	Heavy and complex equipment, warehouses, and factories. A company may need warehouse space in locations different from where existing facilities are located.
Slowdown	Business conditions at peak, with healthy cash flow. Interest rates tend to be higher—aimed at reducing overheating and encouraging investment slowdown.	New orders intended to increase capacity may be an early indicator of the late stage of the expansion phase. Companies continue to place new orders as they operate at or near capacity.	Fiber-optic overinvestment in the late 1990s that peaked with the “technology, media, telecoms bubble.”
Contraction	Companies experience fall in demand, profits, and cash flows.	New orders halted, and some existing orders canceled (no need to expand). Initial cutbacks may be sharp and exaggerate the economy's downturn. As the general cyclical bust matures, cutbacks in spending on heavy equipment further intensify the contraction. Maintenance scaled back.	Technology and light equipment with short lead times get cut first. Cuts in construction and heavy equipment follow.

EXAMPLE 1**Capital Spending**

1. Levels of capacity utilization are one of the factors that determine companies' aggregate need for additional capital expenditure. Which of the following is another factor that affects the capital expenditure decision?
 - A. The rate of unemployment
 - B. The composition of the economy's capacity in relation to how it can satisfy demand
 - C. The ability to reinstate orders canceled during the contraction stage

Solution:

B is correct. The composition of the current productive capacity may not be optimal of the current structure of demand. C is incorrect because the ability to re-instate canceled orders is a matter that is relevant once the decision to increase capital expenditure is made.

2. Orders for technology and light equipment decline before construction projects in a contraction because:
 - A. businesses are uncertain about cyclical directions.
 - B. equipment orders are easier to cancel than large construction projects.
 - C. businesses value light equipment less than structures and heavy machinery.

Solution:

B is correct. Because it usually takes much longer to plan and complete large construction projects than it takes to plan and complete equipment orders, construction projects may be less influenced by business cycles.

Fluctuations in Inventory Levels

The aggregate size of inventories is small relative to the size of the economy, but their accumulation and cutbacks by businesses can occur with substantial speed and frequency. Changing inventories reflect differences between the growth (or decline) in sales and the growth (or decline) in production. A key indicator in this area is the inventory–sales ratio that measures the inventories available for sale to the level of sales. Analysts pay attention to inventories to gauge the position of the economy in the cycle. Exhibit 9 shows how production, sales, and inventories typically move through the phases of a cycle.

Exhibit 9: Inventories throughout the Cycle

Phase of the Cycle	Recovery	Expansion	Slowdown	Contraction
Sales and production	Decline in sales slows. Sales subsequently recover. Production upturn follows but lags behind sales growth. Over time, production approaches normal levels as excess inventories from the downturn are cleared.	Sales increase. Production rises fast to keep up with sales growth and to replenish inventories of finished products. This increases the demand for intermediate products. “Inventory rebuilding or restocking stage.”	Sales slow faster than production; inventories increase. Economic slowdown leads to production cutbacks and order cancellations.	Businesses produce at rates below the sales volumes necessary to dispose of unwanted inventories.
Inventory–sales ratio	Begins to fall as sales recovery outpaces production.	Ratio stable.	Ratio increases. Signals weakening economy.	Ratio begins to fall back to normal.

EXAMPLE 2**Inventory Fluctuation**

1. Although a small part of the overall economy, changes in inventories can influence economic growth measures significantly because they:
- reflect general business sentiment.
 - tend to move forcefully up or down.
 - determine the availability of goods for sale.

Solution:

B is correct. Inventory levels can fluctuate dramatically over the business cycle.

2. Inventories tend to rise when:

- inventory–sales ratios are low.
- inventory–sales ratios are high.
- economic activity begins to rebound.

Solution:

A is correct. When the economy starts to recover, sales of inventories can outpace production, which results in low inventory–sales ratios. Companies then need to accumulate more inventories to restore the ratio to normal level. C is incorrect because in the early stages of a recovery, inventories are likely to fall as sales increase faster than production.

3. Inventories will often fall early in a recovery because:

- businesses need profit.
- sales outstrip production.
- businesses ramp up production because of increased economic activity.

Solution:

B is correct. The companies are slow to increase production in the early recovery phase because they first want to confirm the recession is over. Increasing output also takes time after the downsizing during the recession.

4. In a recession, companies are *most likely* to adjust their stock of physical capital by:
- selling it at fire sale prices.
 - not maintaining equipment.
 - quickly canceling orders for new construction equipment.
- B is correct. Physical capital adjustments to downturns come through aging of equipment plus lack of maintenance.
5. The inventory–sales ratio is *most likely* to be rising:
- as a contraction unfolds.
 - partially into a recovery.
 - near the top of an economic cycle.
- C is correct. Near the top of a cycle, sales begin to slow before production is cut, leading to an increase in inventories relative to sales.

Economic Indicators

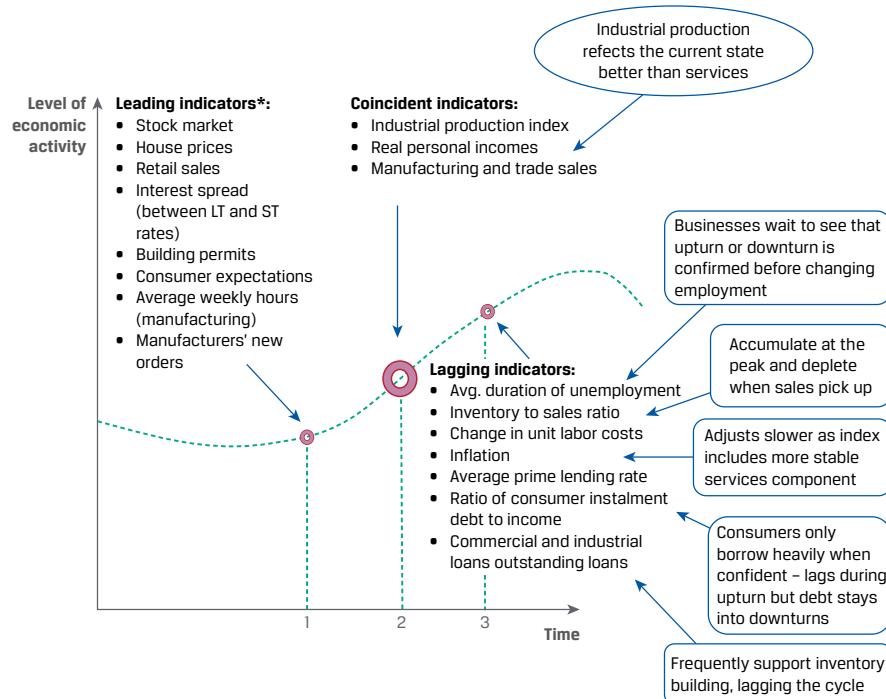
Economic indicators are variables that provide information on the state of the overall economy. They are statistics or data readings that reflect economic circumstances of a country, group of countries, region, or sector. Economic indicators are used by policy makers and analysts to understand and assess the existing condition of the economy and its position in the cycle. They also can be used to help predict or confirm the turning points in the cycle. Such knowledge allows analysts to better predict the financial and market performance of stocks and bonds of issuers operating in different sectors of the economy with different sensitivity to the economic cycle.

Types of Indicators

Economic indicators are often classified according to whether they lag, lead, or coincide with changes in an economy's growth.

- **Leading economic indicators** have turning points that usually precede those of the overall economy. They are believed to have value for predicting the economy's future state, usually near term.
- **Coincident economic indicators** have turning points that are usually close to those of the overall economy. They are believed to have value for identifying the economy's present state.
- **Lagging economic indicators** have turning points that take place later than those of the overall economy. They are believed to have value in identifying the economy's past condition and only change after a trend has been established.

Exhibit 10 provides an illustration of several leading, lagging, and coincident indicators. The leading indicators observed at a point in time labeled as time "1" indicate the direction of the activity (output) at a future point in time, such as time "2." The lagging indicators, released around time "3," refer to and help confirm what the state of the economy was at time "2."

Exhibit 10: Types of Economic Indicators


* Leading indicators will be explored in the subsequent section.

Composite Indicators

An economic indicator either consists of a single variable, like industrial production or the total value of outstanding building permits, or can be a composite of different variables that all tend to move together. The latter are regularly labeled composite indicators. Traditionally, most composite indicators to measure the cyclical state of the economy consist of up to a dozen handpicked variables published by organizations like the OECD or national research institutes. The exact variables combined into these composites vary from one economy to the other. In each case, however, they bring together various economic and financial measures that have displayed a consistently leading, coincident, or lagging relationship to that economy's general cycle.

Leading Indicators

The Conference Board, a US industry research organization, publishes a composite leading indicator known as The Conference Board Leading Economic Index (LEI), which consists of 10 component parts (it uses the classical business cycle as the underlying concept). Exhibit 11 presents the 10 components used in the LEI. In addition to naming the indicators, it offers a general description of why each measure is leading the business cycle.

Exhibit 11: Index of Leading Economic Indicators, United States

Leading indicators	Reason for use
Average weekly hours, manufacturing	Businesses will cut overtime before laying off workers in a downturn and increase it before rehiring in a cyclical upturn. Moves up and down before the general economy.
Average weekly initial claims for unemployment insurance	A very sensitive test of initial layoffs and rehiring.
Manufacturers' new orders for consumer goods and materials	Businesses cannot wait too long to meet demand without ordering. Orders tend to lead at upturns and downturns & captures business sentiment.
<p><i>The Institute of Supply Management (ISM) polls its members to build indexes of manufacturing orders, output, employment, pricing, and comparable gauges for services.</i></p> <p>ISM new order index</p> <p>Survey based</p>	
Manufacturers' new orders for non-defense capital goods excluding aircraft	Reflects the month on month change in new orders for final sales. Decline of new orders can signal weak demand and can lead to recession.
Building permits for new private housing units	Captures business expectations and offers first signal of movement up or down. Important sector.
S&P 500 Index	Stocks tends to anticipate economic turning points; useful early signal.
Leading Credit Index	A vulnerable financial system can amplify the effects of negative shocks, causing widespread recessions.
Interest rate spread between 10-year treasury yields and overnight borrowing rates (federal funds rate)	LT (10 or 30 year) bond yields express market expectations about the direction of short-term interest rates. As rates ultimately follow the economic cycle up and down, a wider spread, by anticipating short rate increases, also anticipates an economic upswing and vice versa.
Average consumer expectations for business conditions	Optimism tends to increase spending. Provides early insight into the direction ahead for the whole economy.

Using Economic Indicators

Exhibit 12 shows a simplified process that an analyst could use to identify business cycle phases. The conclusions then can be used to make investment decisions—for example, to decide in what sectors companies are likely to see improving or deteriorating

cash flows, which could affect the investment performance of the equity and debt securities issued by the companies. Note that the order of the steps does not have to follow this particular sequence.

Exhibit 12: Use of Statistics to Identify Business Cycle Phase**Step 1**

- Data release: Analyst notes an increase in the reported level of consumer instalment debt to income.
- Analysis: The above indicator normally lags cyclical upturns.
- Possible conclusion: Initial evidence that an upturn has been underway.

Step 2

- Data release: Industrial Production Index and non-farm payrolls (employees on non-agricultural payrolls) are moving higher.
- Analysis: These coincident indicators suggest activity is picking up.
- Possible conclusion: Further evidence that expansion is underway.

Step 3

- Observation: Equity market index has been trending higher. Equity index is a leading indicator. Analyst checks the aggregate LEI Index.
- Analysis: If the aggregate LEI is moving higher too, evidence suggests that recovery is underway. Confirmation that output is moving higher.

Or

- If aggregate LEI is not moving higher, analyst cannot draw conclusions about recovery.

Other Composite Leading Indicators

For about 30 countries and several aggregates, such as the EU and G-7, the OECD calculates OECD Composite Leading Indicator (CLI), which gauges the state of the business cycle in the economy using the growth cycle concept. One of the interesting features of OECD CLI is that the underlying methodology is consistent across several countries. Therefore, it can be compared more easily to see how each region is faring. Exhibit 13 shows the eight components of CLI used by the OECD. As is usually the case with leading indicators, some data are based on surveys whereas others are based on reported market or economic data.

Exhibit 13: OECD Euro Area CLI Components**OECD CLI**

*Composite indicator
of economic tendency
survey results*

Economic sentiment index

Residential building permits

Capital goods orders

Euro Stoxx Equity Index

M2 money supply

An interest rate spread

Survey - based

Eurozone/Euro area Manufacturing Purchasing Managers Index (PMI)

Survey - based

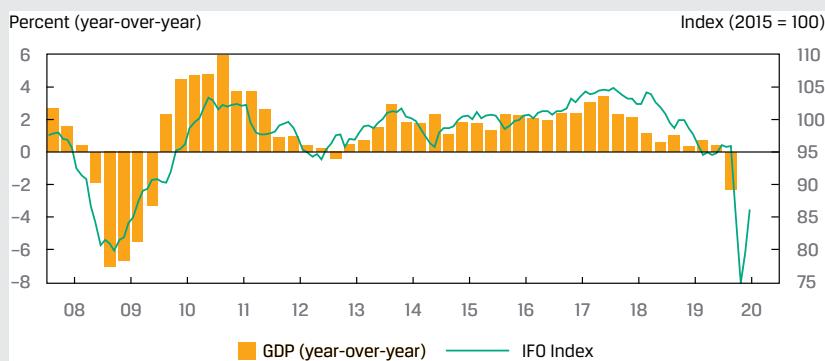
Eurozone/Euro area Service Sector Future Business Activity Expectations Index

Composite leading indicator

The parallels that can be drawn between many of these components and those used in the United States are clear, but the Euro area includes a services component in its business activity measures that the United States lacks. Additionally, the Euro area forgoes many of the overtime and employment gauges that the United States includes. The OECD CLI for Japan is again similar, but it does include labor market indicators (unlike the Euro area) and it adds a measure of business failures not included in the other two.

GERMANY: THE IFO SURVEY

The German ifo survey is a widely used index capturing business climate in Germany and is published monthly. Exhibit 14 shows how the index moved ahead of quarterly-reported year-over-year changes in German GDP. It also shows an uptick in 2018 despite the GDP growth downturn. These indicators are useful, but they are not foolproof.

Exhibit 14: Business Climate Survey and GDP

Sources: The ifo Institute and the Federal Statistical Office of Germany.

Surveys

The composite indicators for region- or country-specific business cycles often make use of economic tendency surveys. These monthly or quarterly surveys carried out by central banks, research institutes, statistical offices, and trade associations are conducted among either businesses, consumers, or experts. The mostly qualitative questions posed are often on the state of their finances, level of activity, and confidence in the future.

CIRET (a forum for leading economists and institutions that conducts and analyzes business and consumer survey data), the United Nations, the OECD, and the European Commission operate at different regional levels to harmonize and exchange knowledge regarding these business tendency and consumer surveys. The Directorate-General for Economic and Financial Affairs (DG ECFIN) publishes fully harmonized results for different sectors of the EU member and applicant economies. The Bank of Japan carries out similar surveys and releases the findings in what is called the "Tankan Report." These diverse sources multiply within and across economies. Over the past decade, so-called purchasing managers indexes along similar lines, albeit with often much smaller survey samples, have been introduced in a wide range of countries and regions, including Europe and China.

These economic tendency surveys are often aggregated into or are a part of composite indicators. IHS Markit publishes a global Purchasing Managers' Index (PMI) indicator using results of its national business surveys (see Exhibit 13). A global consumer confidence indicator is published by The Conference Board. Some other institutions collect national survey data and calculate supranational results. For example, the OECD publishes a business confidence indicator and a consumer confidence indicator for the OECD aggregate in its report "Main Economic Indicators." In addition, the European Commission calculates various survey-based indicators for the EU, reflecting the situation in Europe. Based on economic tendency surveys and using the growth rate cycle concept, the Swiss Economic Institute (*Konjunkturforschungsstelle* or KOF) and *Fundação Getúlio Vargas* (FGV), based in Rio de Janeiro, Brazil, publish a coincident and a leading indicator for the world economy called the Global Economic Barometers. These barometers incorporate hundreds of survey variables from around the world.

The Use of Big Data in Economic Indicators

The vast increase in this information and academic developments regarding the use of big(ger) data have in recent years increased the number of variables that go into these composite indicators. For instance, using a statistical technique called principal components analysis, the Federal Reserve Bank of Chicago computes the Chicago Fed National Activity Index (CFNAI) using 85 monthly macroeconomic series. These series cover industrial production, personal income, capital utilization, employment by sectors, housing starts, retail sales, and so on. Principal components analysis extracts the underlying trend that is common to most of these variables, thus distilling the essence of the US business cycle. Similarly, the Bank of Italy in conjunction with the Centre for Economic Policy Research (CEPR) produces the EuroCOIN statistic, which is also based on principal component analysis. More than 100 macroeconomic series are included in EuroCOIN. EuroCOIN also includes data derived from surveys, interest rates, and other financial variables.

Nowcasting

Policy makers and market practitioners use real-time monitoring of economic and financial variables to continuously assess current conditions. To overcome the publication delays (e.g., GDP numbers are published with a substantial delay) and forecast

the “present,” they make use of a large variety of data—such as financial market transactions, data from the usage of the large amounts of timely internet searches, and electronic payment data—to provide estimates for key low-frequency (monthly or quarterly) economic indicators. This process produces a nowcast, which is an estimate of the current state, and we refer to the process of producing such an estimate as nowcasting. It can be applied to various macroeconomic variables of interest, such as GDP growth, inflation, or unemployment.

GDPNow

Nowcasts are produced by a number of entities in investment banking and asset management for their internal or client use, but they are also published by institutions, such as the Atlanta Fed (the Federal Reserve Bank of Atlanta, one of the 12 Federal Reserve Banks in the United States). According to the Atlanta Fed, “GDPNow” is “best viewed as a running estimate of real GDP growth based on available data for the current measured quarter.” The objective is to forecast GDP for the current quarter (which will not be released until after quarter-end) in real time based on data as they are released throughout the quarter. To do this, the Atlanta Fed attempts to use the same methodology and data as will be used by the US Bureau of Economic Analysis (BEA) to estimate GDP, replacing data that have not yet been released with forecasts based on the data already observed. As the quarter progresses, more of the actual data will have been observed. GDPNow should, at least on average, converge to what will be released by the BEA as their “advance” estimate of quarterly GDP about four weeks after quarter-end.

DIFFUSION INDEX OF ECONOMIC INDICATORS

In the United States, The Conference Board also compiles a monthly diffusion index of the leading, lagging, and coincident indicators. The **diffusion index** reflects the proportion of the index’s components that are moving in a pattern consistent with the overall index. Analysts often rely on these diffusion indexes to provide a measure of the breadth of the change in a composite index.

For example, The Conference Board tracks the growth of each of the 10 constituents of its LEI, assigning a value of 1.0 to each indicator that rises by more than 0.05 percent during the monthly measurement period, a value of 0.5 for each component indicator that changes by less than 0.05 percent, and a value of 0 for each component indicator that falls by more than 0.05 percent. These assigned values, which of course differ in other indexes in other countries, are then summed and divided by 10 (the number of components). To make the overall measure resemble the more familiar indexes, the Board multiplies the result by 100.

A simple numerical example will help explain. Say, for ease of exposition, the indicator has only four component parts: stock prices, money growth, orders, and consumer confidence. In one month, stock prices rise 2.0 percent, money growth rises 1.0 percent, orders are flat, and consumer confidence falls by 0.6 percent. Using The Conference Board’s assigned values, these would contribute respectively: $1.0 + 1.0 + 0.5 + 0$ to create a numerator of 2.5. When divided by four (the number of components) and multiplied by 100, it generates an indicator of 62.5 for that month.

Assume that the following month stock prices fall 0.8 percent, money grows by 0.5 percent, orders pick up 0.5 percent, and consumer confidence grows 3.5 percent. Applying the appropriate values, the components would add to $0 + 1.0 + 1.0 + 1.0 = 3.0$. Divided by the number of components and multiplied by 100, it yields an index value of 75. The 20.0 percent increase in the index value

means more components of the composite index are rising. Given this result, an analyst can be more confident that the higher composite index value actually represents broader movements in the economy. In general, a diffusion index does not reflect outliers in any component (like a straight arithmetic mean would do) but instead tries to capture the overall change common to all components.

QUESTION SET



1. Leading, lagging, and coincident indicators are:

- A. the same worldwide.
- B. based on historical cyclical observations.
- C. based on Keynesian or Monetarist theory.

Solution:

B is correct. The recognition of economic indicators is based on empirical observations for an economy.

2. A diffusion index:

- A. measures growth.
- B. reflects the consensus change in economic indicators.
- C. is roughly analogous to the indexes used to measure industrial production.

Solution:

B is correct. The diffusion indexes are constructed to reflect the common trends embedded in the movements of all the indicators included in such an index.

3. In the morning business news, a financial analyst, Kevin Durbin, learned that average hourly earnings had increased last month. The most appropriate action for Durbin is to:

- A. call his clients to inform them of a good trading opportunity today.
- B. examine other leading indicators to see any confirmation of a possible turning point for the economy.
- C. use the news in his research report as a confirmation for his belief that the economy has recovered from a recession.

Solution:

B is correct. Financial analysts need to synthesize the information from various indicators in order to gather a reliable reading of the economic trends.

4. The following table shows the trends in various economic indicators in the two most recent quarters:

Economic Indicator	Trend
Interest rate spread between long-term government bonds and overnight borrowing rate	Narrowing
New orders for capital goods	Declining
Residential building permits	Declining
Employees on non-agricultural payrolls	Turned from rising to falling

Economic Indicator	Trend
Manufacturing and trade sales	Stable
Average duration of unemployment	Small decline
Change in unit labor costs	Rising

Given the information, this economy is *most likely* experiencing a:

- A. continuing recession.
- B. peak in the business cycle.
- C. strong recovery out of a trough.

Solution:

B is correct. The first three indicators are leading indicators, and all of them are indicating an impending recession, which means the economy has reached the peak in this cycle. Non-agricultural payrolls and manufacturing and trade sales are coincident indicators. The trends in these two variables further indicate that the economy may begin to decline. The trends in the last two indicators—both lagging indicators—indicate that the economy may either continue to grow or it may be close to a peak. Aggregating the signals given by all three groups of economic indicators, it appears the economy may be near the peak of a business cycle.

PRACTICE PROBLEMS

1. Based on typical labor utilization patterns across the business cycle, productivity (output per hours worked) is *most likely* to be highest:
 - A. at the peak of a boom.
 - B. into a maturing expansion.
 - C. at the bottom of a recession.
2. As the expansion phase of the business cycle advances from early stage to late stage, businesses *most likely* experience a decrease in:
 - A. labor costs.
 - B. capital investment.
 - C. availability of qualified workers.
3. An analyst writes in an economic report that the current phase of the business cycle is characterized by accelerating inflationary pressures and borrowing by companies. The analyst is *most likely* referring to the:
 - A. peak of the business cycle.
 - B. contraction phase of the business cycle.
 - C. early expansion phase of the business cycle.
4. The indicator indexes created by various organizations or research agencies:
 - A. include only leading indicators to compute their value.
 - B. are highly reliable signals on the phase of business cycles.
 - C. evolve over time in terms of composition and computation formula.
5. Which one of the following trends in various economic indicators is *most* consistent with a recovery from a recession?
 - A. A declining inventory–sales ratio and stable industrial production index
 - B. A rising broad stock market index and unit labor costs turning from increasing to decreasing
 - C. A decrease in average weekly initial claims for unemployment insurance and an increase in aggregate real personal income
6. Which of the following statements gives the *best* description of nowcasting?
 - A. This method is used to forecast future trends in economic variables based on their past and current values.
 - B. This method is used for real-time monitoring of economic and financial variables to continuously assess current conditions and provide an estimate of the current state.

- C. This method is used to study past relationships between variables to determine which ones have explained the path of a particular variable of interest.
7. Which of the following statements is the *best* description of the characteristics of economic indicators?
- Leading indicators are important because they track the entire economy.
 - Lagging indicators, in measuring past conditions, do not require revisions.
 - A combination of leading and coincident indicators can offer effective forecasts.
8. When the spread between 10-year US Treasury yields and the short-term federal funds rate narrows and at the same time the prime rate stays unchanged, this mix of indicators *most likely* forecasts future economic:
- growth.
 - decline.
 - stability.
9. Current economic statistics indicating little change in services inflation, rising residential building permits, and increasing average duration of unemployment are *best* interpreted as:
- conflicting evidence about the direction of the economy.
 - evidence that a cyclical upturn is expected to occur in the future.
 - evidence that a cyclical downturn is expected to occur in the future.
10. If relative to prior values of their respective indicators, the inventory–sales ratio has risen, unit labor cost is stable, and real personal income has decreased, it is *most likely* that a peak in the business cycle:
- has occurred.
 - is just about to occur.
 - will occur sometime in the future.
11. When aggregate real personal income, industrial output, and the S&P 500 Index all increase in a given period, it is *most accurate* to conclude that a cyclical upturn is:
- occurring.
 - about to end.
 - about to begin.
12. Which of the following is *most likely* to increase after an increase in aggregate real personal income?
- Equity prices
 - Building permits for new private housing units

- C. The ratio of consumer installment debt to income
13. Which of the following indicators is *most* appropriate in predicting a turning point in the economy?
- A. The Industrial Production Index
 - B. The average bank prime lending rate
 - C. Average weekly hours, manufacturing
14. The unemployment rate is considered a lagging indicator because:
- A. new job types must be defined to count their workers.
 - B. multiworker households change jobs at a slower pace.
 - C. businesses are slow to hire and fire due to related costs.
15. During an economic recovery, a lagging unemployment rate is *most likely* attributable to:
- A. businesses quickly rehiring workers.
 - B. new job seekers entering the labor force.
 - C. underemployed workers transitioning to higher-paying jobs.

SOLUTIONS

1. C is correct. At the end of a recession, firms will run “lean production” to generate maximum output with the fewest number of workers.
2. C is correct. When an economy’s expansion is well established, businesses often have difficulty finding qualified workers.
3. A is correct. Accelerating inflation and rapidly expanding capital expenditures typically characterize the peak of the business cycle. During such times, many businesses finance their capital expenditures with debt to expand their production capacity.
4. C is correct. The indicator indexes are constantly updated for their composition and methodology based on the accumulation of empirical knowledge, and they can certainly include more than just leading indicators.
5. C is correct. The improving leading indicator, average weekly initial claims for unemployment insurance, and the improving coincident indicator, aggregate real personal income, are most consistent with an economic recovery. Even though a declining inventory-to-sales ratio, a lagging indicator, is consistent with an early recovery, the coincident indicator, the stable industrial production index, does not support that conclusion. Although a rising stock market index can signal economic expansion, the lagging indicator, the unit labor costs, has peaked, which is more consistent with a recession.
6. B is correct. Nowcasting involves the use of techniques to estimate the present state. A is incorrect because nowcasting aims to estimate the present, not forecast the future. C is incorrect because the focus of nowcasting is to estimate the current, present value of a variable, such as GDP.
7. C is correct. Although no single indicator is definitive, a mix of them—which can be affected by various economic determinants—can offer the strongest signal of performance.
8. B is correct. The narrowing spread of this leading indicator foretells a drop in short-term rates and a fall in economic activity. The prime rate is a lagging indicator and typically moves after the economy turns.
9. B is correct. Rising building permits—a leading indicator—indicate that an upturn is expected to occur or continue. Increasing average duration of unemployment—a lagging indicator—indicates that a downturn has occurred, whereas the lack of any change in services inflation—also a lagging indicator—is neither negative nor positive for the direction of the economy. Taken together, these statistics indicate that a cyclical upturn may be expected to occur.
10. A is correct. Both inventory–sales and unit labor costs are lagging indicators that decline somewhat after a peak. Real personal income is a coincident indicator that, by its decline, shows a slowdown in business activity.
11. A is correct. Aggregate real personal income and industrial output are coincident indicators, whereas the S&P 500 is a leading indicator. An increase in aggregate personal income and industrial output signals that an expansion is occurring, whereas an increase in the S&P 500 signals that an expansion will occur or is expected to continue. Taken together, these statistics indicate that a cyclical upturn is occurring.

12. C is correct. Aggregate real personal income is a coincident indicator of the business cycle, and the ratio of consumer installment debt to income is a lagging indicator. Increases in the ratio of consumer installment debt follow increases in average aggregate income during the typical business cycle.
13. C is correct. Leading economic indicators have turning points that usually precede those of the overall economy. Average weekly hours, manufacturing is a leading economic indicator. The Industrial Production Index is a coincident economic indicator, and the average bank prime lending rate is a lagging economic indicator.
14. C is correct. This effect makes unemployment rise more slowly as recessions start and fall more slowly as recoveries begin.
15. B is correct. In an economic recovery, new job seekers return to the labor force, and because they seldom find work immediately, their return may initially raise the unemployment rate.

LEARNING MODULE

3

Fiscal Policy

by Andrew Clare, PhD, and Stephen Thomas, PhD.

Andrew Clare, PhD, and Stephen Thomas, PhD, are at Cass Business School (UK).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	compare monetary and fiscal policy
<input type="checkbox"/>	describe roles and objectives of fiscal policy as well as arguments as to whether the size of a national debt relative to GDP matters
<input type="checkbox"/>	describe tools of fiscal policy, including their advantages and disadvantages
<input type="checkbox"/>	explain the implementation of fiscal policy and difficulties of implementation as well as whether a fiscal policy is expansionary or contractionary

INTRODUCTION

1

Fiscal policy refers to the government's decisions about taxation and spending, whereas **monetary policy** refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy. Fiscal policy involves the use of government spending and changing tax revenue to affect certain aspects of the economy, such as the overall level of aggregate demand. Government deficits are the difference between government revenues and expenditures over a period of calendar time. The fiscal tools available to a government include transfer payments, current government spending, capital expenditures, and taxes. Economists often examine the **structural budget deficit** as an indicator of a government's fiscal stance.

LEARNING MODULE OVERVIEW



- Fiscal policy refers to the government's decisions about taxation and spending.
- Monetary policy refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy.

- The primary goal of both monetary and fiscal policy is the creation of an economic environment in which growth is stable and positive, and inflation is stable and low.
- Fiscal policy involves the use of government spending and changing tax revenue to affect certain aspects of the economy, including the level of economic activity in an economy, the distribution of income and wealth among different segments of the population, and the allocation of resources between different sectors and economic agents.
- The **budget surplus/deficit** is the difference between government revenue and expenditure for a fixed period of time, such as a fiscal or calendar year.
- There are several strong arguments both for and against being concerned about national debt relative to GDP.
- The fiscal tools available to a government include transfer payments, current government spending, capital expenditures, direct taxes, and indirect taxes.
- Taxes can be justified both in terms of raising revenues to finance expenditures and in terms of income and wealth redistribution policies.
- Fiscal policy tools seek to achieve or maintain an economy on a path of positive, stable growth with low inflation.
- Economists assess the structural (or cyclically adjusted) budget deficit as an indicator of the government's fiscal stance, which is defined as the deficit that would exist if the economy was at full employment.
- Actual government deficits may not be a good measure of fiscal stance because of the distinction between real and nominal interest rates and the role of inflation adjustment when applied to budget deficits.
- Fiscal policy cannot stabilize aggregate demand completely because the difficulties in executing fiscal policy cannot be completely overcome.

2

INTRODUCTION TO MONETARY AND FISCAL POLICY



compare monetary and fiscal policy

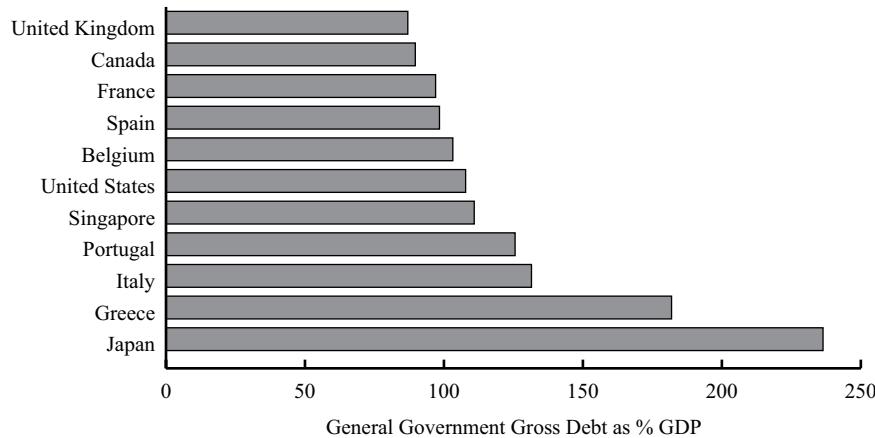
The economic decisions of households can have a significant impact on an economy. For example, a decision on the part of households to consume more and to save less can lead to an increase in employment, investment, and ultimately profits. Equally, the investment decisions made by corporations can have an important impact on the real economy and on corporate profits. But individual corporations can rarely affect large economies on their own; the decisions of a single household concerning consumption will have a negligible impact on the wider economy.

By contrast, the decisions made by governments can have an enormous impact on even the largest and most developed of economies for two main reasons. First, the public sectors of most developed economies normally employ a significant proportion of the population, and they usually are responsible for a significant proportion

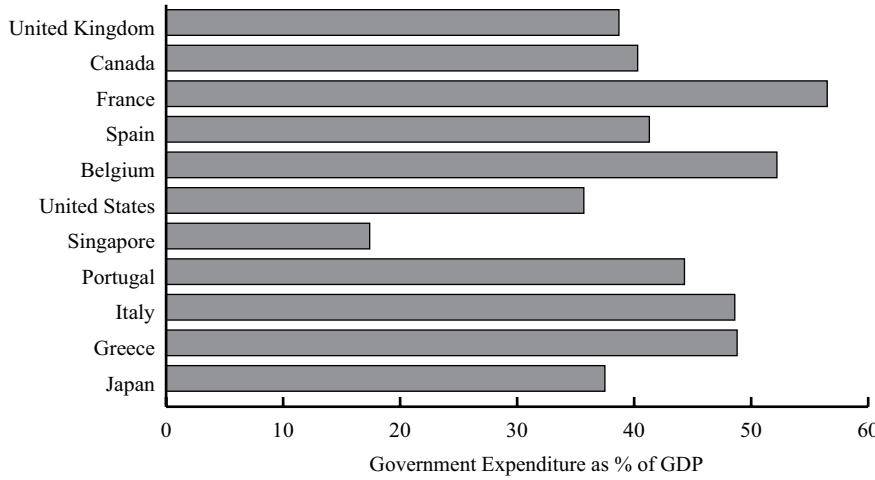
of spending in an economy. Second, governments are also the largest borrowers in world debt markets. Exhibit 1 gives some idea of the scale of government borrowing and spending.

Exhibit 1 Scale of Government Borrowing and Spending

Panel A. Central Government Debt to GDP, 2017



Panel B. Public Sector Spending to GDP, 2017



Source: IMF, World Economic Outlook Database, April 2018.

Government policy is ultimately expressed through its borrowing and spending activities. In this reading, we identify and discuss two types of government policy that can affect the macroeconomy and financial markets: monetary policy and fiscal policy.

Monetary policy refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy. Central banks can implement monetary policy almost completely independent of government interference and influence at one end of the scale or may simply act as the agent of the government at the other end of the scale.

By contrast, fiscal policy refers to the government's decisions about taxation and spending. Both monetary and fiscal policies are used to regulate economic activity over time. They can be used to accelerate growth when an economy starts to slow or to moderate growth and activity when an economy starts to overheat. In addition, fiscal policy can be used to redistribute income and wealth.

The overarching goal of both monetary and fiscal policy is normally the creation of an economic environment in which growth is stable and positive and inflation is stable and low. Crucially, the aim is to steer the underlying economy so that it does not experience economic booms that may be followed by extended periods of low or negative growth and high levels of unemployment. In such a stable economic environment, households can feel secure in their consumption and saving decisions, while corporations can concentrate on their investment decisions, on making their regular coupon payments to their bond holders, and on making profits for their shareholders.

The challenges to achieving this overarching goal are many. Economies frequently are buffeted by shocks (such as oil price jumps), and some economists believe that natural cycles in the economy also exist. Moreover, we can find plenty of examples from history in which government policies—either monetary, fiscal, or both—have exacerbated an economic expansion that eventually led to damaging consequences for the real economy, for financial markets, and for investors.

QUESTION SET



1. Which of the following statements *best* describes monetary policy?

Monetary policy:

- A. involves the setting of medium-term targets for broad money aggregates.
- B. involves the manipulation by a central bank of the government's budget deficit.
- C. seeks to influence the macroeconomy by influencing the quantity of money and credit in the economy.

Solution:

C is correct, as monetary policy involves central bank activities directed toward influencing the quantity of money and credit. Choice A is incorrect because, although the setting of targets for monetary aggregates is a possible *tool* of monetary policy, monetary policy itself is concerned with influencing the overall, or macro, economy.

2. Which of the following statements *best* describes fiscal policy? Fiscal policy:

- A. is used by governments to redistribute wealth and incomes.
- B. is the attempt by governments to balance their budgets from one year to the next.
- C. involves the use of government spending and taxation to influence economy activity.

Solution:

C is correct. Note that governments may wish to use fiscal policy to redistribute income and balance their budgets, but the overriding goal of fiscal policy is usually to influence a broader range of economic activity.

ROLES AND OBJECTIVES OF FISCAL POLICY

3

- describe roles and objectives of fiscal policy as well as arguments as to whether the size of a national debt relative to GDP matters

Fiscal policy involves the use of government spending and changing tax revenue to affect a number of aspects of the economy:

- Overall level of aggregate demand in an economy and hence the level of economic activity.
- Distribution of income and wealth among different segments of the population.
- Allocation of resources between different sectors and economic agents.

The discussion of fiscal policy often focuses on the impact of changes in the difference between government spending and revenue on the aggregate economy, rather than on the actual levels of spending and revenue themselves.

Roles and Objectives of Fiscal Policy

A primary aim for fiscal policy is to help manage the economy through its influence on aggregate national output, that is, real GDP.

Fiscal Policy and Aggregate Demand

Aggregate demand is the amount companies and households plan to spend. We can consider a number of ways that fiscal policy can influence aggregate demand. For example, an **expansionary** policy could take one or more of the following forms:

- Cuts in personal income tax raise disposable income with the objective of boosting aggregate demand.
- Cuts in sales (indirect) taxes to lower prices raise real incomes with the objective of raising consumer demand.
- Cuts in corporation (company) taxes to boost business profits may raise capital spending.
- Cuts in tax rates on personal savings to raise disposable income for those with savings, with the objective of raising consumer demand.
- New public spending on social goods and infrastructure, such as hospitals and schools, boost personal incomes with the objective of raising aggregate demand.

We must stress, however, that the reliability and magnitude of these relationships will vary over time and from country to country. For example, in a recession with rising unemployment, it is not always the case that cuts in income taxes will raise consumer spending because consumers may wish to raise their precautionary (rainy day) saving in anticipation of further deterioration in the economy. Indeed, in very general terms, economists are often divided into two camps regarding the workings of fiscal policy. **Keynesians** believe that fiscal policy can have powerful effects on aggregate demand, output, and employment when there is substantial spare capacity in an economy. **Monetarists** believe that fiscal changes only have a temporary effect on aggregate demand and that monetary policy is a more effective tool for restraining or boosting inflationary pressures. Monetarists tend not to advocate using monetary policy for countercyclical adjustment of aggregate demand. This intellectual division

naturally will be reflected in economists' divergent views on the efficacy of the large fiscal expansions observed in many countries following the 2008–2009 Global Financial Crisis, along with differing views on the possible impact of quantitative easing.

Government Receipts and Expenditure in Major Economies

In Exhibit 2, we present the total government revenues as a percentage of GDP for some major economies. This is the share of a country's output that is gathered by the government through taxes and such related items as fees, charges, fines, and capital transfers. It is often considered as a summary measure of the extent to which a government is involved both directly and indirectly in the economic activity of a country.

Taxes are formally defined as compulsory, unrequited payments to the general government (they are unrequited in the sense that benefits provided by a government to taxpayers usually are not related to payments). Exhibit 2 contains taxes on incomes and profits, social security contributions, indirect taxes on goods and services, employment taxes, and taxes on the ownership and transfer of property.

Exhibit 2: General Government Revenues as Percent of GDP

	1995	2000	2005	2008	2010	2015
Australia	34.5	36.1	36.5	35.3	32.4	34.9
Germany	45.1	46.4	43.6	43.8	43.0	44.5
Japan	31.2	31.4	31.7	34.4	30.6	35.7
United Kingdom	38.2	40.3	40.8	42.2	38.2	38.0
United States	33.8	35.4	33.0	32.3	30.9	33.4
OECD	37.9	39.0	37.7	37.9	39.8	40.9

Source: Organisation for Economic Co-operation and Development (OECD).

Taxes on income and profits have been fairly constant for the member countries of the Organisation for Economic Co-operation and Development (OECD) at around 12.5–13 percent of GDP since the mid-1990s, while taxes on goods and services have been steady at about 11 percent of GDP for that period. Variations between countries can be substantial; taxes on goods and services are around 5 percent of GDP for the United States and Japan but are more than 16 percent for Denmark.

Exhibit 3 shows the percentage of GDP represented by government expenditure in a variety of major economies over time. Generally, these have been fairly constant since 1995, although Germany had a particularly high number at the start of the period because of reunification costs. The impacts of governments' fiscal stimulus programs in the face of the 2008–2009 financial crisis show up as significant increases in government expenditures in Exhibit 3 and increases in government deficits between 2008 and 2010 are evident in Exhibit 4.

Exhibit 3: General Government Expenditures as Percent of GDP

	1995	2000	2005	2008	2010	2015
Australia	38.2	35.2	34.8	34.3	34.4	36.2
Germany	54.8	45.1	46.9	43.8	47.3	43.9
Japan	36.0	39.0	38.4	37.1	39.6	39.4
United Kingdom	44.1	36.6	44.0	47.5	47.6	42.2

	1995	2000	2005	2008	2010	2015
United States	37.1	33.9	36.2	38.8	42.9	37.6
OECD	42.7	38.7	40.5	41.4	45.2	41.8

Source: OECD.

Clearly, the possibility that fiscal policy can influence output means that it may be an important tool for **economic stabilization**. In a recession, governments can raise spending (**expansionary fiscal policy**) in an attempt to raise employment and output. In boom times—when an economy has full employment and wages and prices are rising too fast—government spending may be reduced and taxes raised (**contractionary fiscal policy**).

Hence, a key concept is the budget surplus or deficit, which is the positive or negative difference between government revenue and expenditure for a fixed period of time, such as a fiscal or calendar year. Government revenue includes tax revenues net of transfer payments; government spending includes interest payments on the government debt. Analysts often focus on changes in the budget surplus or deficit from year to year as indicators of whether the fiscal policy is getting tighter or looser. An increase in a budget surplus would be associated with contractionary fiscal policy, while a rise in a deficit is an expansionary fiscal policy. Of course, over the course of a business cycle, the budget surplus will vary automatically in a countercyclical way. For example, as an economy slows and unemployment rises, government spending on social insurance and unemployment benefits will also rise and add to aggregate demand. This is known as an **automatic stabilizer**. Similarly, if boom conditions ensue and employment and incomes are high, then progressive income and profit taxes are rising and also act as automatic stabilizers increasing budget surplus or reducing budget deficit. The great advantage of automatic stabilizers is that they are automatic and do not require the identification of shocks to which policy makers must consider a response. By reducing the responsiveness of the economy to shocks, these automatic stabilizers reduce output fluctuations. Automatic stabilizers should be distinguished from discretionary fiscal policies, such as changes in government spending or tax rates, which are actively used to stabilize aggregate demand. If government spending and revenues are equal, then the budget is **balanced**.

Exhibit 4: General Government Net Borrowing or Lending as Percent of GDP

	1995	2000	2005	2008	2010	2015
Australia	-3.7	0.9	1.7	-3.8	-4.4	-2.2
Germany	-9.7	1.3	-3.3	-0.2	-4.2	0.8
Japan	-4.7	-7.6	-6.7	-4.1	-9.1	-3.6
United Kingdom	-5.8	3.7	-3.3	-5.1	-9.4	-4.2
United States	-3.3	1.5	-3.3	-7.0	-12.0	-4.2
OECD	-4.8	0.2	-2.7	-1.5	-5.1	-1.9

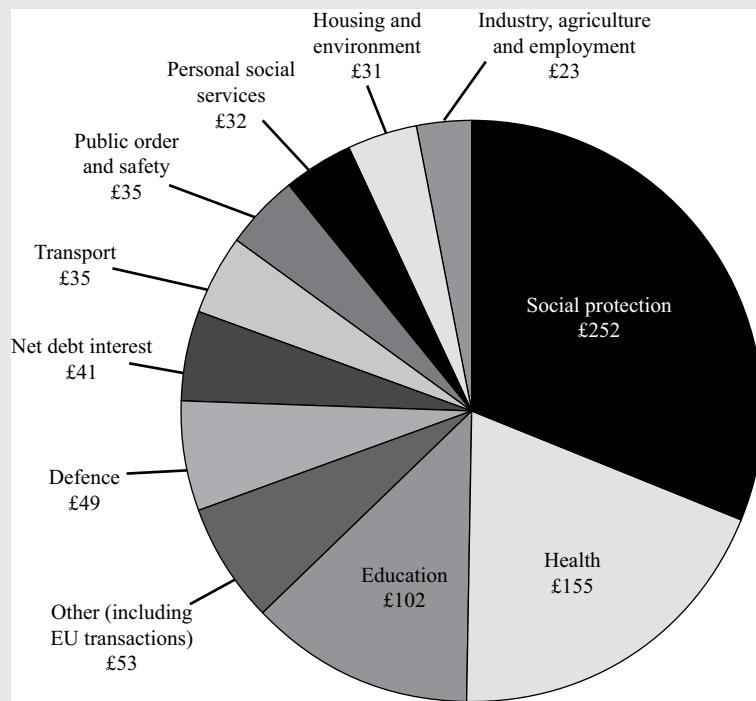
Source: OECD.

EXAMPLE 1

Sources and Uses of Government Cash Flows: The Case of the United Kingdom

The precise components of revenue and expenditure will of course vary over time and between countries. But, as an example of the breakdown of expenditure and revenue, in Exhibit 5 and Exhibit 6 we have presented the budget projections of the United Kingdom for 2018/2019. The budget projected that total spending would come to GBP808 billion, whereas total revenue would be only GBP769 billion. The government was therefore forecasting a budget shortfall of GBP39 billion for the fiscal year, meaning that it had an associated need to borrow GBP39 billion from the private sector in the United Kingdom or the private and public sectors of other economies.

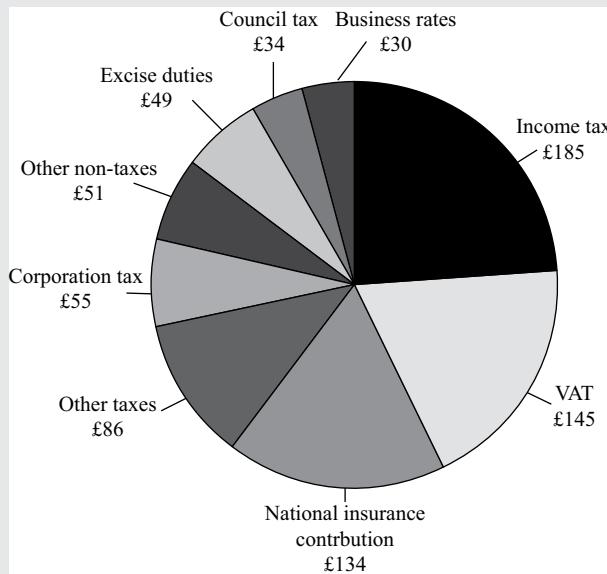
Exhibit 5: Where Does the Money Go? United Kingdom, 2018–2019



Note: All values are in billions of pounds.

Source: HM Treasury, United Kingdom.

Exhibit 6: Where Does the Money Come From? United Kingdom, 2018–2019



Note: All values are in billions of pounds.

Source: HM Treasury, United Kingdom.

QUESTION SET



1. The *least likely* goal of a government's fiscal policy is to:
 - A. redistribute income and wealth.
 - B. influence aggregate national output.
 - C. ensure the stability of the purchasing power of its currency.

Solution:

C is correct. Ensuring stable purchasing power is a goal of monetary rather than fiscal policy. Fiscal policy involves the use of government spending and tax revenue to affect the overall level of aggregate demand in an economy and hence the level of economic activity.

2. Which of the following *best* represents a contractionary fiscal policy?
 - A. Temporary suspension of payroll taxes
 - B. Public spending on a high-speed railway
 - C. Freeze in discretionary government spending

Solution:

C is correct. A freeze in discretionary government spending is an example of a contractionary fiscal policy.

3. A "pay-as-you-go" rule, which requires that any tax cut or increase in entitlement spending be offset by an increase in other taxes or reduction in other entitlement spending, is an example of which fiscal policy stance?
 - A. Neutral

B. Expansionary

C. Contractionary

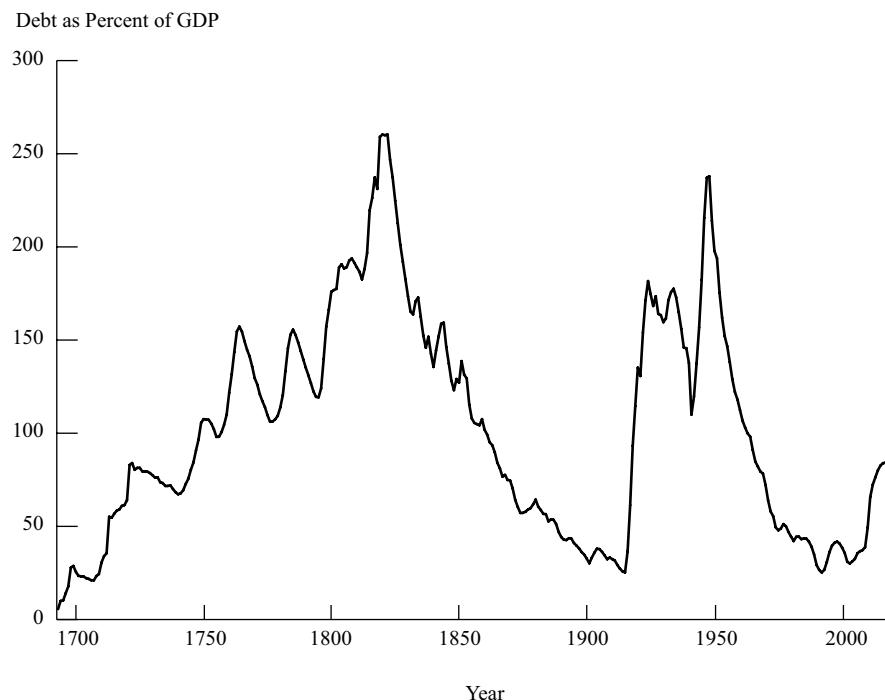
Solution:

A is correct. A “pay-as-you-go” rule is a neutral policy because any increases in spending or reductions in revenues would be offset. Accordingly, there would be no net impact on the budget deficit/surplus.

Deficits and the National Debt

Government deficits are the difference between government revenues and expenditures over a period of calendar time, usually a year. Government (or national) debt is the accumulation over time of these deficits. Government deficits are financed by borrowing from the private sector, often through private pension and insurance fund portfolio investments. We read that governments are more likely to have deficits than surpluses over long periods of time. As a result, a large stock of outstanding government debt may be owned by the private sector. This will vary as the business cycle ebbs and flows. Exhibit 7 shows the time path of the ratio of public debt to GDP for the United Kingdom over several hundred years. It can be clearly seen that the major cause of fluctuations in that ratio through history has been the financing of wars, in particular the Napoleonic Wars of 1799–1815 and the First and Second World Wars of 1914–1918 and 1939–1945.

Exhibit 7: UK National Debt as Percent of GDP, 1692–2018



Source: <http://ukpublicspending.co.uk>.

With the onset of the credit crisis of 2008, governments actively sought to stimulate their economies through increased expenditures without raising taxes and revenues. This led to increased borrowing, shown in Exhibit 8, which became a concern in the

financial markets in 2010 for such countries as Greece. Indeed, between 2008 and 2009, according to the OECD, central government debt rose from USD1.2 trillion to USD1.6 trillion in the United Kingdom and from USD5.8 trillion to USD7.5 trillion for the United States. The fiscal expansion by governments in the face of the financial crisis seems to have significantly raised the general government debt-to-GDP ratio over the long term for many countries, as illustrated in Exhibit 8.

Exhibit 8: General Government Debt as Percent of GDP

	1995	2000	2005	2008	2010	2015
Australia	57.3	41.1	30.0	30.0	41.9	64.1
Germany	54.1	59.5	70.1	68.1	84.5	78.9
Japan	94.7	142.6	176.2	181.6	207.5	237.4
United Kingdom	51.4	48.7	51.3	63.3	88.8	111.7
United States	83.2	61.7	79.0	93.2	117.0	125.3
OECD	65.8	59.9	59.5	60.8	73.0	85.3

Source: www.oecd.org.

Ultimately, if the ratio of debt to GDP rises beyond a certain unknown point, then the solvency of the country comes into question. An additional indicator for potential insolvency is the ratio of interest rate payments to GDP, which is shown for some major economies in Exhibit 9. These represent payments required of governments to service their debts as a percentage of national output and, as such, reflect both the size of debts and the interest charged on them.

Exhibit 9: General Government Net Debt Interest Payments as Percent of GDP

	1995	2000	2005	2008	2010	2015
Australia	3.5	1.7	1.0	-0.5	0.0	0.3
Germany	2.9	2.7	2.4	2.3	2.1	0.9
Japan	1.3	1.5	0.8	0.3	0.6	0.4
United Kingdom	3.1	2.4	1.8	1.7	2.6	2.0
United States	3.5	2.5	1.8	2.6	2.9	2.8
OECD	3.6	2.5	1.8	1.9	2.1	1.9

Source: OECD.

