

2023 CFA[®]
Exam Prep

SchweserNotes[™]
Quantitative Methods and Economics



LEVEL I BOOK 1

KAPLAN[®] SCHWESER

Kaplan Schweser's Path to Success

Level I CFA® Exam

CFA®

Welcome

As the head of Advanced Designations at Kaplan Schweser, I am pleased to have the opportunity to help you prepare for the CFA® exam. Kaplan Schweser has decades of experience in delivering the most effective CFA exam prep products in the market and I know you will find them to be invaluable in your studies.

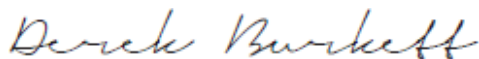
Our products are designed to be an integrated study solution across print and digital media to provide you the best learning experience, whether you are studying with a physical book, online, or on your mobile device.

Our core product, the SchweserNotes™, addresses all of the Topics, Readings, and LOS in the CFA curriculum. Each reading in the SchweserNotes has been broken into smaller, bite-sized modules with Module Quizzes interspersed throughout to help you continually assess your comprehension. After you complete each Topic, take our online Topic Quiz to help you assess your knowledge of the material before you move on to the next section.

All purchasers of the SchweserNotes receive online access to the Kaplan Schweser online platform (our learning management system or LMS) at www.Schweser.com. In the LMS, you will see a dashboard that tracks your overall progress and performance and also includes an Activity Feed, which provides structure and organization to the tasks required to prepare for the CFA exam. You also have access to the SchweserNotes, Module Quizzes, Topic Quizzes, and Mock Exams, as well as the SchweserNotes Videos (if purchased), which contain a short video that complements each module in the SchweserNotes. Look for the icons indicating where video content, Module Quizzes, Topic Quizzes, and Mock Exams are available online. I strongly encourage you to use the dashboard to track your progress and stay motivated.

Again, thank you for trusting Kaplan Schweser with your CFA exam preparation. We're here to help you throughout your journey to become a CFA charterholder.

Regards,



Derek Burkett, CFA, FRM, CAIA

Vice President (Advanced Designations)

Contact us for questions about your study package, upgrading your package, purchasing additional study materials, or for additional information:

888.325.5072 (U.S.) | +1 608.779.8327 (Int'l.)

staff@schweser.com | www.schweser.com/cfa

Book 1: Quantitative Methods and Economics

SchweserNotes™ 2023

Level I CFA®

KAPLAN  **SCHWESER**

SCHWESERNOTES™ 2023 LEVEL I CFA® BOOK 1: QUANTITATIVE METHODS AND ECONOMICS

©2022 Kaplan, Inc. All rights reserved.

Published in 2022 by Kaplan, Inc.

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

ISBN: 978-1-0788-2585-6

These materials may not be copied without written permission from the author. The unauthorized duplication of these notes is a violation of global copyright laws and the CFA Institute Code of Ethics. Your assistance in pursuing potential violators of this law is greatly appreciated.

Required CFA Institute disclaimer: Kaplan Schweser is a CFA Institute Prep Provider. Only CFA Institute Prep Providers are permitted to make use of CFA Institute copyrighted materials which are the building blocks of the exam. We are also required to update our materials every year and this is validated by CFA Institute.

CFA Institute does not endorse, promote, review or warrant the accuracy or quality of the product and services offered by Kaplan Schweser. CFA Institute®, CFA® and “Chartered Financial Analyst®” are trademarks owned by CFA Institute.

Certain materials contained within this text are the copyrighted property of CFA Institute. The following is the copyright disclosure for these materials: “Copyright, 2022, CFA Institute. Reproduced and republished from 2023 Learning Outcome Statements, Level I, II, and III questions from CFA® Program Materials, CFA Institute Standards of Professional Conduct, and CFA Institute’s Global Investment Performance Standards with permission from CFA Institute. All Rights Reserved.”

Disclaimer: The SchweserNotes should be used in conjunction with the original readings as set forth by CFA Institute in their 2023 Level I CFA Study Guide. The information contained in these Notes covers topics contained in the readings referenced by CFA Institute and is believed to be accurate. However, their accuracy cannot be guaranteed nor is any warranty conveyed as to your ultimate exam success. The authors of the referenced readings have not endorsed or sponsored these Notes.

CONTENTS

Learning Outcome Statements (LOS)

QUANTITATIVE METHODS

READING 1

The Time Value of Money

Exam Focus

Module 1.1: EAY and Compounding Frequency

Module 1.2: Calculating PV and FV

Module 1.3: Uneven Cash Flows

Module 1.4: Compounding Frequencies

Key Concepts

Answer Key for Module Quizzes

READING 2

Organizing, Visualizing, and Describing Data

Exam Focus

Module 2.1: Organizing Data

Module 2.2: Visualizing Data

Module 2.3: Measures of Central Tendency

Module 2.4: Measures of Location and Dispersion

Module 2.5: Skewness, Kurtosis, and Correlation

Key Concepts

Answer Key for Module Quizzes

READING 3

Probability Concepts

Exam Focus

Module 3.1: Conditional and Joint Probabilities

Module 3.2: Conditional Expectations and Expected Value

Module 3.3: Portfolio Variance, Bayes, and Counting Problems

Key Concepts

Answers to Module Quiz Questions

READING 4

Common Probability Distributions

Exam Focus

Module 4.1: Uniform and Binomial Distributions

Module 4.2: Normal Distributions

Module 4.3: Lognormal, T, Chi-Square, and F Distributions

Key Concepts

Answer Key for Module Quizzes

READING 5

Sampling and Estimation

Exam Focus

Module 5.1: Sampling Methods, Central Limit Theorem, and Standard Error

Module 5.2: Confidence Intervals, Resampling, and Sampling Biases

Key Concepts

Answer Key for Module Quizzes

READING 6

Hypothesis Testing

Exam Focus

Module 6.1: Hypothesis Tests and Types of Errors

Module 6.2: p -Values and Tests of Means

Module 6.3: Mean Differences and Difference in Means

Module 6.4: Tests of Variance, Correlation, and Independence

Key Concepts

Answer Key for Module Quizzes

READING 7

Introduction to Linear Regression

Exam Focus

Module 7.1: Linear Regression: Introduction

Module 7.2: Goodness of Fit and Hypothesis Tests

Module 7.3: Predicting Dependent Variables and Functional Forms

Key Concepts

Answer Key for Module Quizzes

Topic Quiz: Quantitative Methods

ECONOMICS

READING 8

Topics in Demand and Supply Analysis

Exam Focus

Module 8.1: Elasticity

Module 8.2: Demand and Supply

Key Concepts

Answer Key for Module Quizzes

READING 9

The Firm and Market Structures

Exam Focus

Module 9.1: Perfect Competition

Module 9.2: Monopolistic Competition

Module 9.3: Oligopoly

Module 9.4: Monopoly and Concentration

Key Concepts

Answer Key for Module Quizzes

READING 10

Aggregate Output, Prices, and Economic Growth

Exam Focus

Module 10.1: GDP, Income, and Expenditures

Module 10.2: Aggregate Demand and Supply

Module 10.3: Macroeconomic Equilibrium and Growth

Key Concepts

Answer Key for Module Quizzes

READING 11

Understanding Business Cycles

Exam Focus

Module 11.1: Business Cycle Phases

Module 11.2: Inflation and Indicators

Key Concepts

Answer Key for Module Quizzes

READING 12

Monetary and Fiscal Policy

Exam Focus

Module 12.1: Money and Inflation

Module 12.2: Monetary Policy

Module 12.3: Fiscal Policy

Key Concepts

Answer Key for Module Quizzes

READING 13

Introduction to Geopolitics

Exam Focus

Module 13.1: Geopolitics and Geopolitical Risk

Key Concepts

Answer Key for Module Quizzes

READING 14

International Trade and Capital Flows

Exam Focus

Module 14.1: International Trade Benefits

Module 14.2: Trade Restrictions

Key Concepts

Answer Key for Module Quizzes

READING 15

Currency Exchange Rates

Exam Focus

Module 15.1: Foreign Exchange Rates

Module 15.2: Forward Exchange Rates

Module 15.3: Managing Exchange Rates

Key Concepts

Answer Key for Module Quizzes

Topic Quiz: Economics

Formulas

Appendices

Appendix A: Areas Under The Normal Curve

Cumulative Z-Table

Appendix B: Student's t-Distribution

Appendix C: F-Table at 5% (Upper Tail)

Appendix D: F-Table at 2.5% (Upper Tail)

Appendix E: Chi-Squared Table

Index

LEARNING OUTCOME STATEMENTS (LOS)

1. The Time Value of Money

The candidate should be able to:

- a. interpret interest rates as required rates of return, discount rates, or opportunity costs.
- b. explain an interest rate as the sum of a real risk-free rate and premiums that compensate investors for bearing distinct types of risk.
- c. calculate and interpret the future value (FV) and present value (PV) of a single sum of money, an ordinary annuity, an annuity due, a perpetuity (PV only), and a series of unequal cash flows.
- d. demonstrate the use of a time line in modeling and solving time value of money problems.
- e. calculate the solution for time value of money problems with different frequencies of compounding.
- f. calculate and interpret the effective annual rate, given the stated annual interest rate and the frequency of compounding.

2. Organizing, Visualizing, and Describing Data

The candidate should be able to:

- a. identify and compare data types.
- b. describe how data are organized for quantitative analysis.
- c. interpret frequency and related distributions.
- d. interpret a contingency table.
- e. describe ways that data may be visualized and evaluate uses of specific visualizations.
- f. describe how to select among visualization types.
- g. calculate and interpret measures of central tendency.
- h. evaluate alternative definitions of mean to address an investment problem.
- i. calculate quantiles and interpret related visualizations.
- j. calculate and interpret measures of dispersion.
- k. calculate and interpret target downside deviation.
- l. interpret skewness.
- m. interpret kurtosis.
- n. interpret correlation between two variables.

3. Probability Concepts

The candidate should be able to:

- a. define a random variable, an outcome, and an event.
- b. identify the two defining properties of probability, including mutually exclusive and exhaustive events, and compare and contrast empirical, subjective, and a priori probabilities.
- c. describe the probability of an event in terms of odds for and against the event.
- d. calculate and interpret conditional probabilities.
- e. demonstrate the application of the multiplication and addition rules for probability.
- f. compare and contrast dependent and independent events.
- g. calculate and interpret an unconditional probability using the total probability rule.
- h. calculate and interpret the expected value, variance, and standard deviation of random variables.
- i. explain the use of conditional expectation in investment applications.
- j. interpret a probability tree and demonstrate its application to investment problems.
- k. calculate and interpret the expected value, variance, standard deviation, covariances, and correlations of portfolio returns.
- l. calculate and interpret the covariances of portfolio returns using the joint probability function.
- m. calculate and interpret an updated probability using Bayes' formula.
- n. identify the most appropriate method to solve a particular counting problem and analyze counting problems using factorial, combination, and permutation concepts.

4. Common Probability Distributions

The candidate should be able to:

- a. define a probability distribution and compare and contrast discrete and continuous random variables and their probability functions.
- b. calculate and interpret probabilities for a random variable given its cumulative distribution function.
- c. describe the properties of a discrete uniform random variable, and calculate and interpret probabilities given the discrete uniform distribution function.

- d. describe the properties of the continuous uniform distribution, and calculate and interpret probabilities given a continuous uniform distribution.
- e. describe the properties of a Bernoulli random variable and a binomial random variable, and calculate and interpret probabilities given the binomial distribution function.
- f. explain the key properties of the normal distribution.
- g. contrast a multivariate distribution and a univariate distribution, and explain the role of correlation in the multivariate normal distribution.
- h. calculate the probability that a normally distributed random variable lies inside a given interval.
- i. explain how to standardize a random variable.
- j. calculate and interpret probabilities using the standard normal distribution.
- k. define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion.
- l. explain the relationship between normal and lognormal distributions and why the lognormal distribution is used to model asset prices.
- m. calculate and interpret a continuously compounded rate of return, given a specific holding period return.
- n. describe the properties of the Student's t -distribution, and calculate and interpret its degrees of freedom.
- o. describe the properties of the chi-square distribution and the F -distribution, and calculate and interpret their degrees of freedom.
- p. describe Monte Carlo simulation.

5. Sampling and Estimation

The candidate should be able to:

- a. compare and contrast probability samples with non-probability samples and discuss applications of each to an investment problem.
- b. explain sampling error.
- c. compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling.
- d. explain the central limit theorem and its importance.
- e. calculate and interpret the standard error of the sample mean.
- f. identify and describe desirable properties of an estimator.
- g. contrast a point estimate and a confidence interval estimate of a population parameter.
- h. calculate and interpret a confidence interval for a population mean, given a normal distribution with 1) a known population variance, 2) an unknown population variance, or 3) an unknown population variance and a large sample size.
- i. describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic.
- j. describe the issues regarding selection of the appropriate sample size, data snooping bias, sample selection bias, survivorship bias, look-ahead bias, and time-period bias.

6. Hypothesis Testing

The candidate should be able to:

- a. define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative hypotheses.
- b. compare and contrast one-tailed and two-tailed tests of hypotheses.
- c. explain a test statistic, Type I and Type II errors, a significance level, how significance levels are used in hypothesis testing, and the power of a test.
- d. explain a decision rule and the relation between confidence intervals and hypothesis tests, and determine whether a statistically significant result is also economically meaningful.
- e. explain and interpret the p -value as it relates to hypothesis testing.
- f. describe how to interpret the significance of a test in the context of multiple tests.
- g. identify the appropriate test statistic and interpret the results for a hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed and the variance is (1) known or (2) unknown.
- h. identify the appropriate test statistic and interpret the results for a hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances.
- i. identify the appropriate test statistic and interpret the results for a hypothesis test concerning the mean difference of two normally distributed populations.
- j. identify the appropriate test statistic and interpret the results for a hypothesis test concerning (1) the variance of a normally distributed population and (2) the equality of the variances of two normally distributed populations based on two independent random samples.

- k. compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test.
- l. 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.
- m. explain tests of independence based on contingency table data.

7. Introduction to Linear Regression

The candidate should be able to:

- a. describe a simple linear regression model and the roles of the dependent and independent variables in the model.
- b. describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation.
- c. explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated.
- d. calculate and interpret the coefficient of determination and the F -statistic in a simple linear regression.
- e. 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.
- f. formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance.
- g. 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.
- h. describe different functional forms of simple linear regressions.

8. Topics in Demand and Supply Analysis

The candidate should be able to:

- a. calculate and interpret price, income, and cross-price elasticities of demand and describe factors that affect each measure.
- b. compare substitution and income effects.
- c. contrast normal goods with inferior goods.
- d. describe the phenomenon of diminishing marginal returns.
- e. determine and interpret breakeven and shutdown points of production.
- f. describe how economies of scale and diseconomies of scale affect costs.

9. The Firm and Market Structures

The candidate should be able to:

- a. describe characteristics of perfect competition, monopolistic competition, oligopoly, and pure monopoly.
- b. explain relationships between price, marginal revenue, marginal cost, economic profit, and the elasticity of demand under each market structure.
- c. describe a firm's supply function under each market structure.
- d. describe and determine the optimal price and output for firms under each market structure.
- e. describe pricing strategy under each market structure.
- f. explain factors affecting long-run equilibrium under each market structure.
- g. describe the use and limitations of concentration measures in identifying market structure.
- h. identify the type of market structure within which a firm operates.

10. Aggregate Output, Prices, and Economic Growth

The candidate should be able to:

- a. calculate and explain gross domestic product (GDP) using expenditure and income approaches.
- b. compare the sum-of-value-added and value-of-final-output methods of calculating GDP.
- c. compare nominal and real GDP and calculate and interpret the GDP deflator.
- d. compare GDP, national income, personal income, and personal disposable income.
- e. explain the fundamental relationship among saving, investment, the fiscal balance, and the trade balance.
- f. explain how the aggregate demand curve is generated.
- g. explain the aggregate supply curve in the short run and long run.
- h. explain causes of movements along and shifts in aggregate demand and supply curves.
- i. describe how fluctuations in aggregate demand and aggregate supply cause short-run changes in the economy and the business cycle.
- j. distinguish among the following types of macroeconomic equilibria: long-run full employment, short-run recessionary gap, short-run inflationary gap, and short-run stagflation.
- k. explain how a short-run macroeconomic equilibrium may occur at a level above or below full employment.
- l. analyze the effect of combined changes in aggregate supply and demand on the economy.

- m. describe sources, measurement, and sustainability of economic growth.
- n. describe the production function approach to analyzing the sources of economic growth.
- o. define and contrast input growth with growth of total factor productivity as components of economic growth.

11. Understanding Business Cycles

The candidate should be able to:

- a. describe the business cycle and its phases.
- b. describe credit cycles.
- c. describe how resource use, consumer and business activity, housing sector activity, and external trade sector activity vary as an economy moves through the business cycle.
- d. describe theories of the business cycle.
- e. interpret a set of economic indicators, and describe their uses and limitations.
- f. describe types of unemployment, and compare measures of unemployment.
- g. explain inflation, hyperinflation, disinflation, and deflation.
- h. explain the construction of indexes used to measure inflation.
- i. compare inflation measures, including their uses and limitations.
- j. contrast cost-push and demand-pull inflation.

12. Monetary and Fiscal Policy

The candidate should be able to:

- a. compare monetary and fiscal policy.
- b. describe functions and definitions of money.
- c. explain the money creation process.
- d. describe theories of the demand for and supply of money.
- e. describe the Fisher effect.
- f. describe roles and objectives of central banks.
- g. contrast the costs of expected and unexpected inflation.
- h. describe tools used to implement monetary policy.
- i. describe the monetary transmission mechanism.
- j. explain the relationships between monetary policy and economic growth, inflation, interest, and exchange rates.
- k. describe qualities of effective central banks.
- l. contrast the use of inflation, interest rate, and exchange rate targeting by central banks.
- m. determine whether a monetary policy is expansionary or contractionary.
- n. describe limitations of monetary policy.
- o. describe roles and objectives of fiscal policy.
- p. describe the arguments about whether the size of a national debt relative to GDP matters.
- q. describe tools of fiscal policy, including their advantages and disadvantages.
- r. explain the implementation of fiscal policy and difficulties of implementation.
- s. determine whether a fiscal policy is expansionary or contractionary.
- t. explain the interaction of monetary and fiscal policy.

13. Introduction to Geopolitics

The candidate should be able to:

- a. describe geopolitics from a cooperation versus competition perspective.
- b. describe geopolitics and its relationship with globalization.
- c. describe tools of geopolitics and their impact on regions and economies.
- d. describe geopolitical risk and its impact on investments.