Government spending was far in excess of revenues following the credit crisis of 2007–2010 as governments tried to stimulate their economies; this level of spending raised concerns in some quarters about the scale of governmental debt accumulation. Exhibit 8 shows that government debt relative to GDP for the OECD countries overall rose from 59.5 percent in 2005 to 85.3 percent in 2015. In Japan, where fiscal spending has been used to stimulate the economy from the early 1990s, the ratio has risen from 94.7 percent in 1995 to 237.4 percent in 2015. If an economy grows in real terms, so do the real tax revenues and hence the ability to service a growing real debt at constant tax rate levels. If, however, the real growth in the economy is lower than the real interest rate on the debt, then the debt ratio will worsen even though the economy is growing because the debt burden (i.e., the real interest rate times the debt)

grows faster than the economy. Hence, an important issue for governments and their creditors is whether their additional spending leads to sufficiently higher tax revenues to pay the interest on the debt used to finance the extra spending.

However, within a national economy, the real value of the outstanding debt will fall if the overall price level rises (i.e., inflation, and hence a rise in nominal GDP even if real GDP is static) and thus the ratio of debt to GDP may not be rising. If the general price level falls (i.e., deflation), then the ratio may stay elevated for longer. If net interest payments rise rapidly and investors lose confidence in a government's ability to honor its debts, then financing costs may escalate even more quickly and make the situation unstable.

Should we be concerned about the size of a national debt (relative to GDP)? There are strong arguments both for and against:

- The arguments against being concerned about national debt (relative to GDP) are as follows:
 - The scale of the problem may be overstated because the debt is owed internally to fellow citizens. This is certainly the case in Japan and South Korea, where 93 percent is owned by local residents. Canada is similar with 90 percent of debt is owned by residents. However, other countries have a much lower percentage owned internally. According to data from the Bank for International Settlements (BIS) and International Monetary Fund (IMF), the figures are 53 percent and 73 percent in the United States and United Kingdom, respectively, whereas in Italy, only 46 percent is owned by local residents.
 - A proportion of the money borrowed may have been used for capital investment projects or to enhance human capital (e.g., training, education); these should lead to raised future output and tax revenues.
 - Large fiscal deficits require tax changes that actually may reduce distortions caused by existing tax structures.
 - Deficits may have no net impact because the private sector may act to offset fiscal deficits by increasing saving in anticipation of future increased taxes. This argument is known as "Ricardian equivalence" and is discussed in more detail later.
 - If there is unemployment in an economy, then the debt is not diverting activity away from productive uses (and indeed the debt could be associated with an increase in employment).
- The arguments in favor of being concerned about national debt are as follows:
 - High levels of debt to GDP may lead to higher tax rates in the search for higher tax revenues. This may lead to disincentives to economic activity as the higher marginal tax rates reduce labor effort and entrepreneurial activity, leading to lower growth in the long run.
 - If markets lose confidence in a government, then the central bank may have to print money to finance a government deficit. This ultimately may lead to high inflation, as evidenced by the economic history of Germany in the 1920s and more recently in Zimbabwe.
 - Government borrowing may divert private sector investment from taking place (an effect known as crowding out); if there is a limited amount of savings to be spent on investment, then larger government demands will lead to higher interest rates and lower private sector investing.

An important distinction to make is between long- and short-run effects. Over short periods of time (say, a few years), crowding out may have little effect. If it lasts for a longer time, however, then capital accumulation in an economy may be damaged. Similarly, tax distortions may not be too serious over the short term but will have a more substantial impact over many years.

QUESTION SET

1. Which of the following is *not* associated with an expansionary fiscal policy?

- A. Rise in capital gains taxes
- B. Cuts in personal income taxes
- C. New capital spending by the government on road building

Solution:

A is correct. A rise in capital gains taxes reduces income available for spending and hence reduces aggregate demand, other things being equal. Cutting income tax raises disposable income, while new road building raises employment and incomes; in both cases, aggregate demand rises and hence policy is expansionary.

2. Fiscal expansions will *most likely* have the greatest impact on aggregate output when the economy is in which of the following states?

- A. Full employment
- B. Near full employment
- C. Considerable unemployment

Solution:

C is correct. When an economy is close to full employment, a fiscal expansion raising aggregate demand can have little impact on output because there are few spare unused resources (e.g., labor or idle factories); instead, there will be upward pressure on prices (i.e., inflation). The greatest impact on aggregate output will occur when there is considerable unemployment.

3. Which one of the following is *most likely* a reason to *not* use fiscal deficits as an expansionary tool?

- A. They may crowd out private investment.
- B. They may facilitate tax changes to reduce distortions in an economy.
- C. They may stimulate employment when there is substantial unemployment in an economy.

Solution:

A is correct. A frequent argument against raises in fiscal deficits is that the additional borrowing to fund the deficit in financial markets will displace private sector borrowing for investment (i.e., crowding out).

4. The *most likely* argument against high national debt levels is that:

- A. the debt is owed internally to fellow citizens.
- B. they create disincentives for economic activity.

- C. they may finance investment in physical and human capital.

Solution:

B is correct. The belief is that high levels of debt to GDP may lead to higher future tax rates, which may lead to disincentives to economic activity.

5. Which statement regarding fiscal deficits is *most* accurate?

- A. According to the Ricardian equivalence, deficits have a multiplicative effect on consumer spending.
- B. Higher government spending may lead to higher interest rates and lower private sector investing.
- C. Central bank actions that grow the money supply to address deflationary conditions decrease fiscal deficits.

Solution:

B is correct. Government borrowing may compete with private sector borrowing for investment purposes.

4

FISCAL POLICY TOOLS



describe tools of fiscal policy, including their advantages and disadvantages

We now look at the nature of the fiscal tools available to a government. Government spending can take a variety of forms:

- **Transfer payments** are welfare payments made through the social security system and, depending on the country, include payments for state pensions, housing benefits, tax credits and income support for poorer families, child benefits, unemployment benefits, and job search allowances. Transfer payments exist to provide a basic minimum level of income for low-income households, and they also provide a means by which a government can change the overall income distribution in a society. Note that these payments are not included in the definition of GDP because they do not reflect a reward to a factor of production for economic activity. Also, they are not considered to be part of general government spending on goods and services.
- **Current government spending** involves spending on goods and services that are provided on a regular, recurring basis—including health, education, and defense. Clearly, such spending will have a big impact on a country's skill level and overall labor productivity.
- **Capital expenditure** includes infrastructure spending on roads, hospitals, prisons, and schools. This investment spending will add to a nation's capital stock and affect productive potential for an economy.

Government spending can be justified on both economic and social grounds:

- To provide such services as defense that benefit all citizens equally.
- For infrastructure capital spending (e.g., roads) to help a country's economic growth.

- To guarantee a minimum level of income for poorer people and hence redistribute income and wealth (e.g., welfare and related benefits).
- To influence a government's economic objectives of low inflation and high employment and growth (e.g., management of aggregate demand).
- To subsidize the development of innovative and high-risk new products or markets (e.g., alternative energy sources).

Government revenues can take several forms:

- **Direct taxes** are levied on income, wealth, and corporate profits and include capital gains taxes, national insurance (or labor) taxes, and corporate taxes. They also may include a local income or property tax for both individuals and businesses. Inheritance tax on a deceased's estate will have both revenue-raising and wealth-redistribution aspects.
- **Indirect taxes** are taxes on spending on a variety of goods and services in an economy—such as the excise duties on fuel, alcohol, and tobacco as well as sales (or value-added tax)—and often exclude health and education products on social grounds. In addition, taxes on gambling may be considered to have a social aspect in deterring such activity, while fuel duties will serve an environmental purpose by making fuel consumption and hence travel more expensive.

Taxes can be justified both in terms of raising revenues to finance expenditures and in terms of income and wealth redistribution policies. Economists typically consider four desirable attributes of a tax policy:

- Simplicity: This refers to ease of compliance by the taxpayer and enforcement by the revenue authorities. The final liability should be certain and not easily manipulated.
- Efficiency: Taxation should interfere as little as possible in the choices individuals make in the marketplace. Taxes affect behavior and should, in general, discourage work and investment as little as possible. A major philosophical issue among economists is whether tax policy should deliberately deviate from efficiency to promote "good" economic activities, such as savings, and discourage harmful ones, such as tobacco consumption. Although most would accept a limited role in guiding consumer choices, some will question if policy makers are equipped to decide on such objectives and whether there will be unwanted ancillary effects, such as giving tax breaks for saving among people who already save and whose behavior does not change.
- Fairness: This refers to the fact that people in similar situations should pay the same taxes ("horizontal equity") and that richer people should pay more taxes ("vertical equity"). Of course, the concept of fairness is really subjective. Still, most would agree that income tax rates should be progressive—that is, that households and corporations should pay proportionately more as their incomes rise. However, some people advocate "flat" tax rates, whereby all should pay the same proportion of taxable income.
- Revenue sufficiency: Although revenue sufficiency may seem obvious as a criterion for tax policy, there may be a conflict with fairness and efficiency. For example, one may believe that increasing income tax rates to reduce fiscal deficits reduces labor effort and that tax rate increases are thus an inefficient policy tool.

SOME ISSUES WITH TAX POLICY

1. **Incentives.** Some economists believe that income taxes reduce the incentive to work, save, and invest and that the overall tax burden has become excessive. These ideas are often associated with supply-side economics and the US economist Arthur Laffer. A variety of income tax cuts and simplifications have taken place in the United States since 1981, and despite substantial controversy, some claim that work effort did rise (although tax cuts had little impact on savings). Similarly, some found that business investment did rise, while others claimed it was independent of such cuts.
2. **Fairness.** How do we judge the fairness of the tax system? One way is to calibrate the tax burden falling on different groups of people ranked by their income and to assess how changes in taxes affect these groups. Of course, this imposes huge data demands on investigators and must be considered incomplete. In the United States, the federal system is indeed highly progressive. Many countries use such methods to analyze the impact of tax changes on different income groups when they announce their annual fiscal policy plans.
3. **Tax reform.** There is continuous debate on reforming tax policy. Should there be a flat-rate tax on labor income? Should all investment be immediately deducted for corporate taxes? Should more revenue be sourced from consumption taxes? Should taxes be indexed to inflation? Should dividends be taxed when profits have already been subject to tax? Should estates be taxed at all? Many of these issues are raised in the context of their impact on economic growth.

QUESTION SET

1. Which of the following is *not* a tool of fiscal policy?
 - A. A rise in social transfer payments
 - B. The purchase of new equipment for the armed forces
 - C. An increase in deposit requirements for the buying of houses

Solution:

C is correct. Rises in deposit requirements for house purchases are intended to reduce the demand for credit for house purchases and hence would be considered a tool of monetary policy. This is a policy used actively in several countries and is under consideration by regulators in other countries to constrain house price inflation.

2. Which of the following is *not* an indirect tax?
 - A. Excise duty
 - B. Value-added tax
 - C. Employment taxes

Solution:

C is correct. Both excise duty and value-added tax (VAT) are applied to prices, whereas taxes on employment apply to labor income and hence are not indirect taxes.

3. Which of the following statements is *most* accurate?
- Direct taxes are useful for discouraging alcohol consumption.
 - Because indirect taxes cannot be changed quickly, they are of no use in fiscal policy.
 - Government capital spending decisions are slow to plan, implement, and execute and hence are of little use for short-term economic stabilization.

Solution:

C is correct. Capital spending is much slower to implement than changes in indirect taxes; and indirect taxes affect alcohol consumption more directly than direct taxes.

The Advantages and Disadvantages of Different Fiscal Policy Tools

The different tools used to expedite fiscal policy as a means to try to put or keep an economy on a path of positive, stable growth with low inflation have both advantages and disadvantages:

Advantages

- Indirect taxes can be adjusted almost immediately after they are announced and can influence spending behavior instantly and generate revenue for the government at little or no cost to the government.
- Social policies, such as discouraging alcohol or tobacco use, can be adjusted almost instantly by raising such taxes.

Disadvantages

- Direct taxes are more difficult to change without considerable notice, often many months, because payroll computer systems will have to be adjusted (although the announcement may well have a powerful effect on spending behavior more immediately). The same may be said for welfare and other social transfers.
- Capital spending plans take longer to formulate and implement, typically over a period of years. For example, building a road or hospital requires detailed planning, legal permissions, and implementation. This is often a valid criticism of an active fiscal policy and was widely heard during the US fiscal stimulus in 2009–2010. Such policies, however, do add to the productive potential of an economy, unlike a change in personal or indirect taxes. Of course, the slower the impact of a fiscal change, the more likely other exogenous changes will already be influencing the economy before the fiscal change kicks in.

These tools may have expectational effects at least as powerful as the direct effects. The announcement of future income tax rises a year ahead potentially could lead to reduced consumption immediately. Such delayed tax rises were a feature of the UK fiscal policy of 2009–2010; however, the evidence is anecdotal because spending behavior changed little until the delayed tax changes actually came into force.

We may also consider the relative potency of the different fiscal tools. Direct government spending has a far bigger impact on aggregate spending and output than income tax cuts or transfer increases; however, if the latter are directed at the poorest

in society (basically, those who spend all their income), then this will give a relatively strong boost. Further discussion and examples of these comparisons are given in the section on the interaction between monetary and fiscal policy.

Modeling the Impact of Taxes and Government Spending: The Fiscal Multiplier

The conventional macroeconomic model has government spending, G , adding directly to aggregate demand, AD , and reducing it via taxes, T ; these include both indirect taxes on expenditures and direct taxes on factor incomes. Further government spending is increased through the payment of transfer benefits, B , such as social security payments. Hence, the net impact of the government sector on aggregate demand is as follows:

$$G - T + B = \text{Budget surplus OR deficit.}$$

Net taxes (NT ; taxes less transfers) reduce disposable income (YD) available to individuals relative to national income or output (Y) as follows:

$$YD = Y - NT = (1 - t) Y,$$

where t is the **net tax rate**. Net taxes are often assumed to be proportional to national income, Y , and hence total tax revenue from net taxes is tY . If $t = 20\%$ or 0.2, then for every USD1 rise in national income, net tax revenue will rise by USD0.20 and household disposable income will rise by USD0.80.

The **fiscal multiplier** is important in macroeconomics because it tells us how much output changes as exogenous changes occur in government spending or taxation. The recipients of the increase in government spending will typically save a proportion $(1 - c)$ of each additional dollar of disposable income, where c is the **marginal propensity to consume** (MPC) this additional income. Ignoring income taxes, we can see that $\$c$ will, in turn, be spent by these recipients on more goods and services. The recipients of this $\$c$ also will spend a proportion c of this additional income (i.e., $\$c \times c$, or c -squared). This process continues with income and spending growing at a constant rate of c as it passes from hand to hand through the economy. This is the familiar geometric progression with constant factor c , where $0 < c < 1$. The sum of this geometric series is $1/(1 - c)$.

We define s as the **marginal propensity to save** (MPS), the amount saved out of an additional dollar of disposable income. Because $c + s = 1$, $s = 1 - c$.

Exhibit 10: Disposable Income, Saving, and the MPC

Income	Income Tax	Disposable Income	Consumption	Saving
USD100	USD20	USD80	USD72	USD8

In Exhibit 10, the MPC out of disposable income is 90 percent or 0.9 (72/80). The MPS is therefore $1 - 0.9$ or 0.1.

For every dollar of new (additional) spending, total incomes and spending rises by $\text{USD}1/(1 - c)$. And because $0 < c < 1$, this must be > 1 ; this is the multiplier. If $c = 0.9$ (or individuals spend 90 percent of additions to income), then the multiplier = $1/(1 - 0.9) = 10$.

A formal definition of the multiplier would be the ratio of the change in equilibrium output to the change in autonomous spending that caused the change. This is a monetary measure, but because prices are assumed to be constant in this analysis, real and monetary amounts are identical. Given that fiscal policy is about changes

in government spending, G , net taxes, NT , and tax rates, t , we can see that the multiplier is an important tool for calibrating the possible impact of policy changes on output. How can we introduce tax changes into the multiplier concept? We do this by introducing the idea of disposable income, YD , defined as income less income taxes net of transfers, $Y - NT$.

Households spend a proportion c of disposable income, YD , that is, cYD or $c(Y - NT)$ or $c(1 - t)Y$. The marginal propensity to consume in the presence of taxes is then $c(1 - t)$. If the government increases spending, say on road building, by an amount, G , then disposable income rises by $(1 - t)G$ and consumer spending by $c(1 - t)G$. Provided there are unused sources of capital and labor in the economy, this leads to a rise in aggregate demand and output; the recipients of this extra consumption spending will have $(1 - t)c(1 - t)G$ extra disposable income available and will spend c of it. This cumulative extra spending and income will continue to spread through the economy at a decreasing rate as $0 < c(1 - t) < 1$. The overall final impact on aggregate demand and output will effectively be the sum of this decreasing geometric series with the common ratio $c(1 - t)$, which sums to $1/[1 - c(1 - t)]$. This is known as the fiscal multiplier and is relevant to studies of fiscal policy as changes in G or tax rates will affect output in an economy through the value of the multiplier.

For example, if the tax rate is 20 percent, or 0.2, and the marginal propensity to consume is 90 percent, or 0.9, then the fiscal multiplier will be as follows: $1/[1 - 0.9(1 - 0.2)]$ or $1/0.28 = 3.57$. In other words, if the government raises G by USD1 billion, total incomes and spending rise by USD3.57 billion.

Discretionary fiscal policy will involve changes in these variables with a view to influencing Y .

The Balanced Budget Multiplier

If a government increases G by the same amount as it raises taxes, the aggregate output actually rises. Why is this?

Because the marginal propensity to consume out of disposable income is less than 1, for every dollar less in YD , spending falls only $\$c$. Hence, aggregate spending falls less than the tax rise by a factor of c . A balanced budget leads to a rise in output, which in turn leads to further rises in output and incomes through the multiplier effect.

Suppose an economy has an equilibrium output or income level of USD1,000 consisting of USD900 of consumption and USD100 of investment spending, which is fixed and not related to income. If government spending is set at USD200, financed by a tax rate of 20 percent (giving tax revenue of USD200), what will happen to output? First, additional government spending of USD200 will raise output by that amount. But taxes of USD200 will not reduce output by a similar amount if the MPC is less than 1. Suppose it is 0.9, and hence spending will fall only by 90 percent of USD200, or USD180. The initial impact of the balanced fiscal package on aggregate demand will be to raise it by $\text{USD}200 - \text{USD}180 = \text{USD}20$. This additional output, in turn, will lead to further increases in income and output through the multiplier effect.

Even though this policy involved a combination of government spending and tax increases that initially left the government's budget deficit/surplus unchanged, the induced rise in output will lead to further tax revenue increases and a further change in the budget position. Could the government adjust the initial change in spending to offset exactly the eventual total change in tax revenues? The answer is "yes," and we can ask what the effect will be on output of this genuinely balanced budget change. This balanced budget multiplier always takes the value unity.

EXAMPLE 1**Government Debt, Deficits, and Ricardo Equivalence**

The total stock of government debt is the outstanding stock of IOUs issued by a government and not yet repaid. They are issued when the government has insufficient tax revenues to meet expenditures and has to borrow from the public. The size of the outstanding debt equals the cumulative quantity of net borrowing it has done, and the fiscal or budget deficit is added in the current period to the outstanding stock of debt. If the outstanding stock of debt falls, we have a negative deficit or a surplus.

If a government reduces taxation by USD10 billion one year and replaces that revenue with borrowing of USD10 billion from the public, will it have any real impact on the economy? The important issue is how people perceive that action: Do they recognize what will happen over time as interest and bond principal have to be repaid out of future taxes? If so, they may think of the bond finance as equivalent to delayed taxation finance; thus, the reduction in current taxation will have no impact on spending because individuals save more in anticipation of higher future taxes to repay the bond. This is called **Ricardian equivalence** after the economist David Ricardo. If people do not correctly anticipate all the future taxes required to repay the additional government debt, then they feel wealthier when the debt is issued and may increase their spending, adding to aggregate demand.

Whether Ricardian equivalence holds in practice is ultimately an empirical issue and is difficult to calibrate conclusively given the number of factors that are changing at any time in a modern economy.

QUESTION SET

1. Which of the following is the *most likely* example of a tool of fiscal policy?

- A. Public financing of a power plant
- B. Regulation of the payment system
- C. Central bank's purchase of government bonds

Solution:

A is correct. Public financing of a power plant could be described as a fiscal policy tool to stimulate investment.

5**FISCAL POLICY IMPLEMENTATION**

- explain the implementation of fiscal policy and difficulties of implementation as well as whether a fiscal policy is expansionary or contractionary

We next discuss major issues in fiscal policy implementation.

Deficits and the Fiscal Stance

An important question is the extent to which the budget is a useful measure of the government's fiscal stance. Does the size of the deficit actually indicate whether fiscal policy is expansionary or contractionary? Clearly, such a question is important for economic policy makers insofar as the deficit can change for reasons unrelated to actual fiscal policy changes. For example, the automatic stabilizers mentioned earlier will lead to changes in the budget deficit unrelated to fiscal policy changes; a recession will cause tax revenues to fall and the budget deficit to rise. An observer may conclude that fiscal policy has been loosened and is expansionary and that no further government action is required.

To this end, economists often look at the structural budget deficit as an indicator of the fiscal stance. This is defined as the deficit that would exist *if the economy was at full employment (or full potential output)*. Hence, if we consider a period of relatively high unemployment, such as 2009–2010 with around 9–10 percent of the workforce out of work in the United States and Europe, then the budget deficits in those countries would be expected to be reduced substantially if the economies returned to full employment. At this level, tax revenues would be higher and social transfers lower. Historical data for major countries are given in Exhibit 11, where negative numbers refer to deficits and positive numbers are surpluses.

Exhibit 11: General Government Cyclically Adjusted Balances as Percent of GDP

	1995	2000	2005	2008	2010	2015
Australia	-3.1	0.9	2.0	-0.4	-3.8	-0.1
Germany	-9.5	0.9	-2.6	-0.8	-3.3	0.7
Japan	-4.6	-6.4	-4.1	-4.0	-8.2	-3.6
United Kingdom	-5.6	0.8	-4.5	-5.6	-7.6	-4.3
United States	-2.9	-0.4	-5.4	-7.1	-10.0	-3.5
OECD	-4.6	-1.2	-3.6	-4.5	-6.9	-2.0

Source: OECD Economic Outlook, Volume 2018 Issue 1.

Another reason why actual government deficits may *not* be a good measure of fiscal stance is the distinction between real and nominal interest rates and the role of inflation adjustment when applied to budget deficits. Although national economic statistics treat the cash interest payments on debt as government expenditure, it makes more sense to consider only the inflation-adjusted (or real) interest payments because the real value of the outstanding debt is being eroded by inflation. Automatic stabilizers—such as income tax, VAT, and social benefits—are important because as output and employment fall and reduce tax revenues, *net* tax revenues also fall as unemployment benefits rise. This acts as a fiscal stimulus and serves to reduce the size of the multiplier, dampening the output response of whatever caused the fall in output in the first place. By their very nature, automatic stabilizers do not require policy changes; no policy maker has to decide that an economic shock has occurred and how to respond. Hence, the responsiveness of the economy to shocks is automatically reduced, as are movements in employment and output.

In addition to these automatic adjustments, governments also use discretionary fiscal adjustments to influence aggregate demand. These will involve tax changes and/or spending cuts or increases usually with the aim of stabilizing the economy. A natural question is why fiscal policy cannot stabilize aggregate demand completely, hence ensuring full employment at all times.

Difficulties in Executing Fiscal Policy

Fiscal policy cannot stabilize aggregate demand completely because the difficulties in executing fiscal policy cannot be completely overcome.

First, the policy maker does not have complete information about how the economy functions. It may take several months for policy makers to realize that an economy is slowing, because data appear with a considerable time lag and even then are subject to substantial revision. This is often called the **recognition lag** and has been likened to the problem of driving while looking in the rearview mirror. Then, when policy changes are finally decided on, they may take many months to implement. This is the **action lag**. If a government decides to raise spending on capital projects to increase employment and incomes, for example, these may take many months to plan and put into action. Finally, the result of these actions on the economy will take additional time to become evident; this is the **impact lag**. These types of policy lags also occur in the case of discretionary monetary policy.

A second aspect of time in this process is the uncertainty of where the economy is heading independently of these policy changes. For example, a stimulus may occur simultaneously with a surprise rise in investment spending or in the demand for a country's exports just as discretionary government spending starts to rise. Macroeconomic forecasting models generally do not have a good track record for accuracy and hence cannot be relied on to aid the policy-making process in this context. In addition, when discretionary fiscal adjustments are announced (or are already underway), private sector behavior may well change, leading to rises in consumption or investment, both of which will reinforce the effects of a rise in government expenditure. Again, this will make it difficult to calibrate the required fiscal adjustment to secure full employment.

The following wider macroeconomic issues also are involved:

- If the government is concerned with both unemployment *and* inflation in an economy, then raising aggregate demand toward the full employment level may also lead to a tightening labor market and rising wages and prices. The policy maker may be reluctant to further fine-tune fiscal policy in an uncertain world because it might induce inflation.
- If the budget deficit is already large relative to GDP and further fiscal stimulus is required, then the necessary increase in the deficit may be considered unacceptable by the financial markets when government funding is raised, leading to higher interest rates on government debt and political pressure to tackle the deficit.
- Of course, all this presupposes that we know the level of full employment, which is difficult to measure accurately. Fiscal expansion raises demand, but what if we are already at full employment, which will be changing as productive capacity changes and workers' willingness to work at various wage levels changes?
- If unused resources reflect a low supply of labor or other factors rather than a shortage of demand, then discretionary fiscal policy will not add to demand and will be ineffective, raising the risk of inflationary pressures in the economy.

- The issue of crowding out may occur: If the government borrows from a limited pool of savings, the competition for funds with the private sector may crowd out private firms with subsequently less investing and economic growth. In addition, the cost of borrowing may rise, leading to the cancellation of potentially profitable opportunities. This concept is the subject of continuing empirical debate and investigation.

QUESTION SET

1. Which of the following statements is *least* accurate?
 - A. The economic data available to policy makers have a considerable time lag.
 - B. Economic models always offer an unambiguous guide to the future path of the economy.
 - C. Surprise changes in exogenous economic variables make it difficult to use fiscal policy as a stabilization tool.

Solution:

B is correct. Economic forecasts from models will always have an element of uncertainty attached to them and thus are not unambiguous or precise in their prescriptions. Once a fiscal policy decision has been made and implemented, unforeseen changes in other variables may affect the economy in ways that would lead to changes in the fiscal policy if we had perfect foresight. Note that it is true that official economic data may be available with substantial time lags, making fiscal judgments more difficult.

2. Which of the following statements is *least* accurate?
 - A. Discretionary fiscal changes are aimed at stabilizing an economy.
 - B. Automatic fiscal stabilizers include new plans for additional road building by the government.
 - C. In the context of implementing fiscal policy, the recognition lag is often referred to as “driving while looking in the rearview mirror.”

Solution:

B is correct. New plans for road building are discretionary and not automatic.

3. Which of the following statements regarding a fiscal stimulus is *most* accurate?
 - A. Accommodative monetary policy reduces the impact of a fiscal stimulus.
 - B. Different statistical models will predict different impacts for a fiscal stimulus.
 - C. It is always possible to precisely predict the impact of a fiscal stimulus on employment.

Solution:

B is correct. Different models embrace differing views on how the economy works, including differing views on the impact of fiscal stimuli.

4. Which of the following statements is *most* accurate?

- A. An increase in the budget deficit is always expansionary.
- B. An increase in government spending is always expansionary.
- C. The structural deficit is always larger than the deficit below full employment.

Solution:

A is correct. Note that increases in government spending may be accompanied by even bigger rises in tax receipts and hence may not be expansionary.

PRACTICE PROBLEMS

1. Crowding out refers to a:
 - A. fall in interest rates that reduces private investment.
 - B. rise in private investment that reduces private consumption.
 - C. rise in government borrowing that reduces the ability of the private sector to access investment funds.
2. A contractionary fiscal policy will always involve which of the following?
 - A. Balanced budget
 - B. Reduction in government spending
 - C. Fall in the budget deficit or rise in the surplus
3. Which one of the following statements is *most* accurate?
 - A. Ricardian equivalence refers to individuals having no idea of future tax liabilities.
 - B. Governments do not allow political pressures to influence fiscal policies but do allow voters to affect monetary policies.
 - C. If there is high unemployment in an economy, then easy monetary and fiscal policies should lead to an expansion in aggregate demand.
4. Which statement regarding fiscal policy is *most* accurate?
 - A. Cyclically adjusted budget deficits are appropriate indicators of fiscal policy.
 - B. To raise business capital spending, personal income taxes should be reduced.
 - C. An increase in the budget surplus is associated with expansionary fiscal policy.
5. The *least likely* explanation for why fiscal policy cannot stabilize aggregate demand completely is that:
 - A. private sector behavior changes over time.
 - B. policy changes are implemented very quickly.
 - C. fiscal policy focuses more on inflation than on unemployment.

SOLUTIONS

1. C is correct. A fall in interest rates is likely to lead to a rise in investment. Crowding out refers to government borrowing that reduces the ability of the private sector to invest.
2. C is correct. Note that a reduction in government spending could be accompanied by an even bigger fall in taxation, making it be expansionary.
3. C is correct. Note that governments often allow pressure groups to affect fiscal policy and that Ricardian equivalence involves individuals correctly anticipating future taxes. Thus, A and B are not correct choices.
4. A is correct. Cyclically adjusted budget deficits are appropriate indicators of fiscal policy. These are defined as the deficit that would exist if the economy was at full employment (or full potential output).
5. B is correct. Fiscal policy is subject to recognition, action, and impact lags.

LEARNING MODULE

4

Monetary Policy

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	describe the roles and objectives of central banks
<input type="checkbox"/>	describe tools used to implement monetary policy tools and the monetary transmission mechanism, and explain the relationships between monetary policy and economic growth, inflation, interest, and exchange rates
<input type="checkbox"/>	describe qualities of effective central banks; contrast their use of inflation, interest rate, and exchange rate targeting in expansionary or contractionary monetary policy; and describe the limitations of monetary policy
<input type="checkbox"/>	explain the interaction of monetary and fiscal policy

INTRODUCTION

1

Central banks play several important roles in modern economies. These roles include being the monopoly supplier of the currency, the banker to the government and the bankers' bank, the lender of last resort, the regulator and supervisor of the payments system, the conductor of monetary policy, and the supervisor of the banking system. Central banks have three primary tools available to them: open market operations, the refinancing rate, and reserve requirements. The success of central banks is thought to depend on three key concepts: central bank independence, credibility, and transparency. Both fiscal and monetary policy can alter aggregate demand, but they do so through differing channels with differing impacts on the composition of aggregate demand.

LEARNING MODULE OVERVIEW



- Central banks are the sole supplier of domestic currency, the banker to the government and the bankers' bank, the lender of last resort, the regulator and supervisor of the payments system, the conductor of monetary policy, and the supervisor of the banking system.

- The highest profile role that central banks assume is the operation of a country's monetary policy, which refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy.
- The overarching goal of most central banks in maintaining price stability is the associated goal of controlling inflation.
- Central banks can manipulate the money supply in one of three ways: open market operations, its official policy rate and associated actions in the repo market, and manipulation of official reserve requirements.
- The central bank target **interest rate** (or **policy rate**) is used to influence short- and long-term interest rates and, ultimately, real economic activity.
- The central bank's policy rate works through the economy via the following interconnected channels: short-term interest rates, changes in the values of key asset prices, the exchange rate, and the expectations of economic agents.
- The success of inflation-targeting by central banks depends on three key characteristics: central bank independence, credibility, and transparency.
- Many emerging market economies choose to operate monetary policy by targeting their currency's exchange rate, rather than an explicit level of domestic inflation.
- A major problem for central banks as they try to manage the money supply to influence the real economy is that they cannot control the amount of money that households and corporations put in banks on deposit, nor can they easily control the willingness of banks to create money by expanding credit.
- Both fiscal and monetary policy can alter aggregate demand, but they do so through differing channels with differing impacts on the composition of aggregate demand.
- Both fiscal and monetary policies suffer from a lack of precise current knowledge of the economy, because periodic economic data are released with a time lag and are subject to revision.
- The interaction between monetary and fiscal policies is evident in the Ricardian equivalence because if tax cuts have no impact on private spending due to higher expected future taxes, then this may lead policy makers to favor monetary tools.

2

ROLE OF CENTRAL BANKS



describe the roles and objectives of central banks

Roles of Central Banks and Objectives of Monetary Policy

Central banks play several key roles in modern economies. Generally, a central bank is the sole supplier of the domestic currency, the banker to the government and the bankers' bank, the lender of last resort, the regulator and supervisor of the payments system, the conductor of monetary policy, and the supervisor of the banking system. Let us examine these roles in turn.

In its earliest form, money could be exchanged for a prespecified precious commodity, usually gold, and promissory notes were issued by many private banks. Today, however, state-owned institutions—usually central banks—are designated in law as being the monopoly suppliers of a currency. Initially, these monopolists supplied money that could be converted into a prespecified amount of gold; they adhered to a **gold standard**. For example, until 1931, bank notes issued by Britain's central bank, the Bank of England, could be redeemed at the bank for a prespecified amount of gold. But Britain, like most other major economies, abandoned this convertibility principle in the first half of the twentieth century. Money in all major economies today is not convertible by law into anything else, but it is, in law, **legal tender**. This means that it must be accepted when offered in exchange for goods and services. Money that is not convertible into any other commodity is known as **fiat money**. Fiat money derives its value through government decree and because people accept it for payment of goods and services and for debt repayment.

As long as fiat money is acceptable to everyone as a medium of exchange, and it holds its value over time, then it also will be able to serve as a unit of account. However, once an economy has moved to a system of fiat money, the role of the supplier of that money becomes even more crucial because they could, for example, expand the supply of this money indefinitely should they wish to do so. Central banks, therefore, play a crucial role in modern economies as the suppliers and guardians of the value of their fiat currencies and as institutions charged with the role of maintaining confidence in their currencies. As the sole suppliers of domestic currency, central banks are at the center of economic life. As such, they assume other roles in addition to being the suppliers and guardians of the value of their currencies.

Most central banks act as the banker to the government and to other banks. They also act as a **lender of last resort** to banks. Because the central bank effectively has the capacity to print money, it is in the position to be able to supply the funds to banks that are facing crisis. The facts that economic agents know that the central bank stands ready to provide the liquidity required by any of the banks under its jurisdiction and that they trust government bank deposit insurance help to prevent bank runs in the first place. However, the recent financial crisis has shown that this knowledge is not always sufficient to deter a bank run.

EXAMPLE 1

The Northern Rock Bank Run

In the latter part of the summer of 2007, the fall in US house prices and the related implosion of the US sub-prime mortgage market became the catalyst for a global liquidity crisis. Banks began to hoard cash and refused to lend to other banks at anything other than extremely punitive interest rates through the interbank market. This caused severe difficulties for a UK mortgage bank, Northern Rock. Northern Rock's mortgage book had expanded rapidly in the preceding years as it borrowed aggressively from the money markets. It is now clear that this expansion was at the expense of loan quality. The then-UK regulatory authority, the Financial Services Authority (FSA), later reported in 2008 that Northern Rock's lending practices did not pay due regard to either the credit quality of the mortgagees or the values of the properties on which the mortgages

were secured. Being at the worst end of banking practice and relying heavily on international capital markets for its funding, Northern Rock was therefore susceptible to a global reduction in liquidity. As the liquidity crisis took hold, Northern Rock found that it could not replace its maturing money market borrowings. On 12 September 2007, in desperate need of liquidity, Northern Rock's board approached the UK central bank to ask for the necessary funds.

However, the news of Northern Rock's perilous liquidity position became known by the public and, more pertinently, by Northern Rock's retail depositors. On 14 September, having heard the news, queues began to form outside Northern Rock branches as depositors tried to withdraw their savings. On that day, it was estimated that Northern Rock depositors withdrew around GBP1 billion, representing 5 percent of Northern Rock's deposits. Further panic ensued as investors in "internet-only" Northern Rock accounts could not withdraw their money because of the collapse of Northern Rock's website. A further GBP1 billion was withdrawn over the next two days.

Northern Rock's share price dropped rapidly, as did the share prices of other similar UK banks. The crisis therefore threatened to engulf more than one bank. To prevent contagion, the chancellor of the exchequer announced on 17 September that the UK government would guarantee all Northern Rock deposits. This announcement was enough to stabilize the situation and given that lending to Northern Rock was now just like lending to the government, deposits actually started to rise again.

Eventually Northern Rock was nationalized by the UK government, with the hope that at some time in the future it could be privatized once its balance sheet had been repaired.

Central banks often are charged by the government to supervise the banking system, or at least to supervise those banks that they license to accept deposits. In some countries, this role is undertaken by a separate authority. In other countries, the central bank can be jointly responsible with another body for the supervision of its banks.

Exhibit 1 lists the banking supervisors in the G–10 countries; central banks are underlined. As the exhibit shows, most but not all bank systems have a single supervisor, which is not necessarily a central bank. A few countries, such as Germany and the United States, have more than one supervisor.

Exhibit 1: Banking Supervision in the G–10

Country	Institutions
Belgium	Banking and Finance Commission
Canada	Office of the Superintendent of Financial Institutions
France	Commission Bancaire
Germany	Federal Banking Supervisory Office; Deutsche Bundesbank
Italy	Bank of Italy
Japan	Financial Services Agency
Netherlands	Bank of Netherlands
Sweden	Swedish Financial Supervisory Authority
Switzerland	Federal Commission
United Kingdom	Bank of England
United States	Office of the Comptroller of the Currency; Federal Reserve; Federal Deposit Insurance Corporation

The United Kingdom is an interesting case study in this regard. Until May 1997, the Bank of England had statutory responsibility for banking supervision in the United Kingdom. In May 1997, banking supervision was removed from the Bank of England and assigned to a new agency, the Financial Services Authority (FSA). However, the removal of responsibility for banking supervision from the central bank was seen by some as being a contributory factor in the run on the mortgage bank Northern Rock, and generally as a contributory factor in the recent banking crisis. Because of this perceived weakness in the separation of the central bank from banking supervision, the Bank of England regained responsibility for banking supervision and regulation in 2013.

Perhaps the least appreciated role of a central bank is its role in the **payments system**. Central banks are usually asked to oversee, regulate, and set standards for a country's payments system. For the system to work properly, procedures must be robust and standardized. The central bank will usually oversee the payments system and will also be responsible for the successful introduction of any new processes. Given the international nature of finance, the central bank will also be responsible for coordinating payments systems internationally with other central banks.

Most central banks are responsible for managing their country's **foreign currency reserves** as well as its gold reserves. With regard to the latter, even though countries abandoned the gold standard in the early part of the twentieth century, the world's central bankers still hold large quantities of gold. As such, if central banks were to decide to sell significant proportions of their gold reserves, it could potentially depress gold prices.

Finally, central banks are usually responsible for the operation of a country's **monetary policy**. This is arguably the highest profile role that these important organizations assume. Recall that monetary policy refers to central bank activities that are directed toward influencing the quantity of money and credit in an economy. As the sole supplier of a country's domestic currency, central banks are in the ideal position to implement and determine monetary policy.

To summarize, central banks assume a range of roles and responsibilities. They do not all assume responsibility for the supervision of the banks, but all of the following roles normally are assumed by the central bank:

- Monopoly supplier of the currency;
- Banker to the government and the bankers' bank;
- Lender of last resort;
- Regulator and supervisor of the payments system;
- Conductor of monetary policy; and
- Supervisor of the banking system.

The Objectives of Monetary Policy

Central banks fulfill a variety of important roles, but for what overarching purpose? A perusal of the websites of the world's central banks will reveal a wide range of explanations of their objectives. Their objectives are clearly related to their roles, and so there is frequent mention of objectives related to the stability of the financial system and to the payments systems. Some central banks are charged with doing all they can to maintain full employment and output. Some also have related but less tangible roles, such as *maintaining confidence in the financial system*, or even to *promote understanding of the financial sector*. Most seem to acknowledge explicitly one overarching objective—the objective of maintaining **price stability**.

So, although central banks usually have to perform many roles, most specify an overarching objective. Exhibit 2 lists what we might call the primary objectives of a number of central banks, from both developed market and emerging market economies.

Exhibit 2: The Objectives of Central Banks

The Central Bank of Brazil

Its institutional mission is to “ensure the stability of the currency’s purchasing power and a solid and efficient financial system.”

The European Central Bank

“[T]o maintain price stability is the primary objective of the Euro system and of the single monetary policy for which it is responsible. This is laid down in the Treaty on the Functioning of the European Union, Article 127 (1).”

“Without prejudice to the objective of price stability,” the euro system will also “support the general economic policies in the Community with a view to contributing to the achievement of the objectives of the Community.” These include a “high level of employment” and “sustainable and non-inflationary growth.”

The US Federal Reserve

“The Federal Reserve sets the nation’s monetary policy to promote the objectives of maximum employment, stable prices, and moderate long-term interest rates.”

The Reserve Bank of Australia

“It is the duty of the Reserve Bank Board, within the limits of its powers, to ensure that the monetary and banking policy of the Bank is directed to the greatest advantage of the people of Australia and that the powers of the Bank ... are exercised in such a manner as, in the opinion of the Reserve Bank Board, will best contribute to:

the stability of the currency of Australia;
the maintenance of full employment in Australia; and
the economic prosperity and welfare of the people of Australia.”

The Bank of Korea

“The primary purpose of the Bank, as prescribed by the Bank of Korea Act of 1962, is the pursuit of price stability.”

Source: “Central Bank and Monetary Authority Websites,” Bank for International Settlements, <http://www.bis.org/cbanks.htm>.

QUESTION SET



1. A central bank is normally *not* the:

- A. lender of last resort.
- B. banker to the government and banks.

- C. body that sets tax rates on interest on savings.

Solution:

C is correct. A central bank is normally the lender of last resort and the banker to the banks and government, but the determination of all tax rates is normally the preserve of the government and is a fiscal policy issue.

2. Which of the following *best* describes the overarching, long-run objective of most central banks?

- A. Price stability
- B. Fast economic growth
- C. Current account surplus

Solution:

A is correct. Central banks normally have a variety of objectives, but the overriding one is nearly always price stability.

3. Which role is a central bank *least likely* to assume?

- A. Lender of last resort
- B. Supplier of the currency
- C. Sole supervisor of banks

Solution:

C is correct. The supervision of banks is not a role that all central banks assume. When it is a central bank's role, responsibility may be shared with one or more entities.

As we have discussed, one of the essential features of a monetary system is that the medium of exchange should have a relatively stable value from one period to the next. Arguably then, the overarching goal of most central banks in maintaining price stability is the associated goal of controlling inflation. Before we explore the tools central banks use to control inflation, we should first consider the potential costs of inflation. In other words, we should ask why it is that central bankers believe that it is so important to control a nominal variable.

MONETARY POLICY TOOLS AND MONETARY TRANSMISSION

3



- describe tools used to implement monetary policy tools and the monetary transmission mechanism, and explain the relationships between monetary policy and economic growth, inflation, interest, and exchange rates

Central banks have three primary tools available to them: open market operations, the refinancing rate, and reserve requirements.

Open Market Operations

One of the most direct ways for a central bank to increase or reduce the amount of money in circulation is through **open market operations**. Open market operations involve the purchase and sale of government bonds from and to commercial banks or designated market makers. For example, when the central bank buys government bonds from commercial banks, this increases the reserves of private sector banks on the asset side of their balance sheets. If banks then use these surplus reserves by increasing lending to corporations and households, then broad money growth expands through the money multiplier process. Similarly, the central bank can sell government bonds to commercial banks. In so doing, the reserves of commercial banks decline, reducing their capacity to make loans (i.e., create credit) to households and corporations and thus causing broad money growth to decline through the money multiplier mechanism. In using open market operations, the central bank may target a desired level of commercial bank reserves or a desired interest rate for these reserves.

The Central Bank's Policy Rate

The most obvious expression of a central bank's intentions and views comes through the interest rate it sets. The name of the **official interest rate** (or **official policy rate** or just **policy rate**) varies from central bank to central bank, but its purpose is to influence short- and long-term interest rates and ultimately real economic activity.

The interest rate that a central bank sets and that it announces publicly is normally the rate at which it is willing to lend money to the commercial banks (although practices do vary from country to country). This policy rate can be achieved by using short-term collateralized lending rates, known as repo rates. For example, if the central bank wishes to increase the supply of money, it might buy bonds (usually government bonds) from the banks, with an agreement to sell them back at some time in the future. This transaction is known as a **repurchase agreement**. Normally, the maturity of repo agreements ranges from overnight to two weeks. In effect, this represents a secured loan to the banks, and the lender (in this case the central bank) earns the repo rate.

Suppose that a central bank announces an increase in its official interest rate. Commercial banks normally would increase their **base rates** at the same time. A commercial bank's base rate is the reference rate on which it bases lending rates to all other customers. For example, large corporate clients might pay the base rate plus 1 percent on their borrowing from a bank, whereas the same bank might lend money to a small corporate client at the base rate plus 3 percent. But why would commercial banks immediately increase their base or reference rates just because the central bank's refinancing rate had increased?

The answer is that commercial banks do not want to lend at a rate of interest below that which they are charged by the central bank. Effectively, the central bank can force commercial banks to borrow from it at this rate because it can conduct open market operations that create a shortage of money, forcing the banks to sell bonds to it with an agreed-upon repurchase price (i.e., a repurchase agreement). The repo rate would be such that the central bank earned the official refinancing rate on the transactions.

The name of each central bank's official refinancing rate varies. The Bank of England's refinancing rate is the **two-week repo rate**. In other words, the Bank of England fixes the rate at which it is willing to lend two-week money to the banking sector. The European Central Bank's (ECB's) official policy rate is known as the **refinancing rate**, which defines the rate at which the ECB is willing to lend short-term money to the Euro area banking sector.

The corresponding rate in the United States is the discount rate, which is the rate for member banks borrowing directly from the Federal Reserve System. The most important interest rate used in US monetary policy is the **federal funds rate**. The federal

funds rate (or **fed funds rate**) is the interbank lending rate on overnight borrowings of reserves. The Federal Open Market Committee (FOMC) seeks to move this rate to a target level by reducing or adding reserves to the banking system by means of open market operations. The level of the rate is reviewed by the FOMC at its meetings held every six weeks (although the target can be changed between meetings, if necessary).

Through the setting of a policy rate, a central bank can manipulate the amount of money in the money markets. Generally speaking, the higher the policy rate, the higher the potential penalty that banks will have to pay to the central bank if they run short of liquidity, the greater their willingness will be to reduce lending, and the more likely it will be that broad money growth will shrink.

Reserve Requirements

The third primary way in which central banks can limit or increase the supply of money in an economy is through the **reserve requirement**. We already have seen that the money creation process is more powerful the lower the percentage reserve requirement of banks. So, a central bank could restrict money creation by raising the reserve requirements of banks. However, this policy tool is not used much today in developed market economies. Indeed, some central banks, such as the Bank of England, no longer even set minimum reserve requirements for the banks under their jurisdiction. Changing reserve requirements frequently is disruptive for banks. For example, if a central bank increased the reserve requirements, a bank that was short on reserves might have to cease its lending activities until it had built up the necessary reserves, because deposits would be unlikely to rise quickly enough for the bank to build its reserves in this way. However, reserve requirements are still actively used in many emerging market countries to control lending and remain a potential policy tool for those central banks that do not currently use it.

To summarize, central banks can manipulate the money supply in one of three ways:

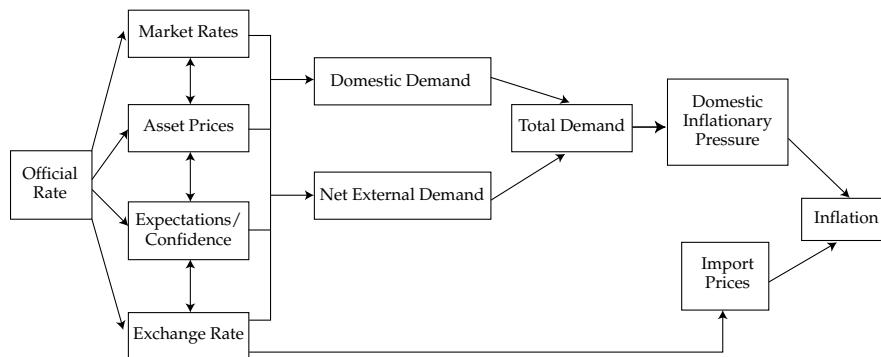
- open market operations;
- official policy rates and associated actions in the repo market; and
- manipulation of official reserve requirements.

The Transmission Mechanism

The overarching goal of a central bank is to maintain price stability. We have demonstrated how a central bank can manipulate the money supply and growth of the money supply. We also indicated how policy rates set and targeted by the central banks are usually very short term in nature; often they target overnight interest rates. However, most businesses and individuals in the real economy borrow and lend over much longer time frames than this. It may not be obvious, then, how changing short-term interest rates can influence the real economy, particularly if money neutrality holds in the long run. The fact that central bankers believe that they can affect real economic variables, in particular economic growth, by influencing broad money growth suggests that they believe that money is not neutral—at least not in the short run.

Exhibit 3 presents a stylized representation of the **monetary transmission mechanism**. This is the process whereby a central bank's interest rate is transmitted through the economy and ultimately affects the rate of increase of prices—that is, inflation.

Exhibit 3: A Stylized Representation of the Monetary Transmission Mechanism



Source: Bank of England, <https://www.bankofengland.co.uk/>.

Suppose that a central bank announces an increase in its official interest rate. The implementation of the policy may begin to work through the economy via four interrelated channels. Those channels include bank lending rates, asset prices, agents' expectations, and exchange rates. First, as described earlier, the base rates of commercial banks and interbank rates should rise in response to the increase in the official rate. Banks, in turn, would increase the cost of borrowing for individuals and companies over both short- and long-term horizons. Businesses and consumers would then tend to borrow less as interest rates rise. An increase in short-term interest rates could also cause the price of such assets as bonds or the value of capital projects to fall as the discount rate for future cash flows rises.

Market participants would then come to the view that higher interest rates will lead to slower economic growth, reduced profits, and reduced borrowing to finance asset purchases. Exporters' profits might decline if the rise in interest rates causes the country's exchange rate to appreciate, because this would make domestic exports more expensive to overseas buyers and dampen demand to purchase them. The fall in asset prices as well as an increase in prices would reduce household financial wealth and therefore lead to a reduction in consumption growth. Expectations regarding interest rates can play a significant role in the economy. Often companies and individuals will make investment and purchasing decisions based on their interest rate expectations, extrapolated from recent events. If the central bank's interest rate move is widely expected to be followed by other interest rate increases, investors and companies will act accordingly. Consumption, borrowing, and asset prices may all decline as a result of the revision in expectations.

A rise in the central bank's policy rate can reduce real domestic demand and net external demand (that is, the difference between export and import consumption) through a wide range of interconnected ways. Weaker total demand would tend to put downward pressure on the rate of domestic inflation—as would a stronger currency, which would reduce the prices of imports. Taken together, these factors might begin to put downward pressure on the overall measure of inflation.

To summarize, the central bank's policy rate works through the economy via one or more of the following interconnected channels:

- Short-term interest rates;
- Changes in the values of key asset prices;
- The exchange rate; and
- The expectations of economic agents.

QUESTION SET

1. Which of the following variables are *most likely* to be affected by a change in a central bank's policy rate?

- A. Asset prices only
- B. Expectations about future interest rates only
- C. Both asset prices and expectations about future interest rates

Solution:

C is correct. The price of equities, for example, might be affected by the expectation of future policy interest rate changes. In other words, a rate change may be taken as a signal of the future stance of monetary policy—contractionary or expansionary.

2. Which of the following does a central bank seek to influence directly via the setting of its official interest rate?

- A. Import prices
- B. Domestic inflation
- C. Inflation expectations

Solution:

C is correct. By setting its official interest rate, a central bank could expect to have a direct influence on inflation expectations—as well as on other market interest rates, asset prices, and the exchange rate (where this is freely floating). If it can influence these factors, it might ultimately hope to influence import prices (via changes in the exchange rate) and also domestically generated inflation (via its impact on domestic or external demand). The problem is that the workings of the transmission mechanism—from the official interest rate to inflation—are complex and can change over time.

3. Monetary policy is *least likely* to include:

- A. setting an inflation rate target.
- B. changing an official interest rate.
- C. enacting a transfer payment program.

Solution:

C is correct. Transfer payment programs represent fiscal, not monetary policy.

4. Which is the *most* accurate statement regarding central banks and monetary policy?

- A. Central bank activities are typically intended to maintain price stability.
- B. Monetary policies work through the economy via four independent channels.

- C. Commercial and interbank interest rates move inversely to official interest rates.

Solution:

A is correct. Central bank activities are typically intended to maintain price stability. B is not correct because the transmission channels of monetary policy are not independent.

4

MONETARY POLICY OBJECTIVES

- describe qualities of effective central banks; contrast their use of inflation, interest rate, and exchange rate targeting in expansionary or contractionary monetary policy; and describe the limitations of monetary policy

Inflation Targeting

Throughout the 1990s, a consensus began to build among both central bankers and politicians that the best way to control inflation and thereby maintain price stability was to target a certain level of inflation and to ensure that this target was met by monitoring a wide range of monetary, financial, and real economic variables. Today, inflation-targeting frameworks are the cornerstone of monetary policy and macroeconomic policy in many economies. Exhibit 4 shows the growth in the number of inflation-targeting monetary policy regimes over time.

The inflation-targeting framework that is now commonly practiced was pioneered in New Zealand. In 1988, the New Zealand Minister of Finance, Roger Douglas, announced that economic policy would focus on bringing inflation down from the prevailing level of around 6 percent to a target range of 0 to 2 percent. This goal was given legal status by the Reserve Bank of New Zealand Act 1989. As part of the Act, the Reserve Bank of New Zealand (RBNZ) was given the role of pursuing this target. The bank was given **operational independence**; it was free to set interest rates in the way that it thought would best meet the inflation target. Although the RBNZ had independent control of monetary policy, it was still accountable to the government and was charged with communicating its decisions in a clear and transparent way. As Exhibit 4 shows, the New Zealand model was widely copied.

Exhibit 4: The Progressive Adoption of Inflation Targeting by Central Banks

1989	New Zealand			
1990	Chile	Canada		
1991	Israel	United Kingdom		
1992	Sweden	Finland	Australia	
1995	Spain			
1998	Czech Republic	South Korea	Poland	
1999	Mexico	Brazil	Colombia	ECB
2000	South Africa	Thailand		

2001	Iceland	Norway	Hungary	Peru	Philippines
2005	Guatemala	Indonesia	Romania		
2006	Turkey	Serbia			
2007	Ghana				

Note: Spain and Finland later joined the EMU.

Sources: For 2001 and earlier, Truman (2003). For 2002 to 2007, Roger (2010).

Although these inflation-targeting regimes vary a little from economy to economy, their success is thought to depend on three key concepts: central bank independence, credibility, and transparency.

Central Bank Independence

In most cases, the central bank that is charged with targeting inflation has a degree of independence from its government. This independence is thought to be important. It is conceivable that politicians could announce an inflation target and direct the central bank to set interest rates accordingly. Indeed, this was the process adopted in the United Kingdom between 1994 and 1997. But politicians have a constant eye on reelection and might be tempted, for example, to keep rates “too low” in the lead-up to an election in the hope that this might help their reelection prospects. As a consequence, this might lead to higher inflation. Thus, it is now widely believed that monetary policy decisions should rest in the hands of an organization that is remote from the electoral process. The central bank is the natural candidate to be the monopoly supplier of a currency.

However, there are degrees of independence. For example, the head of the central bank is nearly always chosen by government officials. The chair of the US Federal Reserve’s Board of Governors is appointed by the president of the United States of America; the head of the ECB is chosen by the committee of Euro area finance ministers; and the governor of the Bank of England is chosen by the chancellor of the exchequer. So, in practice, separating control from political influence completely is probably an impossible (although a desirable) goal.

There are further degrees of independence. Some central banks are both operationally and **target independent**. This means that they not only decide the level of interest rates but also determine the definition of inflation that they target, the rate of inflation that they target, and the horizon over which the target is to be achieved. The ECB has independence of this kind. By contrast, other central banks—including those in New Zealand, Sweden, and the United Kingdom—are tasked to hit a definition and level of inflation determined by the government. Therefore, these central banks are only operationally independent.

Credibility

The independence of the central bank and public confidence in it are key in the design of an inflation-targeting regime.

To illustrate the role of credibility, suppose that instead of the central bank, the government assumes the role of targeting inflation, but the government is heavily indebted. Given that higher inflation reduces the real value of debt, the government would have an incentive to avoid reaching the inflation target or to set a high inflation target such that price stability and confidence in the currency could be endangered. As a result, few would believe the government was really intent on controlling inflation; thus, the government would lack credibility. Many governments have very large levels of debt, especially since the 2008–2009 Global Financial Crisis. In such a situation,

economic agents might expect a high level of inflation, regardless of the actual, stated target. The target might have little credibility if the organization's likelihood of sticking to it is in doubt.

If a respected central bank assumes the inflation-targeting role and if economic agents believe that the central bank will hit its target, this belief could become self-fulfilling. If everyone believes that the central bank will hit an inflation target of 2 percent next year, this expectation might be built into wage claims and other nominal contracts that would make it hit the 2 percent target. For this reason, central bankers pay a great deal of attention to inflation expectations. If these expectations were to rise rapidly, perhaps following a rapid increase in oil prices, unchecked expectations could get embedded into wage claims and eventually cause inflation to rise.

Transparency

One way to establish credibility is for a central bank to be transparent in its decision making. Many, if not all, independent inflation-targeting central banks produce a quarterly assessment of their economies. These **inflation reports**, as they are usually known, give central banks' views on the range of indicators that they watch when they come to their (usually) monthly interest rate decision. They will consider and outline their views on the following subjects, usually in this order:

- Broad money aggregates and credit conditions;
- Conditions in financial markets;
- Developments in the real economy (e.g., the labor market); and
- Evolution of prices.

Consideration of all of these important components of an economy is then usually followed by a forecast of growth and inflation over a medium-term horizon, usually two years.

By explaining their views on the economy and by being transparent in decision making, the independent, inflation-targeting central banks seek to gain reputation and credibility, making it easier to influence inflation expectations and hence ultimately easier to meet the inflation target.

The Target

Whether the target is set by the central bank or by the government for the central bank to hit, the level of the target and the horizon over which the target is to be hit is a crucial consideration in all inflation-targeting frameworks.

Exhibit 5: A Range of Inflation Targets

Country/Region

Australia	Australian Federal Reserve's target is inflation between 2% and 3%.
Canada	Bank of Canada's target is CPI inflation within the 1% to 3% range.
Euro area	ECB's target is CPI inflation close to, but below, a ceiling of 2%.
South Korea	Bank of Korea's target since 2019 has been CPI inflation of 2%.
New Zealand	The Reserve Bank of New Zealand's target is to keep future inflation between 1% and 3% with a focus on the average future inflation rate near 2%.

Country/Region

Sweden	Riksbank's target is CPI inflation within ± 1.0 percentage point of 2%.
United Kingdom	Bank of England's target is CPI inflation within ± 1.0 percentage point of 2%.

Note: CPI, consumer price index.

Source: Central bank websites (<http://www.bis.org/cbanks.htm>).

Exhibit 5 shows that many central banks in developed economies target an inflation rate of 2 percent based on a consumer price index (CPI). Given that the operation of monetary policy is both art and science, the banks are normally allowed a range around the central target of +1 percent or -1 percent. For example, with a 2 percent target, they would be tasked to keep inflation between 1 percent and 3 percent. But why target 2 percent and not 0 percent?

The answer is that aiming to hit 0 percent could result in negative inflation, known as **deflation**. One of the limitations of monetary policy that we will discuss is its ability or inability to deal with periods of deflation. If deflation is something to be avoided, why not target 10 percent? The answer to this question is that levels of inflation that high would not be consistent with price stability; such a high inflation rate would further tend to be associated with high inflation volatility and uncertainty. Central bankers seem to agree that 2 percent is far enough away from the risks of deflation and low enough not to lead to destabilizing inflation shocks.

Finally, we should keep in mind that the headline inflation rate that is announced in most economies every month, and which is the central bank's target, is a measure of how much a basket of goods and services has risen over the previous twelve months. It is history. Furthermore, interest rate changes made today will take some time to have their full effect on the real economy as they make their way through the monetary transmission mechanism. It is for these two reasons that inflation targeters do not target current inflation but instead usually focus on inflation two years ahead.

Although inflation-targeting mandates may vary from country to country, they have common elements: the specification of an explicit inflation target, with permissible bounds, and a requirement that the central bank should be transparent in its objectives and policy actions. This is all usually laid out in legislation that imposes statutory obligations on the central bank. As mentioned earlier, New Zealand pioneered the inflation-targeting approach to monetary policy that has since been copied widely. New Zealand's Policy Targets Agreement, which specifies the inflation-targeting mandate of its central bank, the RBNZ, is included in Example 2.

EXAMPLE 2**New Zealand's Policy Targets Agreement**

"This agreement between the Minister of Finance and the Governor of the Reserve Bank of New Zealand (the Bank) is made under section 9 of the Reserve Bank of New Zealand Act 1989 (the Act). The Minister and the Governor agree as follows:

Price stability

- a. Under Section 8 of the Act the Reserve Bank is required to conduct monetary policy with the goal of maintaining a stable general level of prices.

- b.** The Government's economic objective is to promote a growing, open and competitive economy as the best means of delivering permanently higher incomes and living standards for New Zealanders. Price stability plays an important part in supporting this objective.

Policy target

- a.** In pursuing the objective of a stable general level of prices, the Bank shall monitor prices as measured by a range of price indexes. The price stability target will be defined in terms of the All Groups Consumers Price Index (CPI), as published by Statistics New Zealand.
- b.** For the purpose of this agreement, the policy target shall be to keep future CPI inflation outcomes between 1 per cent and 3 per cent on average over the medium term.

Inflation variations around target

- a.** For a variety of reasons, the actual annual rate of CPI inflation will vary around the medium-term trend of inflation, which is the focus of the policy target. Amongst these reasons, there is a range of events whose impact would normally be temporary. Such events include, for example, shifts in the aggregate price level as a result of exceptional movements in the prices of commodities traded in world markets, changes in indirect taxes,⁹ significant government policy changes that directly affect prices, or a natural disaster affecting a major part of the economy.
- b.** When disturbances of the kind described in clause 3(a) arise, the Bank will respond consistent with meeting its medium-term target.

Communication, implementation, and accountability

- a.** On occasions when the annual rate of inflation is outside the medium-term target range, or when such occasions are projected, the Bank shall explain in Policy Statements made under section 15 of the Act why such outcomes have occurred, or are projected to occur, and what measures it has taken, or proposes to take, to ensure that inflation outcomes remain consistent with the medium-term target.
- b.** In pursuing its price stability objective, the Bank shall implement monetary policy in a sustainable, consistent, and transparent manner and shall seek to avoid unnecessary instability in output, interest rates and the exchange rate.
- c.** The Bank shall be fully accountable for its judgments and actions in implementing monetary policy.”

Source: Reserve Bank of New Zealand, <http://www.rbnz.govt.nz/>.

To summarize, an inflation-targeting framework normally has the following features:

- An independent and credible central bank;
- A commitment to transparency;
- A decision-making framework that considers a wide range of economic and financial market indicators; and
- A clear, symmetric, and forward-looking medium-term inflation target, sufficiently above 0 percent to avoid the risk of deflation but low enough to ensure a significant degree of price stability.

Indeed, independence, credibility, and transparency are arguably the crucial ingredients for an effective central bank, whether or not they target inflation.

The Main Exceptions to the Inflation-Targeting Rule

Although the practice of inflation targeting is widespread, two prominent central banks have not adopted a formal inflation target along the lines of the New Zealand model: the Bank of Japan (BoJ) and the US Federal Reserve System.

The Bank of Japan

Japan's central bank, the BoJ, does not target an explicit measure of inflation. Japan's government and its monetary authorities have been trying to combat deflation for much of the past two decades. However, despite their efforts—including the outright printing of money—deflation has remained very weak. Inflation targeting is seen very much as a way of combating and controlling inflation; as such, it would seem to have no place in an economy that suffers from persistent deflation.

Some economists have argued, however, that an inflation target is exactly what the Japanese economy needs. By announcing that positive inflation of say 3 percent is desired by the central bank, this might become a self-fulfilling prophecy if Japanese consumers and companies factor this target into nominal wage and price contracts. But for economic agents to believe that the target will be achieved, they have to believe that the central bank is capable of achieving it. Given that the BoJ has failed to engineer persistent, positive inflation, it is debatable how much credibility Japanese households and corporations would afford such an inflation-targeting policy.

The US Federal Reserve System

It is perhaps rather ironic that the world's most influential central bank, the US Federal Reserve, which controls the supply of the world's de facto reserve currency, the US dollar, does not have an explicit inflation target. However, it is felt that the single-minded pursuit of inflation might not be compatible with the Fed's statutory goal as laid out in the Federal Reserve Act, which charges the Fed's board to: "promote effectively the goals of maximum employment, stable prices, and moderate long-term interest rates."

In other words, it has been argued that inflation targeting might compromise the goal of "maximum employment." In practice, however, the Fed has indicated that it sees core inflation measured by the personal consumption expenditure (PCE) deflator of about, or just below, 2 percent as being compatible with "stable prices." Financial markets therefore watch this US inflation gauge very carefully to anticipate the rate actions of the Fed.

Monetary Policy in Developing Countries

Developing economies often face significant impediments to the successful operation of any monetary policy—that is, the achievement of price stability. These include:

- the absence of a sufficiently liquid government bond market and developed interbank market through which monetary policy can be conducted;
- a rapidly changing economy, making it difficult to understand what the neutral rate might be and what the equilibrium relationship between monetary aggregates and the real economy might be;
- rapid financial innovation that frequently changes the definition of the money supply;
- a poor track record in controlling inflation in the past, making monetary policy intentions less credible; and

- an unwillingness of governments to grant genuine independence to the central bank.

Taken together, any or all of these impediments might call into question the effectiveness of any developing economy's monetary policy framework, making any related monetary policy goals difficult to achieve.

QUESTION SET



1. The reason some inflation-targeting banks may target low inflation and not zero percent inflation is *best* described by which of the following statements?

- A. Some inflation is viewed as being good for an economy.
- B. It is very difficult to eliminate all inflation from a modern economy.
- C. Targeting zero percent inflation runs a higher risk of a deflationary outcome.

Solution:

C is correct. Inflation targeting is art, not science. Sometimes inflation will be above target and sometimes below. Were central banks to target zero percent, then inflation would almost certainly be negative on some occasions. If a deflationary mindset then sets in among economic agents, it might be difficult for the central bank to respond to this because they cannot cut interest rates much below zero.

2. The degree of credibility that a central bank is afforded by economic agents is important because:

- A. they are the lender of last resort.
- B. they set targets that can become self-fulfilling prophecies.
- C. they are the monopolistic suppliers of the currency.

Solution:

B is correct. If a central bank operates within an inflation-targeting regime and if economic agents believe that it will achieve its target, this expectation will become embedded into wage negotiations, for example, and become a self-fulfilling prophecy. Also, banks need to be confident that the central bank will lend them money when all other sources are closed to them; otherwise, they might curtail their lending drastically, leading to a commensurate reduction in money and economic activity.

3. A central bank that decides the desired levels of interest rates and inflation and the horizon over which the inflation objective is to be achieved is *most* accurately described as being:

- A. target independent and operationally independent.
- B. target independent but not operationally independent.
- C. operationally independent but not target independent.

Solution:

A is correct. The central bank described is target independent because it set its own targets (e.g., the target inflation rate) and operationally independent because it decides how to achieve its targets (e.g., the time horizon).

Exchange Rate Targeting

Many developing economies choose to operate monetary policy by targeting their currency's exchange rate, rather than an explicit level of domestic inflation. Such targeting involves setting a fixed level or band of values for the exchange rate against a major currency, with the central bank supporting the target by buying and selling the national currency in foreign exchange markets. There are recent examples of developed economies using such an approach. In the 1980s, following the failure of its policy of trying to control UK inflation by setting medium-term goals for money supply growth, the UK government decided to operate monetary policy such that the sterling's exchange rate equaled a predetermined value in terms of German deutschemarks. The basic idea is that by tying a domestic economy's currency to that of an economy with a credible policy of maintaining low inflation, the domestic economy would effectively "import" the inflation experience of the low-inflation economy.

Suppose that a developing country wished to maintain the value of its currency against the US dollar. The government or central bank would announce the currency exchange rate that they wished to target. To simplify matters, let us assume that the domestic inflation rates are very similar in both countries and that the monetary authorities of the developing economy have set an exchange rate target that is consistent with relative price levels in the two economies. Under these (admittedly unlikely) circumstances, in the absence of shocks, there would be no reason for the exchange rate to deviate significantly from this target level. As long as domestic inflation closely mirrors US inflation, the exchange rate should remain close to its target (or within a target band). It is in this sense that a successful exchange rate policy imports the inflation of the foreign economy.

Now suppose that economic activity in the developing economy starts to rise rapidly and that domestic inflation in the developing economy rises above the level in the United States. With a freely floating exchange rate regime, the currency of the developing economy would start to fall against the dollar. To arrest this fall, and to protect the exchange rate target, the developing economy's monetary authority sells foreign currency reserves and buys its own currency. This has the effect of reducing the domestic money supply and increasing short-term interest rates. The developing economy experiences a monetary policy tightening that, if expected to bring down inflation, will cause its exchange rate to rise against the dollar.

By contrast, in a scenario in which inflation in the developing country fell relative to the United States, the central bank would need to sell the domestic currency to support the target, tending to increase the domestic money supply and reduce the rate of interest.

In practice, the interventions of the developing economy central bank will simply stabilize the value of its currency, with many frequent adjustments. But this simplistic example should demonstrate one very important fact: *When the central bank or monetary authority chooses to target an exchange rate, interest rates and conditions in the domestic economy must adapt to accommodate this target and domestic interest rates and money supply can become more volatile.*

The monetary authority's commitment to and ability to support the exchange rate target must be credible for exchange rate targeting to be successful. If that is not the case, then speculators may trade against the monetary authority. Speculative attacks forced sterling out of the European Exchange Rate Mechanism in 1992. The fixed exchange rate regime was abandoned, and the United Kingdom allowed its currency to float freely. Eventually, the UK government adopted a formal inflation target in 1997. Similarly, in the Asian financial crisis of 1997–1998, Thailand's central bank tried to defend the Thai baht against speculative attacks for much of the first half of 1997 but

then revealed at the beginning of July that it had no reserves left. The subsequent devaluation triggered a debt crisis for banks and companies that had borrowed in foreign currency, and contagion spread throughout Asia.

Despite these risks, many currencies are pegged to other currencies, most notably the US dollar. Exhibit 6 shows a list of some of the currencies that were pegged to (fixed against) the US dollar at the end of 2018. Other currencies operate under a “managed exchange rate policy,” where they are allowed to fluctuate within a range that is maintained by a monetary authority through market intervention. Dollarization occurs when a country adopts the US dollar as their functional currency. This is stronger than pegging to the dollar because under dollarization, the US dollar replaces the previous national currency. Exhibit 6 breaks out countries that peg their currency to the dollar and those that have adopted the US dollar as their currency.

Exhibit 6: Select Currencies Pegged to the US Dollar, as of December 2018

Pegged to USD

- Bermuda
- Saudi Arabia
- Bahamas
- Qatar
- Lebanon
- United Arab Emirates
- Hong Kong SAR

Dollarized

- Panama (1904)
- El Salvador (2000)
- Ecuador (2000)
- Caribbean Netherlands (2011)
- East Timor (2001)

QUESTION SET



1. When the central bank chooses to target a specific value for its exchange rate:
 - A. it must also target domestic inflation.
 - B. it must also set targets for broad money growth.
 - C. conditions in the domestic economy must adapt to accommodate this target.

Solution:

C is correct. The adoption of an exchange rate target requires that the central bank set interest rates to achieve this target. If the target comes under pressure, domestic interest rates may have to rise, regardless of domestic conditions. It may have a “target” level of inflation in mind as well as “targets” for broad money growth, but as long as it targets the exchange rate, domestic inflation and broad money trends must simply be allowed to evolve.

2. With regard to monetary policy, what is the expected benefit of adopting an exchange rate target?
 - A. Freedom to pursue redistributive fiscal policy
 - B. Freedom to set interest rates according to domestic conditions

- C. Ability to “import” the inflation experience of the economy whose currency is being targeted

Solution:

C is correct. Note that interest rates have to be set to achieve this target and are therefore subordinate to the exchange rate target and partially dependent on economic conditions in the foreign economy.

3. Which of the following is *least* likely to be an impediment to the successful implementation of monetary policy in developing economies?

- A. Fiscal deficits
- B. Rapid financial innovation
- C. Absence of a liquid government bond market

Solution:

A is correct. Note that the absence of a liquid government bond market through which a central bank can enact open market operations and/or repo transactions will inhibit the implementation of monetary policy—as would rapid financial innovation because such innovation can change the relationship between money and economic activity. In contrast, fiscal deficits are not normally an impediment to the implementation of monetary policy, although they could be if they were perceived to be unsustainable.

4. A country that maintains a target exchange rate is *most likely* to have which outcome when its inflation rate rises above the level of the inflation rate in the target country?

- A. Increase in short-term interest rates
- B. Increase in the domestic money supply
- C. Increase in its foreign currency reserves

Solution:

A is correct. Interest rates are expected to rise to protect the exchange rate target.

Contractionary and Expansionary Monetary Policies and Their Limitations

Most central banks will adjust liquidity conditions by adjusting their official policy rate. When they believe that economic activity is likely to lead to an increase in inflation, they might increase interest rates, thereby reducing liquidity. In these cases, market analysts describe such actions as **contractionary** because the policy is designed to cause the rate of growth of the money supply and the real economy to contract (see Exhibit 3 for the possible transmission mechanism). Conversely, when the economy is slowing and inflation and monetary trends are weakening, central banks may increase liquidity by cutting their target rate. In these circumstances, monetary policy is said to be **expansionary**.

Thus, when policy rates are high, monetary policy may be described as contractionary; when low, it may be described as expansionary. But what are they “high” and “low” in comparison to?

The **neutral rate of interest** is often taken as the point of comparison. One way of characterizing the neutral rate is to say that it is that rate of interest that neither spurs on nor slows down the underlying economy. As such, when policy rates are

above the neutral rate, monetary policy is contractionary; when they are below the neutral rate, monetary policy is expansionary. The neutral rate should correspond to the average policy rate over a business cycle.

However, economists' views of the neutral rate for any given economy might differ, and therefore, their view of whether monetary policy is contractionary, neutral, or expansionary might differ too. What economists do agree on is that the neutral policy rate for any economy has two components:

- Real trend rate of growth of the underlying economy, and
- Long-run expected inflation.

The real trend rate of growth of an economy is also difficult to discern, but it corresponds to that rate of economic growth that is achievable in the long run that gives rise to stable inflation. If we are thinking about an economy with a credible inflation-targeting regime, where the inflation target is, say, 2 percent per year and where an analyst believes that the economy can grow sustainably over the long term at a rate of 2.5 percent per year, then they might also estimate the neutral rate to be:

$$\text{Neutral rate} = \text{Trend growth} + \text{Inflation target} = 2.5\% + 2.0\% = 4.5\%$$

The analyst would therefore describe the central bank's monetary policy as being contractionary when its policy rate is above 4.5 percent and expansionary when it is below this level.

In practice, central banks often indicate what they believe to be the neutral rate of interest for their economy too. But determining this "neutral rate" is more art than science. For example, many analysts have recently revised down their estimates of trend growth for many western countries following the collapse of the credit bubble, because in many cases, the governments and private individuals of these economies are now being forced to reduce consumption levels and pay down their debts.

What's the Source of the Shock to the Inflation Rate?

An important aspect of monetary policy for those charged with its conduct is the determination of the source of any shock to the inflation rate. Suppose that the monetary authority sees that inflation is rising beyond its target, or simply in a way that threatens price stability. If this rise was caused by an increase in the confidence of consumers and business leaders, which in turn has led to increases in consumption and investment growth rates, then we could think of it as being a **demand shock**. In this instance, it might be appropriate to tighten monetary policy to bring the inflationary pressures generated by these domestic demand pressures under control.

However, suppose instead that the rise in inflation was caused by a rise in the price of oil (for the sake of argument). In this case, the economy is facing a **supply shock**, and raising interest rates might make a bad situation worse. Consumers are already facing an increase in the cost of fuel prices that might cause profits and consumption to fall and eventually unemployment to rise. Putting up interest rates in this instance might simply exacerbate the oil price-induced downturn, which ultimately might cause inflation to fall sharply.

It is important, then, for the monetary authority to try to identify the source of the shock before engineering a contractionary or expansionary monetary policy phase.

Limitations of Monetary Policy

The limitations of monetary policy include problems in the transmission mechanism and the relative ineffectiveness of interest rate adjustment as a policy tool in deflationary environments.

Problems in the Monetary Transmission Mechanism

In Exhibit 3, we presented a stylized representation of the monetary policy transmission mechanism, including the channels of bank lending rates, asset prices, expectations, and exchange rates. The implication of the diagram is that there are channels through which the actions of the central bank or monetary authority are transmitted to both the nominal and real economy. In some occasions, however, the will of the monetary authority is not transmitted seamlessly through the economy.

Suppose that a central bank raises interest rates because it is concerned about the strength of underlying inflationary pressures. Long-term interest rates are influenced by the path of expected short-term interest rates, so the outcome of the rate hike will depend on market expectations. Suppose that bond market participants think that short-term rates are already too high, that the monetary authorities are risking a recession, and that the central bank will likely undershoot its inflation target. This fall in inflation expectations could cause long-term interest rates to fall. That would make long-term borrowing cheaper for companies and households, which could in turn stimulate economic activity rather than cause it to contract.

Arguably, the more credible the monetary authority, the more stable the long end of the yield curve; moreover, the monetary authority will be more confident that its “policy message” will be transmitted throughout the economy. A term recently used in the marketplace is **bond market vigilantes**. These “vigilantes” are bond market participants who might reduce their demand for long-term bonds, thus pushing up their yields, if they believe that the monetary authority is losing its grip on inflation. That yield increase could act as a brake on any loose monetary policy stance. Conversely, the vigilantes may push long-term rates down by increasing their demand for long-dated government bonds if they expect that tight monetary policy is likely to cause a sharp slowdown in the economy, thereby loosening monetary conditions for long-term borrowers in the economy.

A credible monetary policy framework and authority will tend not to require the vigilantes to do the work for it.

In very extreme instances, there may be occasions in which the demand for money becomes infinitely elastic—that is, where the demand curve is horizontal and individuals are willing to hold additional money balances without any change in the interest rate—so that further injections of money into the economy will not serve to further lower interest rates or affect real activity. This is known as a **liquidity trap**. In this extreme circumstance, monetary policy can become completely ineffective. The economic conditions for a liquidity trap are associated with the phenomenon of deflation.

Interest Rate Adjustment in a Deflationary Environment and Quantitative Easing as a Response

Deflation is a pervasive and persistent fall in a general price index and is more difficult for conventional monetary policy to deal with than inflation. This is because cutting nominal interest rates much below zero to stimulate the economy is difficult. For example, policy interest rates were cut below zero in several European countries in 2014 and subsequently in Japan in 2016. At this point, the economic conditions for a liquidity trap arise.

Deflation raises the real value of debt, while the persistent fall in prices can encourage consumers to put off consumption today, leading to a fall in demand that leads to further deflationary pressure. Thus a deflationary “trap” can develop, which is characterized by weak consumption growth, falling prices, and increases in real debt levels. Japan eventually found itself in such a position following the collapse of its property bubble in the early 1990s.

If conventional monetary policy—the adjustment of short-term interest rates—is no longer capable of stimulating the economy once the zero or even negative nominal interest rate bound has been reached, is monetary policy useless?

In the aftermath of the collapse of the high-tech bubble in November 2002, Federal Reserve Governor Ben Bernanke gave a speech entitled “Deflation: Making Sure ‘It’ Doesn’t Happen Here.” In this speech, Bernanke stated that inflation was always and everywhere a monetary phenomenon, and he expressed great confidence that by expanding the money supply by various means (including dropping it out of a helicopter on the population below), the Federal Reserve as the sole supplier of money could always engineer positive inflation in the US economy. He said:

I am confident that the Fed would take whatever means necessary to prevent significant deflation in the United States and, moreover, that the US central bank, in cooperation with other parts of the government as needed, has sufficient policy instruments to ensure that any deflation that might occur would be both mild and brief.

Following the 2008–2009 Global Financial Crisis, a number of governments along with their central banks cut rates to (near) zero, including those in the United States and the United Kingdom. However, there was concern that the underlying economies might not respond to this drastic monetary medicine, mainly because the related banking crisis had caused banks to reduce their lending drastically. To kick-start the process, both the Federal Reserve and the Bank of England effectively printed money and pumped it into their respective economies. This “unconventional” approach to monetary policy, known as **quantitative easing** (QE), is operationally similar to open market purchase operations but is conducted on a much larger scale.

The additional reserves created by central banks in a policy of quantitative easing can be used to buy any assets. The Bank of England chose to buy **gilts** (bonds issued by the UK government), where the focus was on gilts with three to five years maturity. The idea was that this additional reserve would kick-start lending, causing broad money growth to expand, which would eventually lead to an increase in real economic activity. But there is no guarantee that banks will respond in this way. In a difficult economic climate, it may be better to hold excess reserves rather than to lend to households and businesses that may default.

In the United States, the formal plan for QE mainly involved the purchase of mortgage bonds issued or guaranteed by Freddie Mac and Fannie Mae. Part of the intention was to push down mortgage rates to support the US housing market, as well as to increase the growth rate of broad money. Before implementing this formal program, the Federal Reserve intervened in several other markets that were failing for lack of liquidity, including interbank markets and the commercial paper market. These interventions had a similar effect on the Federal Reserve’s balance sheet and the money supply as the later QE program.

This first round of QE by the Federal Reserve was then followed by another round of QE, known as QE2. In November 2010, the Federal Reserve judged that the US economy had not responded sufficiently to the first round of QE (QE1). The Fed announced that it would create \$600 billion and use this money to purchase long-dated US Treasuries in equal tranches over the following eight months. The purpose of QE2 was to ensure that long bond yields remained low to encourage businesses and households to borrow for investment and consumption purposes, respectively.

The final round of QE, known as QE3, was implemented in September 2012 to provide \$40 billion per month to purchase agency mortgage-backed securities “until the labor market improved substantially.” QE3 lasted until December 2013, when the Federal Reserve announced it was tapering back on these purchases. These purchases, and quantitative easing, ended 10 months later in October 2014.

As long as central banks have the appropriate authority from the government, they can purchase any assets in a quantitative easing program. But the risks involved in purchasing assets with credit risk should be clear. In the end, the central bank is just a special bank. If it accumulates bad assets that then turn out to create losses, it could face a fatal loss of confidence in its main product: fiat money.

Limitations of Monetary Policy: Summary

The ultimate problem for monetary authorities as they try to manipulate the supply of money to influence the real economy is that they cannot control the amount of money that households and corporations put in banks on deposit, nor can they easily control the willingness of banks to create money by expanding credit. Taken together, this also means that they cannot always control the money supply. Therefore, there are definite limits to the power of monetary policy.

EXAMPLE 3

The Limits of Monetary Policy: The Case of Japan

The Background

Between the 1950s and 1980s, Japan's economy achieved faster real growth than any other G-7 economy. But the terrific success of the economy sowed the seeds of the problems that were to follow. The very high real growth rates achieved by Japan over four decades became built into asset prices, particularly equity and commercial property prices. Toward the end of the 1980s, asset prices rose to even higher levels when the BoJ followed a very easy monetary policy as it tried to prevent the Japanese yen from appreciating too much against the US dollar. However, when interest rates went up in 1989–1990 and the economy slowed, investors eventually came to believe that the growth assumptions that were built into asset prices and other aspects of the Japanese economy were unrealistic. This realization caused Japanese asset prices to collapse. For example, the Nikkei 225 stock market index reached 38,915 in 1989; by the end of March 2003, it had fallen by 80 percent to 7,972. The collapse in asset prices caused wealth to decline dramatically. Consumer confidence understandably fell sharply too, and consumption growth slowed. Corporate spending also fell, while bank lending contracted sharply in the weak economic climate. Although many of these phenomena are apparent in all recessions, the situation was made worse when deflation set in. In an environment when prices are falling, consumers may put off discretionary spending today until tomorrow; by doing this, however, they exacerbate the deflationary environment. Deflation also raises the real value of debts; as deflation takes hold, borrowers find the real value of their debts rising and may try to increase their savings accordingly. Once again, such actions exacerbate the recessionary conditions.

The Monetary Policy Response

Faced with such a downturn, the conventional monetary policy response is to cut interest rates to try to stimulate real economic activity. The Japanese central bank, the BoJ, cut rates from 8 percent in 1990 to 1 percent by 1996. By February 2001, the Japanese policy rate was cut to zero where it stayed.

Once rates are at or near zero, there are two broad approaches suggested by theory, though the two are usually complementary. First, the central bank can try to convince markets that interest rates will remain low for a long time, even after the economy and inflation pick up. This will tend to lower interest rates along the yield curve. Second, the central bank can try to increase the money

supply by purchasing assets from the private sector, so-called quantitative easing. The BoJ did both in 2001. It embarked on a program of quantitative easing supplemented by an explicit promise not to raise short-term interest rates until deflation had given way to inflation.

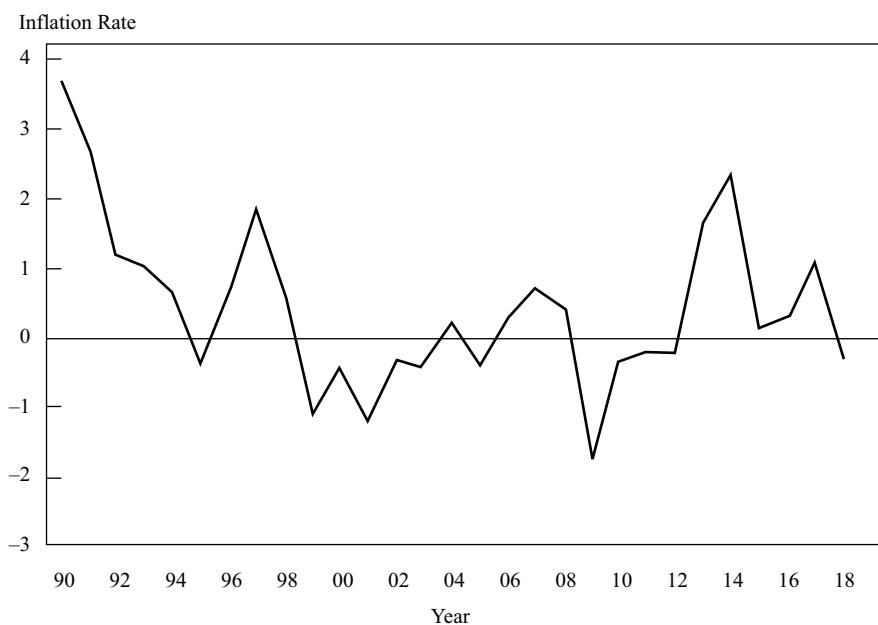
Quantitative easing simply involves the printing of money by the central bank. In practice, this involved the BoJ using open market operations to add reserves to the banking system through the direct purchase of government securities in the open market.

The reserve levels became the new target. The BoJ's monetary policy committee determined the level of reserves and the quantity of bond purchases that should be undertaken, rather than voting on the policy rate.

The success of this policy is difficult to judge. As Exhibit 7 shows, although deflation turned to inflation for a while, it returned to deflation in 2008–2009 when the Japanese economy suffered a sharp recession along with much of the rest of the world. At that time, having reversed its QE policy during 2004–2008 by reducing its bond holdings, the BoJ began to buy again.

The BoJ ramped up its asset purchases starting in 2013, when other central banks began to unwind their QE programs. At the beginning of 2013, BoJ Assets to Japanese GDP were approximately 30 percent, but rose to more than 170 percent in 2020. Economists debate the point, but arguably, even the BoJ's much larger program of QE has not been able to eliminate deflation. The Japanese experience suggests that there may be limits to the power of monetary policy.

Exhibit 7: Inflation and Deflation in Japan



Source: "Japan Annual and Monthly Inflation Tables," StatBureau, www.statbureau.org/en/japan/inflation-tables.

QUESTION SET

1. If an economy's trend GDP growth rate is 3 percent and its central bank has a 2 percent inflation target, which policy rate is *most consistent* with an expansionary monetary policy?
- 4 percent
 - 5 percent
 - 6 percent

Solution:

A is correct. The neutral rate of interest, which in this example is 5 percent, is considered to be that rate of interest that neither spurs on nor slows down the underlying economy. As such, when policy rates are above the neutral rate, monetary policy is contractionary; when they are below the neutral rate, monetary policy is expansionary. It has two components: the real trend rate of growth of the underlying economy (in this example, 3 percent) and long-run expected inflation (in this example, 2 percent).

2. An increase in a central bank's policy rate might be expected to reduce inflationary pressures by:
- reducing consumer demand.
 - reducing the foreign exchange value of the currency.
 - driving up asset prices leading to an increase in personal sector wealth.

Solution:

A is correct. If an increase in the central bank's policy rate is successfully transmitted through the money markets to other parts of the financial sector, consumer demand might decline as the rate of interest on mortgages and other credit rises. This decline in consumer demand should, all other things being equal and among other affects, lead to a reduction in upward pressure on consumer prices.

3. Which of the following statements *best* describes a fundamental limitation of monetary policy? Monetary policy is limited because central bankers:
- cannot control the inflation rate perfectly.
 - are appointed by politicians and are therefore never truly independent.
 - cannot control the amount of money that economic agents put in banks, nor the willingness of banks to make loans.

Solution:

C is correct. Central bankers do not control the decisions of individuals and banks that can influence the money creation process.

4. In theory, setting the policy rate equal to the neutral interest rate should promote:
- stable inflation.
 - balanced budgets.

- C. greater employment.

Solution:

A is correct. The neutral rate of interest is that rate of interest that neither stimulates nor slows down the underlying economy. The neutral rate should be consistent with stable long-run inflation.

5

INTERACTION OF MONETARY AND FISCAL POLICY



explain the interaction of monetary and fiscal policy

The Relationship Between Monetary and Fiscal Policy

Both monetary and fiscal policies can be used to try to influence the macroeconomy. But the impact of monetary policy on aggregate demand may differ depending on the fiscal policy stance. Conversely, the impact of fiscal policy might vary under various alternative monetary policy conditions. Clearly, policy makers need to understand this interaction. For example, they need to consider the impact of changes to the budget when monetary policy is accommodative as opposed to when it is restrictive: Can we expect the same impact on aggregate demand in both situations?

Although both fiscal and monetary policy can alter aggregate demand, they do so through differing channels with differing impact on the composition of aggregate demand. The two policies are not interchangeable. Consider the following cases in which the assumption is made that *wages and prices are rigid*:

- *Easy fiscal policy/tight monetary policy:* If taxes are cut or government spending rises, the expansionary fiscal policy will lead to a rise in aggregate output. If this is accompanied by a reduction in money supply to offset the fiscal expansion, then interest rates will rise and have a negative effect on private sector demand. We have higher output and higher interest rates, and government spending will be a larger proportion of overall national income.
- *Tight fiscal policy/easy monetary policy:* If a fiscal contraction is accompanied by expansionary monetary policy and low interest rates, then the private sector will be stimulated and will rise as a share of GDP, while the public sector will shrink.
- *Easy monetary policy/easy fiscal policy:* If both fiscal and monetary policy are easy, then the joint impact will be highly expansionary—leading to a rise in aggregate demand, lower interest rates (at least if the monetary impact is larger), and growing private and public sectors.
- *Tight monetary policy/tight fiscal policy:* Interest rates rise (at least if the monetary impact on interest rates is larger) and reduce private demand. At the same time, higher taxes and falling government spending lead to a drop in aggregate demand from both public and private sectors.

Factors Influencing the Mix of Fiscal and Monetary Policy

Although governments are concerned about stabilizing the level of aggregate demand at close to the full employment level, they are also concerned with the growth of potential output. To this end, encouraging private investment will be important. It may best be achieved by accommodative monetary policy with low interest rates and a tight fiscal policy to ensure free resources for a growing private sector.

At other times, the lack of a good quality, trained workforce—or perhaps a modern capital infrastructure—will be seen as an impediment to growth; thus, an expansion in government spending in these areas may be seen as a high priority. If taxes are not raised to pay for this, then the fiscal stance will be expansionary. If a loose monetary policy is chosen to accompany this expansionary spending, then it is *possible* that inflation may be induced. Of course, it is an open question as to whether policy makers can judge the appropriate levels of interest rates or fiscal spending levels.

Clearly, the mix of policies will be heavily influenced by the political context. A weak government may raise spending to accommodate the demands of competing vested interests (e.g., subsidies to particular sectors, such as agriculture in the EU), and thus a restrictive monetary policy may be needed to hold back the possibly inflationary growth in aggregate demand through raised interest rates and less credit availability.

Both fiscal and monetary policies suffer from the lack of precise knowledge of where the economy is today, because data initially appear to be subject to revision and to have a time lag. However, fiscal policy suffers from two further issues with regard to its use in the short run.

As we saw earlier, it is difficult to implement quickly because spending on capital projects takes time to plan, procure, and put into practice. In addition, it is politically easier to loosen fiscal policy than to tighten it; in many cases, automatic stabilizers are the source of fiscal tightening, because tax rates are not changing and political opposition is muted. Similarly, the independence of many central banks means that decisions on raising interest rates are outside the hands of politicians and thus can be taken more easily.

The interaction between monetary and fiscal policies was implicit in our discussion of Ricardian equivalence because if tax cuts have no impact on private spending as individuals anticipate future higher taxes, then clearly this may lead policy makers to favor monetary tools.

Ultimately, the interaction of monetary and fiscal policies in practice is an empirical question, which we also touched on earlier. In their detailed research paper using the International Monetary Fund's (IMF'S) Global Integrated Monetary and Fiscal Model, IMF researchers examined four forms of coordinated global fiscal loosening over a two-year period, which will be reversed gradually after the two years are completed. These are:

- an increase in social transfers to all households,
- a decrease in tax on labor income,
- a rise in government investment expenditure, and
- a rise in transfers to the poorest in society.

The two types of monetary policy responses considered are:

- no monetary accommodation, so rising aggregate demand leads to higher interest rates immediately; or
- interest rates are kept unchanged (accommodative policy) for the two years.

The following important policy conclusions from this study emphasize the role of policy interactions:

- *No monetary accommodation:* Government spending increases have a much bigger effect (six times bigger) on GDP than similar size social transfers because the latter are not considered to be permanent, although real interest rates rise as monetary authorities react to rises in aggregate demand and inflation. Targeted social transfers to the poorest citizens have double the effect of the non-targeted transfers, while labor tax reductions have a slightly bigger impact than the latter.
- *Monetary accommodation:* Except for the case of the cut in labor taxes, fiscal multipliers are now much larger than when there is no monetary accommodation. The cumulative multiplier (i.e., the cumulative effect on real GDP over the two years divided by the percentage of GDP, which is a fiscal stimulus) is now 3.9 for government expenditure compared with 1.6 with no monetary accommodation. The corresponding numbers for targeted social transfer payments are 0.5 without monetary accommodation and 1.7 with it. The larger multiplier effects with monetary accommodation result from rises in aggregate demand and inflation, leading to falls in real interest rates and additional private sector spending (e.g., on investment goods). Labor tax cuts are less positive.

Quantitative Easing and Policy Interaction

What about the scenario of zero interest rates and deflation? Fiscal stimulus should still raise demand and inflation, lowering real interest rates and stimulating private sector demand. We saw earlier that quantitative easing was a feature of major economies following the 2008–2009 Global Financial Crisis. This involved the purchase of government or private securities by the central bank from individuals, institutions, or banks and substituting central bank balances for those securities. The ultimate aim was for recipients to subsequently increase expenditures, lending or borrowing in the face of raised cash balances and lower interest rates.

If the central bank purchases government securities on a large scale, it is effectively funding the budget deficit and the independence of monetary policy is an illusion. This so-called printing of money is feared by many economists as the monetization of the government deficit. Note that it is unrelated to the conventional inflation target of central banks, such as the Bank of England. Some economists question whether an independent central bank should engage in such activity.

The Importance of Credibility and Commitment

The IMF model implies that if governments run persistently high budget deficits, real interest rates rise and crowd out private investment, reducing each country's productive potential. As individuals realize that deficits will persist, inflation expectations and long-term interest rates rise: This reduces the effect of the stimulus by half.

Further, if there is a real lack of commitment to fiscal discipline over the longer term, (e.g., because of aging populations) and the ratio of government debt to GDP rose by 10 percentage points permanently in the United States alone, then world real interest rates would rise by 0.14 percent—leading to a 0.6 percent permanent fall in world GDP.

QUESTION SET

1. In a world in which Ricardian equivalence holds, governments would *most likely* prefer to use monetary rather than fiscal policy because under Ricardian equivalence:
- real interest rates have a more powerful effect on the real economy.
 - the transmission mechanism of monetary policy is better understood.
 - the future impact of fiscal policy changes are fully discounted by economic agents.

Solution:

C is correct. If Ricardian equivalence holds, then economic agents anticipate that the consequence of any current tax cut will be future tax rises, which leads them to increase their saving in anticipation of this so that the tax cut has little effect on consumption and investment decisions. Governments would be forced to use monetary policy to affect the real economy on the assumption that money neutrality did not hold in the short term.

2. If fiscal policy is easy and monetary policy tight, then:
- interest rates would tend to fall, reinforcing the fiscal policy stance.
 - the government sector would tend to shrink as a proportion of total GDP.
 - the government sector would tend to expand as a proportion of total GDP.

Solution:

C is correct. With a tight monetary policy, real interest rates should rise and reduce private sector activity, which could be at least partially offset by an expansion in government activity via the loosening of fiscal policy. The net effect, however, would be an expansion in the size of the public sector relative to the private sector.

3. Which of the following has the greatest impact on aggregate demand according to an IMF study? 1 percent of GDP stimulus in:
- government spending
 - rise in transfer benefits
 - cut in labor income tax across all income levels

Solution:

A is correct. The study clearly showed that direct spending by the government leads to a larger impact on GDP than changes in taxes or benefits.

4. Given an independent central bank, monetary policy actions are *more likely* than fiscal policy actions to be:
- implementable quickly.
 - effective when a specific group is targeted.
 - effective when combating a deflationary economy.

Solution:

A is correct. Monetary actions may face fewer delays to taking action than fiscal policy, especially when the central bank is independent.

5. Which policy alternative is *most likely* to be effective for growing both the public and private sectors?

- A. Easy fiscal/easy monetary policy
- B. Easy fiscal/tight monetary policy
- C. Tight fiscal/tight monetary policy

Solution:

A is correct. If both fiscal and monetary policies are “easy,” then the joint impact will be highly expansionary, leading to a rise in aggregate demand, low interest rates, and growing private and public sectors.

PRACTICE PROBLEMS

1. When a central bank announces a decrease in its official policy rate, the desired impact is an increase in:
 - A. investment.
 - B. interbank borrowing rates.
 - C. the national currency's value in exchange for other currencies.

2. Which action is a central bank *least likely* to take if it wants to encourage businesses and households to borrow for investment and consumption purposes?
 - A. Sell long-dated government securities
 - B. Purchase long-dated government treasuries
 - C. Purchase mortgage bonds or other securities

3. A central bank's repeated open market purchases of government bonds:
 - A. decreases the money supply.
 - B. is prohibited in most countries.
 - C. is consistent with an expansionary monetary policy.

4. A prolonged period of an official interest rate very close to zero without an increase in economic growth *most likely* suggests:
 - A. quantitative easing must be limited to be successful.
 - B. the effectiveness of monetary policy may be limited.
 - C. targeting reserve levels is more important than targeting interest rates.

5. Raising the reserve requirement is *most likely* an example of which type of monetary policy?
 - A. Neutral
 - B. Expansionary
 - C. Contractionary

6. Which of the following is a limitation on the ability of central banks to stimulate growth in periods of deflation?
 - A. Ricardian equivalence
 - B. Interaction of monetary and fiscal policy
 - C. Interest rates cannot fall significantly below zero

7. The *least likely* limitation to the effectiveness of monetary policy is that central banks cannot:
accurately determine the neutral rate of interest.

- A. regulate the willingness of financial institutions to lend.
 - B. control amounts that economic agents deposit into banks.
8. Quantitative easing, the purchase of government or private securities by the central banks from individuals or institutions, is an example of which monetary policy stance?
- A. Neutral
 - B. Expansionary
 - C. Contractionary

SOLUTIONS

1. A is correct. Investment is expected to move inversely with the official policy rate.
2. A is correct. Such action would tend to constrict the money supply and increase interest rates, all else being equal.
3. C is correct. The purchase of government bonds through open market operations increases banking reserves and the money supply; it is consistent with an expansionary monetary policy.
4. B is correct. A central bank would decrease an official interest rate to stimulate the economy. The setting in which an official interest rate is lowered to zero (or even slightly below zero) without stimulating economic growth suggests that there are limits to monetary policy.
5. C is correct. Raising reserve requirements should slow money supply growth.
6. C is correct. Deflation poses a challenge to conventional monetary policy because once the central bank has cut nominal interest rates to zero (or slightly less than zero) to stimulate the economy, they cannot cut them further.
7. A is correct. The inability to determine exactly the neutral rate of interest does not necessarily limit the power of monetary policy.
8. B is correct. Quantitative easing is an example of an expansionary monetary policy stance. It attempts to spur aggregate demand by drastically increasing the money supply.

LEARNING MODULE

5

Introduction to Geopolitics

by Goodwin Lauren, Nair-Reichert Usha, PhD, and Witschi Daniel Rober, PhD.

Lauren Goodwin, CFA, is at New York Life Investments (USA). Usha Nair-Reichert, PhD, is at Georgia Institute of Technology (USA). Daniel Robert Witschi, PhD, CFA, is at Dreyfus Sons & Co Ltd. (Switzerland).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	describe geopolitics from a cooperation versus competition perspective
<input type="checkbox"/>	describe geopolitics and its relationship with globalization
<input type="checkbox"/>	describe functions and objectives of the international organizations that facilitate trade, including the World Bank, the International Monetary Fund, and the World Trade Organization
<input type="checkbox"/>	describe geopolitical risk
<input type="checkbox"/>	describe tools of geopolitics and their impact on regions and economies
<input type="checkbox"/>	describe the impact of geopolitical risk on investments

INTRODUCTION

1

Investors study geopolitics and geopolitical risk because they can have a material impact on investment outcomes. These relations affect key drivers of investment performance, including economic growth, business performance, market volatility, and transaction costs. On a portfolio level, geopolitical risk can affect the suitability of a security or strategy for an investor's goals, risk tolerance, and time horizon. In this learning module, we will build a framework by which investors can measure, assess, track, and react to geopolitical risk, with a goal of improving investment outcomes.

LEARNING MODULE OVERVIEW

- Geopolitics is the study of how geography affects politics and international relations. Within the field of geopolitics, analysts study actors—the individuals, organizations, companies, and national governments that carry out political, economic, and financial activities—and how they interact with one another.
- State actors can be cooperative or non-cooperative. A country may want to cooperate with its neighbors or with other state actors for many reasons. These reasons are typically defined by a country's national interest—its goals and ambitions—whether they be military, economic, or cultural.
- The cooperation and engagement among countries is also affected by its resource endowment, standardization of the rules of engagement, and cultural factors and soft power.
- A country's national interest can be viewed as a hierarchy of factors, with those essential for survival at the top of the hierarchy and nice-but-not-essential elements lower in the hierarchy. Governments use the hierarchy of interests to guide their behavior.
- Political cooperation versus non-cooperation is only one lens through which geopolitical actors engage with the world, but it is an important one for understanding countries' priorities.
- Globalization is marked by economic and financial cooperation, including the active trade of goods and services, capital flows, currency exchange, and cultural and information exchange. By contrast, antiglobalization or nationalism is the promotion of a country's own economic interests to the exclusion or detriment of the interests of other nations. Nationalism is marked by limited economic and financial cooperation.
- Globalization provides potential gains, such as:
 - increased profits—through increasing sales and/or reducing costs,
 - access to resources—market access and investment opportunities, and
 - intrinsic gains—an improved quality of life.
- Globalization also has some potential drawbacks, such as:
 - unequal economic and financial gains,
 - interdependence that can lead to supply chain disruption, and
 - possible exploitation of social and environmental resources.
- The International Monetary Fund's (IMF's) main mandate is to ensure the stability of the international monetary system, the system of exchange rates and international payments that enables countries to buy goods and services from each other.
- The World Bank's main objective is to help developing countries fight poverty and enhance environmentally sound economic growth.
- The World Trade Organization (WTO) provides the legal and institutional foundation of the multinational trading system. It regulates cross-border trade relationships among nations on a global scale.

- A geopolitical framework for analysis includes four archetypes of country behavior: autarky, hegemony, multilateralism, and bilateralism. Each archetype has its own costs, benefits, and trade-offs with respect to geopolitical risk.
- Geopolitical risk is the risk associated with tensions or actions between actors (state and non-state) that affect the normal and peaceful course of international relations. Geopolitical risk tends to rise when the geographic and political factors underpinning country relations shift.
- The tools of geopolitics may be separated into the following three types:
 - national security tools,
 - economic tools, and
 - financial tools.
- The most extreme example of a national security tool is that of armed conflict. Espionage is an indirect national security tool. Military alliances are often used either to aid in direct conflict or to deter conflict from arising in the first place.
- Economic tools are used to reinforce a cooperative or non-cooperative stance through economic means. Among state actors, economic tools can include multilateral trade agreements or the global harmonization of tariff rules. Economic tools also can be non-cooperative in nature. Nationalization is a non-cooperative approach to asserting economic control.
- Financial tools are the actions used to reinforce a cooperative or non-cooperative stance through financial mechanisms. Examples of cooperative financial tools include the free exchange of currencies across borders and allowing foreign investment. Examples of non-cooperative financial tools include limiting access to local currency markets and restricting foreign investment.
- There are three basic types of geopolitical risk:
 - event risk,
 - exogenous risk, and
 - thematic risk.
- Event risk evolves around set dates known in advance. Political events, for example, often result in changes to investor expectations related to a country's cooperative stance. Brexit is an example of event risk.
- Exogenous risk is a sudden or unanticipated risk that can affect either a country's cooperative stance, the ability of non-state actors to globalize, or both. Examples include sudden uprisings, invasions, or the aftermath of natural disasters.
- Thematic risks are known risks that evolve and expand over time. Climate change, cyber threats, and the ongoing threat of terrorism fall into this category.
- To make an assessment, an investor considers geopolitical risk in terms of the following three areas:
 - likelihood it will occur,
 - velocity (speed) of its impact, and
 - size and nature of that impact.

- Geopolitical risks seldom develop in linear fashion, making it difficult to monitor and forecast their likelihood, velocity, and size and nature of impact on a portfolio. As a result, many investors deploy an approach that includes scenario building and signposting rather than a single point forecast.
- Investors study geopolitical risk because it has a tangible impact on investment outcomes. On a macroeconomic level, these risks can affect capital markets conditions, such as economic growth, interest rates, and market volatility.
- Changes in capital markets conditions can have an important influence on asset allocation decisions, including an investor's choice of geographic exposures.
- On a portfolio level, geopolitical risk can influence the appropriateness of an investment security or strategy for an investor's goals, risk tolerance, and time horizon.