14. International Trade and Capital Flows

The candidate should be able to:

- a. compare gross domestic product and gross national product.
- b. describe benefits and costs of international trade.
- c. contrast comparative advantage and absolute advantage.
- d. compare the Ricardian and Heckscher–Ohlin models of trade and the source(s) of comparative advantage in each model.
- e. compare types of trade and capital restrictions and their economic implications.
- f. explain motivations for and advantages of trading blocs, common markets, and economic unions.
- g. describe common objectives of capital restrictions imposed by governments.
- h. describe the balance of payments accounts including their components.
- i. explain how decisions by consumers, firms, and governments affect the balance of payments.

- j. describe functions and objectives of the international organizations that facilitate trade, including the World Bank, the International Monetary Fund, and the World Trade Organization.

15. Currency Exchange Rates

The candidate should be able to:

- a. define an exchange rate and distinguish between nominal and real exchange rates and spot and forward exchange rates.
- b. calculate and interpret the percentage change in a currency relative to another currency.
- c. describe functions of and participants in the foreign exchange market.
- d. calculate and interpret currency cross-rates.
- e. calculate an outright forward quotation from forward quotations expressed on a points basis or in percentage terms.
- f. explain the arbitrage relationship between spot rates, forward rates, and interest rates.
- g. calculate and interpret a forward discount or premium.
- h. calculate and interpret the forward rate consistent with the spot rate and the interest rate in each currency.
- i. describe exchange rate regimes.
- j. explain the effects of exchange rates on countries' international trade and capital flows.

READING 1

THE TIME VALUE OF MONEY

EXAM FOCUS

This reading covers time value of money concepts and applications. Procedures are presented for calculating the future value and present value of a single cash flow, an annuity, and a series of uneven cash flows. The impact of different compounding periods is examined, along with the procedures for solving for other variables in time value of money problems. Your main objective in this chapter is to master time value of money mechanics (i.e., learn how to crunch the numbers). Work all the questions and problems found at the end of this review. Make sure you know how to grind out all the time value of money problems on your calculator. The more rapidly you can do them (correctly), the more time you will have for the more conceptual parts of the exam.

MODULE 1.1: EAY AND COMPOUNDING FREQUENCY



Video covering this content is available online.

The concept of **compound interest** or **interest on interest** is deeply embedded in time value of money (TVM) procedures. When an investment is subjected to compound interest, the growth in the value of the investment from period to period reflects not only the interest earned on the original principal amount but also on the interest earned on the previous period's interest earnings—the interest on interest.

TVM applications frequently call for determining the **future value (FV)** of an investment's cash flows as a result of the effects of compound interest. Computing FV involves projecting the cash flows forward, on the basis of an appropriate compound interest rate, to the end of the investment's life. The computation of the **present value (PV)** works in the opposite direction—it brings the cash flows from an investment back to the beginning of the investment's life based on an appropriate compound rate of return.

Being able to measure the PV and/or FV of an investment's cash flows becomes useful when comparing investment alternatives because the value of the investment's cash flows must be measured at some common point in time, typically at the end of the investment horizon (FV) or at the beginning of the investment horizon (PV).

Using a Financial Calculator

It is very important that you be able to use a financial calculator when working TVM problems because the exam is constructed under the assumption that candidates have the ability to do so. There is simply no other way that you will have time to solve TVM problems. *CFA Institute allows only two types of calculators to be used for the exam—the TI BAII Plus[®] (including the*

BaII Plus Professional) and the HP 12C[®] (including the HP 12C Platinum). This reading is written primarily with the TI BAII Plus in mind. If you don't already own a calculator, go out and buy a TI BAII Plus! However, if you already own the HP 12C and are comfortable with it, by all means continue to use it.

The TI BAII Plus comes preloaded from the factory with the periods per year function (P/Y) set to 12. This automatically converts the annual interest rate (I/Y) into monthly rates. While appropriate for many loan-type problems, this feature is not suitable for the vast majority of the TVM applications we will be studying. So prior to using our SchweserNotes™, please set your P/Y key to "1" using the following sequence of keystrokes:

[2nd] [P/Y] "1" [ENTER] [2nd] [QUIT]

As long as you do not change the P/Y setting, it will remain set at one period per year until the battery from your calculator is removed (it does not change when you turn the calculator on and off). If you want to check this setting at any time, press [2nd] [P/Y]. The display should read P/Y = 1.0. If it does, press [2nd] [QUIT] to get out of the "programming" mode. If it doesn't, repeat the procedure previously described to set the P/Y key. With P/Y set to equal 1, it is now possible to think of I/Y as the interest rate per compounding period and N as the number of compounding periods under analysis. Thinking of these keys in this way should help you keep things straight as we work through TVM problems.

Before we begin working with financial calculators, you should familiarize yourself with your TI by locating the TVM keys noted below. These are the only keys you need to know to work virtually on all TVM problems.

- N = Number of compounding periods
- I/Y = Interest rate per compounding period
- PV = Present value
- FV = Future value
- PMT = Annuity payments, or constant periodic cash flow
- CPT = Compute



PROFESSOR'S NOTE

We have provided an online video in the Resource Library on how to use the TI calculator. You can view it by logging in at www.schweser.com.

Time Lines

It is often a good idea to draw a time line before you start to solve a TVM problem. A **time line** is simply a diagram of the cash flows associated with a TVM problem. A cash flow that occurs in the present (today) is put at time zero. Cash outflows (payments) are given a negative sign, and cash inflows (receipts) are given a positive sign. Once the cash flows are assigned to a time line, they may be moved to the beginning of the investment period to calculate the PV through a process called **discounting** or to the end of the period to calculate the FV using a process called **compounding**.

Figure 1.1 illustrates a time line for an investment that costs \$1,000 today (outflow) and will return a stream of cash payments (inflows) of \$300 per year at the end of each of the next five years.

Figure 1.1: Time Line



Please recognize that the cash flows occur at the end of the period depicted on the time line. Furthermore, note that the end of one period is the same as the beginning of the next period. For example, the end of the second year ($t = 2$) is the same as the beginning of the third year, so a cash flow at the beginning of Year 3 appears at time $t = 2$ on the time line. Keeping this convention in mind will help you keep things straight when you are setting up TVM problems.



PROFESSOR'S NOTE

Throughout the problems in this review, rounding differences may occur between the use of different calculators or techniques presented in this document. So don't panic if you are a few cents off in your calculations.

LOS 1.a: Interpret interest rates as required rates of return, discount rates, or opportunity costs.

Interest rates are our measure of the time value of money, although risk differences in financial securities lead to differences in their equilibrium interest rates. Equilibrium interest rates are the **required rate of return** for a particular investment, in the sense that the market rate of return is the return that investors and savers require to get them to willingly lend their funds. Interest rates are also referred to as **discount rates** and, in fact, the terms are often used interchangeably. If an individual can borrow funds at an interest rate of 10%, then that individual should *discount* payments to be made in the future at that rate in order to get their equivalent value in current dollars or other currencies. Finally, we can also view interest rates as the **opportunity cost** of current consumption. If the market rate of interest on 1-year securities is 5%, earning an additional 5% is the opportunity forgone when current consumption is chosen rather than saving (postponing consumption).

LOS 1.b: 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 **real risk-free rate** of interest is a theoretical rate on a single-period loan that has no expectation of inflation in it. When we speak of a real rate of return, we are referring to an investor's increase in purchasing power (after adjusting for inflation). Since expected inflation in future periods is not zero, the rates we observe on U.S. Treasury bills (T-bills), for example, are

risk-free rates but not *real* rates of return. T-bill rates are *nominal risk-free rates* because they contain an *inflation premium*. The approximate relation here is:

$$\text{nominal risk-free rate} = \text{real risk-free rate} + \text{expected inflation rate}$$

Securities may have one or more **types of risk**, and each added risk increases the required rate of return on the security. These types of risk are:

- **Default risk.** The risk that a borrower will not make the promised payments in a timely manner.
- **Liquidity risk.** The risk of receiving less than fair value for an investment if it must be sold for cash quickly.
- **Maturity risk.** As we will cover in detail in the section on debt securities, the prices of longer-term bonds are more volatile than those of shorter-term bonds. Longer maturity bonds have more maturity risk than shorter-term bonds and require a maturity risk premium.

Each of these risk factors is associated with a risk premium that we add to the nominal risk-free rate to adjust for greater default risk, less liquidity, and longer maturity relative to a very liquid, short-term, default risk-free rate such as that on T-bills. We can write:

$$\begin{aligned} \text{nominal rate of interest} &= \text{nominal risk-free rate} \\ &+ \text{default risk premium} \\ &+ \text{liquidity premium} \\ &+ \text{maturity risk premium} \end{aligned}$$



MODULE QUIZ 1.1

1. An interest rate is *best* interpreted as:
 - A. a discount rate or a measure of risk.
 - B. a measure of risk or a required rate of return.
 - C. a required rate of return or the opportunity cost of consumption.
2. An interest rate from which the inflation premium has been subtracted is known as:
 - A. a real interest rate.
 - B. a risk-free interest rate.
 - C. a real risk-free interest rate.

MODULE 1.2: CALCULATING PV AND FV



LOS 1.c: Calculate and interpret the future value (FV) and present value (PV) of a single sum of money, an ordinary annuity, an annuity due, a perpetuity (PV only), and a series of unequal cash flows.

Video covering this content is available online.

Future Value of a Single Sum

Future value is the amount to which a current deposit will grow over time when it is placed in an account paying compound interest. The FV, also called the compound value, is simply an example of compound interest at work.

The formula for the FV of a *single* cash flow is:

$$FV = PV(1 + I/Y)^N$$

where:

PV = amount of money invested today (the present value)

I/Y = rate of return per compounding period

N = total number of compounding periods

In this expression, the investment involves a single cash outflow, PV, which occurs today, at $t = 0$ on the time line. The single sum FV formula will determine the value of an investment at the end of N compounding periods, given that it can earn a fully compounded rate of return, I/Y, over all of the periods.

The factor $(1 + I/Y)^N$ represents the compounding rate on an investment and is frequently referred to as the **future value factor**, or the **future value interest factor**, for a single cash flow at I/Y over N compounding periods. These are the values that appear in interest factor tables, which we will not be using.

EXAMPLE: FV of a single sum

Calculate the FV of a \$200 investment at the end of two years if it earns an annually compounded rate of return of 10%.

Answer:

To solve this problem with your calculator, input the relevant data and compute FV.

$$N = 2; I/Y = 10; PV = -200; CPT \rightarrow FV = \$242$$



PROFESSOR'S NOTE

Note the negative sign on PV. This is not necessary, but it makes the FV come out as a positive number. If you enter PV as a positive number, ignore the negative sign that appears on the FV.

This relatively simple problem could also be solved using the following equation:

$$FV = 200(1 + 0.10)^2 = \$242$$

On the TI calculator, enter 1.10 [x^2] 200 [=].

Present Value of a Single Sum

The PV of a single sum is today's value of a cash flow that is to be received at some point in the future. In other words, it is the amount of money that must be invested today, at a given rate of return over a given period of time, in order to end up with a specified FV. As previously mentioned, the process for finding the PV of a cash flow is known as *discounting* (i.e., future cash flows are "discounted" back to the present). The interest rate used in the discounting process is commonly referred to as the **discount rate** but may also be referred to as the **opportunity cost, required rate of return**, and the **cost of capital**. Whatever you want to call it, it represents the annual compound rate of return that can be earned on an investment.

The relationship between PV and FV can be seen by examining the FV expression stated earlier. Rewriting the FV equation in terms of PV, we get:

$$PV = FV \times \left[\frac{1}{(1 + I/Y)^N} \right] = \frac{FV}{(1 + I/Y)^N}$$

Note that for a single future cash flow, PV is always less than the FV whenever the discount rate is positive.

The quantity $1/(1 + I/Y)^N$ in the PV equation is frequently referred to as the **present value factor**, **present value interest factor**, or **discount factor** for a single cash flow at I/Y over N compounding periods.

EXAMPLE: PV of a single sum

Given a discount rate of 10%, calculate the PV of a \$200 cash flow that will be received in two years.

Answer:

To solve this problem, input the relevant data and compute PV.

$$N = 2; I/Y = 10; FV = 200; CPT \rightarrow PV = -\$165.29 \text{ (ignore the sign)}$$



PROFESSOR'S NOTE

With single sum PV problems, you can either enter FV as a positive number and ignore the negative sign on PV or enter FV as a negative number.

This relatively simple problem could also be solved using the following PV equation:

$$PV = \frac{200}{(1 + 0.10)^2} = \$165.29$$

On the TI, enter 1.10 [y^x] 2 [=] [1/x] [\times] 200 [=].

The PV computed here implies that at a rate of 10%, an investor will be indifferent between \$200 in two years and \$165.29 today. Put another way, \$165.29 is the amount that must be invested today at a 10% rate of return in order to generate a cash flow of \$200 at the end of two years.

Annuities

An **annuity** is a stream of *equal cash flows* that occurs at *equal intervals* over a given period. Receiving \$1,000 per year at the end of each of the next eight years is an example of an annuity. There are two types of annuities: **ordinary annuities** and **annuities due**. The *ordinary annuity* is the most common type of annuity. It is characterized by cash flows that occur at the *end* of each compounding period. This is a typical cash flow pattern for many investment and business finance applications. The other type of annuity is called an *annuity due*, where payments or receipts occur at the beginning of each period (i.e., the first payment is today at $t = 0$).

Computing the FV or PV of an annuity with your calculator is no more difficult than it is for a single cash flow. You will know four of the five relevant variables and solve for the fifth (either PV or FV). The difference between single sum and annuity TVM problems is that instead of solving for the PV or FV of a single cash flow, we solve for the PV or FV of a stream of equal

periodic cash flows, where the size of the periodic cash flow is defined by the payment (PMT) variable on your calculator.

EXAMPLE: FV of an ordinary annuity

What is the future value of an ordinary annuity that pays \$200 per year at the end of each of the next three years, given the investment is expected to earn a 10% rate of return?

Answer:

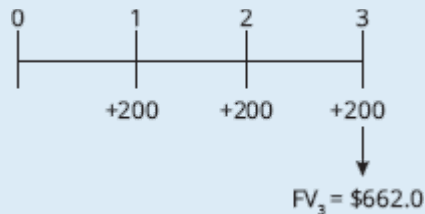
This problem can be solved by entering the relevant data and computing FV.

$$N = 3; I/Y = 10; PMT = -200; CPT \rightarrow FV = \$662$$

Implicit here is that $PV = 0$; clearing the TVM functions sets both PV and FV to zero.

The time line for the cash flows in this problem is depicted in the following figure:

FV of an Ordinary Annuity



As indicated here, the sum of the compounded values of the individual cash flows in this 3-year ordinary annuity is \$662. Note that the annuity payments themselves amounted to \$600, and the balance is the interest earned at the rate of 10% per year.

To find the PV of an ordinary annuity, we use the future cash flow stream, PMT, that we used with FV annuity problems, but we discount the cash flows back to the present (time $t = 0$) rather than compounding them forward to the terminal date of the annuity.

Here again, the PMT variable is a *single* periodic payment, *not* the total of all the payments (or deposits) in the annuity. The PVA_0 measures the collective PV of a stream of equal cash flows received at the end of each compounding period over a stated number of periods, N , given a specified rate of return, I/Y . The following examples illustrate how to determine the PV of an ordinary annuity using a financial calculator:

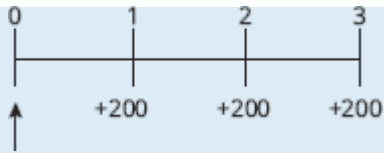
EXAMPLE: PV of an ordinary annuity

What is the PV of an annuity that pays \$200 per year at the end of each of the next three years, given a 10% discount rate?

Answer:

The payments occur at the end of the year, so this annuity is an ordinary annuity. To solve this problem, enter the relevant information and compute PV.

$$N = 3; I/Y = 10; PMT = -200; FV = 0; CPT \rightarrow PV = \$497.37$$



$$PV_0 = \$497.37$$

The \$497.37 computed here represents the amount of money that an investor would need to invest *today* at a 10% rate of return to generate three end-of-year cash flows of \$200 each.

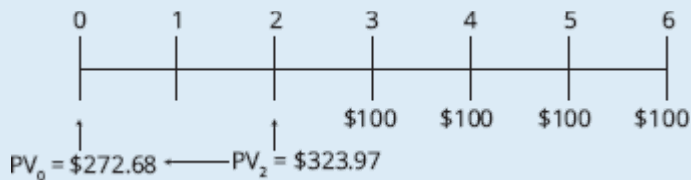
EXAMPLE: PV of an ordinary annuity beginning later than $t = 1$

What is the present value of four \$100 end-of-year payments if the first payment is to be received three years from today and the appropriate rate of return is 9%?

Answer:

The time line for this cash flow stream is shown in the following figure:

PV of an Annuity Beginning at $t = 3$



Step 1: Find the present value of the annuity as of the end of year 2 (PV_2).

Input the relevant data and solve for PV_2 .

$$N = 4; I/Y = 9; PMT = -100; FV = 0; CPT \rightarrow PV = PV_2 = \$323.97$$

Step 2: Find the present value of PV_2 .

Input the relevant data and solve for PV_0 .

$$N = 2; I/Y = 9; PMT = 0; FV = -323.97; CPT \rightarrow PV = PV_0 = \$272.68$$

In this solution, the annuity was treated as an ordinary annuity. The PV was computed one period before the first payment, and we discounted $PV_2 = \$323.97$ over two years. We need to stress this important point. The PV annuity function on your calculator set in “END” mode gives you the value *one period before the annuity begins*. Although the annuity begins at $t = 3$, we discounted the result for only two periods to get the present ($t = 0$) value.

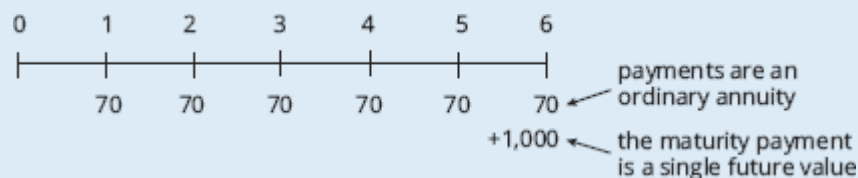
EXAMPLE: PV of a bond's cash flows

A bond will make coupon interest payments of 70 euros (7% of its face value) at the end of each year and will also pay its face value of 1,000 euros at maturity in six years. If the appropriate discount rate is 8%, what is the present value of the bond's promised cash flows?