2

NATIONAL GOVERNMENTS AND POLITICAL COOPERATION



describe geopolitics from a cooperation versus competition perspective

The international environment is constantly evolving. Such trends as the growth of emerging market economies, globalization, and the rise of populism affect the range of opportunities and threats that companies, industries, nations, and regional groups face. **Geopolitics** is the study of how geography affects politics and international relations. Within the field of geopolitics, analysts study actors—the individuals, organizations, companies, and national governments that carry out political, economic, and financial activities—and how they interact with one another. The role of state and non-state actors is discussed in the following section.

State and Non-State Actors

Relationships within and among countries can be complex. To begin, many actors influence international relations, political developments, and economic affairs. In the introduction, we defined actors as the individuals, organizations, companies, and national governments that carry out political, economic, and financial activities. This definition can be divided into two types of actors relevant for geopolitical risk: state actors and non-state actors. **State actors** are typically national governments, political organizations, or country leaders that exert authority over a country's national security and resources. The South African president, sultan of Brunei, Malaysia's parliament, and the British Prime minister are all examples of state actors. **Non-state actors** are those that participate in global political, economic, or financial affairs but do not directly control national security or country resources. Examples of non-state actors are non-governmental organizations (NGOs), multinational companies, charities, and even influential individuals, such as business leaders or cultural icons.

These actors are influenced not only by their relationship with one another but also by factors affecting their other allies and adversaries. For example, if Country A has a cooperative relationship with Country B but Country B attacks Country C, a close ally of Country A, then the relationship between Countries A and B may become strained or even broken. The opposite is also true: A country's cooperation across many political, economic, or financial channels may increase trust over time. As a result, the international system is composed of multifaceted webs of affairs that can shape events and decisions as well as economic and market outcomes. Although a "one-size-fits-all" model does not exist for geopolitical actors, understanding and categorizing the threats and opportunities that a country faces may help us to gauge the likelihood that geopolitical risk will arise.

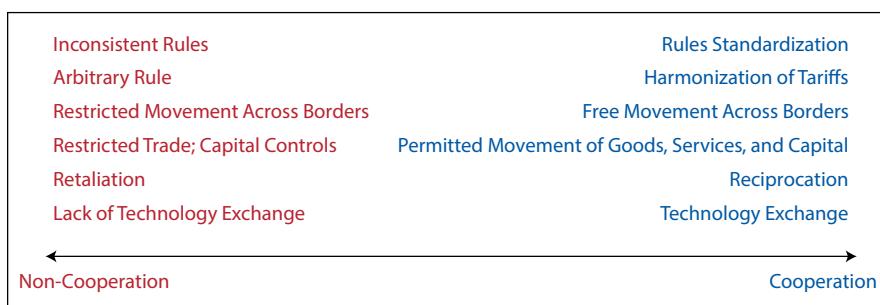
Countries and their governments are heavily influenced by economic, financial, and national security considerations as well as by social and cultural factors and non-state actors. In fact, economic and financial considerations are inextricably linked with a country's national interest and political dynamic. That said, it is useful to isolate one of these areas—national actors and political cooperation—as a starting point in understanding actors' interwoven goals and the way they may affect geopolitical risk. In the next section, we will develop a sense of state actors and their motivations for cooperation (versus non-cooperation) before adding layers of geopolitical risk analysis.

Features of Political Cooperation

At the highest level, relations between countries or national governments (state actors) can be cooperative or competitive in nature. **Cooperation** is the process by which countries work together toward some shared goal or purpose. These goals may, and often do, vary widely—from strategic or military concerns to economic influence or cultural preferences. Given the expansive nature of country goals and interests, their interactions with one another also may be complex, generating the potential for geopolitical risk to arise.

We begin that framework by exploring one specific type of cooperation—political cooperation—which is the degree to which countries work toward agreements on rules and standardization for the activities and interactions between them. Within this context, a **cooperative country** is one that engages and reciprocates in rules standardization; harmonization of tariffs; and international agreements on trade, immigration, or regulation and that also allows for the free flow of information, including technology transfer. In contrast, a **non-cooperative country** is one with inconsistent and even arbitrary rules; restricted movement of goods, services, people, and capital across borders; retaliation; and limited technology exchange. We show the degree of political cooperation in Exhibit 1

Exhibit 1: Political Cooperation and Non-Cooperation



A country may want to cooperate with its neighbors or with other state actors for many reasons. These reasons are typically defined by a country's national interest—its goals and ambitions—whether they be military, economic, or cultural.

National Security or Military Interest

National security or national defense involves protecting a country, including its citizens, economy, and institutions, from external threat. These threats may be broad in nature—from military attacks and terrorism, to crime, cybersecurity, and even natural disasters.

Geographic factors play an important role in shaping a country's approach to national security and the extent to which it will choose a cooperative approach. Landlocked countries, such as Switzerland, rely extensively on their neighbors for access to vital resources. This reliance may make cooperation more important for sustaining international access and growth or even for survival. By the same token, countries highly connected to trade routes, such as Singapore, or countries acting as a conduit for trade, such as Panama, may use their geographic location as a lever of power in broader international dynamics.

Economic Interest

Over time, the concept of national security has expanded to include economic factors, including access to such resources as energy, food, or water. On a domestic level, growing national wealth and limiting income inequality can contribute to social stability, another important component of national security. On the international level, the ability of national firms to operate on a global scale is increasingly important as well. In this context, countries that cooperate in support of their economic interest are likely focused on one of two factors: Either they would like to secure essential resources through trade, or they would like to level the global playing field for their companies or industries through standardization.

EXAMPLE 1

Factors Affecting Cooperation

1. True or false: Geographic factors do not influence the extent to which a country will choose a cooperative approach. Explain your reasoning.

A. True

B. False

Solution:

B is correct. Geographic factors play an important role in shaping a country's approach to national security and the extent to which it will choose a cooperative approach. A country with limited access to resources because of its geographic characteristics might be more inclined to cooperate as it needs access to resources to thrive, whereas a country rich in internal resources might use its resources as leverage to influence international dynamics. Countries with direct access to the sea have an advantage related to trade and transport as well as the natural resources provided by the ocean.

EXAMPLE 2**Factors Affecting Cooperation**

1. Fill in the blank: Two factors on which a cooperative country is likely focused include _____ and _____.

Solution:

Two factors on which a cooperative country is likely focused include *trade* and *standardization*.

Resource Endowment, Standardization, and Soft Power

The cooperation and engagement among countries is also affected by its resource endowment, standardization of the rules of engagement, and cultural factors and soft power. These factors are examined in the section that follows.

Geophysical Resource Endowment

At a basic level, **geophysical resource endowment** includes such factors as livable geography and climate as well as access to food and water, which are necessary for sustainable growth. Geophysical resource endowment is highly unequal among countries. Some countries, such as the United States, Russia, Australia, and China, are relatively self-sufficient in their resource use. Others, such as Western Europe, Japan, and Turkey, are highly reliant on others for key factors of production, such as fossil fuels. Still others, such as Saudi Arabia, have a plentiful endowment of fossil fuels but rely on others for many basic needs.

These different starting points create power dynamics that can affect the terms of engagement between states. A country heavily endowed with a resource may find itself with more political leverage when dealing with another country in desperate need of that resource. At the same time, a resource-rich country may become vulnerable if the use or sale of the resource benefits certain groups more than others, therefore contributing to internal political instability.

Standardization

Economic and financial activities may cross borders with or without explicit government support. As they do, governments have more incentive to cooperate with others in standardizing the rules of engagement. **Standardization** is the process of creating protocols for the production, sale, transport, or use of a product or service. Standardization occurs when relevant parties agree to follow these protocols together. It helps support expanded economic and financial activities across borders, such as trade and capital flows, which support higher economic growth and standards of living. Rules standardization can take many forms—from regulatory cooperation, to process standardization, to operational synchronization. Rules standardization may also be driven by non-state entities, such as industry groups or organizations. Exhibit 2 provides examples of standardization.

Exhibit 2: Types of Rules Standardization

	Regulatory Cooperation	Process Standardization	Operational Synchronization
Challenge	As financial cooperation expands, countries need a comprehensive standard for governance and risk management of the banking sector.	Financial transactions across borders faced higher costs and longer wait times, increasing the burden for cross-border activity.	Increasing international trade created supply chain bottlenecks as containers of different sizes and shapes were sent to ports worldwide.
Solution	Basel Committee on Banking Supervision (BCBS)	Society for Worldwide Interbank Financial Telecommunication (SWIFT)	Containerization
Process	The BCBS was established in 1974 by the G–10 banking authorities. Membership has since expanded to the G–20.	SWIFT was established in 1973 to provide a global financial infrastructure.	Standards set for containers of uniform size and shape using multimodal forms of transport (land, sea, air, rail) and port cranes.
Benefit	Allows for more effective supervision of the global banking sector and international capital flows.	Facilitates global payments in more than 200 countries and territories, servicing more than 11,000 institutions worldwide.	Dramatically reduces the time and cost of shipping goods.

Cultural Considerations and Soft Power

Finally, countries may have cultural reasons for cooperating with others. These could be historical in nature, such as long-standing political ties, immigration patterns, shared experiences, or cultural similarities. In other cases, countries may engage in **soft power**, a means of influencing another country's decisions without force or coercion. Soft power can be built over time through such actions as cultural programs, advertisement, travel grants, and university exchange. For example, the European Union (EU) Erasmus+ program provides funding for exchange programs in education and sport to drive social inclusion and participation between participants from different countries and also to promote such priority policies as the EU's green and digital transitions. In another example, countries like South Korea advertise visiting Seoul, the country's capital city, in subway systems globally. These advertisements use popular Korean-made products, musical acts, and actors to encourage interaction with Korean culture and business.

The Role of Institutions

An **institution** is an established organization or practice in a society or culture. An institution can be a formal structure, such as a university, organization, or process backed by law; or, it can be informal, such as customs or behavioral patterns important to a society. Institutions can, but need not be, formed by national governments. Examples of institutions include NGOs, charities, religious customs, family units, the media, political parties, and educational practice.

Generally, strong institutions contribute to more stable internal and external political forces. That consistency, in turn, gives a country more opportunity to develop cooperative relationships. Countries with strong institutions—including organizations and structures promoting government accountability, rule of law, and property rights—allow them to act with more authority and independence in the international space. Stronger institutions also make cooperative relationships more durable. By integrating a cooperative relationship in multiple layers of society, strong institutions can reduce the likelihood that a country defects from its cooperative roles.

Hierarchy of Interests and Costs of Cooperation

The national interest of a country is its set of goals and ambitions. For some countries, national interest is viewed primarily in geospatial terms—the need for self-determination, survival as a nation state, the need for clear national borders, or expansion of the nation state. For others, national interest incorporates a wide array of interrelated factors, including the economic and social considerations discussed earlier. This broad approach can create conflicts among a country's many important needs, which complicates the assessment of geopolitical actors and their motivations.

We can think of a country's national interest as a hierarchy of factors, with those essential for survival at the top of the hierarchy and nice-but-not-essential elements lower in the hierarchy. Governments use the hierarchy of interests to guide their behavior. They will choose to cooperate where it benefits the nation-state, but when two needs result in conflicting cooperation tactics, those higher on the hierarchy are prioritized. For example, cooperation in the form of tariff harmonization may benefit the country on a stand-alone basis; however, if those countries are in a military conflict, then there is a higher cost to cooperation. In this case, if military determination is higher on the countries' hierarchies, then cooperation may not be in their national interest, despite the potential benefits.

Exhibit 3 shows how countries prioritize their hierarchy of interests. Every country has different resources, goals, and leadership and thus different priorities, as well. In the exhibit, Country A prioritizes access to food and water. In contrast, Country B, prioritizes independence from foreign influence. Additionally, these priorities may shift as political leadership turns over or as global events change. For investors, it is important to understand how these resources, goals, and leadership styles may interact or even conflict with one another over time. When countries' goals misalign or change, it may give rise to geopolitical risk.

Exhibit 3: Hierarchy of Interests



Power of the Decision Maker

Some elements on the hierarchy of national interests may appear clear-cut; securing access to food and water may take precedence over funding a cultural program, for example. However, as basic societal needs are met, the hierarchy of national interests can become more subjective. One government may treat the prioritization of some interests—such as military buildup or providing health care—very differently from its predecessor. How governments weigh those issues will determine the depth and nature of political cooperation.

The length of a country's political cycle has an important impact on priority designation. Many countries have political cycles of just a few years, which means that long-term risks like climate change or addressing income inequality can be difficult

to prioritize against projects or goals that can be achieved in a short-term horizon. Intrinsic in this reality is that governmental decision makers, whether political parties or individuals, have their own set of influences and needs. Although we will not explore this idea in depth, it is important to acknowledge that, for the purpose of geopolitical risk analysis, decision makers' motivations can affect a country's cooperative and non-cooperative choices. This introduces a factor of psychology and non-predictability into choices along the hierarchy of a nation's needs that can shape geopolitical relationships.

EXAMPLE 3

Country's Political Cycle

1. True or false: The length of a country's political cycle does not affect its priority designation.

Explain your reasoning.

- A. True
- B. False

Solution:

B is correct. The length of a country's political cycle has an important impact on priority designation as governments with shorter cycles have little incentive to focus on longer-term priorities, even if those priorities would be beneficial for society. Australia, for example, has shorter than average parliamentary terms, which could potentially affect governmental priorities.

Political Non-Cooperation

We consider political cooperation and non-cooperation as existing along a spectrum. While it is in some countries' interest to be highly politically cooperative, for others it is less essential. Over time, most countries cooperate on standardized rules on an international scale. In some instances of extreme non-cooperation, however, countries' political self-determination is more important than the benefits of any cooperative actions. These extreme cases are rare, not least because the importance of cooperation for other state actors may result in attempts to coerce non-cooperative state actors into participation.

INTERNATIONAL SANCTIONS AGAINST VENEZUELA (2015–)

The United States has had concerns about Venezuelan narcotics trafficking since 2005 and Venezuela's lack of cooperation in combatting terrorism since 2006. With support from Congress and in response to increasing political repression in Venezuela, US President Obama levied additional sanctions for human rights abuses, corruption, and antidemocratic actions. The European Union urged Venezuelan officials to work toward political reconciliation but ultimately joined the United States in targeted sanctions to encourage a credible and meaningful process toward re-starting cooperation.

Throughout this time, sanctions have included targeted restrictions on Venezuelan officials, blocking financial transactions, and an embargo on the oil sector, which is highly important to the country's economy. Some politicians continue to support economic sanctions as a means to pressure the Venezuelan

government to meet international standards on human rights and political cooperation. Others are concerned about the increasing humanitarian cost of those sanctions.

Venezuela's non-cooperative stance in the international arena has resulted in it being subject to substantial international sanctions. Imposing sanctions is in itself a non-cooperative approach by the United States and the European Union meant to influence the behavior of a country or its political leadership. Venezuela's non-cooperative stance indicates that its political self-determination is a priority above that of the humanitarian cost being inflicted on its citizens.

Political cooperation versus non-cooperation is only one lens through which geopolitical actors engage with the world, but it is an important one for understanding countries' priorities. Assessing a state actor's hierarchy of needs and how it may change may help us to understand its motivations and priorities. These factors affect not only countries' political and military actions but also their willingness to support economic and financial cooperation, which we will explore next.

QUESTION SET



1. Which of these is likely lowest on a country's hierarchy of interests?

- A. Tariff harmonization
- B. Military determination
- C. Cultural program development

Solution:

C is correct. Cultural program development is likely lowest on a country's hierarchy of interests. Military determination (B) is often a primary source of national security and key to a country's national interest. Tariff harmonization (A) may improve economic activity and improve cooperation. Cultural programs are important and influential but likely lower priority compared with A and B.

2. Which of the following actions by a country is *most likely* a form of geopolitical cooperation?

- A. Acting as a conduit for trade
- B. Engaging in rules standardization
- C. Opting to use soft power over military retaliation

Solution:

B is correct. Political cooperation is associated with anything related to agreements of rules and standardization, with countries working together toward some shared goal. A cooperative country is one that engages and reciprocates in rules standardization. A is incorrect because acting as a conduit of trade, like Panama, involves non-cooperatively using a country's geographic location as a lever of power in broader international dynamics. C is incorrect because both soft power and military retaliation are examples of non-cooperative behavior, with the former being a less extreme means to influence another country's decisions without force or coercion.

3. Which of the following statements represents an aspect of geopolitical risk?

- A. Modeling geopolitical risk is relatively easy to standardize.

- B. An engaged country can be considered cooperative, even if it does not reciprocate.
- C. The strength of a country's institutions is relevant to the durability of its cooperative relationships.

Solution:

C is correct. The strength of a country's institutions can make cooperative relationships more durable. A is incorrect because modeling geopolitical risk is not easily standardized. B is incorrect because a cooperative country is one that is both engaged and reciprocates.

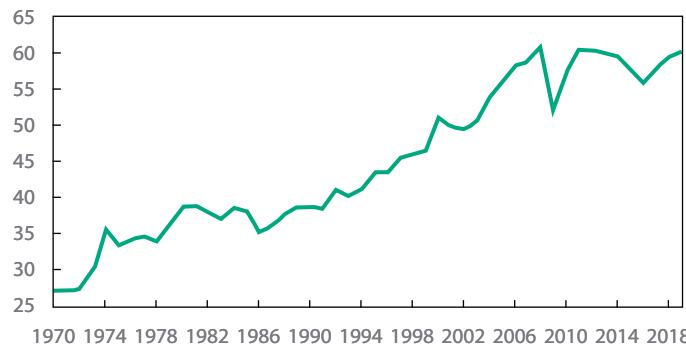
3**FORCES OF GLOBALIZATION**

describe geopolitics and its relationship with globalization

Globalization is the process of interaction and integration among people, companies, and governments worldwide. It is marked by the spread of products, information, jobs, and culture across borders. Indeed, the spread of goods and services across borders has been increasing for decades. The World Bank Openness Index, a key measure of globalization, has risen from 27 percent in 1970 to more than 60 percent in 2019, as shown in Exhibit 4. Since 2008, globalization has experienced headwinds; these include the impact of the Global Financial Crisis, which increased scrutiny of cross-border activity, and the rise in nationalism, which decreased some countries' appetite for using imported products or services. Capacity constraints may also create an important headwind. Global trade cannot make up 100 percent of global economic activity, meaning its expansion may slow even without specific disruptions.

Exhibit 4: Global Trade as Percentage of GDP

Global Trade (as % of GDP)



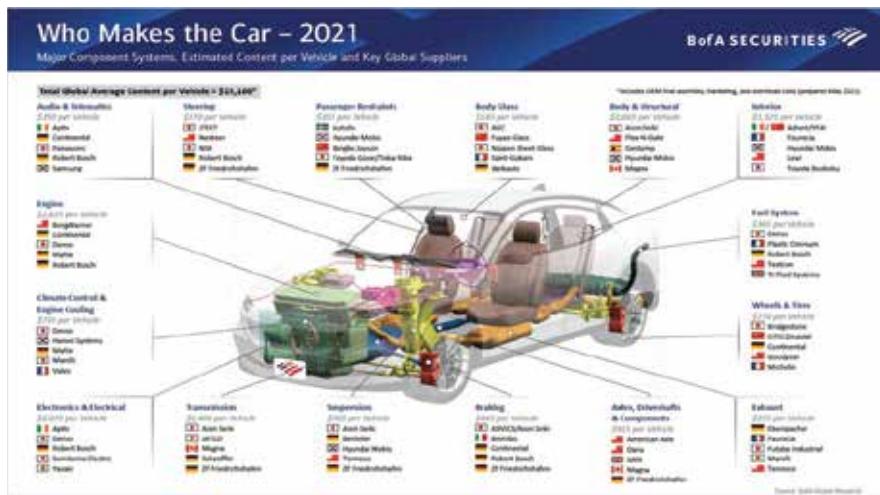
Source: World Bank Open Data, World Bank, <https://data.worldbank.org>. Trade is defined as imports plus exports in the global economy.

In addition to its macroeconomic impact, globalization is also visible at a microeconomic level. Take, for example, the production of an automobile. A car may be designed in Japan, with electrical parts made in Germany, steering systems designed in the United States, seatbelts manufactured in Sweden, climate system made in Belgium,

Forces of Globalization

and vehicles assembled in Mexico. In fact, automobiles are often assembled in one country with parts from all over the world, finished in a second country, and sold in a third (see Exhibit 5). This extensive process provides opportunities for companies to find the best inputs for their product, whether in terms of quality or cost-effectiveness. The process of globalization also opens opportunities for investors worldwide, who may invest in aspects of engineering, production, or even the process of supply chain management and logistics.

Exhibit 5: Production of Automobiles



Notes: Reprinted with permission. Copyright © 2021 Bank of America Corporation ("BAC"). The use of the above in no way implies that BAC or any of its affiliates endorses the views or interpretation or the use of such information or acts as any endorsement of the use of such information. The information is provided "as is" and none of BAC or any of its affiliates warrants the accuracy or completeness of the information.

Source: Bank of America Merrill Lynch Global Research.

Globalization also has cultural and communicative features. Although it can be difficult to measure these features, such as the spread of information or culture, it is not difficult to see them in our daily lives. A grocery store in Warsaw may hold Italian cheese, Moroccan spices, Colombian coffee, and Indian sauces. Social media allows users in South Africa to collaborate on dances with a Japanese music group. Faster and more affordable travel has increased interactions between citizens of countries all over the world. Internet usage allows for the near-instantaneous spread of cultural information and context.

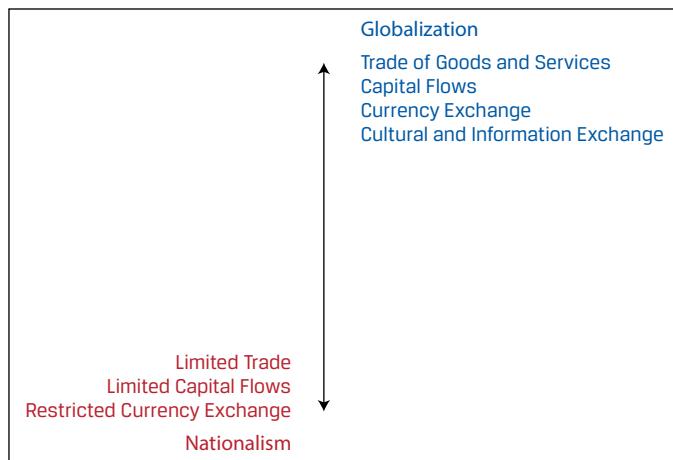
Features of Globalization

In the previous section, we considered political cooperation and non-cooperation as one lens through which to consider geopolitical actors—primarily national or state actors. Globalization, by contrast, is the result of economic and financial cooperation. It is carried out mostly by non-state actors, such as corporations, individuals, or organizations.

Globalization is marked by economic and financial cooperation, including the active trade of goods and services, capital flows, currency exchange, and cultural and information exchange. Actors participating in globalization are likely to reach beyond their national borders for access to new markets, talent, or learning. In contrast,

antiglobalization or **nationalism** is the promotion of a country's own economic interests to the exclusion or detriment of the interests of other nations. Nationalism is marked by limited economic and financial cooperation. These actors may focus on national production and sales, limited cross-border investment and capital flows, and restricted currency exchange. Exhibit 6 shows a continuum from nationalism to globalization.

Exhibit 6: Globalization: Economic and Financial Cooperation



Globalization and cooperation tend to be correlated, meaning that globalization can be facilitated or accelerated by political cooperation, but globalization is also an independent process. Organic private sector forces can drive the exchange of products or ideas even without government support or harmonized rules.

EXAMPLE 4

Features of Globalization

- Fill in the blanks using the following words: *government; private sector; national; subnational*.

Although political cooperation and non-cooperation can be driven by _____ actors, globalization as the result of economic and financial cooperation is carried out mostly by _____ actors, such as corporations, individuals, or organizations. Organic _____ forces can drive the exchange of products or ideas even without _____ support.

Solution:

Although political cooperation and non-cooperation can be driven by *national* actors, globalization as the result of economic and financial cooperation is carried out mostly by *subnational* actors, such as corporations, individuals, or organizations. Organic *private sector* forces can drive the exchange of products or ideas even without *government* support.

EXAMPLE 5**Features of Globalization**

1. True or False: Globalization is primarily carried out by governmental actors.

Explain your reasoning.

- A. True
- B. False

Solution:

B is correct. Globalization is primarily carried out by non-governmental actors, such as corporations, individuals, or organizations, and is the result of economic and financial cooperation. Multinational corporations, for example, want a competitive advantage leading them to seek new markets, talent, and learning as well as trade, capital, currency, and cultural and information exchange. Corporate outsourcing of talent is an example of globalization.

Motivations for Globalization

Non-state actors, such as companies and investors, who choose to participate in globalization consider three potential gains:

- Increasing profits
- Access to resources and markets
- Intrinsic gain

Increasing Profits

Profit can be increased by two means: increasing sales or reducing costs.

Increasing sales

The opportunity to generate higher profits may motivate companies to globalize. The first way to generate profit is to increase sales. Companies may choose to engage in globalization in order to access new customers for their goods and services. This process may require extensive investment, including acquiring or building property, plant, and equipment in the local market. It may also require hiring workers in new markets and offering training directly benefiting the countries involved.

Reducing costs

Another way to increase profits is to reduce costs. Globalization allows companies to access lower tax-operating environments, reduce labor costs, or seek other supply chain efficiency gains, all of which are cost-reduction measures. Of course, increasing sales and reducing costs can be closely intertwined. The consolidation of the global automobile industry is one instructive example. For capital-intensive production processes, companies experience economies of scale, reducing average costs, by producing more. By consolidating, automakers can benefit from a global sales opportunity while reducing costs per auto sold.

Access to Resources and Markets

In “Motivations for Globalization,” we described the economic interest countries may have to cooperate, including access to resources. The same may be true of non-state actors, such as companies, seeking sustainable access to resources, such as talent or raw materials. If those resources are not readily available or affordable in the home country, then the non-state actor may globalize to improve access.

A non-state actor may also seek market access or investment opportunities abroad. For example, a country experiencing low domestic market returns and increasing wealth may experience a higher propensity to accept cross-border risk—a catalyst for seeking returns abroad. For investment professionals, there are two important types of flows. **Portfolio investment flows** are short-term investments in foreign assets, such as stocks or bonds. Alternatively, **foreign direct investments (FDI)** are long-term investments in the productive capacity of a foreign country. These concepts will be covered in more detail elsewhere in the curriculum.

As appetite for cross-border investment has increased, such globalizing actors as companies and organizations (i.e., industry groups, NGOs, or charitable organizations), have established processes for facilitating those needs, including foreign exchange, accounting services, and global investment management services. This globalizing force, while an independent one, has been facilitated and strengthened over time by the harmonization of foreign exchange markets and investment practices.

Intrinsic Gain

Intrinsic gain is a side effect or consequence of an activity that generates a benefit beyond profit itself. It is difficult to measure but contributes to globalization’s momentum. It can also be a stabilizing force, increasing empathy between actors and reducing the likelihood that a geopolitical threat is levied. One example of intrinsic gain is the personal growth or education that individuals may receive from expanding their horizons, experiencing new places, or learning new ideas. Another example is accelerated productivity from learning new methods.

Regardless of the motivation for globalization—profit driven or intrinsically driven—these processes and investments can have multiplicative effects. As globalization deepens, companies develop standards and processes to incorporate multiple cultures into their overall corporate culture. They must cooperate with multiple sets of rules, and they may establish standard procedures based on the groups of companies in which they operate.

While these factors provide important motivating forces for globalization, they are not the only benefits of globalization. The process of reducing barriers between global businesses and organization can also provide aggregate economic benefits, such as increased choice, higher quality goods, increased competition among firms, higher efficiency, and increased labor mobility.

Costs of Globalization and Threats of Rollback

While globalization provides many benefits, costs also may be associated with it for individual economies or sectors, depending on how it is implemented. Some of the potential disadvantages of globalization are discussed in the following sections.

Unequal Accrual of Economic and Financial Gains

Economic theory tells us that aggregate economic activity is improved when all actors seek profit maximization and efficiency. However, improvement on the aggregate does not mean improvement for everyone. If a company moves a factory to another

country, it creates jobs in the new country but reduces them at home, while firms in the new country may have to compete with the foreign firm for labor. Some actors will benefit from this exchange, but others may suffer.

Lower Environmental, Social, and Governance Standards

Companies operating in lower-cost countries often operate in the local standards of those countries. If standards on environmental protection, social benefits, or corporate governance are lower in one country compared with another, and companies ultimately reduce their standards of production in that context, then globalization can create a drain on human, administrative, and environmental resources. The more measurable corporate profit factor may show a gain, but the overall effect of that activity may be negative. For example, many European countries have stricter standards on carbon emissions than those elsewhere in the world. Imagine that a European-based company decides to produce in a different country with lighter environmental regulations and cheaper labor costs. If the company decides to act according to local standards rather than its home country standards, it may make more profit but reduce environmental quality.

Political Consequences

These two costs can create a third cost of globalization: the political consequences of global expansion. While some individuals may enjoy global exposure, others may fear it. While some countries may benefit from improved labor force utilization, others may lose jobs as a company moves abroad. Therefore, globalization can contribute to income and wealth inequality, as well as differences in opportunity, within and between countries. These dynamics can manifest in countries' local politics, resulting in a force not only for reduced political and economic cooperation but also for a rollback in political cooperation.

Interdependence

Through the process of greater economic and financial cooperation, companies may become dependent on other countries' resources for their supply chains. On aggregate, this can result in the nation becoming dependent on other nations for certain resources. If there is a disruption to the supply chain, including through a moment of political non-cooperation, then firms may not be able to produce the goods themselves. Commodities production illustrates this potential risk. Rare earth metals, used for light-emitting diode (LED) lights and most electronic displays, are largely produced in China. Copper, essential for renewable energy construction, is largely produced in Chile. Disruptions to production in these countries, including mining accidents or floods, can disrupt entire industries relying on those resources.

THE COVID-19 PANDEMIC AND SUPPLY CHAIN SHORTAGES

The COVID-19 pandemic, which began in late 2019 and reached global impact in 2020, provides an interesting example of how geopolitical disruptions can impact companies' supply chains and countries' access to important resources. In the early months of 2020, many countries enforced stay-at-home orders in hopes of protecting their citizens against the pandemic's spread. As a result, demand for many goods and services collapsed and manufacturing activity was restricted, resulting in a significant slowdown in production of many goods and services. Even as production resumed, the impact of production slowdown had ripple effects across industries.

Semiconductors were one such product. Semiconductor production is highly concentrated in China and is very important to the automobile industry. The chips are used for fuel-pressure sensors, navigation displays, and speedometers,

among other automotive devices. Gradually, as the pandemic wore on, mobility began to improve, but supply remained constrained. The result was a severe shortage of semiconductor supply, which contributed to high and rising prices for new, used, and rental automobiles in many countries across the world. This dependence of a global industry on one country's production illustrates how interconnected supply chains can create important financial risks. Additionally, the effects of the semiconductor shortage soon moved beyond the automobile industry, as other industrial players struggled to secure chips for their devices.

The experience of the COVID-19 pandemic built upon preexisting challenges to global supply chain management. Supply disruptions revealed the pitfalls of overreliance on any single production location. Corporates found their supply chains were fragile as inputs sourced from foreign countries became unavailable or manufacturing moved offshore suddenly fell victim to lockdown orders.

Threats of Rollback of Globalization

The threat of deglobalization has grabbed headlines since 2018 when the Trump administration began a series of “America first” policies. Rooted in nationalism, isolationism, and concerns for national and economic security, trade wars escalated in fits and starts for several years. Targets evolved to include not only countries traditionally seen as US competitors but also long-standing allies, such as the European Union, through tariffs on imported products, and Canada and Mexico, through a renegotiation of the North America Free Trade Agreement now known as the U.S.-Mexico-Canada Agreement (USMCA). These developments have led investors and policy makers to consider whether globalization is reversing and what it means for the path of geopolitical risk.

Indeed, the impact of political non-cooperation on multinational companies is significant. Global design, manufacturing, and distribution systems are complex, making a change in the rules—such as a restriction on the free movement of goods and services—burdensome, which raises questions about profitability and efficiency. Multinational corporations are accustomed, however, to the political and operational risks of international production. This makes management rightfully slow to change carefully engineered processes over political disputes that eventually may be cleared up.

Despite the deglobalization discussion, completely reversing globalization seems unlikely. Instead, companies are likely to use a combination of the following tactics to fortify their supply chains:

1. *Reshoring the essentials:* Shortages of prescription medication, personal protective equipment, and other essential items during the pandemic highlighted the need for certain “essential” supply chains to be rebuilt domestically for emergency situations. Companies seeking to reduce manufacturing and procurement risk may relocate back to their home countries.
2. *Reglobalizing production:* The same concerns about production disruptions, rising labor costs, or political risk may instead prompt companies to duplicate or fortify their supply chains.
3. *Doubling down on key markets:* China has been the focus of US trade concerns, but the political risk is unlikely to change the need for some globalization. Not all manufacturing would be able to move outside of such key markets as China, Canada, and Mexico, particularly for companies that also seek to sell to those markets. While labor costs in some trading partners have risen over time, so too has the productivity of those workers. Add large market size, physical infrastructure supporting coordination, sophisticated supply chains, and the investment required to rebuild supply chains

elsewhere, and some companies may consider doubling down on key markets. Developing production “In country, for the country,” in combination with external supply chains, may be required.

QUESTION SET

1. Which of these actions would do the most to increase geopolitical risk?

- A. Increase capital flows
- B. Restrict foreign currency exchange
- C. Engage in trade of goods and services

Solution:

B is correct. Restricted foreign currency exchange—a characteristic of anti-globalization—would likely reduce political and economic cooperation and thus increase geopolitical risk. A is incorrect because an increase in capital flows would reduce geopolitical risk. C is incorrect because an increase in trade would reduce geopolitical risk.

INTERNATIONAL TRADE ORGANIZATIONS**4**

- describe functions and objectives of the international organizations that facilitate trade, including the World Bank, the International Monetary Fund, and the World Trade Organization

During the Great Depression in the 1930s, countries attempted to support their failing economies by sharply raising barriers to foreign trade, devaluing their currencies to compete against each other for export markets, and restricting their citizens' freedom to hold foreign exchange. These attempts proved to be self-defeating. World trade declined dramatically and employment and living standards fell sharply in many countries. By the 1940s, it had become a widespread conviction that the world economy needed organizations to help promote international economic cooperation. In July 1944, during the United Nations Monetary and Financial Conference in Bretton Woods, New Hampshire, representatives of 45 governments agreed on a framework for international economic cooperation. Two crucial, multinational organizations emanated from this conference—the World Bank, which was founded during the conference, and the International Monetary Fund (IMF), which came into formal existence in December 1945. Although the IMF was founded with the goal to stabilize exchange rates and assist the reconstruction of the world's international payment system, the World Bank was created to facilitate post-war reconstruction and development.

A third institution, the International Trade Organization (ITO), was to be created to handle the trade side of international economic cooperation, joining the other two Bretton Woods institutions. The draft ITO charter was ambitious, extending beyond world trade regulations to include rules on employment, commodity agreements, restrictive business practices, international investment, and services. The objective was to create the ITO at a United Nations Conference on Trade and Employment in Havana, Cuba in 1947. Meanwhile, 15 countries had begun negotiations in December 1945 to reduce and regulate customs tariffs. With World War II only barely ended,

they wanted to give an early boost to trade liberalization and begin to correct the legacy of protectionist measures that had remained in place since the early 1930s. The group had expanded to 23 nations by the time the deal was signed on 30 October 1947 and the General Agreement on Tariffs and Trade (GATT) was born. The Havana conference began on 21 November 1947, less than a month after GATT was signed. The ITO charter was finally approved in Havana in March 1948, but ratification in some national legislatures proved impossible. The most serious opposition was in the US Congress, even though the US government had been one of the driving forces. In 1950, the US government announced that it would not seek congressional ratification of the Havana Charter, and the ITO was effectively dead. As a consequence, the GATT became the only multilateral instrument governing international trade from 1948 until the World Trade Organization (WTO) was officially established in 1995.

Role of the International Monetary Fund

As we saw earlier, current account deficits reflect a shortage of net savings in an economy and can be addressed by policies designed to rein in domestic demand. This approach, however, could have adverse consequences for domestic employment. The IMF stands ready to lend foreign currencies to member countries to assist them during periods of significant external deficits. A pool of gold and currencies contributed by members provides the IMF with the resources required for these lending operations. The funds are lent only under strict conditions and borrowing countries' macroeconomic policies are continually monitored. The IMF's main mandate is to ensure the stability of the international monetary system, the system of exchange rates and international payments that enables countries to buy goods and services from each other. More specifically, the IMF:

- provides a forum for cooperation on international monetary problems;
- facilitates the growth of international trade and promotes employment, economic growth, and poverty reduction;
- supports exchange rate stability and an open system of international payments; and
- lends foreign exchange to members when needed, on a temporary basis and under adequate safeguards, to help them address balance of payments problems.

The 2008–2009 Global Financial Crisis demonstrated that domestic and international financial stability cannot be taken for granted, even in the world's most developed countries. In light of these events, the IMF has redefined and deepened its operations by:

- *enhancing its lending facilities:* The IMF has upgraded its lending facilities to better serve its members. As part of a wide-ranging reform of its lending practices, it also has redefined the way it engages with countries on issues related to structural reform of their economies. In this context, it has doubled member countries' access to fund resources and streamlined its lending approach to reduce the stigma of borrowing for countries in need of financial help.
- *improving the monitoring of global, regional, and country economies:* The IMF has taken several steps to improve economic and financial surveillance, which is its framework for providing advice to member countries on macroeconomic policies and warning member countries of risks and vulnerabilities in their economies.

- *helping resolve global economic imbalances:* The IMF's analysis of global economic developments provides finance ministers and central bank governors with a common framework for discussing the global economy.
- *analyzing capital market developments:* The IMF is devoting more resources to the analysis of global financial markets and their links with macroeconomic policy. It also offers training to country officials on how to manage their financial systems, monetary and exchange regimes, and capital markets.
- *assessing financial sector vulnerabilities:* Resilient, well-regulated financial systems are essential for macroeconomic stability in a world of ever-growing capital flows. The IMF and the World Bank jointly run an assessment program aimed at alerting countries to vulnerabilities and risks in their financial sectors.

From an investment perspective, the IMF helps to keep country-specific market risk and global systemic risk under control. The Greek sovereign debt crisis, which threatened to destabilize the entire European banking system, is an example. In early 2010, the Greek sovereign debt rating was downgraded to non-investment grade by leading rating agencies as a result of serious concerns about the sustainability of Greece's public sector debt load. Yields on Greek government bonds rose substantially following the downgrading and the country's ability to refinance its national debt was seriously questioned in international capital markets. Bonds issued by some other European governments fell and equity markets worldwide declined in response to spreading concerns of a Greek debt default. The downgrading of Greek sovereign debt was the ultimate consequence of persistent and growing budget deficits the Greek government had run before and after the country had joined the European Monetary Union (EMU) in 2001. Most of the budget shortfalls reflected elevated outlays for public-sector jobs, pensions, and other social benefits as well as persistent tax evasion. Reports that the Greek government had consistently and deliberately misreported the country's official economic and budget statistics contributed to further erosion of confidence in Greek government bonds in international financial markets. Facing default, the Greek government requested that a joint European Union/IMF bailout package be activated, and a loan agreement was reached between Greece, the other EMU member countries, and the IMF. The deal consisted of an immediate EUR45 billion in loans to be provided in 2010, with more funds available later. A total of EUR110 billion was agreed depending on strict economic policy conditions that included cuts in wages and benefits, an increase in the retirement age for public-sector employees, limits on public pensions, increases in direct and indirect taxes, and a substantial reduction in state-owned companies. By providing conditional emergency lending facilities to the Greek government and designing a joint program with the European Union on how to achieve fiscal consolidation, the IMF prevented a contagious wave of sovereign debt crises in global capital markets.

Another example of IMF activities is the East Asian Financial Crisis in the late 1990s. It began in July 1997, when Thailand was forced to abandon its currency's peg with the US dollar. Currency devaluation subsequently hit other East Asian countries that had similar balance of payment problems, including South Korea, Malaysia, the Philippines, and Indonesia. They had run persistent and increasing current account deficits, financed mainly with short-term capital imports, in particular, domestic banks borrowing in international financial markets. External financing was popular because of the combination of lower foreign (especially US) interest rates and fixed exchange rates. Easy money obtained from abroad led to imprudent investment, which contributed to overcapacities in several industries and inflated prices on real

estate and stock markets. The IMF came to the rescue of the affected countries with considerable loans, accompanied by policies designed to control domestic demand, which included fiscal austerity and tightened monetary reins.

World Bank Group and Developing Countries

The World Bank's main objective is to help developing countries fight poverty and enhance environmentally sound economic growth. For developing countries to grow and attract business, they have to

- strengthen their governments and educate their government officials;
- implement legal and judicial systems that encourage business;
- protect individual and property rights and honor contracts;
- develop financial systems robust enough to support endeavors ranging from micro credit to financing larger corporate ventures; and
- combat corruption.

Given these targets, the World Bank provides funds for a wide range of projects in developing countries worldwide and financial and technical expertise aimed at helping those countries reduce poverty.

The World Bank's two closely affiliated entities—the International Bank for Reconstruction and Development (IBRD) and the International Development Association (IDA)—provide low or no-interest loans and grants to countries that have unfavorable or no access to international credit markets. Unlike private financial institutions, neither the IBRD nor the IDA operates for profit. The IBRD is market-based, and uses its high credit rating to pass the low interest it pays for funds on to its borrowers—developing countries. It pays for its own operating costs because it does not look to outside sources to furnish funds for overhead.

IBRD lending to developing countries is primarily financed by selling AAA-rated bonds in the world's financial markets. Although the IBRD earns a small margin on this lending, the greater proportion of its income comes from lending out its own capital. This capital consists of reserves built up over the years and money paid in from the Bank's 185 member country shareholders. IBRD's income also pays for World Bank operating expenses and has contributed to IDA and debt relief. IDA is the world's largest source of interest-free loans and grant assistance to the poorest countries. IDA's funds are replenished every three years by 40 donor countries. Additional funds are regenerated through repayments of loan principal on 35- to 40-year, no-interest loans, which are then available for relending. At the end of September 2010, the IBRD had net loans outstanding of USD125.5 billion, while its borrowings amounted to USD132 billion.

In addition to acting as a financier, the World Bank also provides analysis, advice, and information to its member countries to enable them to achieve the lasting economic and social improvements their people need. Another core function of the World Bank is to increase the capabilities of its partners, people in developing countries, and its own staff. Links to a wide range of knowledge-sharing networks have been set up by the Bank to address the vast need for information and dialogue about development.

From an investment perspective, the World Bank helps to create the basic economic infrastructure that is essential for the creation of domestic financial markets and a well-functioning financial industry in developing countries. Moreover, the IBRD is one of the most important supranational borrowers in the international capital markets. Because of its strong capital position and its very conservative financial, liquidity, and lending policies, it enjoys the top investment-grade rating from the leading agencies,

and investors have confidence in its ability to withstand adverse events. As a result, IBRD bonds denominated in various major currencies are widely held by institutional and private investors.

World Trade Organization and Global Trade

The WTO provides the legal and institutional foundation of the multinational trading system. It is the only international organization that regulates cross-border trade relationships among nations on a global scale. It was founded on 1 January 1995, replacing the General Agreement on Tariffs and Trade (GATT) that had come into existence in 1947. The GATT was the only multilateral body governing international trade from 1947 to 1995. It operated for almost half a century as a quasi-institutionalized, provisional system of multilateral treaties. Several rounds of negotiations took place under the GATT, of which the Tokyo round and the Uruguay round may have been the most far reaching. The Tokyo round was the first major effort to address a wide range of non-tariff trade barriers, whereas the Uruguay round focused on the extension of the world trading system into several new areas, particularly trade in services and intellectual property, but also to reform trade in agricultural products and textiles. The GATT still exists in an updated 1994 version and is the WTO's principal treaty for trade in goods. The GATT and the General Agreement on Trade in Services (GATS) are the major agreements within the WTO's body of treaties that encompasses a total of about 60 agreements, annexes, decisions, and understandings.

In November 2001, the most recent round of negotiations was launched by the WTO in Doha, Qatar. The Doha round was an ambitious effort to enhance globalization by slashing barriers and subsidies in agriculture and addressing a wide range of cross-border services. So far, under GATS, which came into force in January 1995, banks, insurance companies, telecommunication firms, tour operators, hotel chains, and transport companies that want to do business abroad can enjoy the same principles of free and fair trade that previously had applied only to international trade in goods. Although no final agreement was reached in the Doha round, it marked one of the most crucial events in global trade over the past several decades: China's accession to the WTO in December 2001. The inability to reach agreement in the Doha round led to an increasing number of bilateral and multilateral trade agreements, such as the Trans-Pacific Partnership with Japan, Vietnam, and nine other countries.

The WTO's most important functions are the implementation, administration, and operation of individual agreements; acting as a platform for negotiations; and settling disputes. Moreover, the WTO has the mandate to review and propagate its members' trade policies and ensure the coherence and transparency of trade policies through surveillance in a global policy setting. The WTO also provides technical cooperation and training to developing, least-developed, and low-income countries to assist with their adjustment to WTO rules. In addition, the WTO is a major source of economic research and analysis, producing ongoing assessments of global trade in its publications and research reports on special topics. Finally, the WTO works in cooperation with the other two Bretton Woods institutions, the IMF and the World Bank.

From an investment perspective, the WTO's framework of global trade rules provides the major institutional and regulatory base without which today's global multinational corporations would be hard to conceive. Modern financial markets would look different without the large, multinational companies whose stocks and bonds have become key elements in investment portfolios. In the equity universe, for instance, investment considerations focusing on global sectors rather than national markets would make little sense without a critical mass of multinational firms competing with each other in a globally defined business environment.

QUESTION SET

On 10 May 2010, the Greek government officially applied for emergency lending facilities extended by the International Monetary Fund. It sent the following letter to the IMF's Managing Director:

Request for Stand-By Arrangement

This paper was prepared based on the information available at the time it was completed on Monday, May 10, 2010. The views expressed in this document are those of the staff team and do not necessarily reflect the views of the government of Greece or the Executive Board of the IMF. The policy of publication of staff reports and other documents by the IMF allows for the deletion of market-sensitive information.

May 3, 2010
Managing Director
International Monetary Fund
Washington DC

The attached Memorandum of Economic and Financial Policies (MEFP) outlines the economic and financial policies that the Greek government and the Bank of Greece, respectively, will implement during the remainder of 2010 and in the period 2011–2013 to strengthen market confidence and Greece's fiscal and financial position during a difficult transition period toward a more open and competitive economy. The government is fully committed to the policies stipulated in this document and its attachments, to frame tight budgets in the coming years with the aim to reduce the fiscal deficit to below 3 percent in 2014 and achieve a downward trajectory in the public debt-GDP ratio beginning in 2013, to safeguard the stability of the Greek financial system, and to implement structural reforms to boost competitiveness and the economy's capacity to produce, save, and export. (...) The government is strongly determined to lower the fiscal deficit, (...) by achieving higher and more equitable tax collections, and constraining spending in the government wage bill and entitlement outlays, among other items. In view of these efforts and to signal the commitment to effective macroeconomic policies, the Greek government requests that the Fund supports this multi-year program under a Stand-By Arrangement (SBA) for a period of 36 months in an amount equivalent to SDR26.4 billion. (A SDR (special drawing right) is a basket of four leading currencies: Japanese yen (JPY), US dollar (USD), British pound (GBP), and euro (EUR).) A parallel request for financial assistance to euro area countries for a total amount of EUR80 billion has been sent. The implementation of the program will be monitored through quantitative performance criteria and structural benchmarks as described in the attached MEFP and Technical Memorandum of Understanding (TMU). There will be twelve quarterly reviews of the program supported under the SBA by the Fund, (...) to begin with the first review that is expected to be completed in the course of the third calendar quarter of 2010, and then every quarter thereafter until the last quarterly review envisaged to be completed during the second calendar quarter of 2013, to assess progress in implementing the program and reach understandings on any additional measures that may be needed to achieve its objectives. (...) The Greek authorities believe that the policies set forth in the attached memorandum are adequate to achieve the objectives of the economic program, and stand ready to take any further measures that may become appropriate for this purpose. The authorities will consult with the Fund in accordance with its policies on such consultations, (...) and in advance of revisions to the policies contained in the MEFP. All information requested by the Fund (...) to assess implementation of the program will be provided.

(...)

Sincerely,

George Papaconstantinou Minister of Finance	George Provopoulos Governor of the Bank of Greece
--	--

- What is the objective of the IMF's emergency lending facilities?

Solution:

The program seeks to safeguard the stability of the Greek financial system and to implement structural reforms to boost competitiveness and the economy's capacity to produce, save, and export.

- What are the macroeconomic policy conditions under which the IMF provides emergency lending to Greece?

Solution:

The Greek government has to reduce the country's fiscal deficit by achieving higher and more equitable tax collections as well as constrain spending in the government wage bill and entitlement outlays.

- What is the amount Greece requests from the IMF as emergency funds?

Solution:

Greece applied for a standby arrangement in an amount equivalent to SDR26.4 billion (approximately USD39.5 billion, based on the 10 May 2010 exchange rate).

- Which of the following international trade organizations regulates cross-border exchange among nations on a global scale?

- A. World Bank Group (World Bank)
- B. World Trade Organization (WTO)
- C. International Monetary Fund (IMF)

Solution:

B is correct. The WTO provides the legal and institutional foundation of the multinational trading system and is the only international organization that regulates cross-border trade relations among nations on a global scale. The WTO's mission is to foster free trade by providing a major institutional and regulatory framework of global trade rules. Without such global trading rules, today's global transnational corporations would be hard to conceive.

- Which of the following international trade organizations has a mission to help developing countries fight poverty and enhance environmentally sound economic growth?

- A. World Bank Group (World Bank)
- B. World Trade Organization (WTO)
- C. International Monetary Fund (IMF)

Solution:

A is correct. The World Bank's mission is to help developing countries fight poverty and enhance environmentally sound economic growth. The World Bank helps to create the basic economic infrastructure essential for creation and maintenance of domestic financial markets and a well-functioning financial industry in developing countries.

6. Which of the following organizations helps to keep global systemic risk under control by preventing contagion in scenarios such as the 2010 Greek sovereign debt crisis?

- A. World Bank Group (World Bank)
- B. World Trade Organization (WTO)
- C. International Monetary Fund (IMF)

Solution:

C is correct. From an investment perspective, the IMF helps to keep country-specific market risk and global systemic risk under control. The Greek sovereign debt crisis in 2010, which threatened to destabilize the entire European banking system, is a recent example. The IMF's mission is to ensure the stability of the international monetary system—that is, the system of exchange rates and international payments that enables countries to buy goods and services from each other.

7. Which of the following international trade bodies was the only multilateral body governing international trade from 1948 to 1995?

- A. World Trade Organization (WTO)
- B. International Trade Organization (ITO)
- C. General Agreement on Tariffs and Trade (GATT)

Solution:

C is correct. The GATT was the only multilateral body governing international trade from 1948 to 1995. It operated for almost half a century as a quasi-institutionalized, provisional system of multilateral treaties and included several rounds of negotiations.

5

ASSESSING GEOPOLITICAL ACTORS AND RISK

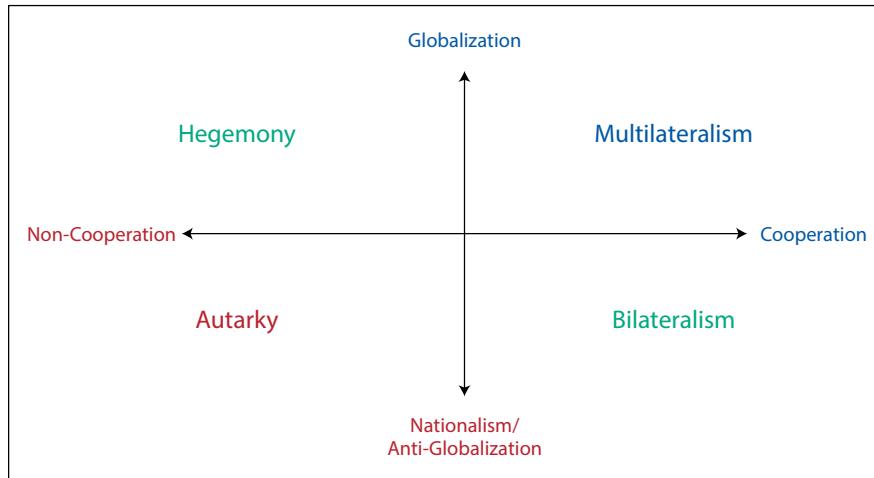
- | describe geopolitical risk

Using the two axes we have discussed—political cooperation versus non-cooperation and globalization versus nationalism—investment analysts can assess geopolitical actors and the likelihood of threat to investment outcomes. Where countries stand in that balance shapes their standing compared with other geopolitical actors, the tools of geopolitics they can use, and the threats and opportunities they face. This lesson provides a framework for analyzing a country's geopolitical risk.

Archetypes of Country Behavior

The framework, using the two axes, shown in Exhibit 7 presents four archetypes of country behavior: **autarky**, **hegemony**, **multilateralism**, and **bilateralism**. Each archetype has its own costs, benefits, and trade-offs with respect to geopolitical risk. In general terms, regions, countries, and industries that are more dependent on cross-border goods and capital flows will have growth rates and investment returns that benefit from greater global cooperation. The interdependent nature of their activities may reduce the likelihood that collaborative countries levy economic, financial, or

political attacks on one another. That same interdependence often makes cooperative actors more vulnerable to geopolitical risk than those less dependent on cooperation and trade. Diversifying the means of production provides some shield against risk but also provides more touchpoints by which risk can occur. In the remainder of the section, we describe the costs and benefits of each archetype, with some examples of each.

Exhibit 7: Archetypes of Globalization and Cooperation


Note that each of the axes explored in this framework—cooperation versus non-cooperation and globalization versus nationalism—represent a spectrum. Rarely do countries represent full extremes of either of these factors. As you can see from the examples provided, a country's place within the framework can be a moving target. As such, the examples provided in no way represent a value judgment of a country's goals or approach.

For geopolitical risk analysis, it matters not only which quadrant a country falls in today but also its stability within that quadrant. A hegemon that is working to build more political cooperation may be less of a threat to investment results than a multilateral actor trying to break them.

The “America first” example given in the section “Threats of Rollback of Globalization” is highly relevant. Adjusting previously agreed-upon rules may be in a country’s national interest, but inconsistent application of those rules presents a risk for companies and economic growth. Investors must be aware of actors’ movement within the framework to assess the likelihood and impact of geopolitical risk more appropriately.

EXAMPLE 6
Geopolitical Risk Analysis

- True or false: The relative permanence of geographic factors implies a fundamental stability in geopolitical risk analysis of the involved parties.

Explain your selection.

A. True

B. False**Solution:**

B is correct. Geopolitical risk analysis is fundamentally dynamic, with analysts needing to be aware of the parties' movement within the relevant framework to assess the likelihood and impact of geopolitical risk.

Geopolitical actors face a delicate balance between harnessing the potential intrinsic and profit gains from globalization while managing the many and sometimes unforeseen consequences. Further complicating the picture is the intricate connection between globalization and cooperation. A country's place within the globalization and cooperation framework can be a moving target, and for geopolitical risk analysis, it matters not only which quadrant a company falls in today but also its stability within that quadrant.

Autarky

Autarky describes countries seeking political self-sufficiency with little or no external trade or finance. State-owned enterprises control strategic domestic industries. The self-sufficiency of autarkic countries allows them to be stronger politically, including the ability to exercise complete control over the supply of technology, goods, and services, as well as media and political messaging. In some cases, periods of autarky can provide a country with swifter economic and political development. For example, for much of the 20th century, China could have been described as autarkic, exercising very little political cooperation or globalization. However, China's autarkic stance resulted in substantial poverty alleviation and an eventual move toward more economic and financial cooperation. That said, an autarkic stance does not come without costs. In other cases, such as North Korea and the earlier example of Venezuela, autarky has resulted in a gradual loss of economic and political development within the country.

Hegemony

Hegemonic countries tend to be regional or even global leaders, and they use their political or economic influence of others to control resources. State-owned enterprises tend to control key export markets. A hegemonic system can provide valuable benefits both to the hegemonic countries and to the international system. For the country, economic and political dominance may provide important influence on global affairs. For the global system, countries aligning with the hegemon's rules and standards may enjoy the rewards provided by the leader, including the stabilizing force of monitoring and enforcing the hegemon's standards. That said, there may be costs to hegemonic systems. As hegemons gain or lose influence in the international system, they may become more competitive, increasing the likelihood of geopolitical risk.

CHINA AND TECHNOLOGY TRANSFER

In 2011, China overtook Japan as the world's second largest economy. Having joined the World Trade Organization (WTO) in 2001, China's economy had globalized meaningfully. At the same time, the importance of China's economy gave it important influence on the rules of play in the international stage. One area where China has exerted its political influence is in technology transfer. For many years, China's stance was, in order to integrate into the Chinese production machine, including access to a skilled and plentiful labor supply, some companies would be required to transfer their technology ideas and processes to Chinese partners. For a smaller or less important economy, these rules might be difficult to enforce, but China's size and importance made it possible. These policies may have some globalizing effects. By having Chinese companies adopt more international technological ideas and rules, they become more globalized

in their approach. At the same time, they can also create investment risks. If China were to become more restrictive, it could disrupt global supply chains and increase costs for global companies.

RUSSIA AND GAS DISPUTES

Russia's economy, and particularly its oil sector, is well-integrated into global supply chains. However, the country is also largely politically autonomous, which can contribute to geopolitical risk and uncertainty. For example, Russia's control of important natural gas pipelines gives it significant political leverage over other countries, particularly those in Europe that rely on it for fossil fuels. As a result, European countries may be less likely to confront Russia in other areas. Between February and March 2014, Russia annexed Crimea, Ukraine. Other countries condemned the annexation and considered it to be a violation of international law, including such previously politically cooperative agreements as the 1991 Belavezha Accords that established the Commonwealth of Independent States. Still, the annexation continues, in part because many countries are unwilling to disrupt an important economic relationship.

In addition to political uncertainty, the annexation contributed to important economic and market developments affecting investors. In the immediate aftermath of the annexation, the Russian ruble depreciated significantly, contributing to higher market volatility and inflation. Other significant economic impacts followed. For example, countries, including Canada, the United States, and European Union member states, imposed sanctions on Russian officials. Ukraine responded to the annexation by cutting off water to the area, contributing to crop failure.

Multilateralism

Multilateralism describes countries that participate in mutually beneficial trade relationships and extensive rules harmonization. Private firms are fully integrated into global supply chains with multiple trade partners. Examples of multilateral countries include Germany and Singapore.

SINGAPORE AS MULTILATERAL ACTOR

Singapore is ranked as the most open economy in the world by the World Economic Forum, the third least corrupt by Transparency International, and the second most pro-business by the World Bank Doing Business Report. It has low tax rates and the second highest per capita GDP in terms of purchasing power of its citizens. The results are shown in Exhibit 8.

Exhibit 8: Select Global Data

World Bank Ease of Doing Business Rank (2019)	Corruption Perceptions Index Rank (2020)	Corporate Tax Rate (2021)	GDP (per capita, in international dollars, 2021)
Singapore	2	3	17.00%
United States	6	25	21.00%
Germany	22	9	30.00%

	World Bank Ease of Doing Business Rank (2019)	Corruption Perceptions Index Rank (2020)	Corporate Tax Rate (2021)	GDP (per capita, in international dollars, 2021)
Australia	18	11	30.00%	51,102
United Kingdom	8	11	19.00%	43,839
Japan	29	19	30.60%	41,507
Mexico	60	124	30.00%	18,867
China	31	78	25.00%	17,624
Brazil	124	94	34.00%	14,563
South Africa	82	69	28.00%	11,582
Philippines	95	115	30.00%	8,436

Sources: World Bank (doingbusiness.org), Transparency International (transparency.org), KPMG, International Monetary Fund (imf.org).

What is behind these impressive statistics? For starters, Singapore's factor endowment makes it highly dependent on cooperation and innovation to survive. The country has limited natural resources, including water and arable land, which means the country must rely on agrotechnology parks and reclaimed land for agricultural production as well as trade partners for the inputs to that production.

Geographic factors also contribute to Singapore's economic openness. The country is located at the intersection of many important global trade routes. Its population is highly ethnically and racially diverse, allowing for English as a key language for global business. In addition, its workforce is highly skilled, giving Singapore an important role as a center for trade and innovation throughout Asia.

Politically, Singapore has highly stable institutions. The government is strategically involved in its economic system to allow for consistent prioritization of economic activity and enforcement of business-friendly institutional governance. Singapore's openness has contributed to its attractiveness as a potential partner.

As a result of its location and factor endowments, Singapore is both capable of and highly dependent on international cooperation for its economic growth. The country generates higher economic growth rates because of this greater global cooperation. However, that same cooperation may leave Singapore more vulnerable to geopolitical risk than those countries that are less dependent on cooperation and trade.

EXAMPLE 7

Key Globalization Factors

- Identify at least two of the factors accounting for Singapore's reliance on globalization to achieve economic success.

Solution:

First, Singapore's factor endowment of limited natural resources makes it dependent on cooperation and innovation to survive.

Second, Singapore's geographic location at the intersection of multiple important global trade routes helps make it an Asian center for world business.

Third, culturally it has a highly ethnically and racially diverse population fluent in the key global business language of English, enhancing its collaborative capabilities.

Finally, it has highly stable political institutions with high governmental priorities on promoting economic activity and enforcing business-friendly institutional governance. These qualities make it an attractive potential business partner and likely sought out for global cooperation.

Bilateralism

Bilateralism is the conduct of political, economic, financial, or cultural cooperation between two countries. Countries engaging in bilateralism may have relations with many different countries, but they are one-at-a-time agreements without multiple partners. Typically, countries exist on a spectrum between bilateralism and multilateralism. In between the two extremes is **regionalism**, in which a group of countries cooperate with one another. Both bilateralism and regionalism can be conducted at the exclusion of other groups. For example, regional blocs may agree to provide trade benefits to one another and increase barriers for those outside of that group.

It is noteworthy that relatively few countries perfectly fit the bilateral mold. Moving toward stronger political cooperation tends to lead organically to globalization; it is common for non-state actors to globalize as long as the path is laid for them. Additionally, innovations, such as the internet and digital transfer, have made it even easier for firms to globalize.

Bilateralism once nicely described Japan; its government engaged in substantial political cooperation for the sake of building a strong export market, but it did not globalize its imports. That said, as the importance of international capital markets has deepened, Japan has been a pioneer of globalizing its equity and bond markets for international players. Today it would be considered a multilateral player.

QUESTION SET



1. For the following contrasting pairs of archetypes of globalization and cooperation, which one reflects the *greatest* differences in country behavior?

- A. Bilateralism versus autarky
- B. Multilateralism versus autarky
- C. Multilateralism versus hegemony

Solution:

B is correct. Multilateralism describes countries that participate in mutually beneficial trade relationships and extensive rules harmonization. Autarky describes countries seeking political self-sufficiency with little or no external trade or finance. In the Exhibit 7 display of these behavior patterns, these choices are most widely separated on both the globalization and cooperation continuums. A is incorrect because bilateral or regional approaches describe those countries leveraging regional trade relationships and may face the world as a group. Bilateralism shares with autarky a bias against globalization. These approaches diverge, however, regarding cooperation, with autarky being more non-cooperative. C is incorrect because multilateralism describes countries that participate in mutually beneficial trade relationships and extensive rules harmonization. Hegemony represents countries exerting political or economic influence of others to control resources.

Multilateralism shares with hegemony an inclination toward globalization,

as shown in Exhibit 7, but diverges from hegemony regarding cooperation; hegemony is more non-cooperative.

6

THE TOOLS OF GEOPOLITICS



describe tools of geopolitics and their impact on regions and economies

Now that we understand the characteristics of geopolitical actors, we can examine the tools these actors use to manifest or reinforce their interests with respect to others. The tools an actor uses are ultimately the source of geopolitical risk as it affects investors—shaping the likelihood of risk as well as the speed and size of impact, as we will explore in Lesson 6. As a result, understanding these tools may be just as important as understanding the motivations for using them.

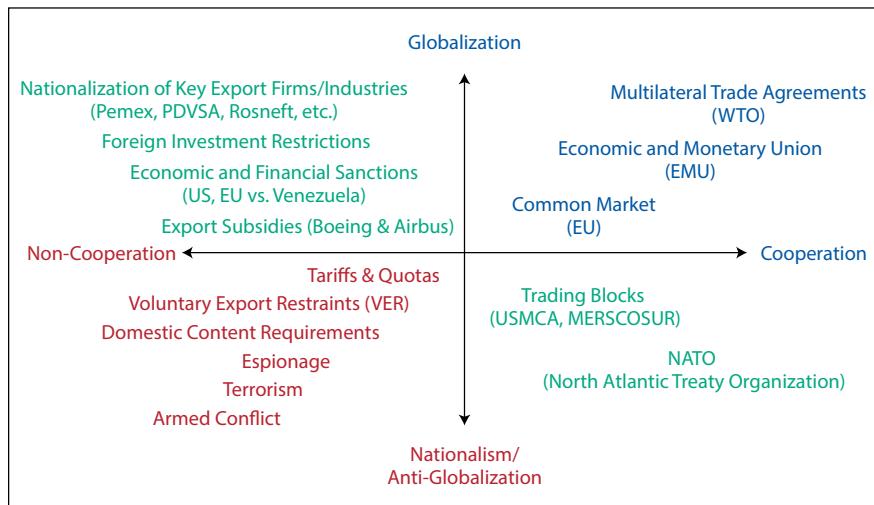
The Tools of Geopolitics

The tools of geopolitics may be separated into three types:

- national security tools,
- economic tools, and
- financial tools.

Among each of these tools, there are choices that reflect or improve cooperation and those that reflect or escalate conflict (non-cooperation). Tools facilitating cooperation are those that increase flows between countries; this can mean an increased flow of goods, services, capital, or labor through treaties, trade agreements, capital provisions, and approved migration. Tools escalating conflict are those that reduce these flows between countries.

We can use the framework provided in the section “Archetypes of Country Behavior” in Lesson 4 to assess the nature of tools used by geopolitical actors and to strengthen our designation of countries within that framework. The extent to which actors use these tools to improve cooperation or escalate conflict influences their position within the framework. As we described earlier, it is important for geopolitical risk analysts to track not only which quadrant best describes an actor but also its stability within that quadrant. Using different national security, economic, or financial tools may indicate that an actor is changing in character, which would increase or decrease geopolitical risk. Exhibit 9 shows how the tools of geopolitics fit in the four archetypes of country behavior (presented in Exhibit 7).

Exhibit 9: Tools of Geopolitics**National Security Tools**

National security tools are those used to influence or coerce a state actor through direct or indirect impact on the country's resources, people, or borders. National security tools may be active, meaning they are being used at the time of analysis, or threatened, meaning they are not currently used but their use is likely enough to warrant concern.

The most extreme example of a national security tool is that of armed conflict. Armed conflict is a direct and active national security tool. It can be internal or external to a country and has two major impacts. The first is the disruption or destruction of physical infrastructure, which can inflict long-term damage on a country's capital stock and ability to rebuild that stock. The second impact is on migration away from areas of armed conflict, which can reshape international flows of goods, services, capital, and labor. It also can affect neighboring countries and states accepting refugees.

SYRIAN REFUGEE CRISIS (SINCE 2011) AND IMPACTS ON GERMANY

The Syrian refugee crisis began in March 2011 after a violent government crackdown on public demonstrations. Protests turned into an armed rebellion and escalated to civil war. As a result, Syria has experienced extreme deterioration of its physical infrastructure, including buildings transportation, and other structures. As conflict prolonged, more than 6.6 million Syrians have fled the country, with another 6.7 million having been internally displaced. Beyond the tragic human toll of this ongoing conflict, Syria has also experienced negative impact on its current and potential economic growth as a result of the deterioration of its capital and labor stock. According to the International Monetary Fund (IMF), Syria's economy grew by 4.5 percent, on average, from 2000 to 2010. By 2016, the Syrian economy was less than half its size before the war. In investment terms, these developments have contributed to a sustained higher discount rate and lower bond and equity returns.

The majority of refugees have found refuge in neighboring countries, such as Turkey, Lebanon, and Jordan. However, many also made their way, often by dangerous routes, to European countries. The refugee crisis attracted significant international attention in 2015 when German Chancellor Angela Merkel announced that the country would admit 1 million Syrian refugees. In the

short-term, the decision was disruptive to domestic politics and international cooperation as not all citizens and neighboring countries were supportive of the approach. The decision also carried economic costs, including constructing housing and establishing resettlement programs. However, in the long-term, Germany appears to be benefiting from the cooperative decision, having attracted a younger and skilled demographic of migrants thus improving the demographic balance of an otherwise-aging country. The resulting improvements in Germany's labor and capital stock may improve Germany's potential economic growth rate. In investment terms, a higher potential economic growth rate contributes to higher expected market returns, all else equal, and may contribute to industry developments or innovations.

Of course, not all tools are used in so direct a nature as armed conflict. For example, espionage, or the practice of using spies to obtain political or military information, is a necessarily indirect national security tool. Military alliances often are used to aid in direct conflict and also to deter conflict from arising in the first place.

Additionally, not all national security tools are used in a non-cooperative way. State actors can combine forces to reduce the likelihood that national security tools are used. For example, the North Atlantic Treaty Organization (NATO), an alliance between the European Union, United States, United Kingdom, and Canada, is used to discuss and deescalate potential conflict among members and between members and outside states. Originally constructed to provide collective security against the Soviet Union, NATO now serves as a collective effort to reduce nuclear proliferation and other common national security goals.

EXAMPLE 8

Tools of Geopolitics

1. True or false: Although geopolitical financial tools can be used for both cooperative and non-cooperative reasons, national security tools are characteristically non-cooperative. Explain your selection.

A. True

B. False

Solution:

B is correct. National security tools are non-cooperative in cases like armed conflict. But collective security agreements, such as NATO, can also be used cooperatively to reduce the possibility of conflict among members and between members and outside states.

Financial tools may be cooperative in reducing geopolitical risk if they encourage cooperation in security, economic, or financial arenas. They may also tend to non-cooperation by creating vulnerabilities in the international system. US dominance is one such example, both promoting financial activity and making other countries vulnerable to US monetary policy changes.

Economic Tools

Economic tools are the actions used to reinforce cooperative or non-cooperative stances through economic means. Among state actors, economic tools can include multilateral trade agreements, such as the Southern Common Market (MERCOSUR), which facilitates trade among member countries in South America, or the global

harmonization of tariff rules, as facilitated by the World Trade Organization (WTO). Highly cooperative economic tools may also include common markets, like the European Union, or a common currency, like the euro.

By contrast, economic tools can also be non-cooperative in nature. Nationalization, the process of transferring an activity or industry from private to state control, is a non-cooperative approach to asserting economic control. Nationalization is most common in sectors perceived as vital to economic security or competitiveness, such as the energy sector. For example, after a period of privatization, Argentina moved in 2012 to renationalize YPF, the nation's largest energy firm. In addition to controlling an important geophysical resource, Argentina's government looked to secure the sale of fossil fuels as an important source of foreign exchange, a scarcity at the time. Countries can engage in voluntary export restraints, meaning they refuse to trade as much of their goods and services as would meet demand. Countries also can impose domestic content requirements, asserting that a certain proportion or type of domestic input be included in an exported good.

Financial Tools

Financial tools are the actions used to reinforce cooperative or non-cooperative stances through financial mechanisms. Examples of cooperative financial tools include the free exchange of currencies across borders and allowing foreign investment. Examples of non-cooperative financial tools include limiting access to local currency markets and restricting foreign investment. Sanctions, described earlier, provide a useful example of countries using financial tools to influence geopolitical outcomes.

Cooperative financial tools may reduce geopolitical risk if they encourage cooperation in security, economic, or financial arenas. However, the same tools may also create vulnerabilities in the international system. The dominance of the US dollar is one such example. The international interbank market, in which banks borrow and lend to one another, hosts transactions heavily denominated in US dollars. The market provides a tool of cooperation, and the free exchange of currency helps facilitate financial activity and cooperation more broadly. That said, the US dollar's importance to exchange also makes other countries vulnerable to changes in US monetary policy. Specifically, tighter US monetary policy can contribute to liquidity shortages in countries that do not or cannot maintain US dollar reserves.

EXAMPLE 9

Financial Tools of Geopolitics

1. Describe the conditions under which cooperative financial tools may decrease geopolitical risk as well as create vulnerabilities in the international system.

Solution:

Cooperative financial tools may reduce geopolitical risk if they encourage cooperation in security, economic, and financial arenas. However, if the system becomes too dependent on a particular financial tool or if the tool becomes too dominant, it may introduce vulnerabilities in the international system that can have far-reaching implications.

Multifaceted Approaches

Just as geopolitics is multifaceted and includes many types of actors and features, so too are the tools of geopolitics. Systems of political, economic, and financial cooperation can be, and often are, intertwined. One interesting example is **cabotage**, or the right to transport passengers or goods within a country by a foreign firm. Many countries, including those with multilateral trade agreements, impose restrictions on cabotage across transportation subsectors—meaning that shippers, airlines, and truck drivers are not allowed to transport goods and services within another country's borders. Allowing cabotage requires coordination on areas like physical security and economic coordination, a highly multilateral process.

International organizations may also make use of multiple tools of geopolitics. For example, the Association of Southeast Asian Nations (ASEAN) is composed of 10 members states and seeks to facilitate economic, political, security, military, educational, and cultural integration between its members. The European Union, which began as a six-country economic bloc, has expanded to include 27 member states (as of 2022) and features substantial financial and national security components. Generally, as actors incorporate more tools of collaboration, they are less likely to initiate conflict or use a non-cooperative tool against associated actors.

EUROPEAN UNION AND BREXIT

As shown in Exhibit 10, one of the most pronounced transitions toward multilateral cooperation and globalization is in the European Union (EU). From 1945 until 2016, European countries exhibited fairly steady progress toward political cooperation—via shared security, free movement of goods and people, and harmonization of rules and regulations—and economic and financial globalization. EU member states offer European citizenship and the right for workers and companies from all EU member states to operate within other EU countries. The strongest manifestation of this multilateralism is exhibited in the creation of the common currency for the euro area, a subset of European countries.

Exhibit 10: Timeline: The Path of European Cooperation and Globalization

1940	1950	1960	1970	1980	1990	2000	2010	2020
1945: World War II Ends	1960s: Customs Duties Removed	1970s: Single European Act is Signed Creating "the Single Market"		2000s: Countries Continue to Join the EU				
1950: European Coal and Steel Community Begins, Uniting Countries Economically and Politically	1970s: EU Regional Policy Begins to Transfer Money to Create Jobs and Infrastructure in Poorer Areas	1990s: Foundational EU Treaties Introduce European Citizenship, Common Foreign and Security Policy, Free Movement of People, and European Parliament	2010s: Croatia Becomes 28th EU Member State, United Kingdom Votes to Leave EU					

In 2016, the European integration process experienced a marked disruption when the United Kingdom voted to exit the EU. Despite the benefits of cooperation and globalization, the costs became increasingly felt, affecting local politics. The United Kingdom's sudden reversal in some of its multilateral momentum was a surprise to many global actors and investors. This geopolitical risk resulted in near-term market volatility, particularly in financial and currency markets. It has also resulted in the dismantling of political cooperation in recent years; the United Kingdom's multilateral approach is, in some areas, shifting toward a bilateral approach.

Geopolitical Risk and Comparative Advantage

The geopolitical tools discussed in this section generate important risks and opportunities for investors, but they do not operate in a vacuum. In fact, geopolitical risk and the tools of geopolitics can shape actors' core priorities. Models of international trade will be explored later in the curriculum. (In this context, we reference them briefly to illustrate how geopolitical tools can shape, destroy, or build upon a country's core approach to other actors.)

In the classic example of international trade, countries are endowed with certain resources, or factors, and technological capabilities. Not every country will be endowed with the same factors and capabilities and thus may benefit from cooperating via trade. For example, Ghana is rich in resources, such as hydrocarbons, gold, agricultural products, and several industrial metals. However, it relies heavily on other countries for both processing and industrial use of those resources. Each country specializes in areas defined by its resources and capabilities and then exchanges with the other in a way that benefits both parties. Through specialization and exchange, industries within each country experience greater economies of scale. Households and firms therefore will have a greater variety of products to choose from. Competition between firms is higher, and resources are allocated more efficiently.

Geopolitical risk and the tools of geopolitics can tilt comparative advantage in one direction or another. For example, countries or regions with limited geopolitical risk exposure may attract more labor and capital. In contrast, those with higher geopolitical risk exposure may suffer a loss of labor and capital. Similarly, a consistent threat of conflict may drive more regular volatility in asset prices, prompting investors to require higher compensation for risk taken (i.e., an increase in the required rate of return or the discount rate used in valuation).

QUESTION SET



1. True or false: Germany's reaction to the Syrian refugee crisis is an example of comparative advantage stemming from geopolitical risk. Explain your selection.

- A. True
- B. False

Solution:

A is correct. Countries with a lower geopolitical risk exposure have the ability to attract resources, such as labor and capital. With its strong economic position in the EU and longstanding stability of political leadership, Germany was able to undertake the resettlement of one million Syrian refugees to it while assuming the short-term relocation costs and disruption of its domestic politics. This improves its long-term demographic balance by adding young and talented migrants, with the resulting increase in labor and capital stock potentially increasing Germany's economic growth rate.

2. In the following table, match the geopolitical tool with the *most* appropriate example of each tool.

Geopolitical Tool	Example
1. Financial	A. Nationalization
2. Economic	B. Espionage
3. National security	C. Free exchange of currency across borders

Solution:

Option 1 (Financial) matches with C (Free exchange of currency across borders).

Option 2 (Economic) matches with A (Nationalization).

Option 3 (National security) matches with B (Espionage).

7

GEOPOLITICAL RISK AND THE INVESTMENT PROCESS



describe the impact of geopolitical risk on investments

There is no shortage of attention to geopolitical risk in financial market analysis. However, the extent to which investors incorporate geopolitical risk into their decision making will vary widely with their investment objectives and risk tolerance. Some investors may be considered *takers* of geopolitical risk. These investors may incorporate geopolitical risk into their analysis only to the extent that it affects the long-term attractiveness of asset classes or strategies. For other investors, geopolitical risk may be a central component of their investment process. For these portfolios, monitoring dislocations is an achievable and meaningful driver of alpha creation, where focused geopolitical risk analysis can reduce the impact and severity of adverse events and enhance the potential for upside growth. For example, an investor that anticipates an important political transition may enact portfolio hedges to shield against market volatility or may use the market volatility as a buying opportunity—in either case, improving investment outcomes.

What follows is a discussion of the impact of geopolitical risk on the investment environment. We begin with a discussion of types of geopolitical risk, followed by a means of assessing those risks and the ways that they can manifest in a portfolio.

Types of Geopolitical Risk

There are three basic types of geopolitical risk: event risk, exogenous risk, and thematic risk.

Event risk evolves around set dates, such as elections, new legislation, or other date-driven milestones, such as holidays or political anniversaries, known in advance. Political events often result in changes to investor expectations related to a country's cooperative stance. As a result, risk analysts often use political calendars as a starting place for assessing event risk.

One example of event risk is the United Kingdom's referendum on European Union membership (Exhibit 11). This was a known event risk; it was planned for 23 June 2016, well in advance. The stakes of the election were well understood, with

many years' worth of politically cooperative steps likely to unwind in the event of a "yes" vote. Most investors expected a "no" vote; when the results proved the opposite, investor expectations related to the United Kingdom's cooperative stance were drastically changed.

Exhibit 11: Market Reaction to Event Risk: United Kingdom's EU Referendum, 2016

	First Day	30 Days	1 Year
FTSE	−3.1%	6.2%	17.1%
GBP	−8.1%	−11.9%	−14.5%
10-year bond yield (Gilt)	−21.0%	−42.0%	−25.0%

The United Kingdom's vote to end its European Union membership came as a surprise to many investors. Several asset classes were immediately affected, including equities, the national currency, and government bonds. In the weeks that followed, investors adjusted to the news but became concerned about what a rollback in political cooperation might mean for long-term economic growth. As a result, equity prices recovered, but the British pound continued on a steady decline. Government bond yields declined precipitously in the month following the vote, recovering somewhat as the year progressed.

It is useful to note that the predictability of an event does not necessarily change its likelihood, its speed of impact, or the size of impact on investors; however, it does give investors more time to prepare a response. We will come back to this example and others in this section when we discuss the assessment of geopolitical threats.

Exogenous risk is a sudden or unanticipated risk that affects either a country's cooperative stance, the ability of non-state actors to globalize, or both (Exhibit 12). Examples include sudden uprisings, invasions, or the aftermath of natural disasters.

Exhibit 12: Market Reaction to Exogenous Risk: Japan's Fukushima Nuclear Disaster, 2011

	First Day	30 Days	1 Year
Nikkei	−6.2%	−5.2%	−3.6%
Yen	−0.3%	3.4%	0.5%
Japanese 10-year bond yield	−3.3%	5.9%	−22.3%

On 11 March 2011, Japan was struck with an earthquake and subsequent tsunami wave, resulting in significant loss of human life, homes, and productive capital. The natural disaster also caused a significant nuclear accident that resulted in further human, property, and environmental damage and also disrupted supply chains. The initial market response reflected market concern: Equities fell, the currency depreciated, and bond prices rose. In the weeks that followed and as the toll of the environmental disaster became more apparent, Japanese equity markets continued to suffer, declining up to 20.4 percent (as of 25 November 2011) from their levels the day of the earthquake.

As the environmental costs of the accident became clear, this event contributed to a shifting stance on political cooperation on environmental issues. Less than three months after the incident, Germany decided to phase out nuclear power entirely by 2022. Belgium confirmed plans to exit nuclear power by 2025, and such countries as Italy, Spain, and Switzerland opted not to reintroduce nuclear energy programs.

Finally, **thematic risks** are known risks that evolve and expand over a period of time. Climate change, pattern migration, the rise of populist forces, and the ongoing threat of terrorism fall into this category.

Cyber threats are another example of thematic risk (Exhibit 13). Cyber risks include any attempt to expose, alter, disable, destroy, steal, or gain information through unauthorized access to or unauthorized use of computer systems. These threats began with the expansion of internet and computer use and have increased in number and scale. Now, the number of records stolen or affected by cyberattacks is in the billions per year. While the basic nature of cyberattacks is consistent, the size, scale, and sophistication of attacks have increased over time.

Exhibit 13: Market Reaction to Thematic Risk: Equifax Data Breach, 2017

	First Day	30 Days	1 Year
Stock price	-13.7%	-22.0%	-4.8%
Financial services industry	0.9%	10.3%	18.6%

In September 2017, the US consumer credit reporting company Equifax announced a data breach that exposed personal information, including names, dates of birth, and personal identification numbers of approximately 147 million people. The initial market reaction was very negative. Equifax's equity price fell by 13.7 percent in one day and by 34.9 percent over the first week, reaching its low on 15 September 2017. Over time, the impact moderated somewhat, but Equifax still underperformed the financial services industry in the year after its data breach.

At first, the data breach affected Equifax the most, including instituting protections and credit monitoring for affected customers, without significant impacts on the broader financial industry.

However, the event also triggered broader impacts over time. Other companies have increased spending on cybersecurity and instituted stronger processes, including software upgrades.

EXAMPLE 10

Geopolitical Risk and the Investor

1. Which of the following types of risks are known in advance? Select all that apply.

- A. Event risk
- B. Exogenous risk
- C. Thematic risk

Solution:

A and C are correct.

Event risk evolves around set dates, and thematic risk is a known risk that evolves and expands over a period of time. Exogenous risk is a sudden or unanticipated risk.

Assessing Geopolitical Threats

Geopolitical risk is always present in the investment environment, and these risks can affect investments in many different ways—from broad macroeconomic levels, to industry impacts, down to individual companies. The question for investors is whether the particular geopolitical risk is relevant for their portfolio management decisions. To make this assessment, an investor considers geopolitical risk in terms of the following three areas:

- *likelihood* it will occur,
- *velocity* (speed) of its impact, and
- size and nature of that *impact*.

Likelihood

The likelihood of a risk is the probability that it will occur. Measuring likelihood is a challenging process. The highly unpredictable nature of many risks—their build over time, the many and conflicting motivations of actors involved—means this exercise can be more art than science. However, we may use the framework from Lesson 4 to assess the basic likelihood of risk occurring. Highly collaborative and globalized countries are, on balance, less likely to experience geopolitical risk because the political, economic, and financial costs of partners inflicting those risks are higher. That same interconnectedness, however, may make multilateral countries more vulnerable to certain risks. Their operation in and cooperation with other countries means a risk posed to any of those countries may also have an impact on itself. Similarly, multiple risk exposures may increase the impact of that risk when it occurs.

Of course, many other factors may increase the likelihood that a risk may occur. Internal political stability, economic need, and the motivations of governmental actors play an important role in increasing the likelihood of disruptive action. In fact, these considerations are so plentiful and intertwined that geopolitical risk monitoring tools have emerged as key components of many research offerings for investors. The numerous scenarios for any given risk make measuring risk likelihood a potentially never-ending task. As a result, investors must balance the time spent on this activity with the relative importance of its input to the investment process.

The examples used in the section “Types of Geopolitical Risk” in Lesson 6, can help us put a finer point on this concept. Of the three risks described, a cyber risk may have been the most likely to occur and to affect a given investment strategy, whereas the United Kingdom’s “yes” vote and Russia’s annexation of Crimea were less likely at the time in which they occurred. Of course, all three of these risks have different potential impacts on investors. As a result, likelihood should be considered only in conjunction with the velocity and impact of the risk.

Velocity

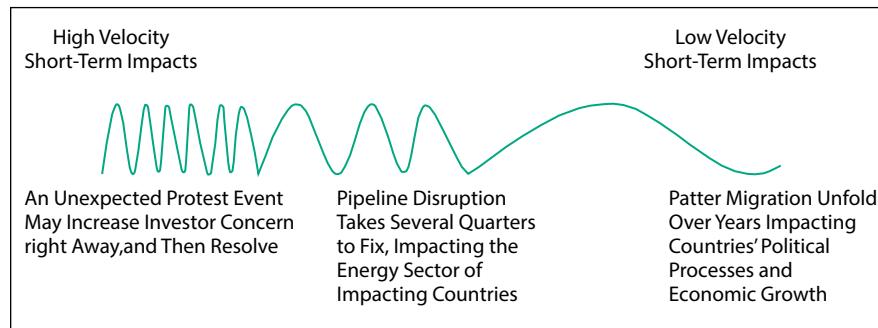
The **velocity** of geopolitical risk is the pace at which it affects an investor portfolio. For the sake of simplicity, we explore short-term or “high-velocity” impacts, medium-term, and long-term or “low-velocity” impacts (see Exhibit 14).

In the short term, we may see volatility in the markets affecting entire industries or even the entire market. Exogenous or “black swan” events tend to fit into this category, causing market volatility and investor flight to quality. A **black swan risk** is an event that is rare and difficult to predict but has an important impact. Investors with the appropriate time horizon and risk tolerance may make tactical changes to their investment choices as a result of these events. Long-term changes are unlikely to be necessary.

Risks with a medium-term impact may begin to impair companies' processes, costs, and investment opportunities, resulting in lower valuations. These risks tend to be distributed toward specific sectors, meaning they will affect some companies much more than others.

Long-term risks may have important environmental, social, governance, and other impacts. This can affect an investor's asset allocation—including choice of asset classes and investment styles—for a long-term horizon; however, the immediate impact on portfolios is likely to be more limited.

Exhibit 14: Risk Velocity



Note that some risks have more than one speed of impact on investments. The United Kingdom's referendum outcome to exit the EU had some immediate impacts, notably a decline in the value of the British pound sterling and other forms of market volatility. Over time, low-velocity impacts have become more apparent and more lasting for investors. Higher transaction costs and the unwind of previous forms of political cooperation—including the freedom of movement between the United Kingdom and the EU—are generating important impacts on investment outcomes and overall economic growth.

EXAMPLE 11

Geopolitical Risk Velocity and Investor Reaction

- Match the potential velocity of a geopolitical risk with the *most likely* investor reaction among those listed in the following table:

Geopolitical Risk Velocity	Investor Reaction
1. Low	A. Adjust investments in specific sectors
2. Medium	B. Flight to quality
3. High	C. Adjust asset allocation

Solution:

Option 1 (Low) matches with C (Adjust asset allocation).

Option 2 (Medium) matches with A (Adjust investments in specific sectors).

Option 3 (High) matches with B (Flight to quality).

EXAMPLE 12**Geopolitical Risk and the Investor**

1. Describe the three areas that an investor should consider when assessing geopolitical risk.

Solution:

An investor should consider the following three areas when assessing geopolitical risk: the likelihood it will occur, the velocity of its impact, and the size and nature of that impact. The likelihood that it will occur considers the probability that geopolitical risk will occur. The velocity of its impact is the pace at which it affects an investor portfolio. The impact can manifest in many ways and can be discrete in size or broad in nature.

Impact of Geopolitical Risk

A risk's impact on investor portfolios can manifest in many different ways. For the sake of this framework for assessing risk, suffice it to say that investors should consider the size of any risk's impact when gauging its importance to the investment process. A high-impact risk may merit extensive study of its drivers and motivations, whereas a low-impact risk may not. In addition, the size of a risk's impact may be compounded by external factors. For example, risk tends to have a greater impact on markets experiencing a general contraction or economic downturn.

Impact may also be discrete or broad in nature. Discrete impacts are those that affect only one company or sector at a time, whereas broad impacts are felt more holistically by a sector, a country, or the global economy. Cyber risks may be considered in this light. In the event of a cyberattack, only the companies and investment strategies exposed to that company will be affected. However, cyber risks may also have a broader impact by increasing monitoring, due diligence, and security costs for all companies and investors seeking to avoid them.

When assessing geopolitical risk for portfolio management, investors should consider all three geopolitical risk factors—likelihood, velocity, and size and nature of impact—together. For example, a highly likely risk with very little impact to the portfolio may not merit extensive analysis and investor attention. However, a highly impactful risk with a low likelihood of occurring may merit building a scenario for response but not regular monitoring and assessment. Between these extremes, investors must consider their goals and risk tolerance to identify high-priority risks.

Scenario Analysis

Geopolitical risks seldom develop in linear fashion, making it difficult to monitor and forecast their likelihood, velocity, and impact on a portfolio as well as difficult to address those changes through portfolio action. As a result, many investors deploy an approach that includes scenario analysis and signposting rather than a single point forecast.

Scenario analysis is the process of evaluating portfolio outcomes across potential circumstances or states of the world. Scenarios help investment teams understand where they stand with respect to a risk that might cause them to change their behavior. Scenario analysis can strengthen a team's conviction about its prioritization and calls to action, thereby helping it make good investment choices at opportune moments.

Scenarios can take the form of qualitative analysis, quantitative measurement, or both. A simple framework for qualitative scenario building begins with developing a base case for the event. What is the most impactful outcome of the risk? How likely

is that risk/outcome to occur in the first place? From there, investors can consider upside and downside scenarios. Is it a persistent tail risk or a short-term shock? How are markets likely to recover once the event has taken place? Considering alternate futures for key risks will drive more precise perspective around what constitutes key developments in the most important risks.

Quantitative scenarios can vary widely by sophistication. One form of a simple quantitative scenario is a stylized scenario in which portfolio sensitivity is measured against one key factor relevant to the portfolio, such as interest rates, asset prices, or exchange rates. Another involves using circumstances from extreme events to help build quantitative tests for portfolio resilience. It is important to have reasonable ambitions, however. Quantitative scenarios can be complicated because of the secondary and linked impacts of geopolitical risks to securities in the portfolio.

Good scenario building can prompt investors to alter their risk prioritization, making it a useful tool not only for tracking risks but also for deciding which portfolio actions may be valuable to take. Good scenario analysis also requires a consistent commitment of investors' time and resources. Teams that read similar research or speak with similar client groups may be affected by **groupthink**, the practice of thinking or making decisions as a group in a way that discourages creativity or individual responsibility. For scenario analysis to be useful in portfolio management, teams must work hard to build creative processes, identify and track scenarios, and assess the need for action on a regular basis.

Tracking Risks According to Signposts

To build a portfolio's resilience to unexpected change, asset managers develop processes in advance that allow for rapid course correction. In other words, by creating plans for addressing priority risks as they occur, investors can help reduce the events' impact on investment outcomes.

One important process is identifying signposts for priority risk. A **signpost** is an indicator, market level, data piece, or event that signals a risk is becoming more or less likely. An analyst can think of signposts like a traffic light. If quantitative and qualitative evidence suggest that a risk is low in likelihood, velocity, or impact, then the signposts are flashing green, or no action needed. If signposts are flashing amber, indicating that a risk is medium in likelihood, velocity, or impact, then higher caution and preparedness against that risk may be warranted. As a risk rises in likelihood, velocity, or impact, an action plan may be necessary.

Identifying signposts should equip a team to differentiate signal from noise and react when signposts flash red. For instance, when the market environment moves to either risk-on or risk-off, it is important to identify what actions should be taken next or communications made. Good signposts are anchored in the key assumptions made up-front around a scenario and mark whether a scenario is materializing.

Let's return to our example of the United Kingdom's Brexit referendum for context. Before 2014, signposts for geopolitical risk in the United Kingdom may have flashed "green." Attitudes toward European Union membership were divided, and while there is always a possibility of disruptive geopolitical change, a reduction in political cooperation with the European Union was not clearly defined. Then, when the referendum was announced in 2015, signposts for disruptive change became more likely and faster in potential velocity, resulting in a higher required level of portfolio manager attention. In May 2016, when phone polls suggested that the "leave" vote was moving to the majority, signposts flashed "red" and attentive portfolio managers prepared action plans for election day.

Identifying the right signposts can require some trial and error. A basic rule of thumb for distinguishing signal from noise is the distinction between politics and policy. For example, there can be a big difference in "politics" between two leaders,

but the “policies” they enact are what create larger or more durable portfolio impacts. Political developments can serve as meaningful signposts, as they can indicate a change in the risk’s likelihood or pace. However, analysts are frequently knocked off course by following developments that do not necessarily indicate a change in real economic or market outcomes. Instead, analysts should look for policy changes to guide their portfolio management decisions.

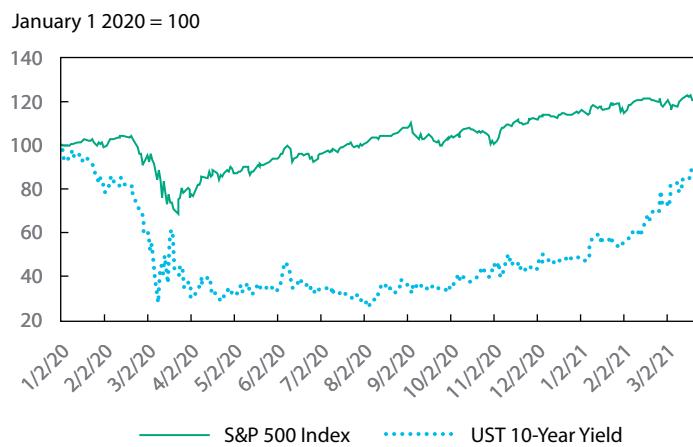
Some combinations of economic and financial market circumstances serve as strong warnings of potential trouble. For example, high inflation and deteriorating employment can signal political unrest. A pegged currency and rapidly declining export value (particularly for commodities exporters) can prompt a change in exchange rate policy. Often, particularly for emerging markets, these signposts will change before official data are released. If a portfolio relies on country-level economic conditions, data screens should be used to help identify any red flags early.

Manifestations of Geopolitical Risk

If geopolitical risk takes many forms, its impact on investor portfolios is just as multifaceted. High-velocity risks are most likely to manifest in market volatility through prompt changes in asset prices. Commonly affected asset prices include commodities, foreign exchange, equities, and bond prices (via changes in interest rates).

One example is the market response to economic shutdowns related to the COVID-19 pandemic (Exhibit 15). Using the United States as an example, the S&P 500 Index fell from a level of 3,386 on 19 February 2020 to 2,237.4 on 23 March, a decline of nearly 34%. Bonds also experienced volatility. Global investors’ “flight to safety” pushed up US bond prices. The US 10-year Treasury yield fell from 1.5661 to a low of 0.5407 during that period, a decline of nearly 68 percent. This volatility was not permanent but created ample risk—and opportunity—for investors during that time.

Exhibit 15: US Market Reaction to the COVID-19 Pandemic, 1 January 2020 to 23 March 2021



Source: Bloomberg Finance LP.

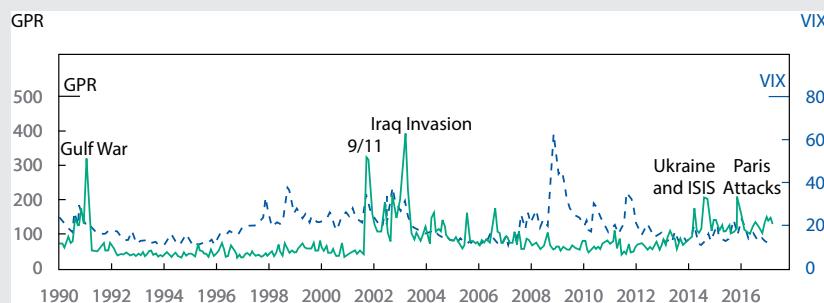
Low-velocity geopolitical risks can have a more prolonged impact on investor inputs. Sustained disruption may result in smaller revenues, higher costs, or both, which can negatively affect a company’s valuation. Here, too, the COVID-19 pandemic is

an instructive example. While risk asset valuations improved over the course of the pandemic, disruptions to mobility and consumption had long-lasting impacts on company revenues and supply chains.

For countries, regions, or sectors perceived to be at more consistent risk of geopolitical disruption, investors may require higher compensation, effectively increasing the discount rate investors use when valuing those securities. Portfolio investment flows face greater volatility because of geopolitical factors, and investors will factor in a higher risk premium. This dynamic is a key reason why asset prices in emerging and frontier markets are typically maintained at a discount to those in developed countries that are perceived to have a lower threat of risk.

GEOPOLITICAL RISK INDEX

In a 2019 paper, two analysts at the US Federal Reserve Board of Governors built the Geopolitical Risk Index (GPR) based on a tally of news articles covering geopolitical tensions and their impact on economic events. The purpose of the index is to measure real-time geopolitical risk as perceived by the press, the public, global investors, and policymakers in a way that is consistent over time.



Through their construction of the GPR, the authors made three important observations. First, they found that high levels of geopolitical risk reduce US investment, employment, and price level of the stock market. Second, and taking this observation deeper, the authors found that individual firm's investment falls more in industries positively exposed to geopolitical risk and that firms reduce investment in the wake of idiosyncratic geopolitical risk events. Finally, they studied the adverse effect of geopolitical events themselves as well as the threat of adverse events, finding that the threat of events had a larger impact over time.

EXAMPLE 13

Geopolitical Risk

- True or False: Higher-velocity geopolitical risks are *most* likely to have a prolonged impact on investor inputs.

Explain your selection.

A. True

B. False

Solution:

B is correct. Higher-velocity risks are most likely to manifest in market volatility via prompt changes in asset prices. Lower-velocity geopolitical risks are likely to have a prolonged impact on investor inputs. The terrorist attacks of 11 September 2001 in the United States are an example of a

higher-velocity geopolitical risk as the market had a sharp downturn and rebounded in a relatively short amount of time.

EXAMPLE 14**Geopolitical Risk**

1. True or false: The probability and impact of geopolitical risks influence relative asset price discount rates across emerging and developed markets.

Explain your selection.

- A. True
- B. False

Solution:

A is correct. For countries, regions, or sectors perceived to be at more consistent risk of geopolitical disruption, investors may require higher compensation, effectively increasing the discount rate used in valuation. When portfolio investment flows face greater volatility because of geopolitical factors, investors will factor in a higher risk premium. This explains why asset prices in emerging markets typically are maintained at a discount to those in developed countries perceived to have a lower threat of geopolitical risk, with the latter more likely to experience lower probability risks with lesser impacts.

Acting on Geopolitical Risk

Determining the likelihood, velocity, and impact of a risk may help an investor to assign priority to which risks might be most important. But if the risk does occur, what, if anything, can be done about it? Even if an investor could anticipate every risk and its impact on their portfolios, they must still consider whether and how to act in the face of such threats. A final step in incorporating geopolitical risk into the portfolio management process thus requires that geopolitical risk analysis be translated into investment action as appropriate for investor goals, risk tolerance, and time horizon.

Taking a top-down approach, asset allocators may consider geopolitical risk in their asset allocation strategy. The likelihood, velocity, and impact of risks may affect key capital markets assumptions as well as an asset allocator's positioning in certain countries or regions. For example, countries with a long history of using a multilateral approach may be considered more reliable investments and see increased investor flows. In contrast, those countries experiencing consistent military threat may have lower economic and investment growth potential because of consistent disruptions. The asset allocator would thus allocate more capital to the countries with lower expected risk profiles.

At the portfolio management level, investors can consider geopolitical risk as a factor in multifactor models. Imagine an analyst assessing global car manufacturers. A company with highly diversified production may be exposed to more risk (i.e., higher likelihood), but production would be less likely to come to a halt given the multiple production alternatives (i.e., lower impact). When making a buy or sell recommendation, the analyst may consider relative geopolitical risk exposure as a factor in their analysis. Disruptive threats may be used as a binary yes-or-no factor or they can affect the confidence intervals around factors related to momentum, valuation, market sentiment, or the economic cycle.

Ultimately, the importance of geopolitical risk to the investment process depends on investor objectives, risk tolerance, and time horizon. For an investor with low risk tolerance, reducing exposure to geopolitical risk may be appropriate, whether through low-volatility investment choices or through hedging.

For an investor with a long time horizon, a geopolitical event like an exogenous shock could be a buying opportunity. For an investor nearing retirement, however, that same exogenous shock can have a major negative impact on their terminal portfolio value.

The extensive political, economic, and financial cooperation in which countries, companies, and organizations participate may raise the stakes of geopolitical risk analysis for global investors. Changes in the style and momentum of international cooperation can have an important impact on capital markets. Global investors ignore those risks at their peril.

PRACTICE PROBLEMS

1. Which of the following statements regarding a country's political cooperation is *most* accurate?
 - A. If a country is engaged in military conflict, there is a higher cost to cooperation.
 - B. A country with few internal resources is not likely to rely on political cooperation.
 - C. Interest prioritization does not determine the depth and nature of political cooperation.
2. A consequence of one of the disadvantages of globalization is that:
 - A. pay differences between countries have narrowed.
 - B. emerging market trade flows have grown more important.
 - C. greater economic and financial cooperation has increased interdependence.
3. Which of the following outcomes is *most* likely a result of globalization?
 - A. Unequal gains
 - B. Increased independence
 - C. Decreased access to talent
4. A US company expanding critical spare part inventories for local customers made at its existing Canadian facility after a supply chain disruption is *most likely* using the coping tactic of:
 - A. reshoring the essentials.
 - B. reglobalizing production.
 - C. doubling down on key markets.
5. An example of a geopolitical multifaceted tool for furthering national interests is:
 - A. cabotage.
 - B. armed conflict.
 - C. nationalization of key export industries.
6. Which of these is most likely to be described as an event risk?
 - A. An earthquake
 - B. An election
 - C. An ongoing civil war
7. Exogenous risks are *best* described as those that:
 - A. are known and evolve and expand over a period of time.

- B. revolve around set dates, such as elections, new legislation, or other date-driven milestones, such as holidays or political anniversaries.
 - C. are sudden or unanticipated and impact either a country's cooperative stance, the ability of non-state actors to globalize, or both.
8. Which of the following statements about geopolitical threats in the investment environment is *most* accurate?
- A. Geopolitical risk is not always present in the investment environment.
 - B. Highly collaborative, interconnected countries are vulnerable to geopolitical risk.
 - C. Geopolitical risk tends to have less of an impact on markets already experiencing a general contraction or economic downturn.
9. An applicable conclusion drawn from the Geopolitical Risk Index (GPR) is that:
- A. high geopolitical risk results in tangible macroeconomic effects.
 - B. recurring geopolitical risk events lead to reduced corporate investment.
 - C. the adverse impact of actual events is greater over time than that of the threat of such events.
10. The basic geopolitical risk type *most likely* in comparison to have the smallest degree of uncertainty is:
- A. exogenous risk.
 - B. event risk.
 - C. thematic risk.

SOLUTIONS

1. A is correct. If a country is engaged in military conflict, there is a higher cost to cooperation. B is incorrect because a country with few internal resources is likely to rely on political cooperation. C is incorrect because interest prioritization does determine the depth and nature of political cooperation.
2. C is correct. Through the process of greater economic and financial cooperation, companies may become dependent on other countries' resources for their supply chains. On aggregate, this can result in the nation itself becoming dependent on other nations for certain resources (such as rare earth metals, cobalt, or copper). If there is a disruption to the supply chain, including via a moment of political non-cooperation, then firms may not be able to produce the good themselves.
A is incorrect because the narrowing of pay differences between countries results from a motivation to pursue globalization, which is related to one of its advantages. It is an important way for companies to increase profitability by reducing their costs. Although wage differentials remain, they are decreasing. B is incorrect because the proportion of flows between developed economies as a share of overall trade continuing to decline as emerging market trade flows rise is a function of increased international investment. This increased investment has provided beneficial aggregate economic benefits, such as increased choice, higher quality goods, increased competition among firms, higher efficiency, and increased labor mobility.
3. A is correct. Unequal accrual of economic and financial gains is a cost of globalization because improvement on the aggregate does not mean improvement for everyone. B is incorrect because globalization leads to *interdependency* as companies may become dependent on other countries' resources for their supply chains. C is incorrect because rather than decreased access to talent, a country might actually globalize to improve access to talent as cooperation and globalization can lead to increased access to resources.
4. A is correct. The COVID-19 pandemic has highlighted the need for certain essential supply chains to be rebuilt domestically for emergency situations, with availability of critical spare parts being an analogy. The close integration of the US and Canadian economies through the revised USMCA agreement effectively makes expanded production at an existing Canadian factory an example of reducing manufacturing risk by relocating to home countries via reshoring. B is incorrect because instead of reducing manufacturing risk by duplicating or fortifying its supply chain, the company is simply continuing to use its existing capacity more intensively. C is incorrect because although the production is intended to better supply its home market, there is no evidence that the company is expanding its presence in the US market or shifting its focus to the exclusion of available opportunities elsewhere.
5. A is correct. Cabotage is the right to transport passengers or goods within a country by a foreign firm. Many countries—including those with multilateral trade agreements—impose restrictions on cabotage across transportation subsectors, meaning that shippers, airlines, and truck drivers are not allowed to transport goods and services within another country's borders. Allowing cabotage requires coordination on areas like physical security and economic coordination, a highly multilateral (multifaceted tool) process. B is incorrect because armed conflict is the most extreme example of a national security tool. It can be either internal or external to a country in taking a direct and active approach to wield-

ing influence. C is incorrect because nationalization of key export industries is an economic tool. This process of transferring an activity or industry from private to state control is a non-cooperative approach to asserting economic control.