Answer:

The six annual coupon payments of 70 euros each can be viewed as an ordinary annuity. The maturity value of 1,000 euros is the future value of the bond at the time the last coupon payment is made. On a time line, the promised payment stream is as shown below.

Cash Flows for a 6-Year, 7%, 1,000 Euro Bond



The PV of the bond's cash flows can be broken down into the PV of a 6-payment ordinary annuity, plus the PV of a 1,000 euro lump sum to be received six years from now.

The calculator solution is:

$$N = 6; PMT = 70; I/Y = 8; FV = 1,000; CPT \rightarrow PV = -953.77$$

With a yield to maturity of 8%, the value of the bond is 953.77 euros.

Note that the PMT and FV must have the same sign, since both are cash flows paid to the investor (paid by the bond issuer). The calculated PV will have the opposite sign from PMT and FV.

Future Value of an Annuity Due

Sometimes it is necessary to find the *FV of an annuity due* (FVA_D), an annuity where the annuity payments (or deposits) occur at the beginning of each compounding period. Fortunately, our financial calculators can be used to do this, but with one slight modification—the calculator must be set to the beginning-of-period (BGN) mode. To switch between the BGN and END modes on the TI, press [2nd] [BGN] [2nd] [SET]. When this is done, “BGN” will appear in the upper right corner of the display window. If the display indicates the desired mode, press [2nd] [QUIT]. You will normally want your calculator to be in the ordinary annuity (END) mode, so remember to switch out of BGN mode after working annuity due problems. Note that nothing appears in the upper right corner of the display window when the TI is set to the END mode. It should be mentioned that while annuity due payments are made or received at the beginning of each period, the FV of an annuity due is calculated as of the end of the last period.

Another way to compute the FV of an annuity due is to calculate the FV of an ordinary annuity, and simply multiply the resulting FV by $[1 + \text{periodic compounding rate } (I/Y)]$. Symbolically, this can be expressed as:

$$FVA_D = FVA_O \times (1 + I/Y)$$

The following examples illustrate how to compute the FV of an annuity due:

EXAMPLE: FV of an annuity due

What is the future value of an annuity that pays \$200 per year at the beginning of each of the next three years, commencing today, if the cash flows can be invested at an annual rate of 10%?

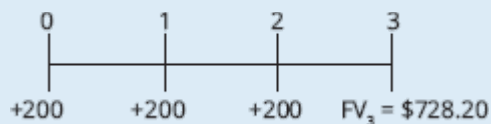
Answer:

Note in the time line in the following figure that the FV is computed as of the end of the last year in the life of the annuity, Year 3, even though the final payment occurs at the beginning of Year 3 (end of Year 2).

To solve this problem, put your calculator in the BGN mode ([2nd] [BGN] [2nd] [SET] [2nd] [QUIT] on the TI or [g] [BEG] on the HP), then input the relevant data and compute FV.

$$N = 3; I/Y = 10; PMT = -200; CPT \rightarrow FV = \$728.20$$

FV of an Annuity Due



Alternatively, we could calculate the FV for an ordinary annuity and multiply it by $(1 + I/Y)$. Leaving your calculator in the END mode, enter the following inputs:

$$N = 3; I/Y = 10; PMT = -200; CPT \rightarrow FVA_0 = \$662.00$$

$$FVA_D = FVA_0 \times (1 + I/Y) = 662 \times 1.10 = \$728.20$$

Present Value of an Annuity Due

While less common than those for ordinary annuities, some problems may require you to find the *PV of an annuity due* (PVA_D). Using a financial calculator, this really shouldn't be much of a problem. With an annuity due, *there is one less discounting period* since the first cash flow occurs at $t = 0$ and thus is already its PV. This implies that, all else equal, the PV of an annuity due will be greater than the PV of an ordinary annuity.

As you will see in the next example, there are two ways to compute the PV of an annuity due. The first is to put the calculator in the BGN mode and then input all the relevant variables (PMT, I/Y , and N) as you normally would. The second, and far easier way, is to treat the cash flow stream as an ordinary annuity over N compounding periods, and simply multiply the resulting PV by $[1 + \text{periodic compounding rate } (I/Y)]$.

Symbolically, this can be stated as:

$$PVA_D = PVA_0 \times (1 + I/Y)$$

The advantage of this second method is that you leave your calculator in the END mode and won't run the risk of forgetting to reset it. Regardless of the procedure used, the computed PV is given as of the beginning of the first period, $t = 0$.

EXAMPLE: PV of an annuity due

Given a discount rate of 10%, what is the present value of an annuity that makes \$200 payments at the beginning of each of the next three years, starting today?

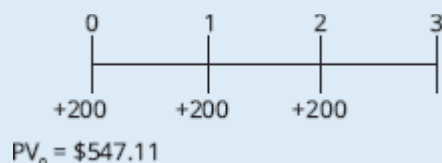
Answer:

First, let's solve this problem using the calculator's BGN mode. Set your calculator to the BGN mode ([2nd] [BGN] [2nd] [SET] [2nd] [QUIT] on the TI or [g] [BEG] on the HP), enter the relevant data, and compute PV.

$$N = 3; I/Y = 10; PMT = -200; CPT \rightarrow PVA_D = \$547.11$$

The time line for this problem is shown in the following figure:

PV of an Annuity Due



Alternatively, this problem can be solved by leaving your calculator in the END mode. First, compute the PV of an ordinary 3-year annuity. Then multiply this PV by $(1 + I/Y)$. To use this approach, enter the relevant inputs and compute PV.

$$N = 3; I/Y = 10; PMT = -200; CPT \rightarrow PVA_O = \$497.37$$

$$PVA_D = PVA_O \times (1 + I/Y) = \$497.37 \times 1.10 = \$547.11$$

Present Value of a Perpetuity

A **perpetuity** is a financial instrument that pays a fixed amount of money at set intervals over an *infinite* period of time. In essence, a perpetuity is a perpetual annuity. Most preferred stocks are examples of perpetuities since they promise fixed interest or dividend payments forever. Without going into all the excruciating mathematical details, the discount factor for a perpetuity is just one divided by the appropriate rate of return (i.e., $1/r$). Given this, we can compute the PV of a perpetuity.

$$PV_{\text{perpetuity}} = \frac{PMT}{I/Y}$$

The PV of a perpetuity is the fixed periodic cash flow divided by the appropriate periodic rate of return.

As with other TVM applications, it is possible to solve for unknown variables in the $PV_{\text{perpetuity}}$ equation. In fact, you can solve for any one of the three relevant variables, given the values for the other two.

EXAMPLE: PV of a perpetuity

Kodon Corporation issues preferred stock that will pay \$4.50 per year in annual dividends beginning next year and plans to follow this dividend policy forever. Given an 8% rate of return, what is the value of Kodon's preferred stock today?

Answer:

Given that the value of the stock is the PV of all future dividends, we have:

$$PV_{\text{perpetuity}} = \frac{4.50}{0.08} = \$56.25$$

Thus, if an investor requires an 8% rate of return, the investor should be willing to pay \$56.25 for each share of Kodon's preferred stock. Note that the PV of a perpetuity is its value one period before its next payment.

EXAMPLE: PV of a deferred perpetuity

Assume the Kodon preferred stock in the preceding examples is scheduled to pay its first dividend in four years, and is non-cumulative (i.e., does not pay any dividends for the first three years). Given an 8% required rate of return, what is the value of Kodon's preferred stock today?

Answer:

As in the previous example, $PV_{\text{perpetuity}} = \frac{4.50}{0.08} = \56.25 , but because the first dividend is paid at $t = 4$, this PV is the value at $t = 3$. To get the value of the preferred stock today, we must discount this value for three periods: $\frac{56.25}{(1.08)^3} = \44.65 .



MODULE QUIZ 1.2

- The amount an investor will have in 15 years if \$1,000 is invested today at an annual interest rate of 9% will be *closest* to:
 - \$1,350.
 - \$3,518.
 - \$3,642.
- How much must be invested today, at 8% interest, to accumulate enough to retire a \$10,000 debt due seven years from today?
 - \$5,835.
 - \$6,123.
 - \$8,794.
- An investor has just won the lottery and will receive \$50,000 per year at the end of each of the next 20 years. At a 10% interest rate, the present value of the winnings is *closest* to:
 - \$425,678.
 - \$637,241.
 - \$2,863,750.
- An investor is to receive a 15-year, \$8,000 annuity, with the first payment to be received today. At an 11% discount rate, this annuity's worth today is *closest* to:
 - \$55,855.
 - \$57,527.
 - \$63,855.
- If \$1,000 is invested today and \$1,000 is invested at the beginning of each of the next three years at 12% interest (compounded annually), the amount an investor will have at the end of the fourth year will be *closest* to:
 - \$4,779.
 - \$5,353.
 - \$6,792.
- Terry Corporation preferred stocks are expected to pay a \$9 annual dividend forever. If the required rate of return on equivalent investments is 11%, a share of Terry preferred should be worth:

- A. \$81.82.
- B. \$99.00.
- C. \$122.22.

MODULE 1.3: UNEVEN CASH FLOWS



Video covering this content is available online.

It is not uncommon to have applications in investments and corporate finance where it is necessary to evaluate a cash flow stream that is not equal from period to period. The time line in Figure 1.2 depicts such a cash flow stream.

Figure 1.2: Time Line for Uneven Cash Flows



This 3-year cash flow series is not an annuity since the cash flows are different every year. In essence, this series of uneven cash flows is nothing more than a stream of annual single sum cash flows. Thus, to find the PV or FV of this cash flow stream, all we need to do is sum the PVs or FVs of the individual cash flows.

EXAMPLE: Computing the FV of an uneven cash flow series

Using a rate of return of 10%, compute the future value of the 3-year uneven cash flow stream described above at the end of the third year.

Answer:

The FV for the cash flow stream is determined by first computing the FV of each individual cash flow, then summing the FVs of the individual cash flows.

$$FV_1: PV = -300; I/Y = 10; N = 2; CPT \rightarrow FV = FV_1 = 363$$

$$FV_2: PV = -600; I/Y = 10; N = 1; CPT \rightarrow FV = FV_2 = 660$$

$$FV_3: PV = -200; I/Y = 10; N = 0; CPT \rightarrow FV = FV_3 = 200$$

$$\mathbf{FV \text{ of cash flow stream} = \Sigma FV_{\text{individual}} = 1,223}$$

EXAMPLE: Computing PV of an uneven cash flow series

Compute the present value of this 3-year uneven cash flow stream described previously using a 10% rate of return.

Answer:

This problem is solved by first computing the PV of each individual cash flow, then summing the PVs of the individual cash flows, which yields the PV of the cash flow stream. Again the signs of the cash flows are preserved.

$$PV_1: FV = 300; I/Y = 10; N = 1; CPT \rightarrow PV = PV_1 = -272.73$$

$$PV_2: FV = 600; I/Y = 10; N = 2; CPT \rightarrow PV = PV_2 = -495.87$$

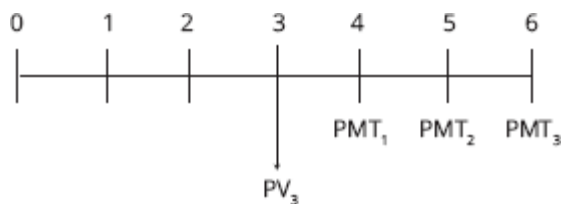
$$PV_3: FV = 200; I/Y = 10; N = 3; CPT \rightarrow PV = PV_3 = -150.26$$

$$\text{PV of cash flow stream} = \Sigma PV_{\text{individual}} = \$918.86$$

LOS 1.d: Demonstrate the use of a time line in modeling and solving time value of money problems.

In most of the PV problems we have discussed, cash flows were discounted back to the current period. In this case, the PV is said to be indexed to $t = 0$, or the time index is $t = 0$. For example, the PV of a 3-year ordinary annuity that is indexed to $t = 0$ is computed at the beginning of Year 1 ($t = 0$). Contrast this situation with another 3-year ordinary annuity that doesn't start until Year 4 and extends to Year 6. It would not be uncommon to want to know the PV of this annuity at the beginning of Year 4, in which case the time index is $t = 3$. The time line for this annuity is presented in Figure 1.3.

Figure 1.3: Indexing Time Line to Other Than $t = 0$



The following examples will illustrate how to compute I/Y , N , or PMT in annuity problems:

EXAMPLE: Computing an annuity payment needed to achieve a given FV

At an expected rate of return of 7%, how much must be deposited at the end of each year for the next 15 years to accumulate \$3,000?

Answer:

To solve this problem, enter the three relevant known values and compute PMT .

$$N = 15; I/Y = 7; FV = +\$3,000; CPT \rightarrow PMT = -\$119.38 \text{ (ignore sign)}$$

EXAMPLE: Computing a loan payment

Suppose you are considering applying for a \$2,000 loan that will be repaid with equal end-of-year payments over the next 13 years. If the annual interest rate for the loan is 6%, how much will your payments be?

Answer:

The size of the end-of-year loan payment can be determined by inputting values for the three known variables and computing PMT .

$$N = 13; I/Y = 6; PV = -2,000; CPT \rightarrow PMT = \$225.92$$

EXAMPLE: Computing the number of periods in an annuity

How many \$100 end-of-year payments are required to accumulate \$920 if the discount rate is 9%?

Answer:

The number of payments necessary can be determined by inputting the relevant data and computing N.

$$I/Y = 9\%; FV = \$920; PMT = -\$100; CPT \rightarrow N = 7 \text{ years}$$

It will take seven annual \$100 payments, compounded at 9% annually, to accrue an investment value of \$920.

**PROFESSOR'S NOTE**

Remember the sign convention. PMT and FV must have opposite signs or your calculator will issue an error message.

EXAMPLE: Computing the number of years in an ordinary annuity

Suppose you have a \$1,000 ordinary annuity earning an 8% return. How many annual end-of-year \$150 withdrawals can be made?

Answer:

The number of years in the annuity can be determined by entering the three relevant variables and computing N.

$$I/Y = 8; PMT = 150; PV = -1,000; CPT \rightarrow N = 9.9 \text{ years}$$

EXAMPLE: Computing the rate of return for an annuity

Suppose you have the opportunity to invest \$100 at the end of each of the next five years in exchange for \$600 at the end of the fifth year. What is the annual rate of return on this investment?

Answer:

The rate of return on this investment can be determined by entering the relevant data and solving for I/Y.

$$N = 5; FV = \$600; PMT = -100; CPT \rightarrow I/Y = 9.13\%$$

EXAMPLE: Computing the discount rate for an annuity

What rate of return will you earn on an ordinary annuity that requires a \$700 deposit today and promises to pay \$100 per year at the end of each of the next 10 years?

Answer:

The discount rate on this annuity is determined by entering the three known values and computing I/Y.

$$N = 10; PV = -700; PMT = 100; CPT \rightarrow I/Y = 7.07\%$$

Funding a Future Obligation

There are many TVM applications where it is necessary to determine the size of the deposit(s) that must be made over a specified period in order to meet a future liability, such as setting up a funding program for future college tuition or a retirement program. In most of these applications, the objective is to determine the size of the payment(s) or deposit(s) necessary to meet a particular monetary goal.

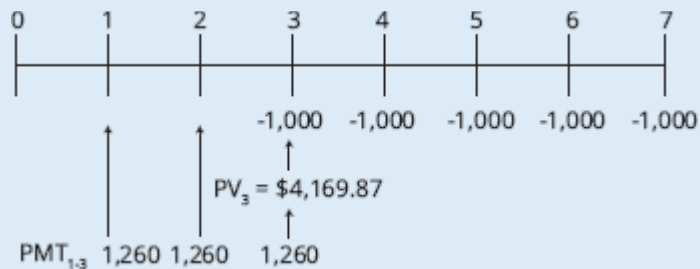
EXAMPLE: Computing the required payment to fund an annuity due

Suppose you must make five annual \$1,000 payments, the first one starting at the beginning of Year 4 (end of Year 3). To accumulate the money to make these payments, you want to make three equal payments into an investment account, the first to be made one year from today. Assuming a 10% rate of return, what is the amount of these three payments?

Answer:

The time line for this annuity problem is shown in the following figure:

Funding an Annuity Due



The first step in this type of problem is to determine the amount of money that must be available at the end of the third year in order to satisfy the payment requirements. This amount is the PV of a 5-year annuity due at the beginning of Year 4 (end of Year 3). To determine this amount, set your calculator to the BGN mode, enter the relevant data, and compute PV.

$$N = 5; I/Y = 10; PMT = -1,000; CPT \rightarrow PV = PV_3 = \$4,169.87$$

PV_3 becomes the FV that you need three years from today from your three equal end-of-year deposits. To determine the amount of the three payments necessary to meet this funding requirement, be sure that your calculator is in the END mode, input the relevant data, and compute PMT.

$$N = 3; I/Y = 10; FV = -4,169.87; CPT \rightarrow PMT = \$1,259.78$$

The second part of this problem is an ordinary annuity. If you changed your calculator to BGN mode and failed to put it back in the END mode, you will get a PMT of \$1,145, which is incorrect.

The Connection Between Present Values, Future Values, and Series of Cash Flows

As we have explained in the discussion of annuities and series of uneven cash flows, the sum of the present values of the cash flows is the present value of the series. The sum of the future values (at some future time = n) of a series of cash flows is the future value of that series of cash flows.

One interpretation of the present value of a series of cash flows is how much would have to be put in the bank today in order to make these future withdrawals and exhaust the account with the final withdrawal. Let's illustrate this with cash flows of \$100 in Year 1, \$200 in Year 2, \$300 in Year 3, and an assumed interest rate of 10%.

Calculate the present value of these three cash flows as:

$$\frac{100}{1.1} + \frac{200}{1.1^2} + \frac{300}{1.1^3} = \$481.59$$

If we put \$481.59 in an account yielding 10%, at the end of the year we would have $481.59 \times 1.1 = \$529.75$. Withdrawing \$100 would leave \$429.75.

Over the second year, the \$429.75 would grow to $429.75 \times 1.1 = \$472.73$. Withdrawing \$200 would leave \$272.73.

Over the third year, \$272.73 would grow to $272.73 \times 1.1 = \$300$, so that the last withdrawal of \$300 would empty the account.

The interpretation of the future value of a series of cash flows is straightforward as well. The FV answers the question, "How much would be in an account when the last of a series of deposits is made?" Using the same three cash flows—\$100, \$200, and \$300—and the same interest rate of 10%, we can calculate the future value of the series as:

$$100(1.1)^2 + 200(1.1) + 300 = \$641$$

This is simply the sum of the $t = 3$ value of each of the cash flows. Note that the $t = 3$ value and the $t = 0$ (present) value of the series are related by the interest rate, $481.59(1.1)^3 = 641$.

The \$100 cash flow (deposit) comes at $t = 1$, so it will earn interest of 10% compounded for two periods (until $t = 3$). The \$200 cash flow (deposit) will earn 10% between $t = 2$ and $t = 3$, and the final cash flow (deposit) of \$300 is made at $t = 3$, so \$300 is the future ($t = 3$) value of that cash flow.

We can also look at the future value in terms of how the account grows over time. At $t = 1$ we deposit \$100, so at $t = 2$ it has grown to \$110 and the \$200 deposit at $t = 2$ makes the account balance \$310. Over the next period, the \$310 grows to $310 \times 1.1 = \$341$ at $t = 3$, and the addition of the final \$300 deposit puts the account balance at \$641. This is, of course, the future value we calculated initially.



PROFESSOR'S NOTE

This last view of the future value of a series of cash flows suggests a quick way to calculate the future value of an uneven cash flow series. The process described previously for the future value of a series of end-of-period payments can be written mathematically as $[(100 \times 1.1) + 200] \times 1.1 + 300 = 641$, and this might be a quick way to do some future value problems. On your TI calculator, you would enter $100 \times 1.1 + 200 = \times 1.1 + 300 =$ to get 641.

Note that questions on the future value of an *annuity due* refer to the amount in the account one period after the last deposit is made. If the three deposits considered here were made at the beginning of each period (at $t = 0, 1, 2$) the amount in the account at the end of three years ($t = 3$) would be 10% higher (i.e., $641 \times 1.1 = \$705.10$).

The **cash flow additivity principle** refers to the fact that present value of any stream of cash flows equals the sum of the present values of the cash flows. There are different applications of this principle in time value of money problems. If we have two series of cash flows, the sum of the present values of the two series is the same as the present values of the two series taken together, adding cash flows that will be paid at the same point in time. We can also divide up a series of cash flows any way we like, and the present value of the “pieces” will equal the present value of the original series.

EXAMPLE: Additivity principle

A security will make the following payments at the end of the next four years: \$100, \$100, \$400, and \$100. Calculate the present value of these cash flows using the concept of the present value of an annuity when the appropriate discount rate is 10%.

Answer:

We can divide the cash flows so that we have:

$t = 1$	$t = 2$	$t = 3$	$t = 4$	
100	100	100	100	cash flow series #1
<u>0</u>	<u>0</u>	<u>300</u>	<u>0</u>	cash flow series #2
\$100	\$100	\$400	\$100	

The additivity principle tells us that to get the present value of the original series, we can just add the present values of series #1 (a 4-period annuity) and series #2 (a single payment three periods from now).

For the annuity: $N = 4$; $PMT = 100$; $FV = 0$; $I/Y = 10$; $CPT \rightarrow PV = -\$316.99$

For the single payment: $N = 3$; $PMT = 0$; $FV = 300$; $I/Y = 10$; $CPT \rightarrow PV = -\$225.39$

The sum of these two values is $316.99 + 225.39 = \$542.38$.

The sum of these two (present) values is identical (except for rounding) to the sum of the present values of the payments of the original series:

$$\frac{100}{1.1} + \frac{100}{1.1^2} + \frac{400}{1.1^3} + \frac{100}{1.1^4} = \$542.38$$



MODULE QUIZ 1.3

1. An analyst estimates that XYZ's earnings will grow from \$3.00 a share to \$4.50 per share over the next eight years. The rate of growth in XYZ's earnings is *closest* to:
 - A. 4.9%.
 - B. 5.2%.
 - C. 6.7%.
2. If \$5,000 is invested in a fund offering a rate of return of 12% per year, approximately how many years will it take for the investment to reach \$10,000?
 - A. 4 years.
 - B. 5 years.
 - C. 6 years.
3. An investment is expected to produce the cash flows of \$500, \$200, and \$800 at the end of the next three years. If the required rate of return is 12%, the present value of this investment is *closest* to:
 - A. \$835.
 - B. \$1,175.
 - C. \$1,235.
4. If \$10,000 is invested today in an account that earns interest at a rate of 9.5%, what is the value of the equal withdrawals that can be taken out of the account at the end of each of the next five years if the investor plans to deplete the account at the end of the time period?
 - A. \$2,453.
 - B. \$2,604.
 - C. \$2,750.
5. Given an 11% rate of return, the amount that must be put into an investment account at the end of each of the next 10 years in order to accumulate \$60,000 to pay for a child's education is *closest* to:
 - A. \$2,500.
 - B. \$3,588.
 - C. \$4,432.
6. An investor will receive an annuity of \$4,000 a year for 10 years. The first payment is to be received five years from today. At a 9% discount rate, this annuity's worth today is *closest* to:
 - A. \$16,684.
 - B. \$18,186.
 - C. \$25,671.

MODULE 1.4: COMPOUNDING FREQUENCIES

LOS 1.e: Calculate the solution for time value of money problems with different frequencies of compounding.

While the conceptual foundations of TVM calculations are not affected by the compounding period, more frequent compounding does have an impact on FV and PV computations. Specifically, since an increase in the frequency of compounding increases the effective rate of interest, it also *increases* the FV of a given cash flow and *decreases* the PV of a given cash flow.

EXAMPLE: The effect of compounding frequency on FV and PV

Compute the FV one year from now of \$1,000 today and the PV of \$1,000 to be received one year from now using a stated annual interest rate of 6% with a range of compounding periods.

Answer:

Compounding Frequency Effect

Compounding Frequency	Interest Rate per Period	Effective Annual Rate	Future Value	Present Value
Annual ($m = 1$)	6.000%	6.00%	\$1,060.00	\$943.396
Semiannual ($m = 2$)	3.000	6.090	1,060.90	942.596
Quarterly ($m = 4$)	1.500	6.136	1,061.36	942.184
Monthly ($m = 12$)	0.500	6.168	1,061.68	941.905
Daily ($m = 365$)	0.016438	6.183	1,061.83	941.769

There are two ways to use your financial calculator to compute PVs and FVs under different compounding frequencies:

1. Adjust the number of periods per year (P/Y) mode on your calculator to correspond to the compounding frequency (e.g., for quarterly, $P/Y = 4$). WE DO NOT RECOMMEND THIS APPROACH!
2. Keep the calculator in the annual compounding mode ($P/Y = 1$) and enter I/Y as the interest rate per compounding period, and N as the number of compounding periods in the investment horizon. Letting m equal the number of compounding periods per year, the basic formulas for the calculator input data are determined as follows:

$$I/Y = \text{the annual interest rate}/m$$

$$N = \text{the number of years} \times m$$

The computations for the FV and PV amounts in the previous example are:

PV_A : $FV = -1,000$; $I/Y = 6/1 = 6$; $N = 1 \times 1 = 1$:
 CPT $\rightarrow PV = PV_A = 943.396$
 PV_S : $FV = -1,000$; $I/Y = 6/2 = 3$; $N = 1 \times 2 = 2$:
 CPT $\rightarrow PV = PV_S = 942.596$
 PV_Q : $FV = -1,000$; $I/Y = 6/4 = 1.5$; $N = 1 \times 4 = 4$:
 CPT $\rightarrow PV = PV_Q = 942.184$
 PV_M : $FV = -1,000$; $I/Y = 6/12 = 0.5$; $N = 1 \times 12 = 12$:
 CPT $\rightarrow PV = PV_M = 941.905$
 PV_D : $FV = -1,000$; $I/Y = 6/365 = 0.016438$; $N = 1 \times 365 = 365$:
 CPT $\rightarrow PV = PV_D = 941.769$
 FV_A : $PV = -1,000$; $I/Y = 6/1 = 6$; $N = 1 \times 1 = 1$:
 CPT $\rightarrow FV = FV_A = 1,060.00$
 FV_S : $PV = -1,000$; $I/Y = 6/2 = 3$; $N = 1 \times 2 = 2$:
 CPT $\rightarrow FV = FV_S = 1,060.90$
 FV_Q : $PV = -1,000$; $I/Y = 6/4 = 1.5$; $N = 1 \times 4 = 4$:
 CPT $\rightarrow FV = FV_Q = 1,061.36$
 FV_M : $PV = -1,000$; $I/Y = 6/12 = 0.5$; $N = 1 \times 12 = 12$:
 CPT $\rightarrow FV = FV_M = 1,061.68$
 FV_D : $PV = -1,000$; $I/Y = 6/365 = 0.016438$; $N = 1 \times 365 = 365$:
 CPT $\rightarrow FV = FV_D = 1,061.83$

EXAMPLE: FV of a single sum using quarterly compounding

Compute the FV of \$2,000 today, five years from today using an interest rate of 12%, compounded quarterly.

Answer:

To solve this problem, enter the relevant data and compute FV:

$$N = 5 \times 4 = 20; I/Y = 12/4 = 3; PV = -\$2,000; CPT \rightarrow FV = \$3,612.22$$

EXAMPLE: Growth with quarterly compounding

John plans to invest \$2,500 in an account that will earn 8% per year with quarterly compounding. How much will be in the account at the end of two years?

Answer:

There are eight quarterly compounding periods in two years, and the effective quarterly rate is $8/4 = 2\%$. The account will grow to $2,500(1.02)^8 = \$2,929.15$. Alternatively, since the EAR is $1.02^4 - 1 = 0.082432$, we can grow the \$2,500 at 8.2432% for two years to get $2,500(1.082432)^2 = \$2,929.15$, which is the same result.

EXAMPLE: Present value with monthly compounding

Alice would like to have \$5,000 saved in an account at the end of three years. If the return on the account is 9% per year with monthly compounding, how much must Alice deposit today in order to reach her savings goal in three years?

Answer:

The effective monthly rate is $9/12 = 0.75\%$, and we can calculate the present value of \$5,000 three years (36 months) from now as $5,000/(1.0075)^{36} = \$3,820.74$. Alternatively, since the EAR is $1.0075^{12} - 1 = 0.093807$, we can calculate the present value by discounting 5,000 at the EAR for three years. $5,000/1.093807^3 = \$3,820.74$, which is the same result.

LOS 1.f: Calculate and interpret the effective annual rate, given the stated annual interest rate and the frequency of compounding.

Financial institutions usually quote rates as stated annual interest rates, along with a compounding frequency, as opposed to quoting rates as periodic rates—the rate of interest earned over a single compounding period. For example, a bank will quote a savings rate as 8%, compounded quarterly, rather than 2% per quarter. The rate of interest that investors actually realize as a result of compounding is known as the **effective annual rate (EAR)** or **effective annual yield (EAY)**. EAR represents the annual rate of return actually being earned *after adjustments have been made for different compounding periods*.

EAR may be determined as follows:

$$\text{EAR} = (1 + \text{periodic rate})^m - 1$$

where:

periodic rate = stated annual rate/ m

m = the number of compounding periods per year

Obviously, the EAR for a stated rate of 8% *compounded annually* is not the same as the EAR for 8% *compounded semiannually*, or *quarterly*. Indeed, whenever compound interest is being used, the stated rate and the actual (effective) rate of interest are equal only when interest is compounded annually. Otherwise, the greater the compounding frequency, the greater the EAR will be in comparison to the stated rate.

The computation of EAR is necessary when comparing investments that have different compounding periods. It allows for an apples-to-apples rate comparison.

EXAMPLE: Computing EAR

Compute EAR if the stated annual rate is 12%, compounded quarterly.

Answer:

Here $m = 4$, so the periodic rate is $\frac{12}{4} = 3\%$.

Thus, $\text{EAR} = (1 + 0.03)^4 - 1 = 1.1255 - 1 = 0.1255 = 12.55\%$.

This solution uses the $[y^x]$ key on your financial calculator. The exact keystrokes on the TI for the above computation are $1.03 [y^x] 4 [=]$. On the HP, the strokes are $1.03 [\text{ENTER}] 4 [y^x]$.

EXAMPLE: Computing EARs for a range of compounding frequencies

Using a stated rate of 6%, compute EARs for semiannual, quarterly, monthly, and daily compounding.

Answer:

EAR with:

$$\begin{array}{l} \text{semiannual} \\ \text{compounding} \end{array} = (1 + 0.03)^2 - 1 = 1.06090 - 1 = 0.06090 = 6.090\%$$

$$\begin{array}{l} \text{quarterly} \\ \text{compounding} \end{array} = (1 + 0.015)^4 - 1 = 1.06136 - 1 = 0.06136 = 6.136\%$$

$$\begin{array}{l} \text{monthly} \\ \text{compounding} \end{array} = (1 + 0.005)^{12} - 1 = 1.06168 - 1 = 0.06168 = 6.168\%$$

$$\begin{array}{l} \text{daily} \\ \text{compounding} \end{array} = (1 + 0.00016438)^{365} - 1 = 1.06183 - 1 = 0.06183 = 6.183\%$$

Notice here that the EAR increases as the compounding frequency increases.



PROFESSOR'S NOTE

The limit of shorter and shorter compounding periods is called continuous compounding, which we will address in a later reading.



MODULE QUIZ 1.4

1. What is the effective annual rate for a credit card that charges 18% compounded monthly?
A. 15.38%.
B. 18.81%.
C. 19.56%.
2. Given daily compounding, the growth of \$5,000 invested for one year at 12% interest will be *closest to*:
A. \$5,600.
B. \$5,628.
C. \$5,637.
3. An investor is looking at a \$150,000 home. If 20% must be put down and the balance is financed at a stated annual rate of 9% over the next 30 years, what is the monthly mortgage payment?
A. \$799.33.
B. \$895.21.
C. \$965.55.

KEY CONCEPTS

LOS 1.a

An interest rate can be interpreted as the rate of return required in equilibrium for a particular investment, the discount rate for calculating the present value of future cash flows, or as the opportunity cost of consuming now, rather than saving and investing.

LOS 1.b

The real risk-free rate is a theoretical rate on a single-period loan when there is no expectation of inflation. Nominal risk-free rate = real risk-free rate + expected inflation rate.

Securities may have several risks, and each increases the required rate of return. These include default risk, liquidity risk, and maturity risk.

The required rate of return on a security = real risk-free rate + expected inflation + default risk premium + liquidity premium + maturity risk premium.

LOS 1.c

Future value: $FV = PV(1 + I/Y)^N$

Present value: $PV = FV/(1 + I/Y)^N$

An annuity is a series of equal cash flows that occurs at evenly spaced intervals over time. Ordinary annuity cash flows occur at the end of each time period. Annuity due cash flows occur at the beginning of each time period.

Perpetuities are annuities with infinite lives (perpetual annuities):

$$PV_{\text{perpetuity}} = \frac{PMT}{I/Y}$$

The present (future) value of any series of cash flows is equal to the sum of the present (future) values of the individual cash flows.

LOS 1.d

Constructing a time line showing future cash flows will help in solving many types of TVM problems. Cash flows occur at the end of the period depicted on the time line. The end of one period is the same as the beginning of the next period. For example, a cash flow at the beginning of Year 3 appears at time $t = 2$ on the time line.

LOS 1.e

For non-annual time value of money problems, divide the stated annual interest rate by the number of compounding periods per year, m , and multiply the number of years by the number of compounding periods per year.

LOS 1.f

The effective annual rate when there are m compounding periods =

$$\left(1 + \frac{\text{stated annual rate}}{m}\right)^m - 1. \text{ Each dollar invested will grow to } \left(1 + \frac{\text{stated annual rate}}{m}\right)^m \text{ in one year.}$$

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 1.1

1. **C** Interest rates can be interpreted as required rates of return, discount rates, or opportunity costs of current consumption. A risk premium can be, but is not always, a component of an interest rate. (LOS 1.a, 1.b)
2. **A** Real interest rates are those that have been adjusted for inflation. (LOS 1.b)

Module Quiz 1.2

1. **C** $N = 15; I/Y = 9; PV = -1,000; PMT = 0; CPT \rightarrow FV = \$3,642.48$ (LOS 1.c)
2. **A** $N = 7; I/Y = 8; FV = -10,000; PMT = 0; CPT \rightarrow PV = \$5,834.90$ (LOS 1.c)
3. **A** $N = 20; I/Y = 10; PMT = -50,000; FV = 0; CPT \rightarrow PV = \$425,678.19$ (LOS 1.c)
4. **C** This is an annuity due. Switch to BGN mode: $N = 15; PMT = -8,000; I/Y = 11; FV = 0; CPT \rightarrow PV = 63,854.92$. Switch back to END mode. (LOS 1.c)
5. **B** The key to this problem is to recognize that it is a 4-year annuity due, so switch to BGN mode: $N = 4; PMT = -1,000; PV = 0; I/Y = 12; CPT \rightarrow FV = 5,352.84$. Switch back to END mode. (LOS 1.c)
6. **A** $9/0.11 = \$81.82$ (LOS 1.c)

Module Quiz 1.3

1. **B** $N = 8; PV = -3; FV = 4.50; PMT = 0; CPT \rightarrow I/Y = 5.1989$ (LOS 1.d)
2. **C** $PV = -5,000; I/Y = 12; FV = 10,000; PMT = 0; CPT \rightarrow N = 6.12$.
Note to HP 12C users: One known problem with the HP 12C is that it does not have the capability to round. In this particular question, you will come up with 7, although the correct answer is 6.1163. CFA Institute is aware of this problem, and hopefully you will not be faced with a situation on exam day where the incorrect solution from the HP is one of the answer choices. (LOS 1.d)
3. **B** Add up the present values of each single cash flow.
 $PV_1 = N = 1; FV = -500; I/Y = 12; CPT \rightarrow PV = 446.43$
 $PV_2 = N = 2; FV = -200; I/Y = 12; CPT \rightarrow PV = 159.44$
 $PV_3 = N = 3; FV = -800; I/Y = 12; CPT \rightarrow PV = 569.42$
Hence, $446.43 + 159.44 + 569.42 = \$1,175.29$. (LOS 1.d)
4. **B** $PV = -10,000; I/Y = 9.5; N = 5; FV = 0; CPT \rightarrow PMT = \$2,604.36$ (LOS 1.d)
5. **B** $N = 10; I/Y = 11; FV = -60,000; PV = 0; CPT \rightarrow PMT = \$3,588.08$ (LOS 1.d)
6. **B** Two steps: (1) Find the PV of the 10-year annuity: $N = 10; I/Y = 9; PMT = -4,000; FV = 0; CPT \rightarrow PV = 25,670.63$. This is the present value as of the end of Year 4; (2) Discount PV of the annuity back four years: $N = 4; PMT = 0; FV = -25,670.63; I/Y = 9; CPT \rightarrow PV = 18,185.72$. (LOS 1.d)

Module Quiz 1.4

1. **C** $EAR = [(1 + (0.18/12))]^{12} - 1 = 19.56\%$ (LOS 1.f)
2. **C** $N = 1 \times 365 = 365; I/Y = 12/365 = 0.0328767; PMT = 0; PV = -5,000; CPT \rightarrow FV = \$5,637.37$ (LOS 1.e)
3. **C** $N = 30 \times 12 = 360; I/Y = 9/12 = 0.75; PV = -150,000(1 - 0.2) = -120,000; FV = 0; CPT \rightarrow PMT = \965.55 (LOS 1.e)

READING 2

ORGANIZING, VISUALIZING, AND DESCRIBING DATA

EXAM FOCUS

Candidates must learn how to interpret the various types of illustrations used to describe data. The various measures of central tendency, dispersion, and risk are used throughout the CFA curriculum and are essential knowledge for candidates. The concepts of skewness, kurtosis, and correlation are also used extensively.