6. B is correct. Event risk evolves around set dates, such as elections and new legislation, or other date-driven milestones, such as holidays or political anniversaries known in advance. The other choices could not be known in advance. An earthquake (A) is an example of an exogenous risk. An ongoing civil war (C) is an example of a thematic risk.
7. C is correct. Exogenous risks are best described as those that are sudden or unanticipated and that affect either a country's cooperative stance, the ability of non-state actors to globalize, or both. A is incorrect because thematic (not exogenous) risks are known risks that evolve and expand over a period of time. B is incorrect because event (not exogenous) risks are those that evolve around set dates, including elections, new legislation, or other date-driven milestones.
8. B is correct. Although highly collaborative and globalized countries are, on balance, less likely to experience geopolitical risk because the political, economic, and financial costs of partners inflicting those risks are higher than less collaborative countries, the same interconnectedness may make them more vulnerable to geopolitical risk. A is incorrect because geopolitical risk is always present in the investment environment. C is incorrect because geopolitical risk tends to have a greater impact on markets already experiencing a general contraction or economic downturn.
9. A is correct. The GPR creators found that high levels of geopolitical risk reduce US investment, employment, and price level of the stock market. B is incorrect because firms reduce investment in the wake of idiosyncratic events, which would be unlikely to repeat. C is incorrect because the threat of an event was shown to have a larger impact over time than that of the actual events themselves.
10. B is correct. *Event risk* evolves around set dates, such as elections or new legislation, or other date-driven milestones, such as holidays or political anniversaries. Analysts can thus look to political calendars as a predictable starting point for determining the occurrence of event risk, with time to devise a suitable response. A is incorrect because exogenous risk is sudden and unanticipated. Examples include sudden uprisings, invasions, or the aftermath of natural disasters. The timing and range of its effects thus have the greatest unknowns. C is incorrect because thematic risks are known risks that evolve and expand over a period of time. Climate change, pattern migration, the rise of populist forces, and the ongoing threat of terrorism fall into this category. These are more foreseeable than exogenous risks, but with their extended interval of exposure, planned responses likely require continual adjustments.

LEARNING MODULE

6

International Trade

by Usha Nair-Reichert, PhD, and Daniel Robert Witschi, PhD, CFA.

Usha Nair-Reichert, PhD, is at Georgia Institute of Technology (USA). Daniel Robert Witschi, PhD, CFA (Switzerland).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	describe the benefits and costs of international trade
<input type="checkbox"/>	compare types of trade restrictions, such as tariffs, quotas, and export subsidies, and their economic implications
<input type="checkbox"/>	explain motivations for and advantages of trading blocs, common markets, and economic unions

INTRODUCTION

1

From an investment perspective, it is important for global investors to understand existing trade policies. Such policies can affect the volume and value of trade and thus can affect the return on investment. Investors need to be aware of potential changes in the government's trade policy. Such changes have important implications for firm profitability and growth by affecting the demand for its products and its pricing. There has been much debate among economists on the role of trade and trade policy and its impact on the overall economy. This learning module examines the benefits and cost of international trade. It then describes trade restrictions and their implications and discusses the motivation for, and advantages of, the different types of trading blocs or trade agreements.

LEARNING MODULE OVERVIEW



- The most compelling arguments supporting international trade are:
 - countries gain from exchange and specialization,
 - industries experience greater economies of scale,
 - households and firms have greater product variety,
 - competition is increased, and
 - resources are allocated more efficiently.

- Newer models of trade focus on the gains from trade that result from economies of scale, greater product variety, and increased competition.
- Opponents of free trade point to the potential for greater income inequality and the loss of jobs in developed market countries as a result of import competition.
- The fact that trade increases overall welfare does not mean that every individual consumer and producer is better off. What it does mean is that the winners could, in theory, compensate the losers and still be better off.
- Trade restrictions (or trade protection) are government policies that limit the ability of domestic households and firms to trade freely with other countries.
- Tariffs are taxes that a government levies on imported goods. The primary objective of tariffs is to protect domestic industries that produce the same or similar goods. They also may aim to reduce a trade deficit.
- The net welfare effect of tariffs is the sum of consumer surplus, producer surplus and government tax revenue. The loss in consumer surplus is greater than the sum of the gain in producer surplus and government revenue and results in a deadweight loss to the country's welfare.
- A quota restricts the quantity of a good that can be imported into a country, generally for a specified period of time. A voluntary export restraint (VER) is a trade barrier under which the exporting country agrees to limit its exports of the good to its trading partners to a specific number of units.
- An export subsidy is a payment by the government to a firm for each unit of a good that is exported. Its goal is to stimulate exports.
- A regional trading bloc is a group of countries that have signed an agreement to reduce and progressively eliminate barriers to trade and the movement of factors of production among the members of the bloc.
- There are many different types of regional trading blocs, depending on the level of integration that takes place. These include free trade areas, customs union, common market, and economic union.
- Trade creation occurs when regional integration results in the replacement of higher-cost domestic production by lower-cost imports from other members. Trade diversion occurs when lower-cost imports from non-member countries are replaced with higher-cost imports from members.

2

BENEFITS AND COSTS OF TRADE



describe the benefits and costs of international trade

Over the past few decades, the global economy has experienced rapid growth in trade and a growing interdependence among countries. This has led to a debate among policy makers over whether the expansion of trade has been helpful for individual national economies. This lesson examines the possible benefits and costs of international trade.

Benefits and Costs of International Trade

The benefits and costs of international trade are widely debated. The most compelling arguments supporting international trade are as follows: countries gain from exchange and specialization, industries experience greater economies of scale, households and firms have greater product variety, competition is increased, and resources are allocated more efficiently.

Gains from exchange occur when trade enables each country to receive a higher price for its exports (and greater profit) or pay a lower price for imported goods instead of producing these goods domestically at a higher cost (i.e., less efficiently). This exchange, in turn, leads to a more efficient allocation of resources by increasing production of the export good and reducing production of the import good in each country (trading partner). This efficiency allows for consumption of a larger bundle of goods, thus increasing overall welfare. The fact that trade increases overall welfare does not mean, of course, that every individual consumer and producer is better off. What it does mean is that the winners could, in theory, compensate the losers and still be better off.

Trade also leads to greater efficiency by fostering specialization based on comparative advantage. Traditional trade models, such as the Ricardian model and the Heckscher–Ohlin model, focus on specialization and trade according to comparative advantage arising from differences in technology and factor endowments, respectively.

Newer models of trade focus on the gains from trade that result from economies of scale, greater product variety, and increased competition. In an open economy, increased competition from foreign firms reduces the monopoly power of domestic firms and forces them to become more efficient, in contrast to a closed economy. Industries that exhibit increasing returns to scale (e.g., the automobile and steel industries) benefit from increased market size as a country starts trading because the average cost of production declines as output increases in these industries. Monopolistically competitive models of trade have been used to explain why there is significant two-way trade (known as *intra-industry trade*) between countries within the same industry. Intra-industry trade occurs when a country exports and imports goods in the same product category or classification.

A monopolistically competitive industry has many firms; each firm produces a unique or differentiated product: it does not have any exit or entry barriers, and long-run economic profits are zero. In such a model, even though countries may be similar, they gain from trade because each country focuses on the production and export of one or more varieties of the good and imports other varieties of the good. For example, the European Union exports and imports different types of cars. Consumers gain from having access to a greater variety of final goods. Firms benefit from greater economies of scale because firms both within and outside the EU are able to sell their goods in both markets. Hence, scale economies allow firms to benefit from the larger market size and experience lower average cost of production as a result of trade.

Research suggests that trade liberalization can lead to increased real (i.e., inflation-adjusted) GDP although the strength of this relationship is still debated. The positive influence of trade on GDP can arise from more efficient allocation of resources, learning by doing, higher productivity, knowledge spillovers, and trade-induced changes in policies and institutions that affect the incentives for innovation. In industries that embrace “learning by doing,” such as the semiconductor industry, the cost of production per unit declines as output increases because of expertise and experience

acquired in the process of production. Trade can lead to increased exchange of ideas, freer flow of technical expertise, and greater awareness of changing consumer tastes and preferences in global markets. It also can contribute to the development of higher quality and more effective institutions and policies that encourage domestic innovation. For example, studies have shown that foreign research and development (R&D) has beneficial effects on domestic productivity, which become stronger the more open an economy is to foreign trade. For example, some estimate that about a quarter of the benefits of R&D investment in a G-7 country accrues to their trading partners. Consider Logitech, a Swiss company that manufactures computer mice. To win original equipment manufacturer (OEM) contracts from IBM and Apple, Logitech needed to develop innovative designs and provide high-volume production at a low cost. So in the late 1980s, the company moved to Taiwan Region, which had a highly qualified labor force, competent parts suppliers, a rapidly expanding local computer industry, and offered Logitech space in a science park at a competitive rate. Soon thereafter, Logitech was able to secure the Apple contract.

Opponents of free trade point to the potential for greater income inequality and the loss of jobs in developed countries as a result of import competition. As a country moves toward free trade, adjustments will be made in domestic industries that are exporters as well as in those that face import competition. Resources (investments) may need to be reallocated into or out of an industry depending on whether that industry is expanding (exporters) or contracting (i.e., facing import competition). As a result of this adjustment process, less-efficient firms may be forced to exit the industry, which, in turn, may lead to higher unemployment and the need for displaced workers to be retrained for jobs in expanding industries. The counter argument is that although there may be short-term and even some medium-term costs, these resources are likely to be more effectively (re-)employed in other industries in the long run. Nonetheless, the adjustment process is virtually certain to impose costs on some groups of stakeholders.

QUESTION SET



Consider two countries that each produce two goods. Suppose the cost of producing cotton relative to lumber is lower in Cottonland than in Lumberland.

1. How would trade between the two countries affect the lumber industry in Lumberland?

Solution:

The lumber industry in Lumberland would benefit from trade. Because the cost of producing lumber relative to producing cotton is lower in Lumberland than in Cottonland (i.e., lumber is relatively cheap in Lumberland), Lumberland will export lumber and the industry will expand.

2. How would trade between the two countries affect the lumber industry in Cottonland?

Solution:

The lumber industry in Cottonland would not benefit from trade, at least in the short run. Because lumber is relatively expensive to produce in Cottonland, the domestic lumber industry will shrink as lumber is imported from Lumberland.

3. What would happen to the lumber industry workers in Cottonland in the long run?

Solution:

The overall welfare effect in both countries is positive. However, in the short run, many lumber producers in Cottonland (and cotton producers in Lumberland) are likely to find themselves without jobs as the lumber industry in Cottonland and the cotton industry in Lumberland contract. Those with skills that also are needed in the other industry may find jobs fairly quickly. Others are likely to do so after some re-training. In the long run, displaced workers should be able to find jobs in the expanding export industry. Those who remain in the import-competing industry, however, may be permanently worse off because their industry-specific skills are now less valuable. Thus, even in the long run, trade does not necessarily make every stakeholder better off. But the winners could compensate the losers and still be better off, so the overall welfare effect of opening trade is positive.

4. What is the meaning of the expression “gains from trade”?

Solution:

Gains from trade imply that the overall benefits of trade outweigh the losses from trade. It does not mean that all stakeholders (producers, consumers, government) benefit (or benefit equally) from trade.

5. What are some of the benefits from trade?

Solution:

Some of the benefits from trade include the following: gains from exchange and specialization based on relative cost advantage; gains from economies of scale as the companies add new markets for their products; greater variety of products available to households and firms; greater efficiency from increased competition; and more efficient allocation of resources.

TRADE RESTRICTIONS AND AGREEMENTS—TARIFFS, QUOTAS, AND EXPORT SUBSIDIES

3



compare types of trade restrictions, such as tariffs, quotas, and export subsidies, and their economic implications

Trade restrictions (or trade protection) are government policies that limit the ability of domestic households and firms to trade freely with other countries. Examples of trade restrictions include tariffs, import quotas, voluntary export restraints (VER), subsidies, embargoes, and domestic content requirements. **Tariffs** are taxes that a government levies on imported goods. **Quotas** restrict the quantity of a good that can be imported into a country, generally for a specified period of time. A voluntary export restraint is similar to a quota but is imposed by the exporting country. An **export subsidy** is paid by the government to the firm when it exports a unit of a good that is being subsidized. The goal is to promote exports, but it reduces welfare by encouraging production and trade that is inconsistent with comparative advantage. **Domestic content provisions** stipulate that some percentage of the value added or

components used in production should be of domestic origin. Trade restrictions are imposed by countries for several reasons, including protecting established domestic industries from foreign competition, protecting new industries from foreign competition until they mature (infant industry argument), protecting and increasing domestic employment, protecting strategic industries for national security reasons, generating revenue from tariffs (especially for developing countries), and retaliation against trade restrictions imposed by other countries.

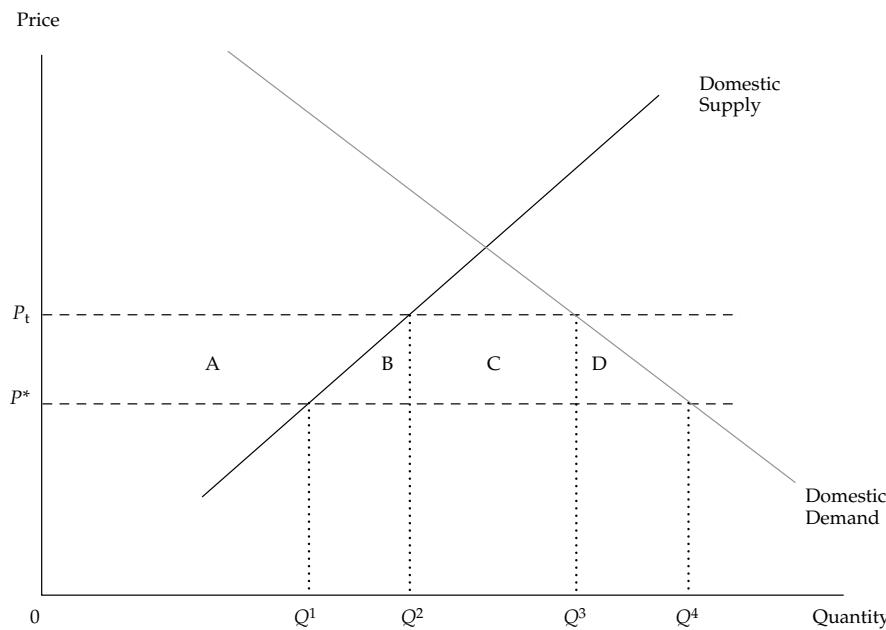
Capital restrictions are defined as controls placed on foreigners' ability to own domestic assets or domestic residents' ability to own foreign assets. Thus, in contrast with trade restrictions, which limit the openness of goods markets, capital restrictions limit the openness of financial markets and will be addressed later.

Tariffs

Tariffs are taxes that a government levies on imported goods. The primary objective of tariffs is to protect domestic industries that produce the same or similar goods. They also may aim to reduce a trade deficit. Tariffs reduce the demand for imported goods by increasing their price above the free trade price. The economic impact of a tariff on imports in a small country is illustrated in Exhibit 1. In this context, a small country is not necessarily small in size, population, or GDP. Instead, a **small country** is one that is a price taker in the world market for a product and cannot influence the world market price. For example, by many measures, Brazil is a large country, but it is a price taker in the world market for cars. A large country, however, is a large importer of the product and can exercise some influence on price in the world market. When a large country imposes a tariff, the exporter reduces the price of the good to retain some of the market share it could lose if it did not lower its price.

This reduction in price alters the terms of trade and represents a redistribution of income from the exporting country to the importing country. So, in theory, it is possible for a large country to increase its welfare by imposing a tariff if (1) its trading partner does not retaliate and (2) the deadweight loss as a result of the tariff (see below) is smaller than the benefit of improving its terms of trade. Global welfare, however, would still experience a net reduction—the large country cannot gain by imposing a tariff unless it imposes an even larger loss on its trading partner.

In Exhibit 1, the world price (free trade price) is P^* . Under free trade, domestic supply is Q^1 , domestic consumption is Q^4 , and imports are Q^1Q^4 . After the imposition of a per-unit tariff t , the domestic price increases to P_t , which is the sum of the world price and the per-unit tariff t . At the new domestic price, domestic production increases to Q^2 and domestic consumption declines to Q^3 , resulting in a reduction in imports to Q^2Q^3 .

Exhibit 1: Welfare Effects of Tariff and Import Quota

The welfare effects of the tariff can be summarized as follows:

- Consumers suffer a loss of consumer surplus because of the increase in price. In Exhibit 1, this effect is represented by areas A + B + C + D.
- Local producers gain producer surplus from a higher price for their output. This effect is represented by area A.
- The government gains tariff revenue on imports Q^2Q^3 . This effect is represented by area C.

The net welfare effect is the sum of these three effects and is summarized in Exhibit 2. The loss in consumer surplus is greater than the sum of the gain in producer surplus and government revenue and results in a deadweight loss to the country's welfare of B + D.

Exhibit 2: Welfare Effects of an Import Tariff or Quota

Importing Country	
Consumer surplus	- (A + B + C + D)
Producer surplus	+A
Tariff revenue or Quota rents	+C
National welfare	- B - D

Tariffs create deadweight loss because they give rise to inefficiencies on both the consumption and production side. B represents inefficiencies in production. Instead of being able to import goods at the world price P^* , tariffs encourage inefficient producers whose cost of production is greater than P^* to enter (or remain in) the market, leading to an inefficient allocation of resources. On the consumption side, tariffs prevent mutually beneficial exchanges from occurring because consumers who were willing to pay more than P^* but less than P_t are now unable to consume the good.

EXAMPLE 1**Analysis of a Tariff**

South Africa manufactures 110,000 tons of paper. However, domestic demand for paper is 200,000 tons. The world price for paper is USD5.00 per ton. South Africa will import 90,000 tons of paper from the world market at free trade prices. If the South African government (a small country) decides to impose a tariff of 20 percent on paper imports, the price of imported paper will increase to USD6.00. Domestic production after the imposition of the tariff increases to 130,000 tons, while the quantity demanded declines to 170,000 tons.

1. Calculate the loss in consumer surplus arising from the imposition of the tariff on imported paper.

Solution:

The loss in consumer surplus = $USD1 \times 170,000 + 1/2 \times USD1 \times 30,000$
 $= USD185,000$. This calculation is represented by areas A + B + C + D in Exhibit 1.

2. Calculate the gain in producer surplus arising from the imposition of the tariff.

Solution:

Gain in producer surplus = $USD1 \times 110,000 + 1/2 \times (USD1 \times 20,000)$
 $= USD120,000$; Area A in Exhibit 1.

3. Calculate the gain in government revenue arising from the imposition of the tariff.

Solution:

Change in government revenue = $USD1 \times 40,000 = USD40,000$; Area C in Exhibit 1.

4. Calculate the deadweight loss arising from the imposition of the tariff.

Solution:

Deadweight loss because of the tariff = $1/2 \times USD1 \times 20,000 + 1/2 \times USD1 \times 30,000 = USD 25,000$; Areas B + D in Exhibit 1.

Quotas

A quota restricts the quantity of a good that can be imported into a country, generally for a specified period of time. An **import license** specifies the quantity that can be imported. For example, the European Union operates a system of annual import quotas for steel producers who are not members of the World Trade Organization. A key difference between tariffs and quotas is that the government is able to collect the revenue generated from a tariff. This effect is uncertain under a quota. With quotas, foreign producers can often raise the price of their goods and earn greater profits than they would without the quota. These profits are called **quota rents**. In Exhibit 1, if the quota is Q^2Q^3 , the equivalent tariff that will restrict imports to Q^2Q^3 is t and the domestic price after the quota is P_t . This is the same as the domestic price after the tariff t was imposed. Area C, however, is now the quota rent or profits that are likely to be captured by the foreign producer rather than tariff revenue that is captured by the domestic government. If the foreign producer or foreign government captures the

quota rent, C, then the welfare loss to the importing country, represented by areas B + D + C in Exhibit 1, under a quota is greater than under the equivalent tariff. If the government of the country that imposes the quota can capture the quota rents by auctioning the import licenses for a fee, then the welfare loss under the quota is similar to that of a tariff, represented by areas B + D.

A VER is a trade barrier under which the exporting country agrees to limit its exports of the good to its trading partners to a specific number of units. The main difference between an import quota and a VER is that the former is imposed by the importer, whereas the latter is imposed by the exporter. The VER allows the quota rent resulting from the decrease in trade to be captured by the exporter (or exporting country), whereas in the case of an import quota, there is ambiguity regarding who captures the quota rents. Hence, a VER results in welfare loss in the importing country. For example, in 1981, the Japanese government imposed VERs on automobile exports to the United States.

Export Subsidies

An export subsidy is a payment by the government to a firm for each unit of a good that is exported. Its goal is to stimulate exports. But it interferes with the functioning of the free market and may distort trade away from comparative advantage. Hence, it reduces welfare. *Countervailing duties* are duties that are levied by the importing country against subsidized exports entering the country. As an example, agricultural subsidies in developed countries, notably the EU, have been a contentious issue in trade negotiations with emerging market countries and developed market countries that are agricultural exporters, such as New Zealand and Australia.

In the case of an export subsidy, the exporter has the incentive to shift sales from the domestic to the export market because it receives the international price plus the per-unit subsidy for each unit of the good exported. This scenario raises the price in the domestic market by the amount of the subsidy in the small country case (price before subsidy plus subsidy). In the large country case, the world price declines as the large country increases exports. The net welfare effect is negative in both the large and small country cases, with a larger decline in the large country case. In the large country case, the decline in world prices implies that a part of the subsidy is transferred to the foreign country, unlike in the small country case.

Exhibit 3 summarizes some of these effects.

Exhibit 3: Summary of Some of the Effects of Trade Restrictions

Panel A. Effects of Alternative Trade Policies

	Tariff	Import Quota	Export Subsidy	VER
Impact on	Importing country	Importing country	Exporting country	Importing country
Producer surplus	Increases	Increases	Increases	Increases
Consumer surplus	Decreases	Decreases	Decreases	Decreases

Panel A. Effects of Alternative Trade Policies

	Tariff	Import Quota	Export Subsidy	VER
Government revenue	Increases	Mixed (depends on whether the quota rents are captured by the importing country through sale of licenses or by the exporters)	Falls (government spending rises)	No change (rent to foreigners)
National welfare	Decreases in small country Could increase in large country	Decreases in small country Could increase in large country	Decreases	Decreases

Panel B. Effects of Alternative Trade Policies on Price, Production, Consumption, and Trade

	Tariff	Import Quota	Export Subsidy	VER
Impact on price	Importing country Increases	Importing country Increases	Exporting country Increases	Importing country Increases
Domestic consumption	Decreases	Decreases	Decreases	Decreases
Domestic production	Increases	Increases	Increases	Increases
Trade	Imports decrease	Imports decrease	Exports increase	Imports decrease

EXAMPLE 2**Analysis of Trade Restrictions**

Thailand, a small country, has to decide whether to impose a tariff or a quota on the import of computers. You are considering investing in a local firm that is a major importer of computers.

- What will be the impact of a tariff on prices, quantity produced, and quantity imported in Thailand (the importing country)?

Solution:

A tariff imposed by a small country, such as Thailand, raises the price of computers in the importing country, reduces the quantity imported, and increases domestic production.

- If Thailand imposes a tariff, what will the impact be on prices in the exporting country?

Solution:

A tariff imposed by a small country would not change the price of computers in the exporting country.

- How would a tariff affect consumer surplus, producer surplus, and government revenue in Thailand?

Solution:

When a small country imposes a tariff, it reduces consumer surplus, increases producer surplus, and increases government revenue in that country.

4. Explain whether the net welfare effect of a tariff is the same as that of a quota.

Solution:

The quota can lead to a greater welfare loss than a tariff if the quota rents are captured by the foreign government or foreign firms.

5. Which policy, a tariff or a quota, would be most beneficial to the local importer in which you may invest and why?

Solution:

A tariff will hurt importers because it will reduce their share of the computer market in Thailand. The impact of a quota depends on whether or not the importers can capture a share of the quota rents. Assuming importers can capture at least part of the rents, they will be better off with a quota.

6. If Thailand were to negotiate a VER with the countries from which it imports computers, would this be better or worse than an import quota for the local importing firm in which you may invest? Why?

Solution:

The VER would not be better for the local importer than the import quota and would most likely be worse. Under the VER, all of the quota rents will be captured by the exporting countries whereas with an import quota at least part of the quota rents may be captured by local importers.

It is important to understand existing trade policies and the potential for policy changes that may affect return on investment. Changes in the government's trade policy can affect the pattern and value of trade and may result in changes in industry structure. These changes may have important implications for firm profitability and growth because they can affect the goods a firm can import or export; change demand for its products; affect its pricing policies; and create delays through increased paperwork, procurement of licenses, approvals, and so on. For example, changes in import policies that affect the ability of a firm to import vital inputs for production may increase the cost of production and reduce firm profitability.

TRADING BLOCS AND REGIONAL INTEGRATION**4**

explain motivations for and advantages of trading blocs, common markets, and economic unions

There has been a proliferation of trading blocs or regional trading agreements (RTA) in recent years. Important examples of regional integration include the United States-Mexico-Canada Agreement (USMCA) and the European Union (EU). A regional trading bloc is a group of countries that have signed an agreement to reduce and progressively eliminate barriers to trade and movement of factors of production among the members of the bloc. It may or may not have common trade barriers against countries that are not members of the bloc.

Types Of Trading Blocs

There are many different types of regional trading blocs, depending on the level of integration that takes place. **Free trade areas** (FTA) are one of the most prevalent forms of regional integration in which all barriers to the flow of goods and services among members have been eliminated. However, each country maintains its own policies against non-members. The USMCA among the United States, Canada, and Mexico is an example of an FTA. A **customs union** extends the FTA by not only allowing free movement of goods and services among members but also by creating a common trade policy against non-members. In 1947, Belgium, the Netherlands, and Luxembourg (Benelux) formed a customs union that became part of the European Community in 1958.

The **common market** is the next level of economic integration that incorporates all aspects of the customs union and extends it by allowing free movement of factors of production among members. The Southern Cone Common Market (MERCOSUR) of Argentina, Brazil, Paraguay, and Uruguay is an example of a common market. An **economic union** requires an even greater degree of integration. It incorporates all aspects of a common market and in addition requires common economic institutions and coordination of economic policies among members. The European Community became the European Union in 1993. If the members of the economic union decide to adopt a common currency, then it is also a **monetary union**. For example, with the adoption of the euro, 19 EU member countries also formed a monetary union.

EXAMPLE 3

Trading Blocs

1. Chile and Australia have a free trade with each other but have separate trade barriers on imports from other countries. Chile and Australia are a part of a(n):
 - A. FTA.
 - B. economic union.
 - C. customs union.
 - D. common market.

Solution:

A is correct. Chile and Australia do not have a customs union because they do not have a common trade policy with respect to other trade partners (C is incorrect). A common market or an economic union entail even more integration (B and D are incorrect).

2. An RTA that removes all tariffs on imports from member countries, and has common external tariffs against all non-members, but does not advance further in deepening economic integration is called a(n):
 - A. FTA.
 - B. economic union.
 - C. customs union.

D. common market.

Solution:

C is correct. A basic FTA does not entail common external tariffs (A is incorrect), whereas a common market and an economic union entail integration beyond common external tariffs (B and D are incorrect).

Regional Integration

Regional integration is popular because eliminating trade and investment barriers among a small group of countries is easier, politically less contentious, and quicker than multilateral trade negotiations under the World Trade Organization (WTO). The WTO is a negotiating forum that deals with the rules of global trade between nations and where member countries can sort out trade disputes. Trade negotiations launched by the WTO have included contentious issues of specific concern to developing countries, such as the cost of implementing trade policy reform in developing countries, market access in developed countries for developing countries' agricultural products, and access to affordable pharmaceuticals in developing countries. Despite decades of negotiations, limited progress has been made on these and other major issues. Hence, it is not surprising to see a renewed interest in bilateral and multilateral trade liberalization on a smaller scale. Policy coordination and harmonization are also easier among a smaller group of countries. Regional integration can be viewed as a movement toward freer trade.

Regional integration results in preferential treatment for members compared with non-members and can lead to changes in the patterns of trade. Member countries move toward freer trade by eliminating or reducing trade barriers against each other, leading to a more efficient allocation of resources. Regional integration also may result in trade and production being shifted from a lower-cost non-member who still faces trade barriers to a higher-cost member who faces no trade barriers. This shift leads to a less-efficient allocation of resources and could reduce welfare. Hence, two static effects are direct results of the formation of the customs union: trade creation and trade diversion.

Trade Creation and Diversion

Trade creation occurs when regional integration results in the replacement of higher-cost domestic production by lower-cost imports from other members. For example, consider two hypothetical countries, Qualor and Vulcan. Qualor produces 10 million shirts annually and imports 2 million shirts from Vulcan, which has a lower cost of production. Qualor has 10 percent tariffs on imports from Vulcan. Qualor and Vulcan then agree to form a customs union. Qualor reduces its production of shirts to 7 million and now imports 11 million shirts from Vulcan. The decline in Qualor's domestic production (from 10 million to 7 million shirts) is replaced by importing 3 million additional shirts from the low-cost producer, Vulcan. This scenario represents trade creation. The rest of the additional imports (6 million shirts) represents increased consumption by Qualor's consumers because the price of shirts declines after formation of the custom union.

Trade diversion occurs when lower-cost imports from non-member countries are replaced with higher-cost imports from members. In the previous example, suppose Qualor initially imposes a 10 percent tariff on imports from both Vulcan and Aurelia. Aurelia is the lowest-cost producer of shirts, so Qualor initially imports 2 million shirts from Aurelia instead of from Vulcan. Qualor and Vulcan then form a customs union, which eliminates tariffs on imports from Vulcan but maintains a 10 percent tariff on imports from Aurelia. Now trade diversion could occur if the free trade price on imports from Vulcan is lower than the price on imports from Aurelia.

Even though Aurelia is the lowest-cost producer, it may be a higher-priced source of imports because of the tariff. If this is the case, then Qualor will stop importing from Aurelia, a non-member, and divert its imports to Vulcan, a member of the RTA. Both trade creation and trade diversion are possible in an RTA. If trade creation is larger than trade diversion, then the net welfare effect is positive. There are concerns, however, that this may not always be the case.

Costs and Benefits of Regional Trading Blocs

The benefits ascribed to free trade—greater specialization according to comparative advantage, reduction in monopoly power because of foreign competition, economies of scale from larger market size, learning by doing, technology transfer, knowledge spillovers, greater foreign investment, and better quality intermediate inputs at world prices—also apply to regional trading blocs. In addition, fostering greater interdependence among members of the regional trading bloc reduces the potential for conflict. Members of the bloc also have greater bargaining power and political clout in the global economy by acting together instead of as individual countries.

Considerable spillover of growth across borders is evident among member countries of the Organisation for Economic Co-operation and Development (OECD), which are highly integrated both as a group and within their own geographic regions. The long-run growth of integrated countries is interconnected because members have greater access to each other's markets. Strong growth in any RTA country could have a positive impact on growth in other RTA member countries. RTAs also enhance the benefits of good policy and lead to convergence in living standards. For example, growth spillovers are likely to be much smaller among Sub-Saharan African countries because of a lack of integration arising from deficiencies in RTAs and inadequate levels of transportation and telecommunications infrastructure. One study estimated what the cumulative loss in real GDP between 1970 and 2000 would have been if Switzerland, which is landlocked and fully integrated with both its immediate neighbors and the world economy, had been subject to the same level of spillovers as the Central African Republic. Under such a scenario, Switzerland's GDP per capita in 2000 would have been 9.3 percent lower. The cumulative GDP loss would have been USD334 billion (constant US dollars, 2000), which was the equivalent of 162 percent of Switzerland's real GDP in 2000.

Although regional integration has many advantages, it may impose costs on some groups. For example, there was significant concern in the United States that fewer trade restrictions and especially low-skilled labor-intensive imports from Mexico could hurt low-skilled workers. Adjustment costs arose as import competition caused inefficient firms to exit the market, and the workers in those firms were at least temporarily unemployed as they sought new jobs. However, the surviving firms experienced an increase in productivity, and US consumers benefited from the increase in product varieties imported from Mexico. One study estimated that the product varieties exported from Mexico to the United States had grown by an average of 2.2 percent a year across all industries. While USMCA imposed estimated private costs of nearly USD5.4 billion a year in the United States during 1994–2002, these costs were offset by an average welfare gain of USD5.5 billion a year accruing from increased varieties imported from Mexico. Consumer gains from more varieties of products continued over time as long as the imports continued, whereas adjustment costs arising from job losses declined over time. In 2003, the gain from increased product varieties from Mexico was USD11 billion, far exceeding the adjustment costs of USD5.4 billion.

It is important to recognize, however, that workers displaced by regional integration may have to bear long-term losses if they are unable to find jobs with wages comparable with the jobs they lost or they remain unemployed for a long period of time. For example, although import competition was certainly not the only factor that led

to a dramatic contraction of the US automobile industry, the impact on employment in that industry is likely to be permanent and many former autoworkers, especially older workers, may never find comparable jobs.

Concerns regarding national sovereignty, especially where big and small nations may be part of the same bloc, also have been an impediment to the formation of FTAs. The proposal for a South Asian regional bloc has faced challenges regarding India's role because it is one of the biggest economies in the region.

Challenges to Deeper Integration

The formation of an RTA and its potential progression from an FTA to deeper integration in the form of a customs union, common market, or economic union face at least two challenges. First, cultural differences and historical considerations—for example, wars and conflicts—may complicate the social and political process of integration. Second, maintaining a high degree of economic integration limits the extent to which member countries can pursue independent economic and social policies. Free trade and mobility of labor and capital tend to thwart policies aimed at controlling relative prices or quantities within a country, while balance-of-payments and fiscal credibility considerations limit the viability of divergent macroeconomic policies. This situation is especially true in the case of a monetary union because monetary policy is not under the control of individual countries and currency devaluation or revaluation is not available as a tool to correct persistent imbalances. When persistent imbalances do arise, they may lead to a crisis that spills over to other countries facing similar problems. The Greek fiscal crisis in 2010 is a case in point. In May 2010, Standard & Poor's reduced the credit ratings on Greece's government from investment grade to junk status. It also downgraded the government debt of Spain and Portugal. These countries were suffering from a combination of high government deficits and slow GDP growth. The credit downgrades increased fears that Greece, in particular, would default on its debt and cause economic turmoil not only among the healthier countries in the EU but also in the United States and Asia. The EU and the International Monetary Fund (IMF) agreed on a USD145 billion (EUR110 billion) bailout for Greece in May 2010 and provided Ireland with a financing package of about USD113 billion (EUR85 billion) in November 2010.

Investment Implications

Regional integration is important from an investment perspective because it offers new opportunities for trade and investment. The cost of doing business in a large, single, regional market is lower and firms can benefit from economies of scale. Note, however, that differences in tastes, culture, and competitive conditions still exist among members of a trading bloc. These differences may limit the potential benefits from investments within the bloc. In addition, depending on the level of integration and the safeguards in place, problems faced by individual member countries in an RTA may quickly spread to other countries in the bloc.

QUESTION SET



Bagopia, Cropland, and Technopia decide to enter into an RTA. In the first stage, they decide to sign an FTA. After several successful years, they decide that it is time to form a common market.

1. Does an FTA make exporting firms in member countries more attractive as investment options?

Solution:

The first stage, where there is free movement of goods and services among RTA members, is called an FTA. It makes exporting firms a more attractive investment proposition because they are able to serve markets in member countries without the additional costs imposed by trade barriers.

2. How does the common market affect firms doing business in these countries compared with an FTA?

Solution:

Unlike an FTA, a common market allows for free movement of factors of production, such as labor and capital, among the member economies. Like an FTA, it provides access to a much larger market and free movement of goods and services. But the common market can create more profitable opportunities for firms than an FTA by allowing them to locate production in and purchase components from anywhere in the common market according to comparative advantage.

PRACTICE PROBLEMS

1. Which of the following statements *best* describes the benefits of international trade?
 - A. Countries gain from exchange and specialization.
 - B. Countries receive lower prices for their exports and pay higher prices for imports.
 - C. Countries gain from trade because all individuals and companies benefit in the long term.
2. Which of the following statements *best* describes the costs of international trade?
 - A. Countries without an absolute advantage in producing a good cannot benefit significantly from international trade.
 - B. Resources may need to be allocated into or out of an industry and less-efficient companies may be forced to exit an industry, which in turn may lead to higher unemployment.
 - C. Loss of manufacturing jobs in developed countries as a result of import competition means that developed countries benefit far less than developing countries from trade.
3. Suppose the cost of producing tea relative to copper is lower in Tealand than in Copperland. With trade, the copper industry in Copperland would *most likely*:
 - A. expand.
 - B. contract.
 - C. remain stable.
4. Which type of trade restriction would most likely increase domestic government revenue?
 - A. Tariff
 - B. Import quota
 - C. Export subsidy
5. Which of the following trade restrictions is likely to result in the greatest welfare loss for the importing country?
 - A. A tariff
 - B. An import quota
 - C. A voluntary export restraint
6. A large country can:
 - A. benefit by imposing a tariff.
 - B. benefit with an export subsidy.

- C. not benefit from any trade restriction.
7. If Brazil and South Africa have free trade with each other, a common trade policy against all other countries, but no free movement of factors of production between them, then Brazil and South Africa are part of a(n):
- A. customs union.
 - B. common market.
 - C. FTA.
8. Which of the following factors *best* explains why regional trading agreements are more popular than larger multilateral trade agreements?
- A. Minimal displacement costs
 - B. Trade diversions benefit members
 - C. Quicker and easier policy coordination

SOLUTIONS

1. A is correct. Countries gain from exchange when trade enables each country to receive a higher price for exported goods or pay a lower price for imported goods. This leads to more efficient resource allocation and allows consumption of a larger variety of goods.
2. B is correct. Resources may need to be reallocated into or out of an industry, depending on whether that industry is an exporting sector or an import-competing sector of that economy. As a result of this adjustment process, less-efficient companies may be forced to exit the industry, which in turn may lead to higher unemployment and the need for retraining so that displaced workers may find jobs in expanding industries.
3. A is correct. The copper industry in Copperland would benefit from trade. Because the cost of producing copper relative to producing tea is lower in Copperland than in Tealand, Copperland will export copper and the industry will expand.
4. A is correct. The imposition of a tariff will most likely increase domestic government revenue. A tariff is a tax on imports collected by the importing country's government.
5. C is correct. With a voluntary export restraint, the price increase induced by restricting the quantity of imports (= quota rent for equivalent quota = tariff revenue for equivalent tariff) accrues to foreign exporters or the foreign government.
6. A is correct. By definition, a large country is big enough to affect the world price of its imports and exports. A large country can benefit by imposing a tariff if its terms of trade improve by enough to outweigh the welfare loss arising from inefficient allocation of resources.
7. A is correct. A customs union extends an FTA by not only allowing free movement of goods and services among members, but also by creating common trade policy against non-members. Unlike a more integrated common market, a customs union does not allow free movement of factors of production among members.
8. C is correct. Regional trading agreements are politically less contentious and quicker to establish than multilateral trade negotiations (e.g., under the World Trade Organization). Policy coordination and harmonization is easier among a smaller group of countries.

LEARNING MODULE

7

Capital Flows and the FX Market

by William A. Barker, PhD, CFA, Paul D. McNelis, and Jerry Nickelsburg, PhD.

William A. Barker, PhD, CFA (Canada). Paul D. McNelis is at Gabelli School of Business, Fordham University (USA). Jerry Nickelsburg, PhD, is at the Anderson School of Management, University of California, Los Angeles (USA).

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	describe the foreign exchange market, including its functions and participants, distinguish between nominal and real exchange rates, and calculate and interpret the percentage change in a currency relative to another currency
<input type="checkbox"/>	describe exchange rate regimes and explain the effects of exchange rates on countries' international trade and capital flows
<input type="checkbox"/>	describe common objectives of capital restrictions imposed by governments

INTRODUCTION

1

The foreign exchange market, which is the largest trading market in the world, facilitates international trade and capital flows. Numerous participants use this market for a wide variety of financial, business, trade, and hedging purposes. As with any trading market, the foreign exchange market uses various terms and conventions that allow participants to understand quoting mechanisms and the factors affecting pricing and then to conduct trades. International capital flows are the primary determinant of short- to medium-term exchange rate movements, and exchange rate movements affect the trade balance between countries. Given the relative economic stability and objectives of different national governments, countries use a range of exchange rate regimes, which are described in this module. These lessons introduce and expand upon these topics to lay the groundwork for a more detailed understanding of the foreign exchange market.

LEARNING MODULE OVERVIEW

- The foreign exchange market is the largest market in the world.
- Nominal spot exchange rates are quoted in the market and are inputs for analysts to determine real exchange rates, which reflect the relationship between domestic and foreign price levels and indicate the relative purchasing power between countries. To track exchange rate movements, calculating the percentage change in a currency relative to another currency enables market participants to understand price changes and use these percentages in market trades.
- Exchange rate regimes can be floating or fully fixed, and various political and economic forces drive countries to use one of a number of intermediate regimes.
- The impact of exchange rates and other factors on a country's trade balance is mirrored by their impact on that country's capital flows.
- Although the free flow of capital between countries is most beneficial economically, governments may restrict capital inflows or outflows to address domestic policy and strategic or defense-related objectives. These restrictions allow governments to avoid capital flight in times of macroeconomic crisis and limit capital inflows, which may hurt the competitiveness of domestic firms.

2**THE FOREIGN EXCHANGE MARKET AND EXCHANGE RATES**

describe the foreign exchange market, including its functions and participants, distinguish between nominal and real exchange rates, and calculate and interpret the percentage change in a currency relative to another currency

Introduction and the Foreign Exchange Market

The foreign exchange (FX) market—the market in which currencies are traded against each other—is by far the world's largest market. Current estimates put daily turnover at approximately USD6.6 trillion for 2019. This is about 10 to 15 times larger than daily turnover in global fixed-income markets and about 50 times larger than global turnover in equities.

The FX market is a truly global market that operates 24 hours a day, each business day. It involves market participants from every time zone connected through electronic communications networks that link players as large as multibillion-dollar investment funds and as small as individuals trading for their own account—all brought together in real time. International trade would be impossible without the trade in currencies that facilitates it and so too would cross-border capital flows that connect all financial markets globally through the FX market.

These factors make FX a key market for investors and market participants to understand. The world economy is increasingly transnational in nature, with both production processes and trade flows often determined more by global factors than by domestic considerations. Likewise, investment portfolio performance increasingly reflects global determinants because pricing in financial markets responds to the array of investment opportunities available worldwide, not just locally. All of these factors funnel through, and are reflected in, the FX market. As investors shed their “home bias” and invest in FX, the exchange rate—the price at which foreign-currency-denominated investments are valued in terms of the domestic currency—becomes an increasingly important determinant of portfolio performance.

Even investors adhering to a purely “domestic” portfolio mandate are increasingly affected by what happens in the FX market. Given the globalization of the world economy, most large companies depend heavily on their foreign operations (e.g., by some estimates about 30 percent of S&P 500 Index earnings are from outside the United States). Almost all companies are exposed to some degree of foreign competition, and the pricing for domestic assets—equities, bonds, real estate, and others—also depend on demand from foreign investors. All of these various influences on investment performance reflect developments in the FX market.

The FX Market

To understand the FX market, it is necessary to become familiar with some of its basic conventions. Individual currencies often are referred to by standardized three-letter codes that the market has agreed upon through the International Organization for Standardization (ISO). Exhibit 1 lists some of the major global currencies and their identification codes.

Exhibit 1: Standard Currency Codes

Three-Letter Currency Code	Currency
AUD	Australian dollar
BRL	Brazilian real
CAD	Canadian dollar
CHF	Swiss franc
CNY	Chinese yuan
EUR	Euro
GBP	British pound sterling
HKD	Hong Kong dollar
INR	Indian rupee
JPY	Japanese yen
KRW	South Korean won
MXN	Mexican peso
NOK	Norwegian krone
NZD	New Zealand dollar
RUB	Russian ruble
SEK	Swedish krona
SGD	Singapore dollar
USD	US dollar
ZAR	South African rand

It is important to understand the difference between referring to an *individual currency* and an *exchange rate*. One can hold an individual currency (e.g., in a EUR100 million deposit); an exchange rate, however, is the price of one currency in terms of another (e.g., the exchange rate between the euro and the US dollar). An individual currency can be singular, but two currencies always are involved in an exchange rate: the price of one currency relative to another. The exchange rate is the number of units of one currency (called the *price currency*) that one unit of another currency (called the *base currency*) will buy. An equivalent way of describing the exchange rate is as the cost of one unit of the base currency in terms of the price currency.

This distinction between individual currencies and exchange rates is important because, as we will see in a later lesson, these three-letter currency codes can be used both ways. For example, when used as an exchange rate in the professional FX market, EUR is understood to be the exchange rate between the euro and US dollar. Be aware of the context (either as a currency or as an exchange rate) in which these three-letter currency codes are being used. To avoid confusion, this lesson will identify exchange rates using the convention of “A/B,” referring to the number of units of currency A that one unit of currency B will buy. For example, a USD/EUR exchange rate of 1.1700 means that 1 euro will buy 1.1700 US dollars (i.e., 1 euro costs 1.1700 US dollars). In this case, the euro is the base currency and the US dollar is the price currency. A decrease in this exchange rate would mean that the euro costs less or that fewer US dollars are needed to buy one euro. In other words, a decline in this exchange rate indicates that the US dollar is *appreciating* against the euro or, equivalently, the euro is *depreciating* against the US dollar.

These exchange rates are referred to as *nominal* exchange rates. In contrast, *real* exchange rates are indexes that often are constructed by economists and other market analysts to assess changes in the relative purchasing power of one currency compared with another. Creating these indexes requires adjusting the nominal exchange rate by using the price levels in each country of the currency pair (hence the name “real exchange rates”) to compare the relative purchasing power between countries.

In a world of homogenous goods and services, and with no market frictions or trade barriers, the relative purchasing power across countries would tend to equalize: Why would you pay more, in real terms, domestically for a “widget” if you could import an identical “widget” from overseas at a cheaper price? This basic concept is the intuition behind a theory known as purchasing power parity (PPP), which describes the long-term equilibrium of nominal exchange rates. PPP asserts that nominal exchange rates adjust so that identical goods (or baskets of goods) will have the same price in different markets. Or, put differently, the purchasing power of different currencies is equalized for a standardized basket of goods.

In practice, the conditions required to enforce PPP are not satisfied: Goods and services are not identical across countries; countries typically have different baskets of goods and services produced and consumed; many goods and services are not traded internationally; there are trade barriers and transaction costs (e.g., shipping costs and import taxes); and capital flows are at least as important as trade flows in determining nominal exchange rates. As a result, nominal exchange rates exhibit persistent deviations from PPP. Moreover, relative purchasing power among countries displays a weak, if any, tendency toward long-term equalization. A simple example of a cross-country comparison of the purchasing power of a standardized good is the Big Mac index produced by the *Economist*, which shows the relative price of this standardized hamburger in different countries. The Big Mac index shows that fast-food hamburger prices can vary widely internationally and that this difference in purchasing power is typical of most goods and services. Hence, movements in real exchange rates provide meaningful information about changes in relative purchasing power among countries.

Consider the case of an individual who wants to purchase goods from a foreign country. The individual would be able to buy fewer of these goods if the nominal spot exchange rate for the foreign currency appreciated or if the foreign price level increased. Conversely, the individual could buy more foreign goods if the individual's domestic income increased. (For this example, we will assume that changes in the individual's income are proportional to changes in the domestic price level.) Hence, in *real* purchasing power terms, the real exchange rate that an individual faces is an increasing function of the nominal exchange rate (quoted in terms of the number of units of domestic currency per one unit of foreign currency) and the foreign price level and a decreasing function of the domestic price level. The *higher* the real exchange rate is, the *fewer* foreign goods, in real terms, the individual can purchase and the *lower* that individual's relative purchasing power will be compared with the other country.

An equivalent way of viewing the real exchange rate is that it represents the relative price levels in the domestic and foreign countries. Mathematically, we can represent the foreign price level in terms of the domestic currency as follows:

$$\text{Foreign price level in domestic currency} = S_{d/f} \times P_f$$

where $S_{d/f}$ is the spot exchange rate (quoted in terms of the number of units of domestic currency per one unit of foreign currency) and P_f is foreign price level quoted in terms of the foreign currency. We can define the domestic price level, in terms of the domestic currency, as P_d . Hence, the ratio between the foreign and domestic price levels is as follows:

$$\text{Real exchange rate}_{(d/f)} = (S_{d/f}P_f)/P_d = S_{d/f} \times (P_f/P_d).$$

For example, for a British consumer wanting to buy goods made in the Eurozone, the real exchange rate (defined in GBP/EUR terms; note that the domestic currency for the United Kingdom is the price currency, not the base currency) will be an increasing function of the nominal spot exchange rate (GBP/EUR) and the Eurozone price level and a decreasing function of the UK price level. This is written as follows:

$$\text{Real exchange rate}_{\frac{\text{GBP}}{\text{EUR}}} = S_{\frac{\text{GBP}}{\text{EUR}}} \times \left(\frac{\text{CPI}_{\text{eur}}}{\text{CPI}_{\text{UK}}} \right).$$

We can examine the effect of movements in the domestic and foreign price levels, and the nominal spot exchange rate, on the real purchasing power of an individual in the United Kingdom wanting to purchase Eurozone goods. Assume that the nominal spot exchange rate (GBP/EUR) increases by 10 percent, the Eurozone price level increases by 5 percent, and the UK price level increases by 2 percent. The change in the real exchange rate is then as follows:

$$\begin{aligned} & \left(1 + \frac{\Delta S_d}{S_d} \right) \times \frac{\left(1 + \frac{\Delta P_f}{P_f} \right)}{\left(1 + \frac{\Delta P_d}{P_d} \right)} - 1 \\ &= (1 + 10\%) \times \frac{1 + 5\%}{1 + 2\%} - 1 \approx 10\% + 5\% - 2\% \approx 13\%. \end{aligned}$$

In this case, the real exchange rate for the UK-based individual has *increased* about 13 percent, meaning that it now costs *more*, in real terms, to buy Eurozone goods. Or put differently, the UK individual's real purchasing power relative to Eurozone goods has *declined* by about 13 percent. An easy way to remember this relationship is to consider the real exchange rate (stated with the domestic currency as the price currency) as representing the real price you face to purchase foreign goods and services: The *higher* the price is (real exchange rate), the *lower* your relative purchasing power will be.

The real exchange rate for a currency can be constructed for the domestic currency relative to a single foreign currency or relative to a basket of foreign currencies. In either case, these real exchange rate indexes depend on the assumptions made by the

analyst creating them. Several investment banks and central banks create proprietary measures of real exchange rates. Note that real exchange rates are *not* quoted or traded in global FX markets: They are only indexes created by analysts to understand the international competitiveness of an economy and the real purchasing power of a currency.

In this context, real exchange rates can be useful for understanding trends in international trade and capital flows and hence can be seen as one of the influences on nominal spot exchange rates. As an example, consider the exchange rate between the Indian rupee and the US dollar. During 2018, the nominal rupee exchange rate against the US dollar (INR/USD) rose by approximately 6.7 percent—meaning that the US dollar appreciated against the rupee. However, the annual inflation rates in the United States and India were different during 2018—approximately 2.5 percent for the United States and 4.7 percent for India. This means that the real exchange rate (in INR/USD terms) was depreciating less rapidly than the nominal INR/USD exchange rate:

$$(1 + \% \Delta S_{\frac{INR}{USD}}) \times \frac{(1 + \% \Delta P_{US})}{(1 + \% \Delta P_{India})} - 1 \approx +6.7\% + 2.5\% - 4.7\% \approx 4.5\%.$$

This combination of a much weaker rupee and a higher Indian inflation rate meant that the real exchange rate faced by India was increasing, thus decreasing Indian purchasing power in US dollar terms.

Movements in real exchange rates can have a similar effect as movements in nominal exchange rates in terms of affecting relative prices and hence trade flows. Even if the nominal spot exchange rate does not move, differences in inflation rates between countries affect their relative competitiveness.

Although real exchange rates can exert some influence on nominal exchange rate movements, they are only one of many factors; it can be difficult to disentangle all of these inter-relationships in a complex and dynamic FX market. As discussed earlier, PPP is a poor guide to predicting future movements in nominal exchange rates because these rates can deviate from PPP equilibrium—and even continue to trend away from their PPP level—for years at a time. Hence, it should not be surprising that real exchange rates, which reflect changes in relative purchasing power, have a poor track record as a predictor of future nominal exchange rate movements.

EXAMPLE 1

Nominal and Real Exchange Rates

An investment adviser located in Sydney, Australia, is meeting with a local client who is looking to diversify her domestic bond portfolio by adding investments in fixed-rate, long-term bonds denominated in the Hong Kong dollar. The client frequently visits Hong Kong SAR, and many of her annual expenses are denominated in the Hong Kong dollar. The client, however, is concerned about the foreign currency risks of offshore investments and whether the investment return on her Hong Kong-dollar-denominated investments will maintain her purchasing power—both domestically (i.e., for her Australian-dollar-denominated expenses) and for her foreign trips (i.e., Hong Kong-dollar-denominated expenses for her visits to Hong Kong SAR). The investment adviser explains the effect of changes in nominal and real exchange rates to the client and illustrates this explanation by making the following statements:

Statement 1 All else equal, an increase in the nominal AUD/HKD exchange rate will lead to an increase in the Australia-dollar-denominated value of your foreign investment.

Statement 2 All else equal, an increase in the nominal AUD/HKD exchange rate means that your relative purchasing power for your Hong Kong SAR trips will increase (based on paying for your trip with the income from your Hong Kong-dollar-denominated bonds).

Statement 3 All else equal, an increase in the Australian inflation rate will lead to an increase in the real exchange rate (AUD/HKD). A higher real exchange rate means that the relative purchasing power of your Australian-dollar-denominated income is higher.

Statement 4 All else equal, a decrease in the nominal exchange rate (AUD/HKD) will decrease the real exchange rate (AUD/HKD) and increase the relative purchasing power of your Australian-dollar-denominated income.

To demonstrate the effects of the changes in inflation and nominal exchange rates on relative purchasing power, the adviser uses the following scenario:

“Suppose that the AUD/HKD exchange rate increases by 5 percent, the price of goods and services in Hong Kong SAR goes up by 5 percent, and the price of Australian goods and services goes up by 2 percent.”

1. Statement 1 is:

- A. correct.
- B. incorrect, because based on the quote convention the investment's value would be decreasing in Australian dollar terms.
- C. incorrect, because the nominal Australian dollar value of the foreign investments will depend on movements in the Australian inflation rate.