MODULE 2.1: ORGANIZING DATA



Video covering this content is available online.

LOS 2.a: Identify and compare data types.

The term *data* encompasses information in any form. For our use as analysts, we may classify data types from three different perspectives:

- Numerical versus categorical.
- Time series versus cross-sectional.
- Structured versus unstructured.

Numerical and Categorical Data

Numerical data, or **quantitative data**, are values that can be counted or measured. Numerical data may be discrete or continuous. **Discrete data** are countable, such as the months, days, or hours in a year. **Continuous data** can take any fractional value (e.g., the annual percentage return on an investment).



PROFESSOR'S NOTE

In our reading on Common Probability Distributions, we will use this concept to distinguish between discrete and continuous random variables.

Categorical data, or **qualitative data**, consist of labels that can be used to classify a set of data into groups. Categorical data may be nominal or ordinal.

Nominal data are labels that cannot be placed in order logically. For example, fixed-income mutual funds may be classified as corporate bond funds, municipal bond funds, international bond funds, and so on. Even if we assign numbers to the categories (such as the number 1 to a

corporate bond fund, the number 2 to a municipal bond fund, and so on), the numbers are arbitrary.

By contrast, **ordinal data** can be ranked in a logical order. Every item is assigned to one of multiple categories based on a specific characteristic, then these categories are ordered with respect to that characteristic. For example, the ranking of 1,000 small-cap growth stocks by performance may be done by assigning the number 1 to the 100 best-performing stocks, the number 2 to the next 100 best-performing stocks, and so on through the number 10 for the 100 worst-performing stocks. Based on this type of measurement, we can say a stock ranked 3 performed better than a stock ranked 4. However, we cannot conclude that the difference between a 3 and a 4 is the same as the difference between a 4 and a 5.

The key distinction between numerical data and categorical data is that we can perform mathematical operations only on numerical data.

Time Series and Cross-Sectional Data

A **time series** is a set of observations taken periodically, most often at equal intervals over time. Daily closing prices of a stock over the past year and quarterly earnings per share of a company over a five-year period are examples of time series data.

Cross-sectional data refers to a set of comparable observations all taken at one specific point in time. Today's closing prices of the 30 stocks in the Dow Jones Industrial Average and fourth-quarter earnings per share for 10 health care companies are examples of cross-sectional data.

Time series and cross-sectional data may be combined to form **panel data**. Panel data are often presented in tables. Figure 2.1 is an example of panel data for an economic indicator. In this table, each row represents cross-sectional data and each column represents time series data.

Figure 2.1: OECD Composite Leading Indicators, Year-on-Year Growth Rate

	Canada	United States	Japan	France	Germany	Italy	United Kingdom
January 2019	-1.47	-0.90	-0.36	-1.49	-1.34	-1.45	-1.41
February 2019	-1.46	-1.15	-0.45	-1.39	-1.47	-1.51	-1.40
March 2019	-1.43	-1.34	-0.51	-1.27	-1.57	-1.51	-1.36
April 2019	-1.39	-1.48	-0.58	-1.14	-1.67	-1.46	-1.31
May 2019	-1.36	-1.58	-0.67	-1.01	-1.78	-1.40	-1.24
June 2019	-1.32	-1.66	-0.75	-0.85	-1.90	-1.33	-1.12
July 2019	-1.27	-1.71	-0.83	-0.65	-2.02	-1.24	-0.96
August 2019	-1.18	-1.70	-0.91	-0.43	-2.05	-1.15	-0.75
September 2019	-1.03	-1.58	-0.97	-0.23	-1.99	-1.05	-0.49
October 2019	-0.83	-1.35	-1.01	-0.07	-1.82	-0.94	-0.18
November 2019	-0.57	-1.02	-1.00	0.05	-1.57	-0.83	0.16
December 2019	-0.27	-0.64	-0.92	0.11	-1.27	-0.70	0.48

Source: www.oecd.org

Structured and Unstructured Data

Time series, cross-sectional, and panel data are examples of **structured data**—they are organized in a defined way. *Market data*, such as security prices; *fundamental data*, such as

accounting values; and *analytical data*, such as analysts' earnings forecasts, are typically presented as structured data.

Unstructured data refers to information that is presented in a form with no defined structure. Management's commentary in company financial statements is an example of unstructured data. One way of classifying unstructured data is according to how it is generated. Data may be *generated by individuals*, such as posts on social media; *generated by business processes*, such as deposits, withdrawals, and transfers of cash; or *generated by sensors*, such as satellites or traffic cameras. While unstructured data often contains useful information, it usually must be transformed into structured data for analysis.



PROFESSOR'S NOTE

Technologies such as artificial intelligence can be used to analyze unstructured data. We address some of these in our reading on Fintech in Investment Management, in the Portfolio Management topic area.

LOS 2.b: Describe how data are organized for quantitative analysis.

Data are typically organized into arrays for analysis. A time series is an example of a **one-dimensional array** in that it represents a single variable. A key feature of a time series is that new data can be added without affecting the existing data. Sequentially ordered data are used to identify trends, cycles, and other patterns in the data that can be useful for forecasting.

The panel data in Figure 2.1 are an example of a **two-dimensional array**, or a **data table**. While data tables are not limited to this structure, organizing data sequentially with a cross section of observations for each measurement date is often useful for analysis.

LOS 2.c: Interpret frequency and related distributions.

A **frequency distribution** is a tabular presentation of statistical data that aids the analysis of large data sets. Frequency distributions summarize statistical data by assigning them to specified groups, or intervals.



PROFESSOR'S NOTE

Intervals are also known as *classes*.

The following procedure describes how to construct a frequency distribution:

Step 1: Define the intervals. The first step in building a frequency distribution is to define the intervals to which data measurements (observations) will be assigned. An interval is the set of values that an observation may take on. The range of values for each interval must have a lower and upper limit and be all-inclusive and non-overlapping. Intervals must be *mutually exclusive* so that each observation can be placed in only one interval, and the total set of intervals should cover the total range of values for the entire population. The number of intervals used is an important consideration. If too few intervals are used, the

data may be too broadly summarized and important characteristics may be lost. On the other hand, if too many intervals are used, the data may not be summarized enough.

Step 2: Tally the observations. After the intervals have been defined, the observations must be tallied or assigned to their appropriate interval.

Step 3: Count the observations. Having tallied the data set, the number of observations that are assigned to each interval must be counted. The *absolute frequency*, or simply the *frequency*, is the actual number of observations that fall within a given interval.

EXAMPLE: Constructing a frequency distribution

Use the data in Table A to construct a frequency distribution for the returns on Intelco's common stock.

Table A: Annual Returns for Intelco, Inc., Common Stock

10.4%	22.5%	11.1%	-12.4%
9.8%	17.0%	2.8%	8.4%
34.6%	-28.6%	0.6%	5.0%
-17.6%	5.6%	8.9%	40.4%
-1.0%	-4.2%	-5.2%	21.0%

Answer:

Step 1: Defining the interval. For Intelco's stock, the range of returns is 69.0% (-28.6% to 40.4%). Using a return interval of 1% would result in 69 separate intervals, which in this case is too many. So let's use eight non-overlapping intervals with a width of 10%. The lowest return intervals will be $-30\% \leq R_t < -20\%$, and the intervals will increase to $40\% \leq R_t \leq 50\%$.

Step 2: Tally the observations and count the observations within each interval. The tallies and counts of the observations are presented in Table B.

Table B: Tally and Interval Count for Returns Data

Interval	Tallies	Absolute Frequency
$-30\% \leq R_t < -20\%$	/	1
$-20\% \leq R_t < -10\%$	//	2
$-10\% \leq R_t < 0\%$	///	3
$0\% \leq R_t < 10\%$	////// //	7
$10\% \leq R_t < 20\%$	///	3
$20\% \leq R_t < 30\%$	//	2
$30\% \leq R_t < 40\%$	/	1
$40\% \leq R_t \leq 50\%$	/	1
Total		20

Tallying and counting the observations generates a frequency distribution that summarizes the pattern of annual returns on Intelco common stock. Notice that the interval with the greatest (absolute) frequency is the $(0\% \leq R_t < 10\%)$ interval, which includes seven return

observations. For any frequency distribution, the interval with the greatest frequency is referred to as the **modal interval**.

The **relative frequency** is another useful way to present data. The relative frequency is calculated by dividing the absolute frequency of each return interval by the total number of observations. Simply stated, relative frequency is the percentage of total observations falling within each interval. Continuing with our example, the relative frequencies are presented in Figure 2.2.

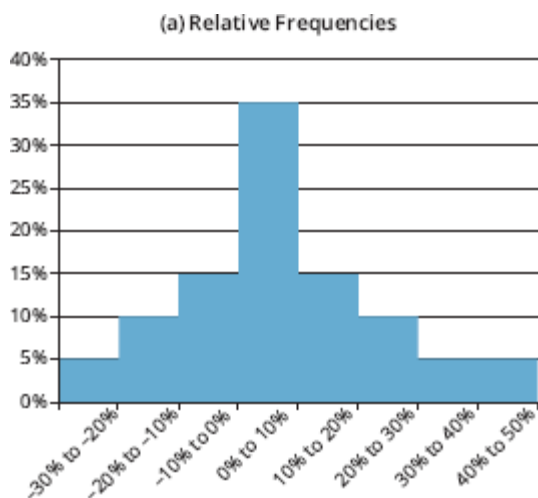
Figure 2.2: Absolute and Relative Frequencies of Intelco Returns

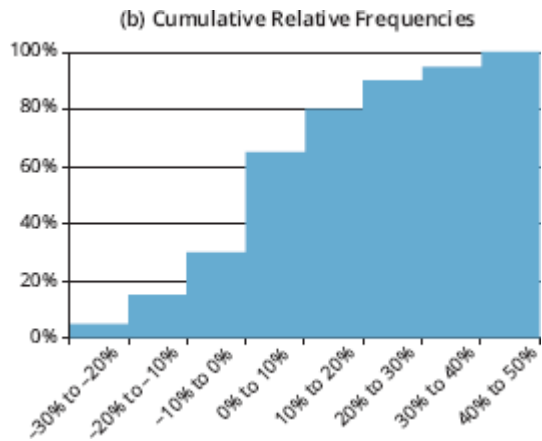
Interval	Absolute Frequency	Relative Frequency
$-30\% \leq R_t < -20\%$	1	$1/20 = 0.05$, or 5%
$-20\% \leq R_t < -10\%$	2	$2/20 = 0.10$, or 10%
$-10\% \leq R_t < 0\%$	3	$3/20 = 0.15$, or 15%
$0\% \leq R_t < 10\%$	7	$7/20 = 0.35$, or 35%
$10\% \leq R_t < 20\%$	3	$3/20 = 0.15$, or 15%
$20\% \leq R_t < 30\%$	2	$2/20 = 0.10$, or 10%
$30\% \leq R_t < 40\%$	1	$1/20 = 0.05$, or 5%
$40\% \leq R_t \leq 50\%$	1	$1/20 = 0.05$, or 5%
Total	20	100%

It is also possible to compute the **cumulative absolute frequency** and **cumulative relative frequency** by summing the absolute or relative frequencies starting at the lowest interval and progressing through the highest. The relative and cumulative relative frequencies for the Intelco stock returns example are presented in Figure 2.3.

The cumulative absolute frequency or cumulative relative frequency for any given interval is the sum of the absolute or relative frequencies up to and including the given interval. For example, the cumulative absolute frequency for $R_t < 10\%$ is $13 = 1 + 2 + 3 + 7$ and the cumulative relative frequency over this range is $5\% + 10\% + 15\% + 35\% = 65\%$.

Figure 2.3: Relative and Cumulative Frequencies of Intelco Returns





LOS 2.d: Interpret a contingency table.

A **contingency table** is a two-dimensional array with which we can analyze two variables at the same time. The rows represent attributes of one of the variables and the columns represent attributes of the other variable. These attributes can be defined using nominal or ordinal data, but there must be a finite number of them.

The data in each cell show the frequency with which we observe two attributes simultaneously. These are known as **joint frequencies** and they can be absolute or relative frequencies. The total of frequencies for a row or a column is termed the **marginal frequency** for that attribute.

For example, Figure 2.4 displays the number of traffic accidents in one year on weekdays at four intersections of a highway. From this table, we can see that accidents occur most frequently on Mondays and Fridays (marginal frequencies of 19 and 18) and that the Front Street intersection has the most accidents (marginal frequency of 25). The Front Street intersection on Mondays experiences the greatest number of accidents (joint frequency of 7).

Figure 2.4: Accidents by Intersection and Day of Week

Intersection	Monday	Tuesday	Wednesday	Thursday	Friday	Total
Palace Street	5	2	1	2	4	14
National Drive	3	2	3	1	3	12
Front Street	7	5	4	3	6	25
Jay Street	4	3	2	3	5	17
Total	19	12	10	9	18	68

One kind of contingency table is a 2-by-2 array called a **confusion matrix**. For each of two possible outcomes, a confusion matrix displays the number of occurrences predicted and the number actually observed.

Figure 2.5 illustrates a confusion matrix for a model that predicted the outcomes of 1,000 events. We can read from this table that the model predicted the outcome would occur 662 times, and it actually occurred 681 times. On 28 occasions, the model predicted the event would occur but it did not, while on 47 occasions, the model predicted the event would not occur when it actually did occur.

Figure 2.5: Confusion Matrix

	Actual Yes	Actual No	
Predicted Yes	634	28	662
Predicted No	<u>47</u>	<u>291</u>	<u>338</u>
	681	319	1,000

Another use of a contingency table is to use the values to determine whether two variables (characteristics), such as firm size and risk, are independent based on a chi-square test statistic.



PROFESSOR'S NOTE

In our reading on Hypothesis Testing, we address statistical tests, including the chi-square test for independence.



MODULE QUIZ 2.1

- To perform meaningful mathematical analysis, an analyst must use data that are:
 - discrete.
 - numerical.
 - continuous.
- Which of the following types of data would *most likely* be organized as a two-dimensional array?
 - Panel.
 - Time series.
 - Cross sectional.
- The intervals in a frequency distribution should always be:
 - truncated.
 - open-ended.
 - non-overlapping.
- Consider the following contingency table from a political opinion poll:

	Supports Johnson	Supports Williams	Total
Supports Smith	42%	14%	56%
Supports Jones	10%	34%	44%
Total	52%	48%	100%

In this table, the value 34% represents:

- a joint frequency.
- a marginal frequency.
- an absolute frequency.

MODULE 2.2: VISUALIZING DATA



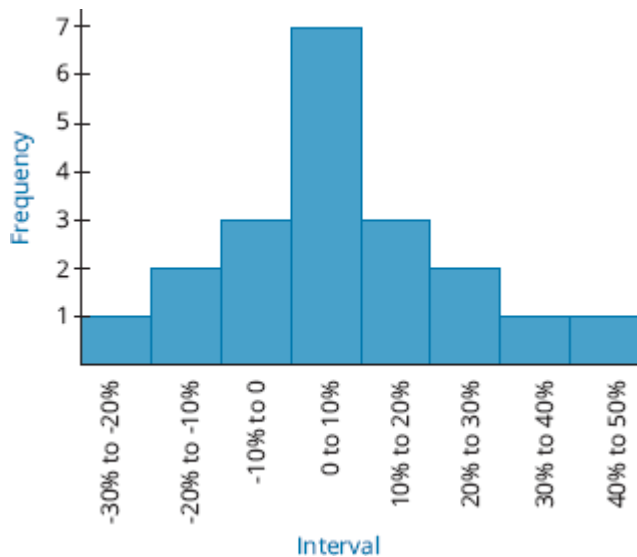
LOS 2.e: Describe ways that data may be visualized and evaluate uses of specific visualizations.

Video covering this content is available online.

A **histogram** is the graphical presentation of the absolute frequency distribution. A histogram is simply a bar chart of continuous data that has been classified into a frequency distribution. The attractive feature of a histogram is that it allows us to quickly see where most of the observations are concentrated.

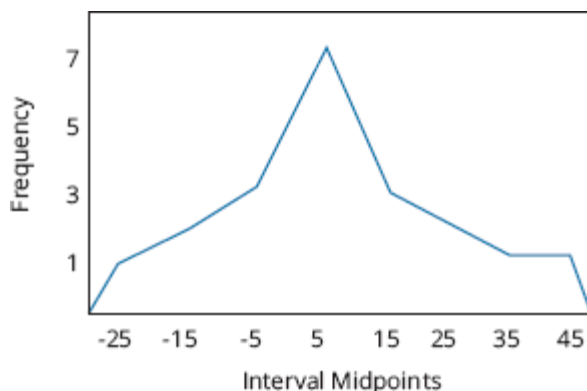
To construct a histogram, the intervals are shown on the horizontal axis and the absolute frequencies are shown on the vertical axis. The histogram for the Intelco returns data from the example presented earlier is provided in Figure 2.6.

Figure 2.6: Histogram of Intelco Stock Return Data



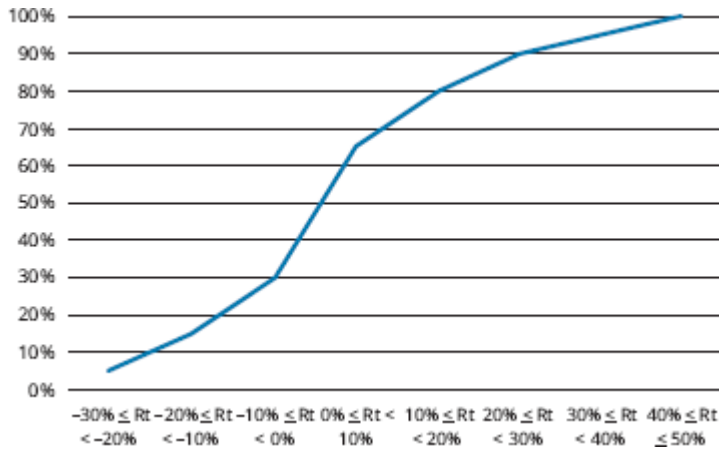
To construct a **frequency polygon**, successive frequencies at the midpoints of the intervals are joined with line segments. A frequency polygon for the Intelco returns data presented previously is illustrated in Figure 2.7.