Solution:

A is correct. Given the quoting convention, an increase in the AUD/HKD rate means that the base currency (Hong Kong dollar) is appreciating (one Hong Kong dollar will buy more Australian dollars). This increases the nominal value of the Hong Kong-dollar-denominated investments when measured in Australian dollar terms.

2. Statement 2 is:

- A. correct.
- B. incorrect, because purchasing power is not affected in this case.
- C. incorrect, because based on the quote convention, the client's relative purchasing power would be decreasing.

Solution:

B is correct. When paying for Hong Kong-dollar-denominated expenses with Hong Kong-dollar-denominated income, the value of the AUD/HKD spot exchange rate (or any other spot exchange rate) would not be relevant. In fact, this is a basic principle of currency risk management: reducing FX risk exposures by denominating assets and liabilities (or income and expenses) in the same currency.

3. Statement 3 is:

- A. correct.
- B. incorrect with respect to the real exchange rate only.

- C. incorrect with respect to both the real exchange rate and the purchasing power of Australian dollar -denominated income.

Solution:

C is correct. An increase in the Australian (i.e., domestic) inflation rate means that the real exchange rate (measured in domestic/foreign, or AUD/HKD, terms) would be decreasing, not increasing. Moreover, an increase in the real exchange rate ($R_{AUD/HKD}$) would be equivalent to a reduction of the purchasing power of the Australian client: Goods and services denominated in Hong Kong dollar would cost more.

4. Statement 4 is:

- A. correct.
- B. incorrect with respect to the real exchange rate.
- C. incorrect with respect to the purchasing power of Australian dollar-denominated income.

Solution:

A is correct. As the spot AUD/HKD exchange rate decreases, the Hong Kong dollar is depreciating against the Australian dollar; or equivalently, the Australian dollar is appreciating against the Hong Kong dollar. This is reducing the real exchange rate ($R_{AUD/HKD}$) and increasing the Australian client's purchasing power.

5. Based on the adviser's scenario and assuming that the Hong Kong dollar value of the Hong Kong dollar bonds remained unchanged, the nominal Australian dollar value of the client's Hong Kong dollar investments would:

- A. decrease by about 5 percent.
- B. increase by about 5 percent.
- C. remain about the same.

Solution:

B is correct. As the AUD/HKD spot exchange rate increases by 5 percent, the Hong Kong dollar is appreciating against the Australian dollar by 5 percent and, all else equal, the value of the Hong Kong-dollar-denominated investment is increasing by 5 percent in Australian dollar terms.

6. Based on the adviser's scenario, the change in the relative purchasing power of the client's Australian-dollar-denominated income is *closest* to:

- A. -8 percent.
- B. +8 percent.
- C. +12 percent.

Solution:

A is correct. The real exchange rate ($R_{AUD/HKD}$) is expressed as follows:

$$R_{\frac{AUD}{HKD}} = S_{\frac{AUD}{HKD}} \times \frac{P_{HKD}}{P_{AUD}}.$$

The information in the adviser's scenario can be expressed as follows:

$$\% \Delta R_{\frac{AUD}{HKD}} \approx \% \Delta S_{\frac{AUD}{HKD}} + \% \Delta P_{HKD} - \% \Delta P_{AUD} \approx +5\% + 5\% - 2\% \approx +8\%.$$

Because the real exchange rate (expressed in AUD/HKD terms) has gone up by about 8 percent, the real purchasing power of the investor based in Australia has declined by about 8 percent. This can be seen from the fact that

Hong Kong dollar has appreciated against the Australian dollar in nominal terms, and the Hong Kong SAR price level has also increased. This increase in the cost of Hong Kong SAR goods and services (measured in Australian dollars) is only partially offset by the small (2 percent) increase in the investor's income (assumed equal to the change in the Australian price level).

QUESTION SET



1. A decrease in the real exchange rate (quoted in terms of domestic currency per unit of foreign currency) is *most likely* to be associated with an increase in which of the following?
 - A. Foreign price level.
 - B. Domestic price level.
 - C. Nominal exchange rate.

Solution:

B is correct. The real exchange rate (quoted in terms of domestic currency per unit of foreign currency) is given as follows:

$$\text{Real exchange rate}_{(d/f)} = S_{d/f} \times (P_f/P_d).$$

An increase in the domestic price level (P_d) *decreases* the real exchange rate because it implies an *increase* in the relative purchasing power of the domestic currency.

Market Participants

We now turn to the counterparties that participate in FX markets. As mentioned previously, an extremely diverse universe of market participants ranges in size from multi-billion-dollar investment funds to individuals trading for their own account (including foreign tourists exchanging currencies at airport kiosks).

To understand the various market participants, it is useful to separate them into broad categories. One broad distinction is between what the market refers to as the *buy side* and the *sell side*. The sell side generally consists of large FX trading banks (such as Citigroup, UBS, and Deutsche Bank); the buy side consists of clients who use these banks to undertake FX transactions (i.e., buy FX products) from the sell-side banks.

The buy side can be further broken down into several categories:

- *Corporate accounts:* Corporations of all sizes undertake FX transactions during cross-border purchases and sales of goods and services. Many of their FX flows also are related to cross-border investment flows—such as international mergers and acquisitions (M&A) transactions, investment of corporate funds in foreign assets, and foreign currency borrowing.
- *Real money accounts:* These are investment funds managed by insurance companies, mutual funds, pension funds, endowments, exchange-traded funds (ETFs), and other institutional investors. These accounts are referred to as real money because they usually are restricted in their use of leverage or financial derivatives. This feature distinguishes them from leveraged accounts (discussed next), although many institutional investors often engage in some form of leverage, either directly through some use of borrowed funds or indirectly using financial derivatives.

- *Leveraged accounts:* This category, often referred to as the professional trading community, consists of hedge funds, proprietary trading shops, commodity trading advisers (CTAs), high-frequency algorithmic traders, and the proprietary trading desks at banks—and, indeed, almost any active trading account that accepts and manages FX risk for profit. The professional trading community accounts for a large and growing proportion of daily FX market turnover. These active trading accounts also have a wide diversity of trading styles. Some are macro-hedge funds that take long-term FX positions based on their views of the underlying economic fundamentals of a currency. Others are high-frequency algorithmic traders that use technical trading strategies (such as those based on moving averages or Fibonacci levels) and whose trading cycles and investment horizons are sometimes measured in milliseconds.
- *Retail accounts:* The simplest example of a retail account is the archetypical foreign tourist exchanging currency at an airport kiosk. However, as electronic trading technology has reduced the barriers to entry into FX markets and the costs of FX trading, there has been a huge surge in speculative trading activity by retail accounts—consisting of individuals trading for their own accounts as well as smaller hedge funds and other active traders. This category also includes households using electronic trading technology to move their savings into foreign currencies (e.g., this is relatively widespread among households in Japan). It is estimated that retail trading accounts for as much as 10 percent of all spot transactions in some currency pairs and that this proportion is growing.
- *Governments:* Public entities of all types often have FX needs, ranging from relatively small (e.g., maintaining consulates in foreign countries) to large (e.g., military equipment purchases or maintaining overseas military bases). Sometimes these flows are purely transactional—the business simply needs to be done—and sometimes government FX flows reflect, at least in part, the public policy goals of the government. Some government FX business resembles that of investment funds, although sometimes with a public policy mandate as well. In some countries, public sector pension plans and public insurance schemes are run by a branch of the government. One example is the Caisse de dépôt et placement du Québec, which was created by the Québec provincial government in Canada to manage that province's public sector pension plans. The Caisse, as it is called, is a relatively large player in financial markets, with nearly CAD420 billion of assets under management at the end of 2021. Although it has a mandate to invest these assets for optimal return, it is also called upon to help promote the economic development of Québec. Many governments—both at the federal and provincial/state levels—issue debt in foreign currencies; this, too, creates FX flows. Such supranational agencies as the World Bank and the African Development Bank issue debt in a variety of currencies as well.
- *Central banks:* These entities sometimes intervene in FX markets to influence either the level or trend in the domestic exchange rate. This often occurs when the central banks judge their domestic currency to be too weak and when the exchange rate has overshot any concept of equilibrium level (e.g., because of a speculative attack) to the degree that the exchange rate no longer reflects underlying economic fundamentals. Alternatively, central banks also intervene when the FX market has become so erratic and dysfunctional that end-users such as corporations can no longer transact necessary FX business. Conversely, central banks sometimes intervene when they believe that their domestic currency has become too strong, to the

point that it undercuts that country's export competitiveness. The Bank of Japan intervened against yen strength versus the US dollar in 2004 and again in March 2011 after the massive earthquake and nuclear disaster. Similarly, in 2010, 2013, and again in 2015, the Swiss National Bank intervened against strength in the Swiss franc versus the euro by selling the Swiss franc on the euro–Swiss (CHF/EUR) cross-rate. Central bank reserve managers are frequent participants in FX markets to manage their country's FX reserves. In this context, they act much like real money investment funds—although generally with a cautious, conservative mandate to safeguard the value of their country's FX reserves. The FX reserves of some countries are enormous, and central bank participation in FX markets can sometimes have a material impact on exchange rates even when these reserve managers are not intervening for public policy purposes. Total FX reserves reached nearly USD13 trillion at the end of 2021.

This largely reflects the rapid growth in foreign reserves held by Asian central banks, because these countries typically run large current account surpluses with the United States and other developed market economies. Reserve accumulation by energy-exporting countries in the Middle East and elsewhere is also a factor. As of the end of 2021, nearly 60 percent of global allocated currency reserves were held in US dollars, while just over 20 percent was held in euros, the second most widely held currency in central bank FX reserves.

- *Sovereign wealth funds (SWFs)*: Many countries with large current account surpluses have diverted some of the resultant international capital flows into SWFs rather than into FX reserves managed by central banks. Although SWFs are government entities, their mandate is usually more oriented toward purely investment purposes rather than public policy purposes. As such, SWFs can be thought of as akin to real money accounts, although some SWFs can employ derivatives or engage in aggressive trading strategies. It generally is understood that SWFs use their resources to help fulfill the public policy mandate of their government owners. The SWFs of many current account surplus countries (such as exporting countries in East Asia (e.g., Singapore) or oil-exporting countries (e.g., Kuwait) are enormous, and their FX flows can be an important determinant of exchange rate movements in almost all of the major currency pairs.

The sell side generally consists of the FX dealing banks that sell FX products to the buy side. The following sell-side distinctions can be made.

- A large and growing proportion of the daily FX turnover is accounted for by the very largest money center dealing banks, such as Deutsche Bank, Citigroup, UBS, HSBC, and a few other multinational banking behemoths. Maintaining a competitive advantage in FX requires huge fixed-cost investments in the electronic technology that connects the FX market, and it also requires a broad, global client base. As a result, only the largest first-tier banks are able to compete successfully in providing competitive price quotes to clients across the broad range of FX products. In fact, among the largest FX dealing banks, a large proportion of their business is crossed internally, meaning that these banks are able to connect buyers and sellers within their own extremely diverse client base and have no need to show these FX flows outside of the bank.
- All other banks fall into the second and third tier of the FX market sell side. Many of these financial institutions are regional or local banks with well-developed business relationships, but they lack the economies of scale,

broad global client base, or information technology (IT) expertise required to offer competitive pricing across a wide range of currencies and FX products. In many cases, these are banks in emerging markets that do not have the business connections or credit lines required to access the FX market on a cost-effective basis on their own. As a result, these banks often outsource FX services by forming business relationships with the larger tier-one banks; otherwise, they depend on the deep, competitive liquidity provided by the largest FX market participants.

The categories presented are based on functions that are closely associated with the named groups. However, in some cases, functions typifying a group also may be assumed by or shared with another group. For example, sell-side banks provide FX price quotes. Hedge funds and other large players, however, may access the professional FX market on equal terms with the dealing banks and effectively act as market makers.

One of the most important ideas to draw from this categorization of market participants is that there is an extremely wide variety of FX market participants, reflecting a complex mix of trading motives and strategies that can vary with time. Most market participants reflect a combination of hedging and speculative motives in tailoring their FX risk exposures. Among public sector market participants, public policy motives also may be a factor. The dynamic, complex interaction of FX market participants and their trading objectives make it difficult to analyze or predict movements in FX rates with any precision or to describe the FX market adequately with simple characterizations.

Market Composition

In this section, we first describe components of the FX market, then present a descriptive overview of the global FX market drawn from the 2016 Triennial Survey undertaken by the Bank for International Settlements (BIS).

In addition to spot transactions, the FX market includes forward transactions and **FX swaps**. Forwards are transactions made using forward exchange rates with settlement at agreed-upon future dates, and forward rates can be used to manage FX risk. The combination of spot and forward transactions can be used to create FX swaps. These instruments are used for hedging purposes and to raise foreign currency at more favorable rates, and their trading constitutes the largest daily volume of the FX market.

The BIS is an umbrella organization for the world's central banks. Every three years, participating central banks undertake a survey of the FX market in their jurisdictions, the results of which are aggregated and compiled at the BIS. The survey, taken in April 2016, gives a broad indication of the current size and distribution of global FX market flows.

As of April 2016, the BIS estimates that average daily turnover in the traditional FX market (composed of spot, outright forward, and FX swap transactions) totaled approximately USD5.1 trillion. Exhibit 2 shows the approximate percentage allocation among FX product types, including both traditional FX products and exchange-traded FX derivatives. Note that this table of percentage allocations adds exchange-traded derivatives to the BIS estimate of average daily turnover of USD5.1 trillion; the "Spot" and "Outright forwards" categories include only transactions that are not executed as part of a swap transaction.

Exhibit 2: FX Turnover by Instrument

Instrument	FX Turnover (%)
Spot	33
Outright forwards	14
Swaps ^a	49
FX options	5
Total	100

Note: Swaps includes FX and currency swaps. An "FX swap" is not the same as a "currency swap"; a currency swap is generally used for multiple periods and payments. May not add to 100% due to rounding.

^a Includes both FX and currency swaps.

The survey also provides a percentage breakdown of the average daily flows between sell-side banks (called the interbank market), between banks and financial customers (all non-bank financial entities, such as real money and leveraged accounts, SWFs, and central banks), and between banks and non-financial customers (such as corporations, retail accounts, and governments). The breakdown is provided in Exhibit 3. It bears noting that the proportion of average daily FX flow accounted for by financial clients is much larger than that for non-financial clients. The BIS also reports that the proportion of financial client flows has been growing rapidly, and in 2010, it exceeded interbank trading volume for the first time. This underscores the fact that only a minority of the daily FX flow is accounted for by corporations and individuals buying and selling foreign goods and services. Huge investment pools and professional traders account for a large and growing proportion of the FX business.

Exhibit 3: FX Flows by Counterparty

Counterparty	FX Flows (%)
Interbank	42
Financial clients	51
Non-financial clients	8

Note: May not add to 100% due to rounding.

The 2016 BIS survey also identified the top five currency pairs in terms of their percentage share of average daily global FX turnover. These are shown in Exhibit 4. Note that each of these most active pairs includes the US dollar (USD).

Exhibit 4: FX Turnover by Currency Pair

Currency Pair	Percent of Market (%)
USD/EUR	23.1
JPY/USD	17.8
USD/GBP	9.3
USD/AUD	5.2
CAD/USD	4.3

The largest proportion of global FX trading occurs in London, followed by New York. This means that FX markets are most active between approximately 8:00 a.m. and 11:30 a.m. New York time, when banks in both cities are open. (The official London close is at 11:00 a.m. New York time, but London markets remain relatively active for a period after that.) Tokyo is the third-largest FX trading hub.

EXAMPLE 2

Market Participants and Composition of Trades

The investment adviser based in Sydney, Australia, makes the following statements to her client when describing some of the basic characteristics of the FX market:

Statement 1 “FX transactions for spot settlement see the most trade volume in terms of average daily turnover because the FX market is primarily focused on settling international trade flows.”

Statement 2 “The most important FX market participants on the buy side are corporations engaged in international trade; on the sell side they are the local banks that service their FX needs.”

1. Statement 1 is:

- A. correct.
- B. incorrect with respect to the importance of spot settlements.
- C. incorrect both with respect to the importance of spot settlements and international trade flows.

Solution:

C is correct. Although the media generally focus on the spot market when discussing FX, the majority of average daily trade volume involves the FX swap market as market participants either roll over or modify their existing hedging and speculative positions (or engage in FX swap financing).

Although it is true that all international trade transactions eventually result in some form of spot settlement, this typically generates a great deal of hedging (and speculative) activity in advance of spot settlement. Moreover, an important group of FX market participants engages in purely speculative positioning with no intention of ever delivering/receiving the principal amount of the trades. Most FX trading volume is not related to international trade: Portfolio flows (cross-border capital movements) and speculative activities dominate.

2. Statement 2 is:

- A. correct.
- B. incorrect with respect to corporations engaged in international trade.
- C. incorrect with respect to both corporations and the local banks that service their trade needs.

Solution:

C is correct. The most important FX market participants in terms of average daily turnover are found not among corporations engaged in international trade but among huge investment managers, both private (e.g., pension funds) and public (e.g., central bank reserve managers or sovereign wealth funds). A large and growing amount of daily turnover is also being generated by high-frequency traders who use computer algorithms to automatically execute extremely high numbers of speculative trades (although their

individual ticket sizes are generally small, they add up to large aggregate flows). On the sell side, the largest money center banks (e.g., Deutsche Bank, Citigroup, HSBC, UBS) are increasingly dominating the amount of trading activity routed through dealers. Regional and local banks are increasingly being marginalized in terms of their share of average daily turnover in FX markets.

QUESTION SET



1. Which of the following counterparties is *most likely* to be considered a sell-side FX market participant?
 - A. A large corporation that borrows in foreign currencies
 - B. A sovereign wealth fund that influences cross-border capital flows
 - C. A multinational bank that trades FX with its diverse client base

Solution:

C is correct. The sell-side parties generally consist of large banks that sell FX and related instruments to buy-side clients. These banks act as market makers, quoting exchange rates at which they will buy (the bid price) or sell (the offer price) the base currency.

Exchange Rate Quotations

Exchange rates represent the relative price of one currency in terms of another. This price can be represented in two ways: (1) currency A buys how many units of currency B; or (2) currency B buys how many units of currency A. Of course, these two prices are simply the inverse of each other.

To distinguish between these two prices, market participants sometimes distinguish between *direct* and *indirect* exchange rates. In the quoting convention A/B (where there are a certain number of units of currency A per one unit of currency B), we refer to currency A as the *price currency* (or quote currency); currency B is referred to as the *base currency*. (The reason for this choice of names will become clear.) The base currency is always set at a quantity of one. A *direct* currency quote takes the domestic country as the price currency and the foreign country as the base currency. For example, for a Paris-based trader, the domestic currency would be the euro (EUR) and a foreign currency would be the UK pound (GBP). For this Paris-based trader, a *direct* quote would be EUR/GBP. An exchange rate quote of EUR/GBP = 1.1211 means that GBP1 costs EUR1.1211. For this Paris-based trader, an *indirect* quote has the domestic currency—the euro—as the base currency. An indirect quote of GBP/EUR = 0.8920 means that EUR1 costs GBP0.8920. *Direct and indirect quotes are just the inverse (reciprocal) of each other.*

The professional FX market does not use the convention of describing exchange rates as either being direct or indirect because determining the domestic currency and the foreign currency depends on where one is located. For a London-based market participant, the UK pound (GBP) is the domestic currency and the euro (EUR) is a foreign currency. For a Paris-based market participant, it would be the other way around.

To avoid confusion, the FX market has developed a set of market conventions that all market participants typically adhere to when making and asking for FX quotes. Exhibit 5 displays some of these for the major currencies: the currency code used for obtaining exchange rate quotes, how the market lingo refers to this currency pair, and the actual ratio—price currency per unit of base currency—represented by the quote.

Exhibit 5: Exchange Rate Quote Conventions

FX Rate Quote Convention	Name Convention	Actual Ratio (Price currency/Base currency)
EUR	Euro	USD/EUR
JPY	Dollar–yen	JPY/USD
GBP	Sterling	USD/GBP
CAD	Dollar–Canada	CAD/USD
AUD	Aussie	USD/AUD
NZD	Kiwi	USD/NZD
CHF	Swiss franc	CHF/USD
EURJPY	Euro–yen	JPY/EUR
EURGBP	Euro–sterling	GBP/EUR
EURCHF	Euro–Swiss	CHF/EUR
GBPJPY	Sterling–yen	JPY/GBP
EURCAD	Euro–Canada	CAD/EUR
CADJPY	Canada–yen	JPY/CAD

Several things should be noted in this exhibit. First, the three-letter currency codes in the first column (for FX rate quotes) refer to what are considered to be the major exchange rates. Remember that an exchange rate is the price of one currency in terms of another: Two currencies are always involved in the price. This is different from referring to a single currency in its own right. For example, one can refer to the euro (EUR) as a *currency*; but if we refer to a euro *exchange rate* (EUR), it is always the price of the euro in terms of another currency, in this case the US dollar. This is because in the professional FX market, the three-letter code EUR is always taken to refer to the euro–US dollar exchange rate, which is quoted in terms of the number of US dollars per euro (USD/EUR). Second, the six-letter currency codes in the first column refer to some of the major *cross-rates*. This topic will be covered in the next section, but generally these are secondary exchange rates and are not as common as the main exchange rates. (Note that three-letter codes are always in terms of an exchange rate involving the US dollar, but the six-letter codes are not.) Third, when both currencies are mentioned in the code or the name convention, *the base currency is always mentioned first, the opposite order of the actual ratio (price currency/base currency)*. Thus, the code for “Sterling–yen” is “GBPJPY,” but the actual number quoted is the number of yen per sterling (JPY/GBP). Note also that *the codes may appear in a variety of formats that all mean the same thing*. For example, GBPJPY might instead appear as GBP:JPY or GBP–JPY. Fourth, regardless of where a market participant is located, there is always a mix of direct and indirect quotes in common market usage. For example, a trader based in Toronto, Canada, will typically refer to the euro–Canada and Canada–yen exchange rates—a mixture of direct (CAD/EUR) and indirect (JPY/CAD) quotes for that Canada-based trader. No overall consistency is observed in this mixture of direct and indirect quoting conventions in the professional FX market; a market participant must get familiar with how the conventions are used. In general, however, there is a hierarchy for quoting conventions. For quotes involving the euro, it serves as the base currency (e.g., GBP/EUR). Next in the priority sequence, for quotes involving the British pound (but not the euro), it serves as the base currency (e.g., USD/GBP). Finally, for quotes involving the US dollar (but not the GBP or EUR), it serves as the base currency (e.g., CAD/USD). Exceptions among the major currencies are the Australian and New Zealand dollars: they serve as the base currency when quoted against the US dollars (i.e., USD/AUD, USD/NZD).

Another concept involving exchange rate quotes in professional FX markets is that of a *two-sided price*. When a client asks a bank for an exchange rate quote, the bank will provide a “*bid*” (the price at which the bank is willing to buy the currency) and an “*offer*” (the price at which the bank is willing to sell the currency). But *two* currencies are referenced in an exchange rate quote, which is always the price of one currency relative to the other. So, which currency is being bought and which is being sold in this two-sided price quote? In this situation, the lingo involving the price currency (or quote currency) and the base currency, explained earlier, becomes useful. *The two-sided price quoted by the dealer is in terms of buying/selling the base currency.* It shows the number of units of the *price* currency that the client will receive from the dealer for one unit of the base currency (the bid) and the number of units of the price currency that the client must sell to the dealer to obtain one unit of the base currency (the offer). Consider the case of a client that is interested in a transaction involving the Swiss franc (CHF) and the euro (EUR). As we have read, the market convention is to quote this as euro–Swiss (CHF/EUR). The euro is the base currency, and the two-sided quote (price) shows the number of units of the price currency (CHF) that must be paid or will be received for EUR1. For example, a two-sided price in euro–Swiss (CHF/EUR) might look like: 1.1583–1.1585. The client will receive CHF1.1583 for selling EUR1 to the dealer and must pay CHF1.1585 to the dealer to buy EUR1. Note that *the price is shown in terms of the price currency* and that *the bid is always less than the offer*: The bank buys the base currency (euro, in this case) at the low price and sells the base currency at the high price. Buying low and selling high is profitable for banks, and spreading clients—trying to widen the bid/offer spread—is how dealers try to increase their profit margins. Note, however, that the electronic dealing systems currently used in professional FX markets are extremely efficient in connecting buyers and sellers globally. Moreover, this worldwide competition for business has compressed most bid/offer spreads to very tight levels. For simplicity, in the remainder of this reading, we will focus on exchange rates as a single number (with no bid/offer spread).

One last point in exchange rate quoting conventions is that most major spot exchange rates are typically quoted to four decimal places. One exception among the major currencies involves the yen, for which spot exchange rates are usually quoted to two decimal places. (For example, a USD/EUR quote would be expressed as 1.1701, whereas a JPY/EUR quote would be expressed as 130.98.) This difference involving the yen comes from the fact that the units of yen per unit of other currencies typically is relatively large, and hence extending the exchange rate quote to four decimal places is viewed as unnecessary.

Regardless of which quoting convention is used, changes in an exchange rate can be expressed as a percentage appreciation of one currency against the other: One simply has to be careful in identifying which currency is the price currency and which is the base currency. For example, suppose the exchange rate for the euro (USD/EUR) increases from 1.1500 to 1.2000. This represents an (unannualized) percentage change of the following:

$$\frac{1.2000}{1.1500} - 1 = 4.35\%.$$

This represents a 4.35 percent appreciation in the euro against the US dollar (and not an appreciation of the US dollar against the euro) because the USD/EUR exchange rate is expressed with the US dollar as the price currency.

Note that this appreciation of the euro against the US dollar also can be expressed as a depreciation of the US dollar against the euro; but in this case, the depreciation is not equal to 4.35 percent. Inverting the exchange rate quote from USD/EUR to EUR/USD, so that the euro is now the price currency, leads to the following:

$$\frac{\left(\frac{1}{1.2000}\right)}{\left(\frac{1}{1.1500}\right)} - 1 = \frac{1.1500}{1.2000} - 1 = -4.17\%.$$

Note that the US dollar depreciation is not the same, in percentage terms, as the euro appreciation. Mathematically, these percentages will always be different.

EXAMPLE 3**Exchange Rate Conventions**

A dealer based in New York City provides a spot exchange rate quote of 18.8590 MXN/USD to a client in Mexico City. The inverse of 18.8590 is 0.0530.

1. From the perspective of the Mexican client, the *most* accurate statement is that the:

- A. direct exchange rate quotation is equal to 0.0530.
- B. direct exchange rate quotation is equal to 18.8590.
- C. indirect exchange rate quotation is equal to 18.8590.

Solution:

B is correct. A direct exchange rate uses the domestic currency as the price currency and the foreign currency as the base currency. For an MXN/USD quote, the Mexican peso is the price currency; therefore, the direct quote for the Mexican client is 18.8590 (it costs MXN18.8590 to purchase USD1). Another way of understanding a *direct* exchange rate quote is that it is the price of one unit of foreign currency in terms of your own currency. This purchase of a unit of foreign currency can be thought of as a purchase much like any other you might make; think of the unit of foreign currency as just another item that you might be purchasing with your domestic currency. For example, for someone based in Canada, a liter of milk might cost about CAD1.25 at a time when USD1 costs about CAD1.30. This *direct* currency quote uses the *domestic* currency (the Canadian dollar, in this case) as the *price* currency and simply gives the price of a unit of foreign currency that is being purchased.

2. If the bid/offer quote from the dealer was 18.8580–18.8600 MXN/USD, then the bid/offer quote in USD/MXN terms would be *closest* to:

- A. 0.05302–0.05303.
- B. 0.05303–0.05302.
- C. 0.053025–0.053025.

Solution:

A is correct. An MXN/USD quote is the amount of Mexican pesos the dealer is bidding (offering) to buy (sell) USD1. The dealer's bid to buy USD1 at MXN18.8580 is equivalent to the dealer paying MXN18.8580 to buy USD1. Dividing both terms by 18.8580 means the dealer is paying (i.e., selling) MXN1 to buy USD0.05303. This is the offer in USD/MXN terms: The dealer offers to sell MXN1 at a price of USD0.05303. In USD/MXN terms, the dealer's bid for MXN1 is 0.05302, calculated by inverting the offer of 18.8600 in MXN/USD terms ($1/18.8600 = 0.05302$). Note that in any bid/offer quote, no matter which base or price currencies are used, the bid is always lower than the offer.

EXCHANGE RATE REGIMES: IDEALS AND HISTORICAL PERSPECTIVE

3



describe exchange rate regimes and explain the effects of exchange rates on countries' international trade and capital flows

FX rates affect international capital flows and trading relationships. These rate movements are based on the relative economic stability and efficiency of the trading countries involved. As a result, countries having higher or lower economic or trading volatility use different exchange regimes to address their economic objectives. This lesson describes these factors and concludes with an example from Malaysia.

Highly volatile exchange rates create uncertainty that undermines the efficiency of real economic activity and the financial transactions required to facilitate that activity. Exchange rate volatility also has a direct impact on investment decisions because it is a key component of the risk inherent in foreign (i.e., foreign-currency-denominated) assets. Exchange rate volatility is also a critical factor in selecting hedging strategies for foreign currency exposures.

The amount of FX rate volatility will depend, at least in part, on the institutional and policy arrangements associated with trade in any given currency. Virtually every exchange rate is managed to some degree by central banks. The policy framework that each central bank adopts is called an *exchange rate regime*. Although there are many potential variations, these regimes fall into a few general categories. Before describing each of these types, we consider the possibility of an ideal regime and provide some historical perspective on the evolution of currency arrangements.

The Ideal Currency Regime

The ideal currency regime would have three properties. First, the exchange rate between any two currencies would be credibly fixed. This would eliminate currency-related uncertainty with respect to the prices of goods and services as well as real and financial assets. Second, all currencies would be fully convertible (i.e., currencies could be freely exchanged for any purpose and in any amount). This condition ensures the unrestricted flow of capital. Third, each country would be able to undertake fully independent monetary policy in pursuit of domestic objectives, such as growth and inflation targets.

Unfortunately, these three conditions are not consistent. If the first two conditions were satisfied—credibly fixed exchange rates and full convertibility—then there would really be only one currency in the world. Converting from one national currency to another would have no more significance (indeed less) than deciding whether to carry coins or paper currency in your wallet. Any attempt to influence interest rates, asset prices, or inflation by adjusting the supply of one currency versus another would be futile. Thus, independent monetary policy is not possible if exchange rates are credibly fixed and currencies are fully convertible. *There can be no ideal currency regime.*

The impact of the currency regime on a country's ability to exercise independent monetary policy is a recurring theme in open-economy macroeconomics. It will be covered in more detail in other readings; however, it is worth emphasizing the basic point by considering what would happen in an idealized world of perfect capital mobility. If the exchange rate were credibly fixed, then any attempt to decrease default-free interest rates in one country below those in another—that is, to undertake independent, expansionary monetary policy—would result in a potentially unlimited outflow of capital because funds would seek the higher return. The central bank would be forced to

sell foreign currency and buy domestic currency to maintain the fixed exchange rate. The loss of reserves and reduction in the domestic money supply would put upward pressure on domestic interest rates until rates were forced back to equality, negating the initial expansionary policy. Similarly, a contractionary monetary policy (higher interest rates) would be thwarted by an inflow of capital.

The situation is quite different, however, with a floating exchange rate. A decrease in the domestic interest rate would make the domestic currency less attractive. The resulting depreciation of the domestic currency would shift demand toward domestically produced goods (i.e., exports rise and imports fall), reinforcing the expansionary impact of the initial decline in the interest rate. Similarly, a contractionary increase in the interest rate would be reinforced by appreciation of the domestic currency.

In practice, of course, capital is not perfectly mobile, and the impact on monetary policy is not so stark. The fact remains, however, that fixed exchange rates limit the scope for independent monetary policy and that national monetary policy regains potency and independence, at least to some degree, if the exchange rate is allowed to fluctuate or restrictions are placed on convertibility. In general, the more freely the exchange rate is allowed to float and the more tightly convertibility is controlled, the more effective the central bank can be in addressing domestic macroeconomic objectives. The downside, of course, is the potential distortion of economic activity caused by exchange rate risk and inefficient allocation of financial capital.

Historical Perspective on Currency Regimes

How currencies exchange for one another has evolved over the centuries. At any point in time, different exchange rate systems may coexist; still, the world economy tends to have one dominant system. Throughout most of the 19th century and the early 20th century until the start of World War I, the US dollar and the British pound sterling operated on the “classical gold standard.” The price of each currency was fixed in terms of gold. Gold was the numeraire (the unit in terms of which other goods, services, and assets were priced) for each currency; therefore, it was indirectly the numeraire for all other prices in the economy. Many countries (e.g., the colonies of the United Kingdom) fixed their currencies relative to sterling and were therefore implicitly also operating on the classical gold standard.

The classical gold standard operated by what is called the *price-specie-flow mechanism*. This mechanism operated through the impact of trade imbalances on capital flows, namely gold. As countries experienced a trade surplus, they accumulated gold as payment, their domestic money supply expanded by the amount dictated by the fixed parity, prices rose, and exports fell. Similarly, when a country ran a trade deficit, there was an automatic outflow of gold, a contraction of the domestic money supply, and a fall in prices leading to increased exports.

In this system, national currencies were backed by gold. A country could print only as much money as its gold reserve allowed. The system was limited by the amount of gold, but it was self-adjusting and inspired confidence. With a fixed stock of gold, the price-specie-flow mechanism would work well. Still, new gold discoveries as well as more efficient methods of refining gold would enable a country to increase its gold reserves and increase its money supply apart from the effect of trade flows. In general, however, trade flows drove changes in national money supplies.

Economic historians disagree about the effect of the classical gold standard on overall macroeconomic stability. Was it destabilizing? On the one hand, monetary policy was tied to trade flows, so a country could not engage in expansionary policies when there was a downturn in the non-traded sector. On the other hand, tying monetary policy to trade flows kept inflation in check.

During the 1930s, the use of gold as a clearing device for settlement of trade imbalances, combined with increasing protectionism on the part of economies struggling with depression as well as episodes of deflation and hyperinflation, created a chaotic environment for world trade. As a consequence of these factors, world trade dropped by more than 50 percent and the gold standard was abandoned.

In the latter stages of World War II, a new system of fixed exchange rates with periodic realignments was devised by John Maynard Keynes and Harry Dexter White, representing the UK and US Treasuries, respectively. The Bretton Woods system, named after the town where it was negotiated, was adopted by 44 countries in 1944. From the end of the war until the collapse of the system in the early 1970s, the United States, Japan, and most of the industrial countries of Europe maintained a system of fixed parities for exchange rates between currencies. When the parities were significantly and persistently out of line with the balancing of supply and demand, there would be a realignment of currencies with some appreciating in value and others depreciating in value. These periodic realignments were viewed as a part of standard monetary policy.

By 1973, with chronic inflation taking hold throughout the world, most nations abandoned the Bretton Woods system in favor of a flexible exchange rate system under what are known as the Smithsonian Agreements. Milton Friedman had called for such a system as far back as the 1950s. His argument was that the fixed parity system with periodic realignments would become unsustainable. When the inevitable realignments were imminent, large speculative profit opportunities would appear. Speculators would force the hand of monetary policy authorities, and their actions would distort the data needed to ascertain appropriate trade-related parities. It is better, he argued, to let the market, rather than central bank governors and treasury ministers, determine the exchange rate.

After 1973, most of the industrial world changed to a system of flexible exchange rates. The original thinking was that the forces that caused exchange rate chaos in the 1930s—poor domestic monetary policy and trade barriers—would not be present in a flexible exchange rate regime, and therefore exchange rates would move in response to the exchange of goods and services among countries. However, exchange rates moved around much more than anyone expected. Academic economists and financial analysts alike soon realized that the high degree of exchange rate volatility was the manifestation of a highly liquid, forward-looking asset market. Investment-driven FX transactions—for both long-term investment and short-term speculation—mattered much more in setting the spot exchange rate than anyone had previously imagined.

There are costs, of course, to a high degree of exchange rate volatility. These include difficulty planning without hedging exchange rate risks—a form of insurance cost, domestic price fluctuations, uncertain costs of raw materials, and short-term interruptions in financing transactions. For these reasons, in 1979 the European Economic Community opted for a system of limited flexibility, the European Exchange Rate Mechanism (ERM).

Initially, the system called for European currency values to fluctuate within a narrow band called “the snake,” but this system did not last long. The end of the Cold War and the reunification of Germany created conditions ripe for speculative attack. In the early 1990s, the United Kingdom was in a recession and the government’s monetary policy leaned toward low interest rates to stimulate economic recovery. Germany was issuing large amounts of debt to pay for reunification, and the German central bank (the Deutsche Bundesbank) opted for high interest rates to ensure price stability. Capital began to flow from sterling to Deutsche marks to obtain the higher interest rate. The Bank of England tried to lean against these flows and maintain the exchange rate within the ERM, but eventually it began to run out of marks to sell. Because it was almost certain that devaluation would be required, holders of sterling rushed to purchase marks at the old rate, and the speculative attack forced the United Kingdom out of the ERM in September 1992, only two years after it finally had joined the system.

Despite these difficulties, 1999 saw the creation of a common currency for most Western European countries, without Switzerland or the United Kingdom, called the euro. The hope was that the common currency would increase transparency of prices across borders in Europe, enhance market competition, and facilitate more efficient allocation of resources. The drawback, of course, is that each member country lost the ability to manage its exchange rate and therefore to engage in independent monetary policy.

QUESTION SET

1. An exchange rate:
 - A. is most commonly quoted in real terms.
 - B. is the price of one currency in terms of another.
 - C. between two currencies ensures that they are fully convertible.

Solution:

B is correct. The exchange rate is the number of units of the price currency that one unit of the base currency will buy. Equivalently, it is the number of units of the price currency required to buy one unit of the base currency.

2. Which of the following is *not* a condition of an ideal currency regime?
 - A. Fully convertible currencies
 - B. Fully independent monetary policy
 - C. Independently floating exchange rates

Solution:

C is correct. An ideal currency regime would have credibly fixed exchange rates among all currencies. This would eliminate currency-related uncertainty with respect to the prices of goods and services as well as real and financial assets.

A Taxonomy of Currency Regimes

Although the pros and cons of fixed and flexible exchange rate regimes continue to be debated, regimes have been adopted that lie somewhere between these polar cases. In some cases, the driving force is the lack of credibility with respect to sound monetary policy. An economy with a history of hyperinflation may be forced to adopt a form of fixed-rate regime because its promise to maintain a sound currency with a floating rate regime would not be credible. This has been a persistent issue in Latin America. In other cases, the driving force is as much political as it is economic. The decision to create the euro was strongly influenced by the desire to enhance political union within the European Community, whose members had been at war with each other twice in the 20th century.

As of April 2008, the International Monetary Fund (IMF) classified exchange rate regimes into the eight categories shown in Exhibit 6.

Exhibit 6: Exchange Rate Regimes for Selected Economies as of 30 April 2008

Type of Regime	Currency Anchor		
	US Dollar	Euro	Basket/None
No separate legal tender			
Dollarized	Ecuador, El Salvador, Marshall Islands, Micronesia, Palau, Panama, Timor-Leste, Zimbabwe	Kosovo, Montenegro, San Marino	Kiribati, Tuvalu
Monetary union		<i>EMU:</i> Austria, Belgium, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Luxembourg, Lithuania, Malta, Netherlands, Portugal, Slovak Rep., Slovenia, Spain	
Currency board	Djibouti, Hong Kong SAR, Antigua and Barbuda	Bosnia and Herzegovina, Bulgaria	Brunei Darussalam
Fixed parity	Aruba, The Bahamas, Bahrain, Barbados, Belize, Curaçao and Saint Maarten, Eritrea, Jordan, Oman, Qatar, Saudi Arabia, South Sudan, Turkmenistan, UAE, Venezuela	Cabo Verde, Comoros, Denmark, São Tomé and Príncipe WAEMU: Benin, Burkina Faso, Côte d'Ivoire, Guinea-Bissau, Mali, Niger, Senegal, Togo CEMAC: Cameroon, Central African Rep., Chad, Rep. of Congo, Equatorial Guinea, Gabon	Fiji, Kuwait, Libya, Morocco, Samoa, Bhutan, Lesotho, Namibia, Nepal, Swaziland
Target zone		Slovak Republic	Syria
Crawling peg	Nicaragua		Botswana
Crawling band	Honduras, Jamaica	Croatia	China, Ethiopia, Uzbekistan, Armenia, Dominican Republic, Guatemala, Argentina, Belarus, Haiti, Switzerland, Tunisia
Managed float	Cambodia, Liberia		Algeria, Iran, Syria, The Gambia, Myanmar, Nigeria, Rwanda, Czech Rep., Costa Rica, Malaysia, Mauritania, Pakistan, Russia, Sudan, Vanuatu
Independent float	Australia, Canada, Chile, Japan, Mexico, Norway, Poland, Sweden, United Kingdom, Somalia, United States	Albania, Brazil, Colombia, Georgia, Ghana, Hungary, Iceland, Indonesia, Israel, Korea, Moldova, New Zealand, Paraguay, Peru	Philippines, Romania, Serbia, South Africa, Thailand, Turkey, Uganda

Global financial markets are too complex and diverse to be fully captured by this (or any other) classification system. A government's control over the domestic currency's exchange rate will depend on many factors; for example, the degree of capital controls used to prevent the free flow of funds in and out of the economy. Also, even under an "independent float" regime, monetary authorities will occasionally intervene in FX markets to influence the value of their domestic currency. Additionally, the specifics of exchange rate policy implementation are subject to change.

This means that the classifications in Exhibit 6 are somewhat arbitrary and subject to interpretation, and they change over time. The important point is that the prices and flows in FX markets will, to varying degrees, reflect the legal and regulatory framework imposed by governments, not just "pure" market forces. Governments have a variety of motives and tools to attempt to manage exchange rates. The taxonomy in Exhibit

6 can be used to help understand the main distinctions among currency regimes and the rationales for adopting them, but the specific definitions should not be interpreted too rigidly. Instead, the focus should be on the diversity of FX markets globally as well as the implications of these various currency regimes for market pricing.

Arrangements with No Separate Legal Tender

The IMF identifies two types of arrangements for countries that do not have their own legal tender. In the first, known as *dollarization*, the country uses the currency of another nation as its medium of exchange and unit of account. In the second, the country participates in a monetary union whose members share the same legal tender. In either case, the country gives up the ability to conduct its own monetary policy.

In principle, a country could adopt any currency as its medium of exchange and unit of account, but the main reserve currency, the US dollar, is an obvious choice—hence the name dollarization. Many countries are dollarized: East Timor, El Salvador, Ecuador, and Panama, for example. By adopting another country's currency as legal tender, a dollarized country inherits that country's currency credibility, but not its creditworthiness. For example, although local banks may borrow, lend, and accept deposits in US dollars, they are not members of the US Federal Reserve System, nor are they backed by deposit insurance from the Federal Deposit Insurance Corporation. Thus, interest rates on US dollars in a dollarized economy need not be, and generally are not, the same as on dollar deposits in the United States.

Dollarization imposes fiscal discipline by eliminating the possibility that the central bank will be induced to monetize government debt (i.e., to persistently purchase government debt with newly created local currency). For countries with a history of fiscal excess or a lack of monetary discipline, dollarizing the economy can facilitate growth of international trade and capital flows if it creates an expectation of economic and financial stability. In the process, however, it removes another potential source of stabilization—domestic monetary policy.

The European Economic and Monetary Union (EMU) is the most prominent example of the second type of arrangement lacking separate legal tender. Each EMU member country uses the euro as its currency. Although member countries cannot have their own monetary policies, they jointly determine monetary policy through their representation at the European Central Bank (ECB). As with dollarization, a monetary union confers currency credibility on members with a history of fiscal excess or a lack of monetary discipline. As shown by the 2010 EMU sovereign debt crisis, however, a monetary union alone cannot confer creditworthiness.

Currency Board System

The IMF defines a *currency board system* (CBS) as follows:

A monetary regime based on an explicit legislative commitment to exchange domestic currency for a specified foreign currency at a fixed exchange rate, combined with restrictions on the issuing authority to ensure fulfillment of its legal obligation. This implies that domestic currency will be issued only against foreign exchange and it remains fully backed by foreign assets.

Hong Kong SAR has the leading example of a long-standing (since 1983) currency board. US dollar reserves are held to cover, at the fixed parity, the entire *monetary base*—essentially bank reserves plus all Hong Kong dollar notes and coins in circulation. Note that Hong Kong-dollar-denominated bank deposits are not fully collateralized by US dollar reserves; to do so would mean that banks could not lend against their deposits. The Hong Kong Monetary Authority (HKMA) does not function as a traditional central bank under this system because the obligation to maintain 100

percent foreign currency reserves against the monetary base prevents it from acting as a lender-of-last-resort for troubled financial institutions. It can, however, provide short-term liquidity by lending against foreign currency collateral.

A CBS works much like the classical gold standard in that expansion and contraction of the monetary base are directly linked to trade and capital flows. As with the gold standard, a CBS works best if domestic prices and wages are very flexible, non-traded sectors of the domestic economy are relatively small, and the global supply of the reserve asset grows at a slow, steady rate consistent with long-run real growth with stable prices. The first two of these conditions are satisfied in Hong Kong SAR. Until and unless Hong Kong SAR selects a new reserve asset, however, the third condition depends on US monetary policy.

In practice, the HKD exhibits modest fluctuations around the official parity of HKD/USD = 7.80 because the HKMA buys (sells) US dollars at a pre-announced level slightly below (above) the parity. Persistent flows on one side of this convertibility zone or the other result in interest rate adjustments rather than exchange rate adjustments. Inside the zone, however, the exchange rate is determined by the market and the HKMA is free to conduct limited monetary operations aimed at dampening transitory interest rate movements.

One of the advantages of a CBS as opposed to dollarization is that the monetary authority can earn a profit by paying little or no interest on its liability—the monetary base—and can earn a market rate on its asset—the foreign currency reserves. This profit is called *seigniorage*. Under dollarization, the seigniorage goes to the monetary authority whose currency is used.

Fixed Parity

A simple fixed-rate system differs from a CBS in two important respects. First, there is no legislative commitment to maintaining the specified parity. Thus, market participants know that the country may choose to adjust or abandon the parity rather than endure other, potentially more painful, adjustments. Second, the target level of FX reserves is discretionary; it bears no particular relationship to domestic monetary aggregates. Thus, although monetary independence is ultimately limited as long as the exchange peg is maintained, the central bank can carry out traditional functions, such as serving as lender of last resort.

In the conventional fixed-rate system, the exchange rate may be pegged to a single currency—for example, the US dollar—or to a basket index of the currencies of major trading partners. There is a band of up to ± 1 percent around the parity level within which private flows are allowed to determine the exchange rate. The monetary authority stands ready to spend its foreign currency reserves, or buy foreign currency, to maintain the rate within these bands.

The credibility of the fixed parity depends on the country's willingness and ability to offset imbalances in private sector demand for its currency. Both excess and deficient private demand for the currency can exert pressure to adjust or abandon the parity. Excess private demand for the domestic currency implies a rapidly growing stock of FX reserves, expansion of the domestic money supply, and potentially accelerating inflation. Deficient demand for the currency depletes FX reserves and exerts deflationary pressure on the economy. If market participants believe the FX reserves are insufficient to sustain the parity, then that belief may be self-fulfilling because the resulting speculative attack will drain reserves and may force an immediate devaluation. Thus, the level of reserves required to maintain credibility is a key issue for a simple fixed exchange rate regime.

Target Zone

A target zone regime has a fixed parity with fixed horizontal intervention bands that are somewhat wider—up to ± 2 percent around the parity—than in the simple fixed parity regime. The wider bands provide the monetary authority with greater scope for discretionary policy.

Active and Passive Crawling Pegs

Crawling pegs for the exchange rate—usually against a single currency, such as the US dollar—were common in the 1980s in Latin America, particularly Brazil, during the high inflation periods. To prevent a run on the US dollar reserves, the exchange rate was adjusted frequently (weekly or daily) to keep pace with the inflation rate. Such a system was called a passive crawl. An adaptation used in Argentina, Chile, and Uruguay was the active crawl: The exchange rate was pre-announced for the coming weeks with changes taking place in small steps. The aim of the active crawl was to manipulate expectations of inflation. Because the domestic prices of many goods were directly tied to import prices, announced changes in the exchange rate would effectively signal future changes in the inflation rate of these goods.

Fixed Parity with Crawling Bands

A country can also have a fixed central parity with crawling bands. Initially, a country may fix its rates to a foreign currency to anchor expectations about future inflation, but then the country may gradually permit increasing flexibility in the form of a pre-announced widening band around the central parity. Such a system has the desirable property of allowing a gradual exit strategy from the fixed parity. A country might want to introduce greater flexibility and greater scope for monetary policy, but it may not yet have the credibility or financial infrastructure for full flexibility. In this case, it can maintain a fixed parity with slowly widening bands.

Managed Float

A country may simply follow an exchange rate policy based on either internal or external policy targets—intervening or not to achieve trade balance, price stability, or employment objectives. Such a policy, often called *dirty floating*, invites trading partners to respond likewise with their exchange rate policy and potentially decreases stability in FX markets as a whole. The exchange rate target, in terms of either a level or a rate of change, typically is not explicit.

Independently Floating Rates

In this case, the exchange rate is left to market determination and the monetary authority is able to exercise independent monetary policy aimed at achieving such objectives as price stability and full employment. The central bank also has latitude to act as a lender of last resort to troubled financial institutions, if necessary.

It should be clear from recent experience that the concepts of float, managed float, crawl, and target zone are not hard and fast rules. Central banks do occasionally engage, implicitly or explicitly, in regime switches—even in countries nominally following an independently floating exchange rate regime. For example, when the US dollar appreciated in the mid-1980s with record US trade deficits, then-US Treasury Secretary James Baker engineered the Plaza Accord, in which Japan and Germany implemented an appreciation of their currencies against the US dollar. (The Plaza Accord is so named because it was negotiated at the Plaza Hotel in New York City.) This 1985 policy agreement involved a combination of fiscal and monetary policy measures by the countries involved as well as direct intervention in FX markets. The Plaza Accord was a clear departure from a pure independently floating exchange rate system.

There are more recent examples of government intervention in FX markets. In September 2000, the ECB, the Federal Reserve Board, the Bank of Japan, the Bank of England, and the Bank of Canada engaged in “concerted” intervention to support the value of the euro, a “freely floating” currency that was then under pressure within FX markets. (This intervention was described as “concerted” because it was pre-arranged and coordinated among the central banks involved.) During 2010, many countries engaged in unilateral intervention to prevent the rapid appreciation of their currencies against the US dollar. Several of these countries also employed various fiscal and regulatory measures (e.g., taxes on capital inflows) to further affect exchange rate movements.

The important point to draw from this discussion is that exchange rates not only reflect private sector market forces but also will be influenced, to varying degrees, by the legal and regulatory framework (currency regimes) within which FX markets operate. Moreover, they will occasionally be influenced by government policies (fiscal, monetary, and intervention) intended to manage exchange rates. All of these can vary widely among countries and are subject to change with time.

Nonetheless, the most widely traded currencies in FX markets (the US dollar, yen, euro, UK pound, Swiss franc, and the Canadian and Australian dollars) are typically considered to be free floating, although subject to relatively infrequent intervention.

EXAMPLE 4

Exchange Rate Regimes

An investment adviser in Los Angeles, United States, is meeting with a client who wishes to diversify her portfolio by including more international investments. To evaluate the suitability of international diversification for the client, the adviser attempts to explain some of the characteristics of FX markets. The adviser points out that exchange rate regimes affect the performance of domestic economies as well as the amount of FX risk posed by international investments.

The client and her adviser discuss potential investments in Hong Kong SAR, Panama, and Canada. The adviser notes that the currency regimes of Hong Kong SAR, Panama, and Canada are a currency board, dollarization, and a free float, respectively. The adviser tells his client that these regimes imply different degrees of FX risk for her portfolio.

The discussion between the investment adviser and his client then turns to potential investments in other markets with different currency regimes. The adviser notes that some markets are subject to fixed parity regimes against the US dollar. The client asks whether a fixed parity regime would imply less foreign currency risk for her portfolio than would a currency board. The adviser replies: “Yes, a fixed parity regime means a constant exchange rate and is more credible than a currency board.”

The adviser goes on to explain that in some markets, exchange rates are allowed to vary, although with different degrees of FX market intervention to limit exchange rate volatility. Citing examples, he notes that mainland China has a crawling peg regime with reference to the US dollar, but the average daily percentage changes in the mainland China/US exchange rate are very small compared with the average daily volatility for a freely floating currency. The adviser also indicates that Denmark has a target zone regime with reference to the euro, and South Korea usually follows a freely floating currency regime but sometimes switches to a managed float regime. The currencies of mainland China, Denmark, and South Korea are the yuan renminbi (CNY), krone (DKK), and won (KRW), respectively.

1. Based solely on the exchange rate risk the client would face, what is the correct ranking (from most to least risky) of the following investment locations?

- A. Panama, Canada, Hong Kong SAR
- B. Canada, Hong Kong SAR, Panama
- C. Hong Kong SAR, Panama, Canada

Solution:

B is correct. The CAD/USD exchange rate is a floating exchange rate, and Canadian investments would therefore carry exchange rate risk for a US-based investor. Although Hong Kong SAR follows a currency board system, the HKD/USD exchange rate nonetheless does display some variation, albeit much less than in a floating exchange rate regime. In contrast, Panama has a dollarized economy (i.e., it uses the US dollar as the domestic currency); therefore, there is no FX risk for a US investor.

2. Based solely on their FX regimes, which investment location is least likely to import inflation or deflation from the United States?

- A. Canada
- B. Panama
- C. Hong Kong SAR

Solution:

A is correct. The Canadian dollar floats independently against the US dollar leaving the Bank of Canada able to adjust monetary policy to maintain price stability. Neither Hong Kong SAR (currency board) nor Panama (dollarized) can exercise independent monetary policy to buffer its economy from the inflationary or deflationary consequences of US monetary policy.