Figure 2.7: Frequency Polygon of Intelco Stock Return Data



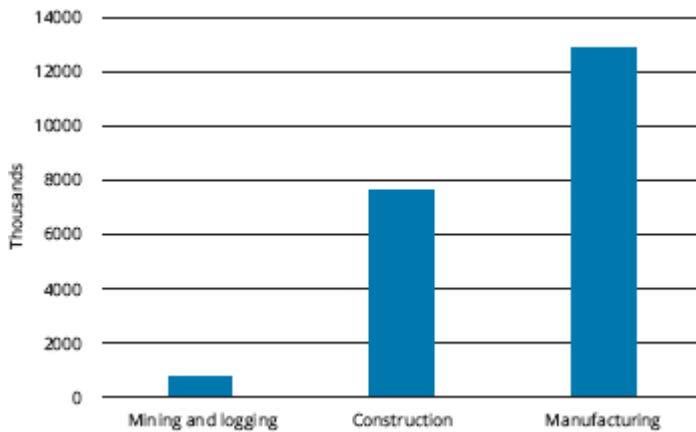
A **cumulative frequency distribution chart** displays either the cumulative absolute frequency or the cumulative relative frequency. Earlier, we showed the cumulative relative frequencies for Intelco as columns. They can also be displayed in a line chart, as in Figure 2.8.

Figure 2.8: Cumulative Relative Frequency Distribution



The histogram shown earlier is an example of a **bar chart**. In general, bar charts are used to illustrate relative sizes, degrees, or magnitudes. The bars can be displayed vertically or horizontally. Figure 2.9 shows a bar chart of employment in goods-producing industry groups in the United States. From this chart, we can see that the construction industries employ about 10 times as many people as the mining and logging industries and that manufacturing payrolls are a bit less than twice as large as construction payrolls.

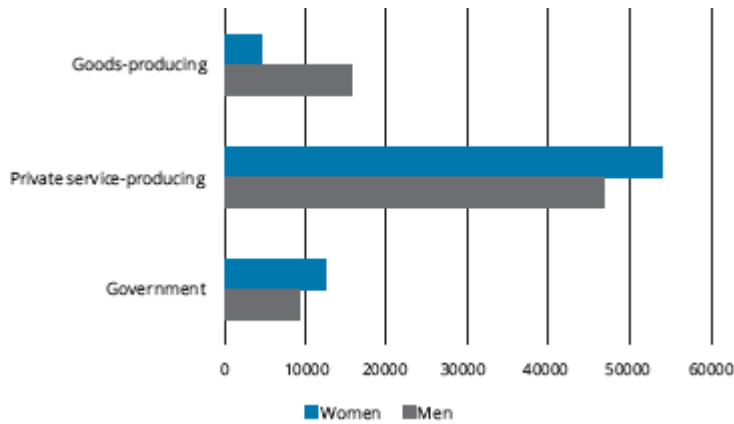
Figure 2.9: Employees on Payrolls in Goods-Producing Industries, January 2020



Source: Bureau of Labor Statistics, stats.bls.gov

A **grouped bar chart** or **clustered bar chart** can illustrate two categories at once, much like a data table. Figure 2.10 displays the number of men and women employed in three segments of the U.S. economy. Here we can see that more men than women are employed in the goods-producing industries, but more women than men are employed in the service-producing industries and government.

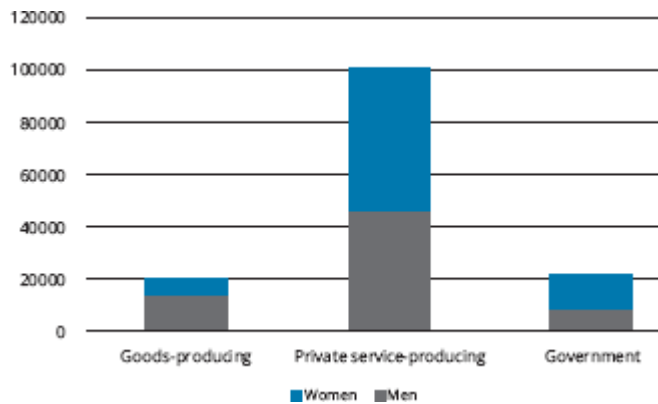
Figure 2.10: Grouped Bar Chart of Employment by Sector, December 2020



Source: Bureau of Labor Statistics, stats.bls.gov, with government payrolls estimated as the difference between total service-producing and private service-producing.

Another way to present two categories at once is with a **stacked bar chart**, as shown in Figure 2.11. In a stacked bar chart, the height of each bar represents the cumulative frequency for a category (such as goods-producing industries) and the colors within each bar represent joint frequencies (such as women employed in government). From this stacked bar chart, we can see the size of the private service-producing sector relative to the other two sectors.

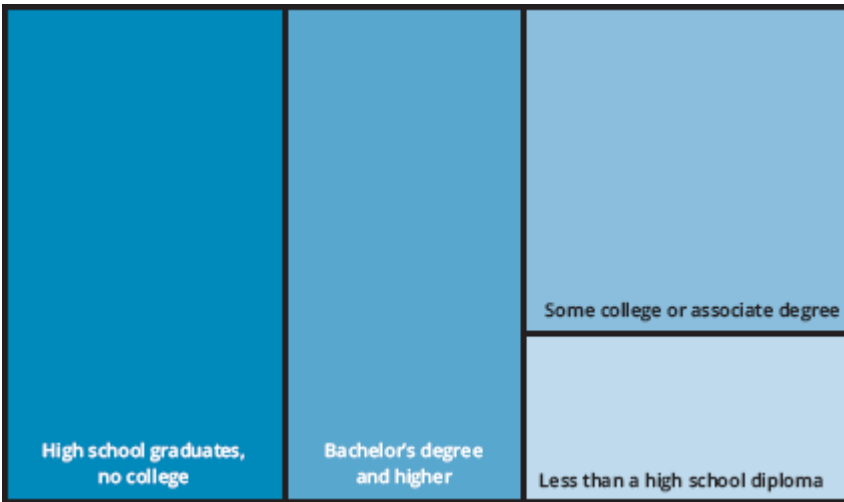
Figure 2.11: Stacked Bar Chart of Employment by Sector, December 2020



Source: Bureau of Labor Statistics, stats.bls.gov, with government payrolls estimated as the difference between total service-producing and private service-producing.

A **tree map** is another method for visualizing the relative sizes of categories. Figure 2.12 is a tree map of labor force categories for the United States by level of education. The filled areas of a tree map may each be divided into subcategories that are displayed in different colors or shades.

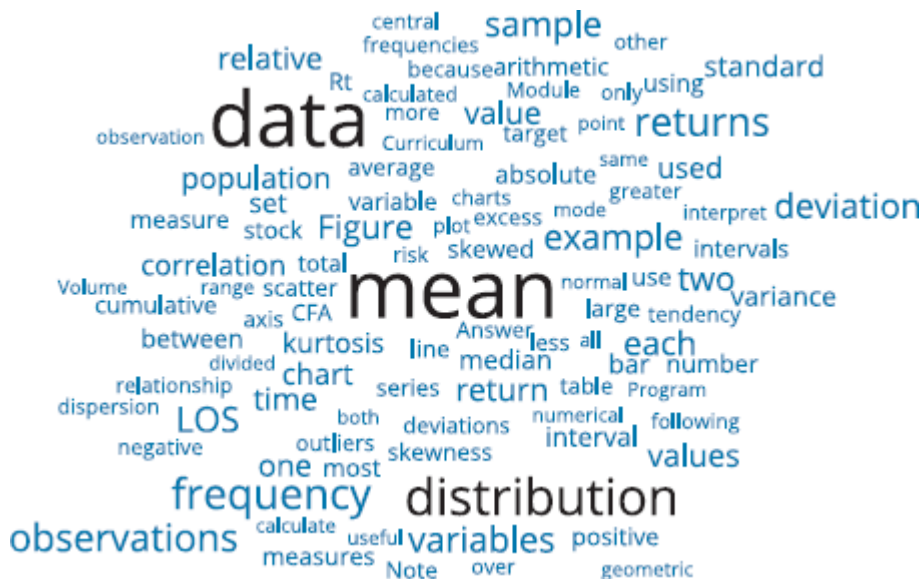
Figure 2.12: Tree Map of Labor Force by Educational Attainment, January 2020



Source: Bureau of Labor Statistics, stats.bls.gov

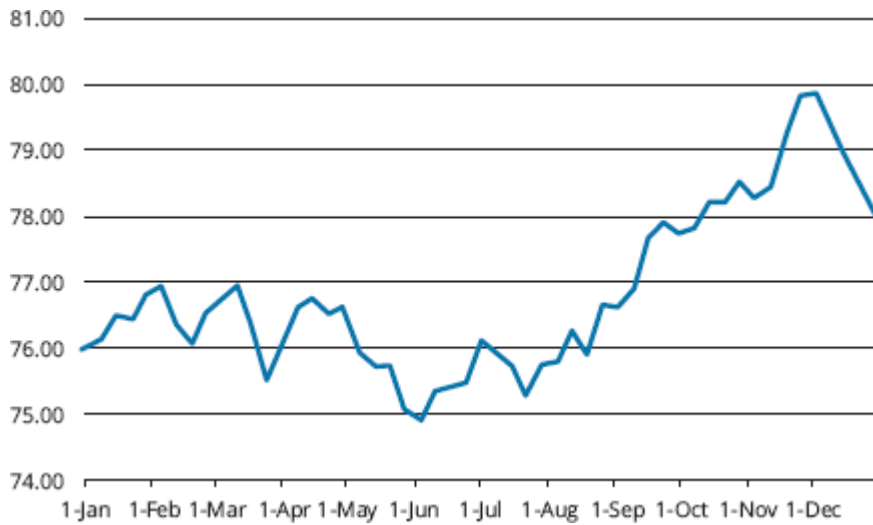
When analyzing text, a useful visualization technique is a **word cloud**. A word cloud is generated by counting the uses of specific words in text data. It displays frequently occurring words, in type sizes that are scaled to the frequency of their use. Figure 2.13 is an example of a word cloud generated from this reading. From this word cloud, we can easily see two of the major concepts this reading addresses: types of data and definitions of the mean.

Figure 2.13: Word Cloud



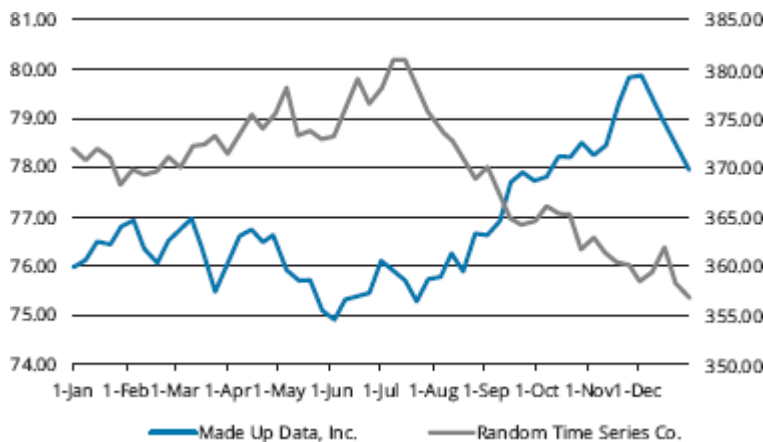
We have already seen some examples of **line charts**. Line charts are particularly useful for illustrating time series data, such as securities prices. Figure 2.14 is a line chart of weekly closing prices for a hypothetical stock.

Figure 2.14: Line Chart



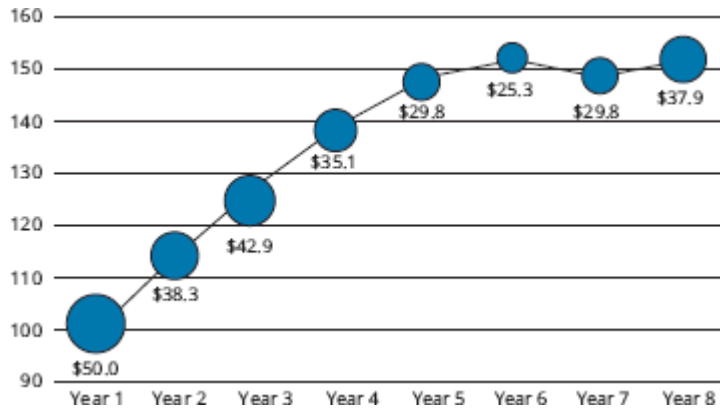
Multiple time series can be displayed on a line chart if their scales are comparable. It is also possible to display two time series on a line chart if their scales are different, by using left and right vertical axes as shown in Figure 2.15. This is one way of showing changes in two variables over time relative to each other.

Figure 2.15: Dual-Scale Line Chart



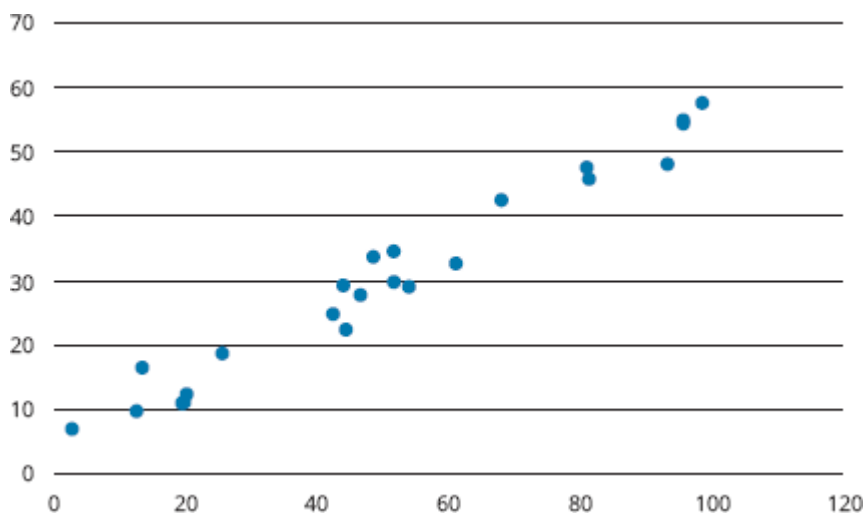
Another technique for adding a dimension to a line chart is to create a **bubble line chart**. For example, in Figure 2.16, a line chart shows total revenues for a company, and at each point, the different-sized bubbles represent revenues per salesperson. From this chart, we can see that revenues per salesperson were declining in years 1 to 6 while total revenues were increasing, but revenues per salesperson started increasing again in years 7 and 8, even though revenues were flat. This suggests the company was adding salespeople during the growth period but has been reducing its sales force recently.

Figure 2.16: Bubble Line Chart



A **scatter plot** is a way of displaying how two variables tend to change in relation to each other. The vertical axis represents one value of a variable and the horizontal axis represents the value of a second variable. Each point in the scatter plot shows the values of both variables at a point in time. Figure 2.17 is a scatter plot for two variables that have a fairly strong positive linear relationship.

Figure 2.17: Scatter Plot



While the relationship in the previous figure appears to be linear, scatter plots can also be useful for identifying nonlinear relationships that are not apparent when using a measure of the strength of a linear relationship, such as the correlation coefficient.

To analyze three variables at the same time, an analyst can create a **scatter plot matrix** that consists of three scatter plots of these variables, each presenting two of the three variables.

A **heat map** uses color and shade to display data frequency. Figure 2.18 is a heat map that uses data from the contingency table example we used previously to examine traffic accidents at several highway intersections. The darker shades indicate more accidents.

Figure 2.18: Heat Map

Intersection	Monday	Tuesday	Wednesday	Thursday	Friday
Palace Street	5	2	1	2	4
National Drive	3	2	3	1	3
Front Street	7	5	4	3	6
Jay Street	4	3	2	3	5

LOS 2.f: Describe how to select among visualization types.

Given the variety of charts to choose from, it can be useful to have a framework for choosing which to use in a specific circumstance. In general, we want to use the simplest chart that will clearly communicate the information to be presented.

We may need a chart to illustrate a relationship between two or more variables, compare two or more variables, or show the distribution of a single variable. Typically the most effective chart types for these purposes are as follows:

- **Relationships.** Scatter plots, scatter plot matrices, and heat maps.
- **Comparisons.** Bar charts, tree maps, and heat maps for comparisons among categories; line charts, dual-scale line charts, and bubble line charts for comparisons over time.
- **Distributions.** Histograms, frequency polygons, and cumulative distribution charts for numerical data; bar charts, tree maps, and heat maps for categorical data; and word clouds for text data.

When creating any chart, an analyst must take care to avoid misrepresentations. Selecting a chart type that is effective for visualizing the underlying data is a good first step. Beyond that, we must avoid potentially misleading practices, such as showing only a time period that supports our analysis while leaving out periods that illustrate the opposite, or choosing the scale of the axes so as to obscure meaningful variations or exaggerate non-meaningful variations in the data.



PROFESSOR'S NOTE

As we will see in our review of Ethical and Professional Standards, presenting selective data to mislead investors is a violation of Standard I(C) Misrepresentation.



MODULE QUIZ 2.2

1. The vertical axis of a histogram shows:
 - A. the frequency with which observations occur.
 - B. the range of observations within each interval.
 - C. the intervals into which the observations are arranged.
2. In which type of bar chart does the height or length of a bar represent the cumulative frequency for its category?
 - A. Stacked bar chart.
 - B. Grouped bar chart.
 - C. Clustered bar chart.

3. An analyst who wants to illustrate the relationships among three variables should *most appropriately* construct:
- A. a bubble line chart.
 - B. a scatter plot matrix.
 - C. a frequency polygon.

MODULE 2.3: MEASURES OF CENTRAL TENDENCY



Video covering this content is available online.

LOS 2.g: Calculate and interpret measures of central tendency.

Measures of central tendency identify the center, or average, of a data set. This central point can then be used to represent the typical, or expected, value in the data set.

To compute the **population mean**, all the observed values in the population are summed (ΣX) and divided by the number of observations in the population, N . Note that the population mean is unique in that a given population only has one mean. The population mean is expressed as:

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

The **sample mean** is the sum of all the values in a sample of a population, ΣX , divided by the number of observations in the sample, n . It is used to make *inferences* about the population mean. The sample mean is expressed as:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Note the use of n , the sample size, versus N , the population size.

EXAMPLE: Population mean and sample mean

You have calculated the stock returns for AXZ Corporation over the last five years as 25%, 34%, 19%, 54%, and 17%. Given this information, estimate the mean of the distribution of returns.

Answer:

The sample mean can be used as an estimate of the mean of the distribution:

$$\bar{X} = \text{sample mean} = \frac{25 + 34 + 19 + 54 + 17}{5} = 29.8\%$$

The population mean and sample mean are both examples of **arithmetic means**. The arithmetic mean is the sum of the observation values divided by the number of observations. It is the most widely used measure of central tendency and has the following properties:

- All interval and ratio data sets have an arithmetic mean.
- All data values are considered and included in the arithmetic mean computation.
- A data set has only one arithmetic mean (i.e., the arithmetic mean is unique).

- The sum of the deviations of each observation in the data set from the mean is always zero. The arithmetic mean is the only measure of central tendency for which the sum of the deviations from the mean is zero. Mathematically, this property can be expressed as follows:

$$\text{sum of mean deviations} = \sum_{i=1}^n (X_i - \bar{X}) = 0$$

Unusually large or small values, **outliers** can have a disproportionate influence on the arithmetic mean. The mean of 1, 2, 3, and 50 is 14 and is not a good indication of what the individual data values really are. On the positive side, the arithmetic mean uses all the information available about the observations. The arithmetic mean of a sample from a population is the best estimate of both the true mean of the sample and of the value of a single future observation.

In some cases, a researcher may decide that outliers should be excluded from a measure of central tendency. One technique for doing so is to use a **trimmed mean**. A trimmed mean excludes a stated percentage of the most extreme observations. A 1% trimmed mean, for example, would discard the lowest 0.5% and the highest 0.5% of the observations.

Another technique is to use a **winsorized mean**. Instead of discarding the highest and lowest observations, we substitute a value for them. To calculate a 90% winsorized mean, for example, we would determine the 5th and 95th percentile of the observations, substitute the 5th percentile for any values lower than that, substitute the 95th percentile for any values higher than that, and then calculate the mean of the revised data set.



PROFESSOR'S NOTE

Percentiles are explained later in this reading.

The computation of a **weighted mean** (or **weighted average**) recognizes that different observations may have a disproportionate influence on the mean. The weighted mean of a set of numbers is computed with the following equation:

$$\bar{X}_w = \sum_{i=1}^n w_i X_i = (w_1 X_1 + w_2 X_2 + \dots + w_n X_n)$$

where:

X_1, X_2, \dots, X_n = observed values

w_1, w_2, \dots, w_n = corresponding weights associated with each of the observations
such that $\sum w_i = 1$

EXAMPLE: Weighted mean as a portfolio return

A portfolio consists of 50% common stocks, 40% bonds, and 10% cash. If the return on common stocks is 12%, the return on bonds is 7%, and the return on cash is 3%, what is the portfolio return?

Answer:

$$\bar{X}_w = w_{\text{stock}} R_{\text{stock}} + w_{\text{bonds}} R_{\text{bonds}} + w_{\text{cash}} R_{\text{cash}}$$

$$\bar{X}_w = (0.50 \times 0.12) + (0.40 \times 0.07) + (0.10 \times 0.03) = 0.091, \text{ or } 9.1\%$$

The example illustrates an extremely important investments concept: *the return for a portfolio is the weighted average of the returns of the individual assets in the portfolio*. Asset weights are market weights, the market value of each asset relative to the market value of the entire portfolio.

The **median** is the midpoint of a data set when the data is arranged in ascending or descending order. Half the observations lie above the median and half are below. To determine the median, arrange the data from the highest to the lowest value, or lowest to highest value, and find the middle observation.

The median is important because the arithmetic mean can be affected by extremely large or small values (outliers). When this occurs, the median is a better measure of central tendency than the mean because it is not affected by extreme values that may actually be the result of errors in the data.

EXAMPLE: The median using an odd number of observations

What is the median return for five portfolio managers with a 10-year annualized total returns record of 30%, 15%, 25%, 21%, and 23%?

Answer:

First, arrange the returns in descending order.

30%, 25%, 23%, 21%, 15%

Then, select the observation that has an equal number of observations above and below it—the one in the middle. For the given data set, the third observation, 23%, is the median value.

EXAMPLE: The median using an even number of observations

Suppose we add a sixth manager to the previous example with a return of 28%. What is the median return?

Answer:

Arranging the returns in descending order gives us:

30%, 28%, 25%, 23%, 21%, 15%

With an even number of observations, there is no single middle value. The median value in this case is the arithmetic mean of the two middle observations, 25% and 23%. Thus, the median return for the six managers is $24.0\% = 0.5(25 + 23)$.

Consider that while we calculated the mean of 1, 2, 3, and 50 as 14, the median is 2.5. If the data were 1, 2, 3, and 4 instead, the arithmetic mean and median would both be 2.5.

The **mode** is the value that occurs most frequently in a data set. A data set may have more than one mode or even no mode. When a distribution has one value that appears most frequently, it is said to be **unimodal**. When a set of data has two or three values that occur most frequently, it is said to be **bimodal** or **trimodal**, respectively.

EXAMPLE: The mode

What is the mode of the following data set?

Data set: [30%, 28%, 25%, 23%, 28%, 15%, 5%]

Answer:

The mode is 28% because it is the value appearing most frequently.

The **geometric mean** is often used when calculating investment returns over multiple periods or when measuring compound growth rates. The general formula for the geometric mean, G , is as follows:

$$G = \sqrt[n]{X_1 \times X_2 \times \dots \times X_n} = (X_1 \times X_2 \times \dots \times X_n)^{1/n}$$

Note that this equation has a solution only if the product under the radical sign is nonnegative.

When calculating the geometric mean for a returns data set, it is necessary to add 1 to each value under the radical and then subtract 1 from the result. The geometric mean return (R_G) can be computed using the following equation:

$$1 + R_G = \sqrt[n]{(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_n)}$$

where:

R_t = the return for period t

EXAMPLE: Geometric mean return

For the last three years, the returns for Acme Corporation common stock have been -9.34%, 23.45%, and 8.92%. Compute the compound annual rate of return over the three-year period.

Answer:

$$1 + R_G = \sqrt[3]{(1 - 0.0934) \times (1 + 0.2345) \times (1 + 0.0892)}$$

$$1 + R_G = \sqrt[3]{0.9066 \times 1.2345 \times 1.0892} = \sqrt[3]{1.21903}$$

$$= (1.21903)^{1/3} = 1.06825$$

$$R_G = 1.06825 - 1 = 6.825\%$$

Solve this type of problem with your calculator as follows:

- On the TI, enter 1.21903 [y^x] 3 [$1/x$] [=]
- On the HP, enter 1.21903 [ENTER] 3 [$1/x$] [y^x]



PROFESSOR'S NOTE

The geometric mean is always less than or equal to the arithmetic mean, and the difference increases as the dispersion of the observations increases. The only time the arithmetic and geometric means are equal is when there is no variability in the observations (i.e., all observations are equal).

A **harmonic mean** is used for certain computations, such as the average cost of shares purchased over time. The harmonic mean is calculated as $\frac{N}{\sum_{i=1}^N \frac{1}{X_i}}$, where there are N values of X_i .

EXAMPLE: Calculating average cost with the harmonic mean

An investor purchases \$1,000 of mutual fund shares each month, and over the last three months, the prices paid per share were \$8, \$9, and \$10. What is the average cost per share?

Answer:

$$\bar{X}_H = \frac{3}{\frac{1}{8} + \frac{1}{9} + \frac{1}{10}} = \$8.926 \text{ per share}$$

To check this result, calculate the total shares purchased as:

$$\frac{1,000}{8} + \frac{1,000}{9} + \frac{1,000}{10} = 336.11 \text{ shares}$$

The average price is $\frac{\$3,000}{336.11} = \8.926 per share.

The previous example illustrates the interpretation of the harmonic mean in its most common application. Note that the average price paid per share (\$8.93) is less than the arithmetic average of the share prices, $\frac{8+9+10}{3} = 9$.

For values that are not all equal, harmonic mean < geometric mean < arithmetic mean. This mathematical fact is the basis for the claimed benefit of purchasing the same dollar amount of mutual fund shares each month or each week. Some refer to this practice as cost averaging.

LOS 2.h: Evaluate alternative definitions of mean to address an investment problem.

Appropriate uses for the various definitions of the mean are as follows:

- **Arithmetic mean.** Estimate the next observation, expected value of a distribution.
- **Geometric mean.** Compound rate of returns over multiple periods.
- **Trimmed mean.** Estimate the mean without the effects of a given percentage of outliers.
- **Winsorized mean.** Decrease the effect of outliers on the mean.
- **Harmonic mean.** Calculate the average share cost from periodic purchases in a fixed dollar amount.



MODULE QUIZ 2.3

1. XYZ Corp. Annual Stock Returns

2015	2016	2017	2018	2019	2020
22%	5%	-7%	11%	2%	11%

What is the arithmetic mean return for XYZ stock?

- A. 7.3%.

- B. 8.0%.
- C. 11.0%.

2. XYZ Corp. Annual Stock Returns

2015	2016	2017	2018	2019	2020
22%	5%	-7%	11%	2%	11%

What is the median return for XYZ stock?

- A. 7.3%.
 - B. 8.0%.
 - C. 11.0%.
3. A data set has 100 observations. Which of the following measures of central tendency will be calculated using a denominator of 100?
- A. The winsorized mean, but not the trimmed mean.
 - B. Both the trimmed mean and the winsorized mean.
 - C. Neither the trimmed mean nor the winsorized mean.
4. The harmonic mean of 3, 4, and 5 is:
- A. 3.74.
 - B. 3.83.
 - C. 4.12.

5. XYZ Corp. Annual Stock Returns

2015	2016	2017	2018	2019	2020
22%	5%	-7%	11%	2%	11%

The mean annual return on XYZ stock is *most appropriately* calculated using:

- A. the harmonic mean.
- B. the arithmetic mean.
- C. the geometric mean.

MODULE 2.4: MEASURES OF LOCATION AND DISPERSION



Video covering this content is available online.

LOS 2.i: Calculate quantiles and interpret related visualizations.

Quantile is the general term for a value at or below which a stated proportion of the data in a distribution lies. Examples of quantiles include the following:

- **Quartile.** The distribution is divided into quarters.
- **Quintile.** The distribution is divided into fifths.
- **Decile.** The distribution is divided into tenths.
- **Percentile.** The distribution is divided into hundredths (percents).

Note that any quantile may be expressed as a percentile. For example, the third quartile partitions the distribution at a value such that three-fourths, or 75%, of the observations fall below that value. Thus, the third quartile is the 75th percentile. The difference between the third quartile and the first quartile (25th percentile) is known as the **interquartile range**.

The formula for the position of the observation at a given percentile, y , with n data points sorted in ascending order is:

$$L_y = (n + 1) \frac{y}{100}$$

Quantiles and measures of central tendency are known collectively as **measures of location**.

EXAMPLE: Quartiles

What is the third quartile for the following distribution of returns?

8%, 10%, 12%, 13%, 15%, 17%, 17%, 18%, 19%, 23%

Answer:

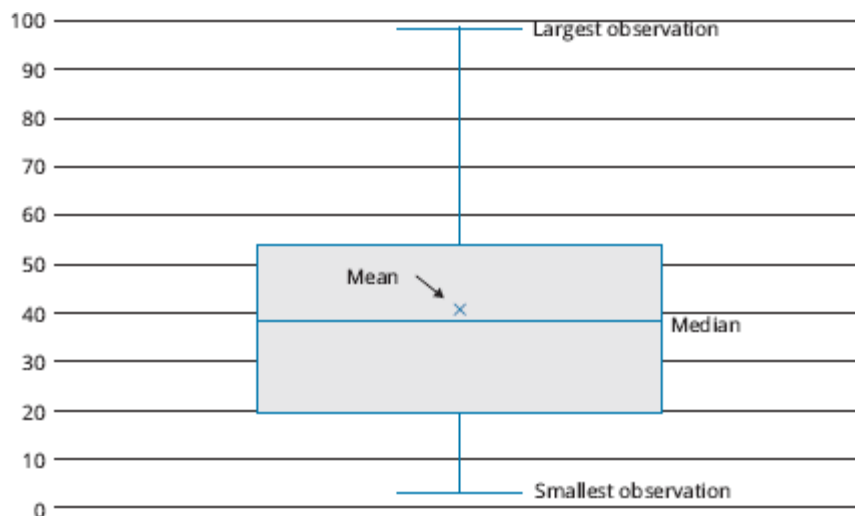
The third quartile is the point below which 75% of the observations lie. Recognizing that there are 10 observations in the data set, the third quartile can be identified as:

$$L_y = (10 + 1) \times \frac{75}{100} = 8.25$$

When the data are arranged in ascending order, the third quartile is a fourth (0.25) of the way from the eighth data point (18%) to the ninth data point (19%), or 18.25%. This means that 75% of all observations lie below 18.25%.

To visualize a data set based on quantiles, we can create a **box and whisker plot**, as shown in Figure 2.19. In a box and whisker plot, the box represents the central portion of the data, such as the interquartile range. The vertical line represents the entire range. In Figure 2.19, we can see that the largest observation is farther away from the center than the smallest observation is. This suggests that the data might include one or more outliers on the high side.

Figure 2.19: Box and Whisker Plot



LOS 2.j: Calculate and interpret measures of dispersion.

Dispersion is defined as the *variability around the central tendency*. The common theme in finance and investments is the tradeoff between reward and variability, where the central tendency is the measure of the reward and dispersion is a measure of risk.

The **range** is a relatively simple measure of variability, but when used with other measures, it provides extremely useful information. The range is the distance between the largest and the smallest value in the data set, or:

$$\text{range} = \text{maximum value} - \text{minimum value}$$

EXAMPLE: The range

What is the range for the 5-year annualized total returns for five investment managers if the managers' individual returns were 30%, 12%, 25%, 20%, and 23%?

Answer:

$$\text{range} = 30 - 12 = 18\%$$

The **mean absolute deviation (MAD)** is the average of the absolute values of the deviations of individual observations from the arithmetic mean:

$$\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

The computation of the MAD uses the absolute values of each deviation from the mean because the sum of the actual deviations from the arithmetic mean is zero.

EXAMPLE: MAD

What is the MAD of the investment returns for the five managers discussed in the preceding example? How is it interpreted?

Answer:

annualized returns: [30%, 12%, 25%, 20%, 23%]

$$\bar{X} = \frac{[30 + 12 + 25 + 20 + 23]}{5} = 22\%$$

$$\text{MAD} = \frac{[|30 - 22| + |12 - 22| + |25 - 22| + |20 - 22| + |23 - 22|]}{5}$$

$$\text{MAD} = \frac{[8 + 10 + 3 + 2 + 1]}{5} = 4.8\%$$

This result can be interpreted to mean that, on average, an individual return will deviate $\pm 4.8\%$ from the mean return of 22%.

The **sample variance**, s^2 , is the measure of dispersion that applies when we are evaluating a sample of n observations from a population. The sample variance is calculated using the following formula:

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

The denominator for s^2 is $n - 1$, one less than the sample size n . Based on the mathematical theory behind statistical procedures, the use of the entire number of sample observations, n , instead of $n - 1$ as the divisor in the computation of s^2 , will systematically *underestimate* the population variance, particularly for small sample sizes. This systematic underestimation causes the sample variance to be a **biased estimator** of the population variance. Using $n - 1$ instead of n in the denominator, however, improves the statistical properties of s^2 as an estimator of the population variance.

EXAMPLE: Sample variance

Assume that the 5-year annualized total returns for the five investment managers used in the preceding examples represent only a sample of the managers at a large investment firm. What is the sample variance of these returns?

Answer:

$$\bar{X} = \frac{[30 + 12 + 25 + 20 + 23]}{5} = 22\%$$

$$s^2 = \frac{[(30 - 22)^2 + (12 - 22)^2 + (25 - 22)^2 + (20 - 22)^2 + (23 - 22)^2]}{5 - 1} = 44.5(\%^2)$$

Thus, the sample variance of $44.5(\%^2)$ can be interpreted to be an unbiased estimator of the population variance. Note that 44.5 “percent squared” is 0.00445 and you will get this value if you put the percentage returns in decimal form [e.g., $(0.30 - 0.22)^2$].

A major problem with using variance is the difficulty of interpreting it. The computed variance, unlike the mean, is in terms of squared units of measurement. How does one interpret squared percentages, squared dollars, or squared yen? This problem is mitigated through the use of the *standard deviation*. The units of standard deviation are the same as the units of the data (e.g., percentage return, dollars, euros). The **sample standard deviation** is the square root of the sample variance. The sample standard deviation, s , is calculated as:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

EXAMPLE: Sample standard deviation

Compute the sample standard deviation based on the result of the preceding example.

Answer:

Because the sample variance for the preceding example was computed to be $44.5(\%^2)$, the sample standard deviation is:

$$s = [44.5(\%^2)]^{1/2} = 6.67\%, \text{ or } \sqrt{0.00445} = 0.0667$$

The results shown here mean that the sample standard deviation, $s = 6.67\%$, can be interpreted as an unbiased estimator of the population standard deviation, σ .

A direct comparison between two or more measures of dispersion may be difficult. For instance, suppose you are comparing the annual returns distribution for retail stocks with a mean of 8% and an annual returns distribution for a real estate portfolio with a mean of 16%. A direct comparison between the dispersion of the two distributions is not meaningful because of the relatively large difference in their means. To make a meaningful comparison, a relative measure of dispersion must be used. **Relative dispersion** is the amount of variability in a distribution relative to a reference point or benchmark. Relative dispersion is commonly measured with the **coefficient of variation (CV)**, which is computed as:

$$CV = \frac{s_x}{\bar{X}} = \frac{\text{standard deviation of } x}{\text{average value of } x}$$

CV measures the amount of dispersion in a distribution relative to the distribution's mean. It is useful because it enables us to make a direct comparison of dispersion across different sets of data. In an investments setting, the CV is used to measure the risk (variability) per unit of expected return (mean). A lower CV is better.

EXAMPLE: Coefficient of variation

You have just been presented with a report that indicates that the mean monthly return on T-bills is 0.25% with a standard deviation of 0.36%, and the mean monthly return for the S&P 500 is 1.09% with a standard deviation of 7.30%. Your unit manager has asked you to compute the CV for these two investments and to interpret your results.

Answer:

$$CV_{\text{T-bills}} = \frac{0.36}{0.25} = 1.44$$
$$CV_{\text{S\&P 500}} = \frac{7.30}{1.09} = 6.70$$

These results indicate that there is less dispersion (risk) per unit of monthly return for T-bills than for the S&P 500 (1.44 versus 6.70).