3. The adviser's reply about fixed parity regimes is incorrect with regard to:

- A. credibility.
- B. a constant exchange rate.
- C. both a constant exchange rate and credibility.

Solution:

C is correct. A fixed exchange rate regime does not mean that the exchange rate is rigidly fixed at a constant level. In practice, both a fixed-rate regime and a currency board allow the exchange rate to vary within a band around the stated parity level. Thus, both regimes involve at least a modest amount of exchange rate risk. The fixed parity regime exposes the investor to the additional risk that the parity may not be maintained. In a fixed parity regime, the level of foreign currency reserves is discretionary and typically only a small fraction of the domestic money supply. With no legal obligation to maintain the parity, the monetary authority may adjust the parity (devalue or revalue its currency) or allow its currency to float if doing so is deemed to be less painful than other adjustment mechanisms (e.g., fiscal restraint). In contrast, a currency board entails a legal commitment to maintain the parity and to fully back the domestic currency with reserve currency assets. Hence, there is little risk that the parity will be abandoned.

4. Based on the adviser's categorization of mainland China's currency regime, if the US dollar is depreciating against the South Korean won, then it is *most* likely correct that the Chinese yuan is:

- A. fixed against the South Korean won.
- B. appreciating against the South Korean won.
- C. depreciating against the South Korean won.

Solution:

C is correct. If the Chinese yuan is subject to a crawling peg with very small daily adjustments versus the US dollar, and the US dollar is depreciating against the South Korean won, then the Chinese yuan *most likely* would be depreciating against the South Korean won as well. In fact, this was an important issue in FX markets through the latter part of 2010: As the US dollar depreciated against most Asian currencies (and less so against the Chinese yuan), many Asian countries felt that they were losing their competitive export advantage because the Chinese yuan was so closely tied to the US dollar. This led many Asian countries to intervene in FX markets against the strength of their domestic currencies to avoid losing an export pricing advantage against the Chinese mainland.

5. Based on the adviser's categorization of Denmark's currency regime, it would be *most* correct to infer that the:

- A. krone is allowed to float against the euro within fixed bands.
- B. Danish central bank will intervene if the exchange rate strays from its target level.
- C. target zone will be adjusted periodically to manage inflation expectations.

Solution:

A is correct. A target zone means that the exchange rate between the euro and Danish krone (DKK) will be allowed to vary within a fixed band (as of 2010, the target zone for the DKK/EUR is a ±2.5 percent band). This does not mean that the DKK/EUR rate is fixed at a certain level (B is incorrect) or that the target zone will vary to manage inflation expectations (this is a description of a crawling peg, which makes C incorrect).

6. Based on the adviser's categorization of South Korea's currency policy, it would be *most* correct to infer that the Korean:

- A. central bank is engineering a gradual exit from a fixed-rate regime.
- B. government is attempting to peg the exchange rate within a predefined zone.
- C. won is allowed to float, but with occasional intervention by the Korean central bank.

Solution:

C is correct. Similar to the monetary authorities responsible for many of the world's major currencies, the South Korean policy typically involves letting market forces determine the exchange rate (an independent floating rate regime). But this approach does not mean that market forces are the sole determinant of the won exchange rate. As with most governments, the South Korean policy is to intervene in FX markets when movements in the exchange rate are viewed as undesirable (a managed float). For example, during the latter part of 2010, South Korea and many other countries

intervened in FX markets to moderate the appreciation of their currencies against the US dollar. Answer A describes a fixed parity with a crawling bands regime, and B describes a target zone regime: Both answers are incorrect.

QUESTION SET



1. In practice, both a fixed parity regime and a target zone regime allow the exchange rate to float within a band around the parity level. The *most likely* rationale for the band is that the band allows the monetary authority to:
 - A. be less active in the currency market.
 - B. earn a spread on its currency transactions.
 - C. exercise more discretion in monetary policy.

Solution:

C is correct. Fixed exchange rates impose severe limitations on the exercise of independent monetary policy. With a rigidly fixed exchange rate, domestic interest rates, monetary aggregates (e.g., money supply), and credit conditions are dictated by the requirement to buy/sell the currency at the rigid parity. Even a narrow band around the parity level allows the monetary authority to exercise some discretionary control over these conditions. In general, the wider the band, the more independent control the monetary authority can exercise.

2. A fixed exchange rate regime in which the monetary authority is legally required to hold FX reserves backing 100 percent of its domestic currency issuance is best described as:
 - A. dollarization.
 - B. a currency board.
 - C. a monetary union.

Solution:

B is correct. With a currency board, the monetary authority is legally required to exchange domestic currency for a specified foreign currency at a fixed exchange rate. It cannot issue domestic currency without receiving foreign currency in exchange, and it must hold that foreign currency as a 100 percent reserve against the domestic currency issued. Thus, the country's monetary base (bank reserves plus notes and coins in circulation) is fully backed by FX reserves.

Exchange Rates and the Trade Balance: Introduction

Just as a family that spends more than it earns must borrow or sell assets to finance the excess, a country that imports more goods and services than it exports must either borrow from or sell assets to foreign entities to finance the trade deficit. Conversely, a country that exports more goods and services than it imports must invest the excess either by lending to foreigners or by buying assets from foreigners. Thus, a trade deficit (surplus) must be exactly matched by an offsetting *capital account* surplus (deficit). This implies that any factor that affects the trade balance must have an equal

and opposite impact on the capital account, and vice versa. To put this differently, the *impact of exchange rates and other factors on the trade balance must be mirrored by their impact on capital flows*: They cannot affect one without affecting the other.

Using a fundamental identity from macroeconomics, the relationship between the trade balance and expenditure/saving decisions can be expressed as follows:

$$X - M = (S - I) + (T - G),$$

where X represents exports, M is imports, S is private savings, I is investment in plant and equipment, T is taxes net of transfers, and G is government expenditure. From this relationship, we can see that a trade surplus ($X > M$) must be reflected in a fiscal surplus ($T > G$), an excess of private saving over investment ($S > I$), or both. Because a fiscal surplus can be viewed as government saving, we can summarize this relationship more simply by saying that a trade surplus means the country saves more than enough to fund its investment (I) in plant and equipment. The excess saving is used to accumulate financial claims on the rest of the world. Conversely, a trade deficit means the country does not save enough to fund its investment spending (I) and must reduce its net financial claims on the rest of the world.

Although this identity provides a key link between real expenditure and saving decisions and the aggregate flow of financial assets into or out of a country, it does not tell us what type of financial assets will be exchanged or in what currency they will be denominated. All that can be said is that asset prices and exchange rates at home and abroad must adjust so that all financial assets are willingly held by investors.

If investors anticipate a significant change in an exchange rate, they will try to sell the currency that is expected to depreciate and buy the currency that is expected to appreciate. This implies an incipient (i.e., potential) flow of capital from one country to the other, which must either be accompanied by a simultaneous shift in the trade balance or be discouraged by changes in asset prices and exchange rates. Because expenditure/saving decisions and prices of goods change much more slowly than financial investment decisions and asset prices, most of the adjustment usually occurs within the financial markets. That is, *asset prices and exchange rates adjust so that the potential flow of financial capital is mitigated and actual capital flows remain consistent with trade flows*. In a fixed exchange rate regime, the central bank offsets the private capital flows in the process of maintaining the exchange rate peg and the adjustment occurs in other asset prices, typically interest rates, until and unless the central bank is forced to allow the exchange rate to adjust. In a floating exchange rate regime, the main adjustment is often a rapid change in the exchange rate that dampens an investor's conviction that further movement will be forthcoming. Thus, *capital flows—potential and actual—are the primary determinant of exchange rate movements in the short to intermediate term*. Trade flows become increasingly important in the long term as expenditure and saving decisions as well as the prices of goods and services adjust.

QUESTION SET



1. A country with a trade deficit will *most likely*:
 - A. have an offsetting capital account surplus.
 - B. save enough to fund its investment spending.
 - C. buy assets from foreigners to fund the imbalance.

Solution:

A is correct. A trade deficit must be exactly matched by an offsetting capital account surplus to fund the deficit. A capital account surplus reflects borrowing from foreigners (an increase in domestic liabilities) and/or selling assets to foreigners (a decrease in domestic assets). A capital account sur-

plus is often referred to as a “capital inflow” because the net effect is foreign investment in the domestic economy.

4

CAPITAL RESTRICTIONS



describe common objectives of capital restrictions imposed by governments

Governments restrict inward and outward flow of capital for many reasons. For example, the government may want to meet some objective regarding employment or regional development, or it may have a strategic or defense-related objective. Many countries require approval for foreigners to invest in their country and for citizens to invest abroad. Control over inward investment by foreigners results in restrictions on how much can be invested, and on the type of industries in which capital can be invested. For example, such strategic industries as defense and telecommunications are often subject to ownership restrictions. Outflow restrictions can include restrictions on repatriation of capital, interest, profits, royalty payments, and license fees. Citizens are often limited in their ability to invest abroad, especially in FX-scarce economies, and there can be deadlines for repatriation of income earned from any investments abroad.

Economists consider free movement of financial capital to be beneficial because it allows capital to be invested where it will earn the highest return. Inflows of capital also allow countries to invest in productive capacity at a rate that is higher than could be achieved with domestic savings alone, and it can enable countries to achieve a higher rate of growth. Long-term investments by foreign firms that establish a presence in the local economy can bring in not only much needed capital but also new technology, skills, and advanced production and management practices as well as create spillover benefits for local firms. Investment by foreign firms can create a network of local suppliers if they source some of their components locally. Such suppliers may receive advanced training and spillover benefits from a close working relationship with the foreign firms. On the one hand, increased competition from foreign firms in the market may force domestic firms to become more efficient. On the other hand, it is possible that the domestic industry may be hurt because domestic firms that are unable to compete are forced to exit the market.

In times of macroeconomic crisis, capital mobility can result in capital flight out of the country, especially if most of the inflow reflects short-term portfolio flows into stocks, bonds, and other liquid assets rather than foreign direct investment in productive assets. In such circumstances, capital restrictions are often used in conjunction with other policy instruments, such as fixed exchange rate targets. Capital restrictions and fixed exchange rate targets are complementary instruments because in a regime of perfect capital mobility, governments cannot achieve domestic and external policy objectives simultaneously using only standard monetary and fiscal policy tools. By limiting the free flow of capital, capital controls provide a way to exercise control over a country's external balance, whereas more traditional macro-policy tools are used to address other objectives.

Modern capital controls were developed by the belligerents in World War I as a method to finance the war effort. At the start of the war, all major powers restricted capital outflows (i.e., the purchase of foreign assets or loans abroad). These restrictions raised revenues by keeping capital in the domestic economies, facilitating the taxation

of wealth, and producing interest income. Moreover, capital controls helped to maintain a low level of interest rates, reducing the governments' borrowing costs on their liabilities. Since World War I, controls on capital outflows have been used similarly in other countries, mostly developing nations, to generate revenue for governments or to permit them to allocate credit in their domestic economies without risking capital flight. In broad terms, a capital restriction is any policy designed to limit or redirect capital flows. Such restrictions may take the form of taxes, price or quantity controls, or outright prohibitions on international trade in assets. Price controls may take the form of special taxes on returns to international investment, taxes on certain types of transactions, or mandatory reserve requirements—that is, a requirement forcing foreign parties wishing to deposit money in a domestic bank account to deposit some percentage of the inflow with the central bank for a minimum period at zero interest. Quantity restrictions on capital flows may include rules imposing ceilings or requiring special authorization for new or existing borrowing from foreign creditors. Or administrative controls may have an impact on cross-border capital movements in which a government agency must approve transactions for certain types of assets.

Effective implementation of capital restrictions may entail non-trivial administration costs, particularly if the measures have to be broadened to close potential loopholes. Protecting the domestic financial markets by capital restrictions also may postpone necessary policy adjustments or impede private-sector adaptation to changing international circumstances. Most important, controls may give rise to negative market perceptions, which may, in turn, make it more costly and difficult for the country to access foreign funds.

In a study on the effectiveness of capital controls, the International Monetary Fund considered restrictions on capital outflows and inflows separately. The authors concluded that for restrictions on capital inflows to be effective (i.e., not circumvented), the coverage needs to be comprehensive and the controls need to be implemented forcefully. Considerable administrative costs are incurred in continuously extending, amending, and monitoring compliance with the regulations. Although controls on inflows appeared to be effective in some countries, it was difficult to distinguish the impact of the controls from the impact of other policies, such as strengthening of prudential regulations, increased exchange rate flexibility, and adjustment of monetary policy. In the case of capital outflows, the imposition of controls during episodes of financial crisis seems to have produced mixed results, providing only temporary relief of varying duration to some countries, while successfully shielding others (e.g., Malaysia) and providing them with sufficient time to restructure their economies.

EXAMPLE 5

Historical Example—Capital Restrictions: Malaysia's Capital Controls in 1998–2001

After the devaluation of the Thai baht in July 1997, Southeast Asia suffered from significant capital outflows that led to falling local equity and real estate prices and declining exchange rates. To counter the outflows of capital, the IMF urged many of the countries in the region to increase interest rates, thus making their assets more attractive to foreign investors. Higher interest rates, however, weighed heavily on the domestic economies. In response to this dilemma, Malaysia imposed capital controls on 1 September 1998. These controls prohibited transfers between domestic and foreign accounts, eliminated credit facilities to offshore parties, prevented repatriation of investment until 1 September 1999, and fixed the exchange rate of the Malaysian ringgit at 3.8 per US dollar. In February 1999, a system of taxes on capital flows replaced the prohibition on repatriation of capital. Although the details were complex, the

net effect was to discourage short-term capital flows while permitting long-term transactions. By imposing capital controls, Malaysia hoped to regain monetary independence and to be able to cut interest rates without provoking a fall in the value of its currency as investors avoided Malaysian assets. The imposition of outflow controls indeed curtailed speculative capital outflows and allowed interest rates to be reduced substantially. At the same time, under the umbrella of the capital controls, the authorities pursued bank and corporate restructuring and achieved a strong economic recovery in 1999 and 2000. With the restoration of economic and financial stability, administrative controls on portfolio outflows were replaced by a two-tier, price-based exit system in February 1999, which was finally eliminated in May 2001. Although Malaysia's capital controls did contribute to a stabilization of its economy, they came with long-term costs associated with the country's removal from the MSCI developed equity market index, an important benchmark in the institutional asset management industry, and its relegation to the emerging market universe. The Malaysian market was no longer seen as on par with developed equity markets whose institutional and regulatory frameworks provide a higher standard of safety for investors. As a result, a number of market analysts suggested that it became more difficult for Malaysia to attract net long-term capital inflows.

1. Under what economic circumstances were Malaysia's capital restrictions imposed?

Solution:

As a result of the Southeast Asian crisis, Malaysia suffered substantial net capital outflows pushing up the domestic interest rate level.

2. What was the ultimate objective of Malaysia's capital restrictions?

Solution:

The restrictions were designed to limit and redirect capital flows to allow the government to reduce interest rates and pursue bank and corporate restructurings.

3. How successful were the country's capital restrictions?

Solution:

Although the capital controls helped stabilize Malaysia's economy, they contributed to a change in investors' perception of Malaysian financial markets and the removal of the Malaysian equity market from the MSCI benchmark universe of developed equity markets. This situation undermined international demand for Malaysian equities and made it more difficult to attract net long-term capital inflows.

PRACTICE PROBLEMS

1. What will be the effect on a direct exchange rate quote if the domestic currency appreciates?
 - A. Increase
 - B. Decrease
 - C. No change
2. An executive from Switzerland checks into a hotel room in Spain and is told by the manager that EUR1 will buy CHF1.2983. From the executive's perspective, an indirect exchange rate quote would be:
 - A. EURO0.7702 per CHF1.
 - B. CHF0.7702 per EUR1.
 - C. EUR1.2983 per CHF1.
3. Over the past month, the Swiss franc (CHF) has depreciated 12 percent against the British pound (GBP). How much has the pound sterling appreciated against the Swiss franc?
 - A. 12 percent
 - B. Less than 12 percent
 - C. More than 12 percent
4. An exchange rate between two currencies has increased to 1.4500. If the base currency has appreciated by 8 percent against the price currency, the initial exchange rate between the two currencies was *closest* to:
 - A. 1.3340.
 - B. 1.3426.
 - C. 1.5660.

SOLUTIONS

1. B is correct. In the case of a direct exchange rate, the domestic currency is the price currency (the numerator) and the foreign currency is the base currency (the denominator). If the domestic currency appreciates, then fewer units of the domestic currency are required to buy one unit of the foreign currency, and the exchange rate (domestic per foreign) declines. For example, if British pound sterling (GBP) appreciates against the euro (EUR), then euro–sterling (GBP/EUR) might decline.
2. A is correct. An indirect quote takes the foreign country as the price currency and the domestic country as the base currency. To get Swiss francs—which is the executive’s domestic currency—as the base currency, the quote must be stated as EUR/CHF. Using the manager’s information, the indirect exchange rate is $(1/1.2983) = 0.7702$.
3. C is correct. The appreciation of the British pound against the Swiss franc is the inverse of the 12 percent depreciation of the Swiss franc against the pound sterling: $[1/(1 - 0.12)] - 1 = (1/0.88) - 1 = 0.1364$, or 13.64%.
4. B is correct. The percentage appreciation of the base currency can be calculated by dividing the appreciated exchange rate by the initial exchange rate. In this case, the unknown is the initial exchange rate. The initial exchange is the value of X that satisfies the formula:

$$1.4500/X = 1.08$$

Solving for X leads to $1.45/1.08 = 1.3426$.

LEARNING MODULE

8

Exchange Rate Calculations

LEARNING OUTCOMES

Mastery	<i>The candidate should be able to:</i>
<input type="checkbox"/>	calculate and interpret currency cross-rates
<input type="checkbox"/>	explain the arbitrage relationship between spot and forward exchange rates and interest rates, calculate a forward rate using points or in percentage terms, and interpret a forward discount or premium

INTRODUCTION

1

The foreign exchange market facilitates international currency and trade flows, and it is important to understand how currency exchange rates are calculated. Market participants can also derive cross-rates to expand trading opportunities by determining quotes for currencies not directly traded. Understanding the concept of arbitrage relationships in the foreign exchange market provides a basis for understanding the interrelationships between four key market inputs. Global entities trade currencies for a wide variety of purposes and understanding the relationships between the market factors affecting spot and forward rates is crucial. These interactions are reinforced by the calculations in the second lesson.

LEARNING MODULE OVERVIEW



- An exchange rate between two currencies that are not expressly quoted on the market is known as a cross-rate and can be calculated using conventional currency quotes.
- Three conventional currency market quotes can be used with one inversion to calculate a cross-rate.
- Discrepancies in exchange rates can create arbitrage opportunities but they are rare due to market efficiencies.
- The premium of a forward exchange rate over a spot rate is quoted in terms of forward points, which are also called swap points.
- Forward rates are directly proportional to currency spot rates, the interest rate differential, and the maturity of the forward contract.

- As a result of the interrelationship among these four variables, any variable can be calculated by using the other three as inputs.

2

CROSS-RATE CALCULATIONS



calculate and interpret currency cross-rates

Global currencies are bought, sold, and exchanged in the foreign exchange (FX) market. In this decentralized market, participants trade currencies utilizing exchange rates, which typically reflect an efficient market. This section will cover the use of cross exchange rate relationships (cross-rates) to calculate exchange rates between two currencies using a third currency. It also will introduce calculations used in the FX market to trade currencies.

Given two exchange rates involving three currencies, it is possible to back out the cross-rate. For example, as we have seen in a prior lesson, the FX market convention is to quote the exchange rate between the US dollar and the euro as euro–dollar (USD/EUR). The FX market also quotes the exchange rate between the Canadian dollar and US dollar as dollar–Canada (CAD/USD). Given these two exchange rates, it is possible to back out the cross-rate between the euro and the Canadian dollar, which according to market convention is quoted as euro–Canada (CAD/EUR). This calculation is shown as follows:

$$\frac{\text{CAD}}{\text{USD}} \times \frac{\text{USD}}{\text{EUR}} = \frac{\text{CAD}}{\text{USD}} \times \frac{\text{USD}}{\text{EUR}} = \frac{\text{CAD}}{\text{EUR}}.$$

Hence, to get a euro–Canada (CAD/EUR) quote, we must multiply the dollar–Canada (CAD/USD) quote by the euro–dollar (USD/EUR) quote. For example, assume the exchange rate for dollar–Canada is 1.3020 and the exchange rate for euro–dollar is 1.1701. Using these spot exchange rates, the euro–Canada cross-rate equals:

$$1.3020 \times 1.1701 = 1.5235 \text{ CAD per EUR.}$$

The professional FX market does not use the convention of direct or indirect quotes because these conventions depend on one's location to determine the domestic versus foreign currencies. Instead, the market uses rate quotes on defined conventional currency pairs. Sometimes, to get a cross-rate using several currency quotes, it is necessary to invert a quote to get an intermediary currency that can be canceled out in the equation to obtain the cross-rate. For example, to get a Canada–yen (JPY/CAD) quote, one typically uses the dollar–Canada (CAD/USD) rate and dollar–yen (JPY/USD) rate, which are the market conventions. This Canada–yen calculation requires that the dollar–Canada rate (CAD/USD) be inverted to a Canada–dollar (USD/CAD) quote for the calculations to work, as follows:

$$\left(\frac{\text{CAD}}{\text{USD}}\right)^{-1} \times \frac{\text{JPY}}{\text{USD}} = \frac{\text{USD}}{\text{CAD}} \times \frac{\text{JPY}}{\text{USD}} = \frac{\text{USD}}{\text{CAD}} \times \frac{\text{JPY}}{\text{USD}} = \frac{\text{JPY}}{\text{CAD}}.$$

Hence, to get a Canada–yen (JPY/CAD) quote, we must first invert the dollar–Canada (CAD/USD) quote before multiplying by the dollar–yen (JPY/USD) quote. Market quotes for most currencies are quoted to four decimal places; however, the Japanese yen exchange rate is quoted to two decimal places. For example, assume that we have spot exchange rates of 1.3020 for dollar–Canada (CAD/USD) and 111.94 for dollar–yen (JPY/USD). The dollar–Canada rate of 1.3020 inverts to 0.7680; multiplying this value by the dollar–yen quote of 111.94 gives the following Canada–yen quote:

$$0.7680 \times 111.94 = 85.97 \text{ JPY per CAD.}$$

Market participants asking for a quote in a cross-rate currency pair typically will not need to do this calculation themselves: Either the dealer or the electronic trading platform will provide a quote in the specified currency pair. (For example, a client asking for a quote in Canada–yen will receive that quote from the dealer; he will not be given separate dollar–Canada and dollar–yen quotes to do the calculation.) Dealers providing the quotes often have to do this calculation themselves if only because the dollar–Canada and dollar–yen currency pairs often trade on different trading desks and involve different traders. Electronic dealing machines used in both the interbank market and bank-to-client markets often provide this mathematical operation to calculate cross-rates automatically.

Because market participants can receive both a cross-rate quote (e.g., Canada–yen) as well as the component underlying exchange rate quotes (e.g., dollar–Canada and dollar–yen), these cross-rate quotes must be consistent with the previous equation; otherwise, the market will arbitrage the mispricing. Extending our example, we calculate a Canada–yen (JPY/CAD) rate of 85.97 based on underlying dollar–Canada (CAD/USD) and dollar–yen (JPY/USD) rates of 1.3020 and 111.94, respectively. Now suppose that at the same time a misguided dealer quotes a Canada–yen rate of 86.20. This is a different price in Canada–yen for an identical service—that is, converting yen into Canadian dollars. Hence, any trader could buy CAD1 at the lower price of JPY85.97 and then turn around and sell CAD1 at JPY86.20 (recall our earlier discussion of how price and base currencies are defined). The riskless arbitrage profit is JPY0.23 per CAD1. The arbitrage—called *triangular arbitrage* (we use “tri–” because it involves three currencies)—would continue until the price discrepancy was removed.

In reality, however, these discrepancies in cross-rates rarely occur because both human traders and automatic trading algorithms are constantly on alert for any pricing inefficiencies. In practice, and for the purposes of this lesson, we can consider cross-rates as being consistent with their underlying exchange rate quotes and can assume that given any two exchange rates involving three currencies, we can back out the third cross-rate.

EXAMPLE 1

Cross-Rates and Percentage Changes

A research report produced by a dealer includes the following spot rate quotes:

Currency	Spot Rate	Expected Spot Rate in One Year
USD/EUR	1.1701	1.1619
CHF/USD	0.9900	0.9866
USD/GBP	1.3118	1.3066

1. The spot CHF/EUR cross-rate is *closest* to:

- A. 0.8461.
- B. 0.8546.
- C. 1.1584.

Solution:

C is correct:

$$\frac{\text{CHF}}{\text{EUR}} = \frac{\text{USD}}{\text{EUR}} \times \frac{\text{CHF}}{\text{USD}} = 1.1701 \times 0.9900 = 1.1584$$

2. The spot GBP/EUR cross-rate is *closest* to:

- A. 0.8920.
- B. 1.1211.
- C. 1.4653.

Solution:

A is correct:

$$\frac{\text{GBP}}{\text{EUR}} = \frac{\text{USD}}{\text{EUR}} \times \left(\frac{\text{USD}}{\text{GBP}} \right)^{-1} = \frac{\text{USD}}{\text{EUR}} \times \frac{\text{GBP}}{\text{USD}} = \frac{1.1701}{1.3118} = 0.8920$$

3. Based on the research report, the euro is expected to appreciate by how much against the US dollar over the next year?

- A. -0.7 percent
- B. +0.7 percent
- C. +1.0 percent

Solution:

A is correct. The euro is the base currency in the USD/EUR quote, and the expected decrease in the USD/EUR rate indicates that the euro is depreciating. In one year, it will cost less, in US dollars, to buy one euro. Mathematically:

$$\frac{1.1619}{1.1701} - 1 = -0.7\%$$

4. Based on the research report, how much is the US dollar expected to appreciate against the British pound sterling over the next year?

- A. +0.6 percent
- B. -0.4 percent
- C. +0.4 percent

Solution:

C is correct. The British pound is the base currency in the USD/GBP quote, and the expected decrease in the USD/GBP rate means that the British pound is expected to depreciate against the US dollar. Or equivalently, the US dollar is expected to appreciate against the British pound. Mathematically:

$$\left(\frac{1.3066}{1.3118} \right)^{-1} - 1 = \frac{1.3118}{1.3066} - 1 = +0.4\%$$

5. Over the next year, the Swiss franc is expected to:

- A. depreciate against the British pound.
- B. depreciate against the euro.
- C. appreciate against the British pound, euro, and US dollar.

Solution:

C is correct: Because the question does not require calculating the magnitude of the appreciation or depreciation, we can use the Swiss franc as either the price currency or the base currency. In this case, it is easier to use the Swiss franc as the price currency. CHF/USD is expected to decline from 0.9900 to 0.9866, so the Swiss franc is expected to be stronger (i.e., it should appreciate against the US dollar). CHF/EUR is currently 1.1584 (see the

solution to problem 1) and is expected to be 1.1463 ($= 0.9866 \times 1.1619$), so the Swiss franc is expected to appreciate against the euro. CHF/GBP is currently 1.2987 ($= 0.9900 \times 1.3118$) and is expected to be 1.2891 ($= 0.9866 \times 1.3066$), so the Swiss franc is also expected to appreciate against the British pound.

Alternatively, we can derive this answer intuitively. According to the research report, the CHF/USD rate is expected to decline: That is, the US dollar is expected to depreciate against the Swiss franc, or alternatively, the Swiss franc is expected to appreciate against the US dollar. The USD/EUR and USD/GBP rates are also decreasing, meaning that the euro and British pound are expected to depreciate against the US dollar, or alternatively, the US dollar is expected to appreciate against the euro and British pound. If the Swiss franc is expected to appreciate against the US dollar and the US dollar is expected to appreciate against both the euro and British pound, it follows that the Swiss franc is expected to appreciate against both the euro and British pound.

6. Based on the research report, which of the following lists the three currencies from strongest to weakest over the next year?

- A. US dollar, British pound, euro
- B. US dollar, euro, British pound
- C. Euro, US dollar, British pound

Solution:

A is correct. USD/EUR is expected to decline from 1.1701 to 1.1619, while USD/GBP is expected to decline from 1.3118 to 1.3066. So, the US dollar is expected to be stronger than both the euro and British pound. GBP/EUR is currently 0.8920 [$= (1.3118)^{-1} \times 1.1701$] and is expected to be 0.8893 [$= (1.3066)^{-1} \times 1.1619$], so the British pound is expected to be stronger than the euro.

7. Based on the research report, which of the following lists the three currencies in order of appreciating the most to appreciating the least (in percentage terms) against the US dollar over the next year?

- A. British pound, Swiss franc, euro
- B. Swiss franc, British pound, euro
- C. Euro, Swiss franc, British pound

Solution:

B is correct. The USD/EUR rate depreciates by -0.7 percent ($= [1.1619/1.1701] - 1$), which is the depreciation of the base currency euro against the US dollar. The USD/GBP rate declines -0.4 percent ($= [1.3066/1.3118] - 1$), which is the depreciation of the British pound against the US dollar. Inverting the CHF/USD rate to a USD/CHF convention shows that the base currency Swiss franc appreciates by $+0.35$ percent against the US dollar ($= [1.0136/1.0101] - 1$).

3**FORWARD RATE CALCULATIONS**

explain the arbitrage relationship between spot and forward exchange rates and interest rates, calculate a forward rate using points or in percentage terms, and interpret a forward discount or premium

This lesson continues the previous discussion of the FX market by considering the interactions between spot and forward rates, interest rates, and maturities, which exist because of arbitrage relationships. The relationships among these four factors are maintained because of market efficiencies, and any one factor can be determined using the other three as inputs. In addition, this lesson covers the methods of calculating forward rates in point and percentage terms as well as forward discounts and premiums for these rate relationships.

In professional FX markets, forward exchange rates typically are quoted in terms of points (also sometimes referred to as “pips”). The points on a forward rate quote are simply the difference between the forward exchange rate quote and the spot exchange rate quote, with the points scaled so that they can be related to the last decimal in the spot quote. When the forward rate is higher than the spot rate, the points are positive and the base currency is said to be trading at a *forward premium*. Conversely, if the forward rate is less than the spot rate, the points (forward rate minus spot rate) are negative and the base currency is said to be trading at a *forward discount*. Of course, if the base currency is trading at a forward premium, then the price currency is trading at a forward discount, and vice versa.

This can best be explained by means of an example. Assume the spot euro–dollar exchange rate (USD/EUR) is 1.15885 and the one-year forward rate is 1.19532. Hence, the forward rate is trading at a premium to the spot rate (the forward rate is larger than the spot rate) and the one-year forward points are quoted as +364.7. This +364.7 comes from the following calculation:

$$1.19532 - 1.15885 = +0.03647.$$

Recall that most non-yen exchange rates are quoted to four decimal places. In this case, we would scale up by four decimal places (multiply by 10,000) so that this +0.03647 would be represented as +364.7 points. Notice that the points are scaled to the size of the last digit in the spot exchange rate quote—usually the fourth decimal place. Notice as well that points typically are quoted to one (or more) decimal places, meaning that the forward rate will typically be quoted to five or more decimal places. The exception among the major currencies is the yen, which is typically quoted to two decimal places for spot rates. Here, forward points are scaled up by two decimal places—the last digit in the spot rate quote—by multiplying the difference between forward and spot rates by 100.

Typically, quotes for forward rates are shown as the number of forward points at each maturity, the time between spot settlement and the settlement of the forward contract. These forward points are also called *swap points* because an FX swap consists of simultaneous spot and forward transactions. In our example, a trader would have faced a spot rate and forward points in the euro–dollar (USD/EUR) currency pair similar to those in Exhibit 1,

Exhibit 1: Sample Spot and Forward Quotes

Maturity	Spot Rate or Forward Points
Spot	1.15885
One week	+5.6
One month	+27.1
Three months	+80.9
Six months	+175.6
Twelve months	+364.7

Notice that the absolute number of points generally increases with maturity. This is because the number of points is proportional to the yield differential between the two countries (the Eurozone and the United States, in this case) scaled by the term to maturity. Given the interest rate differential, the longer the term to maturity, the greater the absolute number of forward points. Similarly, given the term to maturity, a wider interest rate differential implies a greater absolute number of forward points. (This relationship will be explained and demonstrated in more detail later in this lesson.)

To convert any of these quoted forward points into a forward rate, one would divide the number of points by 10,000 (to scale down to the fourth decimal place, the last decimal place in the spot quote) and then add the result to the spot exchange rate quote. (As mentioned previously, exchange rates for the Japanese yen, such as the JPY/USD exchange rate, are quoted to two decimal places only, so forward points for the dollar–yen currency pair are divided by 100.) For example, using the data in Exhibit 1 for USD/EUR, the three-month forward rate in this case would be as follows:

$$1.15885 + \left(\frac{+80.9}{10,000} \right) = 1.15885 + 0.00809 = 1.16694.$$

Occasionally, one will see the forward rate or forward points represented as a percentage of the spot rate rather than as an absolute number of points. Continuing the previous example, the three-month forward rate for USD/EUR can be represented as follows:

$$\frac{1.15885 + 0.00809}{1.15885} - 1 = \left(\frac{1.16694}{1.15885} \right) - 1 = +0.698\%.$$

This shows that either the forward rate or the forward points can be used to calculate the percentage discount (or premium) in the forward market—in this case, +0.698 percent rounding to three decimal places. To convert a spot quote into a forward quote when the points are shown as a percentage, one simply multiplies the spot rate by one plus the percentage premium or discount:

$$1.15885 \times (1 + 0.698\%) = 1.15885 \times (1.0000 + 0.00698) \approx 1.16694.$$

Note that, rounded to the fifth decimal place, this is equal to our previous calculation. However, it is typically the case in professional FX markets that forward rates will be quoted in terms of pips rather than percentages.

Arbitrage Relationships

We now turn to the interaction between spot rates, forward rates, and interest rates and how their relationship is derived. Forward exchange rates are based on an arbitrage relationship that equates the investment return on two alternative but equivalent investments. Consider the case of an investor with funds to invest. For simplicity, we will assume that one unit of the investor's domestic currency will be invested for one period. One alternative is to invest for one period at the domestic risk-free rate (r_d); at the end of the period, the amount of funds held is equal to $(1 + r_d)$. An alternative

investment is to convert this one unit of domestic currency to foreign currency using the spot rate of $S_{f/d}$ (number of units of foreign currency per one unit of domestic currency). This can be invested for one period at the foreign risk-free rate; at the end of the period, the investor would have $S_{f/d}(1 + r_f)$ units of foreign currency. These funds must then be converted back to the investor's domestic currency. If the exchange rate to be used for this end-of-period conversion was pre-contracted at the start of the period (i.e., a forward rate was used), it would eliminate any FX risk from converting at a future, unknown spot rate. Given the assumed exchange rate convention (foreign/domestic), the investor would obtain $(1/F_{f/d})$ units of the domestic currency for each unit of foreign currency sold forward. Note that this process of converting domestic funds in the spot FX market, investing at the foreign risk-free rate, and then converting back to the domestic currency with a forward rate is termed "swap financing."

Hence, we have two alternative investments—both are risk free because both are invested at risk-free interest rates and because any FX risk was eliminated (hedged) by using a forward rate. Because these two investments are equal in risk characteristics, they must have the same return. Bearing in mind that the currency quoting convention is the number of foreign currency units per single domestic unit (f/d), this relationship can be stated as follows:

$$(1 + r_d) = S_{f/d}(1 + r_f)\left(\frac{1}{F_{f/d}}\right).$$

This is an arbitrage relationship because it describes two alternative investments (one on either side of the equal sign) that should have equal returns. If they do not, a riskless arbitrage opportunity exists because an investor can sell short the investment with the lower return and invest the funds in the investment with the higher return; the difference between the two returns is pure profit. It is because of this arbitrage relationship that the all-in financing rate using swap financing is close to the domestic interest rate.

This formula is perhaps the easiest and most intuitive way to remember the formula for the forward rate because this formula is based directly on the underlying intuition (the arbitrage relationship of two alternative but equivalent investments, one on either side of the equal sign). Also, the right-hand side of the equation, for the hedged foreign investment alternative, is arranged in proper time sequence: (1) convert domestic to foreign currency; then (2) invest the foreign currency at the foreign interest rate; and finally (3) convert the foreign currency back to the domestic currency. Recall that this equation is based on an f/d exchange rate quoting convention. If the exchange rate data were presented in d/f form, one could either invert these quotes back to f/d form and use the previous equation or use the following equivalent equation:

$$(1 + r_d) = (1/S_{d/f})(1 + r_f)F_{d/f}$$

If this latter equation were used, remember that forward and spot exchange rates are now being quoted on a d/f convention.

This arbitrage equation can be rearranged as needs require. For example, to get the formula for the forward rate, the previous equation can be restated as follows:

$$F_{f/d} = S_{f/d}\left(\frac{1 + r_f}{1 + r_d}\right).$$

Given the spot exchange rate and the domestic and foreign risk-free interest rates, the forward rate is the value that completes this equation and eliminates any arbitrage opportunity. For example, let's assume that the spot exchange rate ($S_{f/d}$) is 1.6535, the domestic 12-month risk-free rate is 3.50 percent, and the foreign 12-month risk-free rate is 5.00 percent. The 12-month forward rate ($F_{f/d}$) must then be equal to:

$$1.6535\left(\frac{1.0500}{1.0350}\right) = 1.6775.$$

Suppose instead that, with the spot exchange rate and interest rates unchanged, you were given a quote on the 12-month forward rate ($F_{f/d}$) of 1.6900. Because this misquoted forward rate does not agree with the arbitrage equation, it would present a riskless arbitrage opportunity. This can be calculated by using the arbitrage equation to compute the return on the two alternative investment strategies. The return on the domestic-only investment approach is the domestic risk-free rate (3.50 percent). In contrast, the return on the hedged foreign investment when this misquoted forward rate is put into the arbitrage equation equals:

$$S_{f/d}(1 + r_f) \left(\frac{1}{F_{f/d}} \right) = 1.6535(1.05) \left(\frac{1}{1.6900} \right) = 1.0273.$$

This results in a return of 2.73 percent. Hence, the investor could make riskless arbitrage profits by borrowing at the higher foreign risk-free rate, selling the foreign currency at the spot exchange rate, hedging the currency exposure (buying the foreign currency back) at the misquoted forward rate, investing the funds at the lower domestic risk-free rate, and thereby getting a profit of 77 basis points (3.50% – 2.73%) for each unit of domestic currency involved—all with no upfront commitment of the investor's own capital. Any such opportunity in real-world financial markets would be quickly “arbed” away. In this example, the investor actually borrows at the higher of the two interest rates but makes a profit because the foreign currency is underpriced in the forward market.

The underlying arbitrage equation can also be rearranged to show the forward rate as a percentage of the spot rate:

$$\frac{F_{f/d}}{S_{f/d}} = \left(\frac{1 + r_f}{1 + r_d} \right).$$

This shows that, given an *f/d* quoting convention, the forward rate will be higher than (be at a premium to) the spot rate if foreign interest rates are higher than domestic interest rates. More generally, and regardless of the quoting convention, *the currency with the higher (lower) interest rate will always trade at a discount (premium) in the forward market*.

One context in which forward rates are quoted as a percentage of spot rates occurs when forward rates are interpreted as expected future spot rates, as follows:

$$F_t = S_{t+1}.$$

Substituting this expression into the previous equation and doing some rearranging leads to the following:

$$\frac{S_{t+1}}{S_t} - 1 = \% \Delta S_{t+1} = \left(\frac{r_f - r_d}{1 + r_d} \right).$$

This shows that if forward rates are interpreted as expected future spot rates, the expected percentage change in the spot rate is proportional to the interest rate differential ($r_f - r_d$).

It is intuitively appealing to see forward rates as expected future spot rates. However, this interpretation of forward rates should be used cautiously. The direction of the expected change in spot rates is somewhat counterintuitive. All else being equal, an increase in domestic interest rates (e.g., the central bank tightens monetary policy) would typically be expected to lead to an increase in the value of the domestic currency. In contrast, the previous equation indicates that, all else equal, a higher domestic interest rate implies slower expected appreciation (or greater expected depreciation) of the domestic currency (recall that this equation is based on an *f/d* quoting convention).

More important, historical data show that forward rates are poor predictors of future spot rates. Although various econometric studies suggest that forward rates may be unbiased predictors of future spot rates (i.e., they do not systematically over- or under-estimate future spot rates), this is not particularly useful information because

the margin of error for these forecasts is so large. As mentioned in the Introduction, the FX market is far too complex and dynamic to be captured by a single variable, such as the level of the yield differential between countries. Moreover, according to the formula for the forward rate, forward rates are based on domestic and foreign interest rates. This means that anything that affects the level and shape of the yield curve in either the domestic or foreign market will also affect the relationship between spot and forward exchange rates. In other words, FX markets do not operate in isolation but rather reflect almost all factors affecting other markets globally; anything that affects expectations or risk premia in these other markets will reverberate in forward exchange rates as well. Although the level of the yield differential is one factor that the market may look at in forming spot exchange rate expectations, it is only one of many factors. (Many traders look to the trend in the yield differential rather than the level of the differential.) Moreover, a lot of noise in FX markets makes almost any model—no matter how complex—a relatively poor predictor of spot rates at any given point in the future. In practice, FX traders and market strategists do *not* base either their currency expectations or trading strategies solely on forward rates.

For the purposes of this lesson, *it is best to understand forward exchange rates simply as a product of the arbitrage equation outlined earlier and forward points as being related to the (time-scaled) interest rate differential between the two countries.* Reading any more than that into forward rates or interpreting them as the “market forecast” can be potentially misleading.

Forward Discounts and Premiums

We now continue our discussion of forward discounts and premiums based on spot and interest rates and add the impact of maturity. To understand the relationship between maturity and forward points, we need to generalize our arbitrage formula slightly. Suppose the investment horizon is a fraction, τ , of the period for which the interest rates are quoted. Then the interest earned in the domestic and foreign markets would be $(r_d \tau)$ and $(r_f \tau)$, respectively. Substituting this into our arbitrage relationship and solving for the difference between the forward and spot exchange rates gives the following:

$$F_{f/d} - S_{f/d} = S_{f/d} \left(\frac{r_f - r_d}{1 + r_d \tau} \right) \tau.$$

This equation shows that forward points (appropriately scaled) are proportional to the spot exchange rate and to the interest rate differential and approximately (but not exactly) proportional to the horizon of the forward contract.

For example, suppose that we wanted to determine the 30-day forward exchange rate given a 30-day domestic risk-free interest rate of 2.00 percent per year, a 30-day foreign risk-free interest rate of 3.00 percent per year, and a spot exchange rate ($S_{f/d}$) of 1.6555. The risk-free assets used in this arbitrage relationship are typically bank deposits quoted using the London Interbank Offered Rate (Libor) for the currencies involved. The day count convention for Libor deposits is actual/360. Incorporating the fractional period (τ) and inserting the data into the forward rate equation leads to the following 30-day forward rate:

$$F_{f/d} = S_{f/d} \left(\frac{1 + r_f \tau}{1 + r_d \tau} \right) = 1.6555 \left(\frac{1 + 0.0300 \left[\frac{30}{360} \right]}{1 + 0.0200 \left[\frac{30}{360} \right]} \right) = 1.6569.$$

This means that, for a 30-day term, forward rates are trading at a premium of 14 pips (1.6569 – 1.6555). This can also be calculated using the previous formula for swap points:

$$F_{f/d} - S_{f/d} = S_{f/d} \left(\frac{r_f - r_d}{1 + r_d \tau} \right) \tau = 1.6555 \left(\frac{0.0300 - 0.0200}{1 + 0.0200 \left[\frac{30}{360} \right]} \right) \left[\frac{30}{360} \right] = 0.0014.$$

As should be clear from this expression, the absolute number of swap points will be closely related to the term of the forward contract (i.e., approximately proportional to $\tau = \text{actual}/360$). For example, leaving the spot exchange rate and interest rates unchanged, and setting the term of the forward contract to 180 days, we obtain the following:

$$F_{fd} - S_{fd} = 1.6555 \left(\frac{0.0300 - 0.0200}{1 + 0.0200 \left[\frac{180}{360} \right]} \right) \left[\frac{180}{360} \right] = 0.0082.$$

This leads to the forward rate trading at a premium of 82 pips. The increase in the number of forward points is approximately proportional to the increase in the term of the contract (from 30 days to 180 days). Note that although the term of the 180-day forward contract is six times longer than that of a 30-day contract, the number of forward points is not exactly six times larger: $6 \times 14 = 84$.

Similarly, the number of forward points is proportional to the spread between foreign and domestic interest rates ($r_f - r_d$). For example, with reference to the original 30-day forward contract, let's set the foreign interest rate to 4.00 percent leaving the domestic interest rate and spot exchange rate unchanged. This doubles the interest rate differential ($r_f - r_d$) from 1.00 percent to 2.00 percent; it also doubles the forward points (rounding to four decimal places), as follows:

$$F_{fd} - S_{fd} = 1.6555 \left(\frac{0.0400 - 0.0200}{1 + 0.0200 \left[\frac{30}{360} \right]} \right) \left[\frac{30}{360} \right] = 0.0028.$$

EXAMPLE 2

Forward Rate Calculations

A French company recently finalized a sale of goods to a UK-based client and expects to receive a payment of GBP50 million in 32 days. The corporate treasurer at the French company wants to hedge the FX risk of this transaction and receives the following exchange rate information from a dealer:

GBP/EUR spot rate	0.8752
One-month forward points	-1.4

1. According to the exchange rate information, the treasurer could hedge the FX risk by:

- A. buying euro (selling British pounds) at a forward rate of 0.87380.
- B. buying euro (selling British pounds) at a forward rate of 0.87506.
- C. selling euro (buying British pounds) at a forward rate of 0.87506.

Solution:

B is correct. The French company would want to convert the British pound to its domestic currency, the euro (it wants to sell British pounds and buy euros). The forward rate would be equal to: $0.8752 + (-1.4/10,000) = 0.87506$.

2. According to the exchange rate information, the *best* interpretation of the forward discount shown is that:

- A. the euro is expected to depreciate over the next 30 days.
- B. one-month UK interest rates are higher than those in the Eurozone.

- C. one-month Eurozone interest rates are higher than those in the United Kingdom.

Solution:

C is correct. A forward discount indicates that interest rates in the base currency country (France, in this case, which uses the euro) are higher than those in the price currency country (the United Kingdom).

3. According to the exchange rate information, if the 12-month forward rate is 0.87295 GBP/EUR, then the 12-month forward points are *closest* to:

- A. -22.5.
- B. -2.25.
- C. -0.00225.

Solution:

A is correct. The number of forward points is equal to the scaled difference between the forward rate and the spot rate. In this case: $0.87295 - 0.87520 = -0.00225$. This is then multiplied by 10,000 to convert to the number of forward points.

4. If a second dealer quotes GBP/EUR at a 12-month forward discount of 0.30 percent on the same spot rate, the French company could:

- A. trade with either dealer because the 12-month forward quotes are equivalent.
- B. lock in a profit in 12 months by buying euros from the second dealer and selling it to the original dealer.
- C. lock in a profit in 12 months by buying euros from the original dealer and selling it to the second dealer.

Solution:

B is correct. A 0.30 percent discount means that the second dealer will sell euros 12 months forward at $0.8752 \times (1 - 0.0030) = 0.87257$, a lower price per euro than the original dealer's quote of 0.87295. Buying euros at the cheaper 12-month forward rate (0.87257) and selling the same amount of euros 12 months forward at the higher 12-month forward rate (0.87295) means a profit of $(0.87295 - 0.87257 = \text{GBP}0.00038)$ per euro transacted, receivable when both forward contracts settle in 12 months.

5. If the 270-day Libor rates (annualized) for the euro and British pound are 1.370 percent and 1.325 percent, respectively, and the spot GBP/EUR exchange rate is 0.8489, then the number of forward points for a 270-day forward rate ($F_{\text{GBP/EUR}}$) is *closest* to:

- A. -22.8.
- B. -3.8.
- C. -2.8.

Solution:

C is correct, because the forward rate is calculated as:

$$F_{\frac{GBP}{EUR}} = S_{\frac{GBP}{EUR}} \left(\frac{1 + r_{GBP} \left[\frac{\text{Actual}}{360} \right]}{1 + r_{EUR} \left[\frac{\text{Actual}}{360} \right]} \right) = 0.8489 \left(\frac{1 + 0.01325 \left[\frac{270}{360} \right]}{1 + 0.01370 \left[\frac{270}{360} \right]} \right) = 0.84862.$$

This shows that the forward points are at a discount of: $0.84862 - 0.84890 = -0.00028$, or -2.8 points. This can also be seen using the swap points formula:

$$F_{\frac{GBP}{EUR}} - S_{\frac{GBP}{EUR}} = 0.8489 \left(\frac{0.01325 - 0.01370}{1 + 0.01370 \left[\frac{270}{360} \right]} \right) \left[\frac{270}{360} \right] = -0.00028.$$

The calculation of -3.8 points omits the day count ($270/360$), and -22.8 points gets the scaling wrong.

PRACTICE PROBLEMS

The following information relates to questions 1-2

A dealer provides spot rate quotes for the following currencies:

Currency	Spot rate
CNY/HKD	0.8422
CNY/ZAR	0.9149
CNY/SEK	1.0218

1. The spot ZAR/HKD cross-rate is *closest* to:
 - A. 0.9205.
 - B. 1.0864.
 - C. 1.2978.
2. Another dealer is quoting the ZAR/SEK cross-rate at 1.1210. The arbitrage profit that can be earned is *closest* to:
 - A. ZAR3671 per million Swedish krona traded.
 - B. SEK4200 per million South African rand traded.
 - C. ZAR4200 per million Swedish krona traded.
3. A BRL/MXN spot rate is listed by a dealer at 0.1378. The six-month forward rate is 0.14193. The six-month forward points are *closest* to:
 - A. -41.3.
 - B. +41.3.
 - C. +299.7.
4. A three-month forward exchange rate in CAD/USD is listed by a dealer at 1.0123. The dealer also quotes three-month forward points as a percentage at 6.8 percent. The CAD/USD spot rate is *closest* to:
 - A. 0.9478.
 - B. 1.0550.
 - C. 1.0862.
5. If the base currency in a forward exchange rate quote is trading at a forward discount, which of the following statements is *most* accurate?
 - A. The forward points will be positive.

- B. The forward percentage will be negative.
 - C. The base currency is expected to appreciate versus the price currency.
6. A forward premium indicates:
- A. an expected increase in demand for the base currency.
 - B. the interest rate is higher in the base currency than in the price currency.
 - C. the interest rate is higher in the price currency than in the base currency.
7. The JPY/AUD spot exchange rate is 82.42, the Japanese yen interest rate is 0.15 percent, and the Australian dollar interest rate is 4.95 percent. If the interest rates are quoted on the basis of a 360-day year, the 90-day forward points in JPY/AUD would be *closest* to:
- A. -377.0.
 - B. -97.7.
 - C. 98.9.

SOLUTIONS

1. A is correct. To get to the ZAR/HKD cross-rate, it is necessary to take the inverse of the CNY/ZAR spot rate and then multiply by the CNY/HKD exchange rate:

$$\begin{aligned}\text{ZAR/HKD} &= (\text{CNY/ZAR})^{-1} \times (\text{CNY/HKD}) \\ &= (1/0.9149) \times 0.8422 = 0.9205\end{aligned}$$

2. C is correct. The ZAR/SEK cross-rate from the original dealer is $(1.0218/0.9149) = 1.1168$, which is lower than the quote from the second dealer. To earn an arbitrage profit, a currency trader would buy Swedish krona (sell South African rand) from the original dealer and sell Swedish krona (buy South African rand) to the second dealer. On SEK1 million, the profit would be:

$$\text{SEK}1,000,000 \times (1.1210 - 1.1168) = \text{ZAR}4,200$$

3. B is correct. The number of forward points equals the forward rate minus the spot rate, or $0.14193 - 0.1378 = 0.00413$, multiplied by 10,000: $10,000 \times 0.00413 = 41.3$ points. By convention, forward points are scaled so that ±1 forward point corresponds to a change of ±1 in the last decimal place of the spot exchange rate.
4. A is correct. Given the forward rate and forward points as a percentage, the unknown in the calculation is the spot rate. The calculation is as follows:

$$\text{Spot rate} \times (1 + \text{Forward points as a percentage}) = \text{Forward rate}$$

$$\text{Spot rate} \times (1 + 0.068) = 1.0123$$

$$\text{Spot} = 1.0123/1.068 = 0.9478$$

5. B is correct. The base currency trading at a forward discount means that 1 unit of the base currency costs less for forward delivery than for spot delivery (i.e., the forward exchange rate is less than the spot exchange rate). The forward points, expressed either as an absolute number of points or as a percentage, are negative.
6. C is correct. To eliminate arbitrage opportunities, the spot exchange rate (S), the forward exchange rate (F), the interest rate in the base currency (r_d), and the interest rate in the price currency (r_f) must satisfy:

$$Ff/d / Sf/d = (1+rft / 1+rdt).$$

According to this formula, the base currency will trade at forward premium ($F > S$) if, and only if, the interest rate in the price currency is higher than the interest rate in the base currency ($r_f > r_d$).

7. B is correct. The forward exchange rate is given by:

$$F_{AUD}^{JPY} = S_{AUD}^{JPY} \left(\frac{1 + r_{JPY}\tau}{1 + r_{AUD}\tau} \right) = 82.42 \left(\frac{1 + 0.0015 \left[\frac{90}{360} \right]}{1 + 0.0495 \left[\frac{90}{360} \right]} \right)$$

$$= 82.42 \times 0.98815 = 81.443.$$

The forward points are as follows:

$$100 \times (F \times S) = 100 \times (81.443 - 82.42) = 100 \times (-0.977) = -97.7.$$

Because the spot exchange rate is quoted with two decimal places, the forward points are scaled by 100.