PROFESSOR'S NOTE

To remember the formula for CV, remember that the coefficient of variation is a measure of variation, so standard deviation goes in the numerator. CV is variation per unit of return.

LOS 2.k: Calculate and interpret target downside deviation.

When we use variance or standard deviation as risk measures, we calculate risk based on outcomes both above and below the mean. In some situations, it may be more appropriate to consider only outcomes less than the mean (or some other specific value) in calculating a risk measure. In this case, we are measuring **downside risk**.

One measure of downside risk is **target downside deviation**, which is also known as **target semideviation**. Calculating target downside deviation is similar to calculating standard deviation, but in this case, we choose a target value against which to measure each outcome and only include deviations from the target value in our calculation if the outcomes are below that target.

The formula for target downside deviation is stated as:

$$s_{\text{target}} = \sqrt{\frac{\sum_{\text{all } X_i < B} (X_i - B)^2}{n - 1}}, \text{ where } B \text{ is the target. Note that the denominator}$$

remains the sample size n minus one, even though we are not using all the observations in the numerator.

EXAMPLE: Target downside deviation

Calculate the target downside deviation based on the data in the preceding examples, for a target return equal to the mean (22%), and for a target return of 24%.

Answer:

Return	Deviation From Mean	Deviation From Target Return
30%	30% - 22% = 8%	30% - 24% = 6%
12%	12% - 22% = -10%	12% - 24% = -12%
25%	25% - 22% = 3%	25% - 24% = 1%
20%	20% - 22% = -2%	20% - 24% = -4%
23%	23% - 22% = 1%	23% - 24% = -1%

$$s_{22\%} = \sqrt{\frac{(-10)^2 + (-2)^2}{5 - 1}} = 5.10\%$$

$$s_{24\%} = \sqrt{\frac{(-12)^2 + (-4)^2 + (-1)^2}{5 - 1}} = 6.34\%$$



MODULE QUIZ 2.4

1. Given the following observations:

2, 4, 5, 6, 7, 9, 10, 11

The 65th percentile is *closest* to:

- A. 5.85.
- B. 6.55.
- C. 8.70.

2. XYZ Corp. Annual Stock Returns

20x1	20x2	20x3	20x4	20x5	20x6
22%	5%	-7%	11%	2%	11%

What is the sample standard deviation?

- A. 9.8%.
- B. 72.4%.
- C. 96.3%.

3. XYZ Corp. Annual Stock Returns

20x1	20x2	20x3	20x4	20x5	20x6
22%	5%	-7%	11%	2%	11%

Assume an investor has a target return of 11% for XYZ stock. What is the stock's target downside deviation?

- A. 9.4%.
- B. 12.1%.
- C. 14.8%.

MODULE 2.5: SKEWNESS, KURTOSIS, AND CORRELATION



Video covering this content is available online.

LOS 2.I: Interpret skewness.

A distribution is **symmetrical** if it is shaped identically on both sides of its mean. Distributional symmetry implies that intervals of losses and gains will exhibit the same frequency. For example, a symmetrical distribution with a mean return of zero will have losses in the -6% to -4% interval as frequently as it will have gains in the $+4\%$ to $+6\%$ interval. The extent to which a returns distribution is symmetrical is important because the degree of symmetry tells analysts if deviations from the mean are more likely to be positive or negative.

Skewness, or skew, refers to the extent to which a distribution is not symmetrical. Nonsymmetrical distributions may be either positively or negatively skewed and result from the occurrence of outliers in the data set. **Outliers** are observations extraordinarily far from the mean, either above or below:

- A *positively skewed* distribution is characterized by outliers greater than the mean (in the upper region, or right tail). A positively skewed distribution is said to be skewed right because of its relatively long upper (right) tail.
- A *negatively skewed* distribution has a disproportionately large amount of outliers less than the mean that fall within its lower (left) tail. A negatively skewed distribution is said to be skewed left because of its long lower tail.

Skewness affects the location of the mean, median, and mode of a distribution:

- For a symmetrical distribution, the mean, median, and mode are equal.
- For a positively skewed, unimodal distribution, the mode is less than the median, which is less than the mean. The mean is affected by outliers; in a positively skewed distribution, there are large, positive outliers, which will tend to pull the mean upward, or more positive. An example of a positively skewed distribution is that of housing prices. Suppose you live in a neighborhood with 100 homes; 99 of them sell for \$100,000 and one sells for \$1,000,000. The median and the mode will be \$100,000, but the mean will be \$109,000. Hence, the mean has been pulled upward (to the right) by the existence of one home (outlier) in the neighborhood.
- For a negatively skewed, unimodal distribution, the mean is less than the median, which is less than the mode. In this case, there are large, negative outliers that tend to pull the mean downward (to the left).



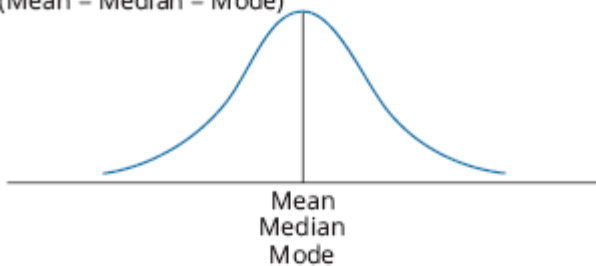
PROFESSOR'S NOTE

The key to remembering how measures of central tendency are affected by skewed data is to recognize that skew affects the mean more than the median and mode, and the mean is pulled in the direction of the skew. The relative location of the mean, median, and mode for different distribution shapes is shown in Figure 2.20. Note the median is between the other two measures for positively or negatively skewed distributions.

Figure 2.20: Effect of Skewness on Mean, Median, and Mode

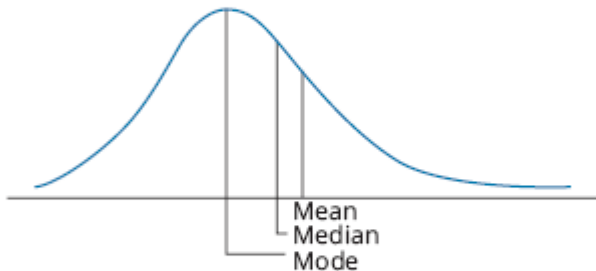
Symmetrical

(Mean = Median = Mode)



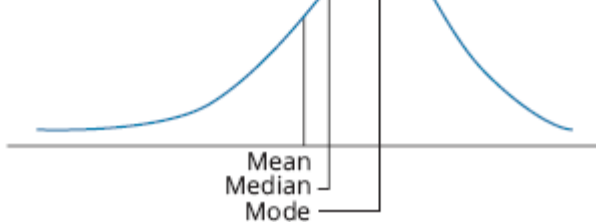
Positive (right) skew

(Mean > Median > Mode)



Negative (left) skew

(Mean < Median < Mode)



Sample skewness is equal to the sum of the cubed deviations from the mean divided by the cubed standard deviation and by the number of observations. Sample skewness for large samples is computed as:

$$\text{sample skewness} = \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{s^3}$$

where:

s = sample standard deviation

Note that the denominator is always positive but that the numerator can be positive or negative depending on whether observations above the mean or observations below the mean tend to be farther from the mean on average. When a distribution is right skewed, sample skewness is positive because the deviations above the mean are larger on average. A left-skewed distribution has a negative sample skewness.

Dividing by standard deviation cubed standardizes the statistic and allows **interpretation of the skewness measure**. If relative skewness is equal to zero, the data is not skewed. Positive levels of relative skewness imply a positively skewed distribution, whereas negative values of relative skewness imply a negatively skewed distribution. Values of sample skewness in excess of 0.5 in absolute value are considered significant.



PROFESSOR'S NOTE

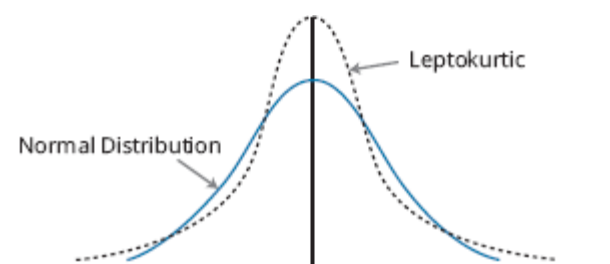
The LOS related to skewness and kurtosis require us to “interpret” these measures, but not to calculate them.

LOS 2.m: Interpret kurtosis.

Kurtosis is a measure of the degree to which a distribution is more or less peaked than a normal distribution. **Leptokurtic** describes a distribution that is more peaked than a normal distribution, whereas **platykurtic** refers to a distribution that is less peaked, or flatter than a normal distribution. A distribution is **mesokurtic** if it has the same kurtosis as a normal distribution.

As indicated in Figure 2.21, a leptokurtic return distribution will have more returns clustered around the mean and more returns with large deviations from the mean (fatter tails). Relative to a normal distribution, a leptokurtic distribution will have a greater percentage of small deviations from the mean and a greater percentage of extremely large deviations from the mean. This means that there is a relatively greater probability of an observed value being either close to the mean or far from the mean. With regard to an investment returns distribution, a greater likelihood of a large deviation from the mean return is often perceived as an increase in risk.

Figure 2.21: Kurtosis



A distribution is said to exhibit **excess kurtosis** if it has either more or less kurtosis than the normal distribution. The computed kurtosis for all normal distributions is three. Statisticians, however, sometimes report excess kurtosis, which is defined as kurtosis minus three. Thus, a normal distribution has excess kurtosis equal to zero, a leptokurtic distribution has excess kurtosis greater than zero, and platykurtic distributions will have excess kurtosis less than zero.

Kurtosis is critical in a risk management setting. Most research about the distribution of securities returns has shown that returns are not normally distributed. Actual securities returns tend to exhibit both skewness and kurtosis. Skewness and kurtosis are critical concepts for risk management because when securities returns are modeled using an assumed normal distribution, the predictions from the models will not take into account the potential for extremely large, negative outcomes. In fact, most risk managers put very little emphasis on the mean and standard deviation of a distribution and focus more on the distribution of returns in the tails of the distribution—that is where the risk is. In general, greater positive kurtosis and more negative skew in returns distributions indicates increased risk.

Sample kurtosis is measured using deviations raised to the *fourth power*:

$$\text{sample kurtosis} = \frac{1}{n} \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{s^4}$$

where:

s = sample standard deviation

To interpret kurtosis, note that it is measured relative to the kurtosis of a normal distribution, which is 3. Positive values of excess kurtosis indicate a distribution that is leptokurtic (more peaked, fat tails), whereas negative values indicate a platykurtic distribution (less peaked, thin tails). We can calculate kurtosis relative to that of a normal distribution as:

$$\text{excess kurtosis} = \text{sample kurtosis} - 3$$

LOS 2.n: Interpret correlation between two variables.

Covariance is a measure of how two variables move together. The calculation of the **sample covariance** is based on the following formula:

$$s_{X,Y} = \frac{\sum_{i=1}^n \{ [X_i - \bar{X}] [Y_i - \bar{Y}] \}}{n - 1}$$

where:

X_i = an observation of variable X

Y_i = an observation of variable Y

\bar{X} = mean of variable X

\bar{Y} = mean of variable Y

n = number of periods

In practice, the covariance is difficult to interpret. The value of covariance depends on the units of the variables. The covariance of daily price changes of two securities priced in yen will be much greater than the covariance they will be if the securities are priced in dollars. Like the variance, the units of covariance are the square of the units used for the data.

Additionally, we cannot interpret the relative strength of the relationship between two variables. Knowing that the covariance of X and Y is 0.8756 tells us only that they tend to move together because the covariance is positive. A standardized measure of the linear relationship between two variables is called the **correlation coefficient**, or simply correlation. The correlation between two variables, X and Y , is calculated as:

$$\rho_{XY} = \frac{s_{XY}}{s_X s_Y} \text{ which implies,}$$

$$s_{XY} = \rho_{XY} s_X s_Y$$

The properties of the correlation of two random variables, X and Y , are summarized here:

- Correlation measures the strength of the linear relationship between two random variables.
- Correlation has no units.
- The correlation ranges from -1 to $+1$. That is, $-1 \leq \rho_{XY} \leq +1$.
- If $\rho^{XY} = 1.0$, the random variables have perfect positive correlation. This means that a movement in one random variable results in a proportional positive movement in the other relative to its mean.
- If $\rho^{XY} = -1.0$, the random variables have perfect negative correlation. This means that a movement in one random variable results in an exact opposite proportional movement in the other relative to its mean.
- If $\rho^{XY} = 0$, there is no linear relationship between the variables, indicating that prediction of Y cannot be made on the basis of X using linear methods.

EXAMPLE: Correlation

The variance of returns on stock A is 0.0028, the variance of returns on stock B is 0.0124, and their covariance of returns is 0.0058. Calculate and interpret the correlation of the returns for stocks A and B.

Answer:

First, it is necessary to convert the variances to standard deviations:

$$s_A = (0.0028)^{1/2} = 0.0529$$

$$s_B = (0.0124)^{1/2} = 0.1114$$

Now, the correlation between the returns of stock A and stock B can be computed as follows:

$$\rho_{AB} = \frac{0.0058}{(0.0529)(0.1114)} = 0.9842$$

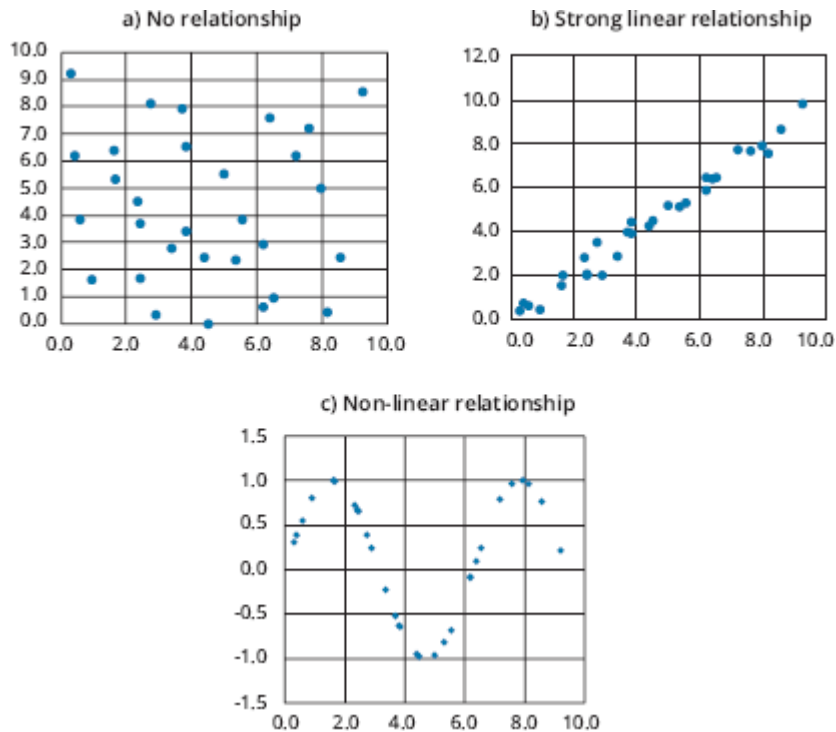
The fact that this value is close to $+1$ indicates that the linear relationship is not only positive but very strong.

Scatter plots are a method for displaying the relationship between two variables. With one variable on the vertical axis and the other on the horizontal axis, their paired observations can each be plotted as a single point. For example, in panel a of Figure 2.22, the point farthest to the upper right shows that when one of the variables (on the horizontal axis) equaled 9.2, the other variable (on the vertical axis) equaled 8.5.

The scatter plot in panel a is typical of two variables that have no clear relationship. Panel b shows two variables that have a strong linear relationship—that is, a high correlation coefficient.

A key advantage of creating scatter plots is that they can reveal nonlinear relationships, which are not described by the correlation coefficient. Panel c illustrates such a relationship. Although the correlation coefficient for these two variables is close to zero, their scatter plot shows clearly that they are related in a predictable way.

Figure 2.22: Scatter plots



Care should be taken when drawing conclusions based on correlation. Causation is not implied just from significant correlation. Even if it were, which variable is causing change in the other is not revealed by correlation. It is more prudent to say that two variables exhibit positive (or negative) association, suggesting that the nature of any causal relationship is to be separately investigated or based on theory that can be subject to additional tests.

One question that can be investigated is the role of outliers (extreme values) in the correlation of two variables. If removing the outliers significantly reduces the calculated correlation, further inquiry is necessary into whether the outliers provide information or are caused by noise (randomness) in the data used.

Spurious correlation refers to correlation that is either the result of chance or present due to changes in both variables over time that is caused by their association with a third variable. For example, we can find instances where two variables that are both related to the inflation rate exhibit significant correlation but for which causation in either direction is not present.

In his book *Spurious Correlation*¹, Tyler Vigen presents the following examples. The correlation between the age of each year's Miss America and the number of films Nicholas Cage appeared in that year is 87%. This seems a bit random. The correlation between the U.S. spending on science, space, and technology and suicides by hanging, strangulation, and suffocation over the 1999–2009 period is 99.87%. Impressive correlation, but both variables increased in an approximately linear fashion over the period.



MODULE QUIZ 2.5

1. Which of the following is *most accurate* regarding a distribution of returns that has a mean greater than its median?
 - A. It is positively skewed.
 - B. It is a symmetric distribution.
 - C. It has positive excess kurtosis.
2. A distribution of returns that has a greater percentage of small deviations from the mean and a greater percentage of extremely large deviations from the mean compared with a normal distribution:
 - A. is positively skewed.
 - B. has positive excess kurtosis.
 - C. has negative excess kurtosis.
3. The correlation between two variables is +0.25. The *most appropriate* way to interpret this value is to say:
 - A. a scatter plot of the two variables is likely to show a strong linear relationship.
 - B. when one variable is above its mean, the other variable tends to be above its mean as well.
 - C. a change in one of the variables usually causes the other variable to change in the same direction.

KEY CONCEPTS

LOS 2.a

We may classify data types from three different perspectives: numerical versus categorical, time series versus cross sectional, and structured versus unstructured.

Numerical, or quantitative, data are values that can be counted or measured and may be discrete or continuous. Categorical, or qualitative, data are labels that can be used to classify a set of data into groups and may be nominal or ordinal.

A time series is a set of observations taken at a sequence of points in time. Cross-sectional data are a set of comparable observations taken at one point in time. Time series and cross-sectional data may be combined to form panel data.

Unstructured data refers to information that is presented in forms that are not regularly structured and may be generated by individuals, business processes, or sensors.

LOS 2.b

Data are typically organized into arrays for analysis. A time series is an example of a one-dimensional array. A data table is an example of a two-dimensional array.

LOS 2.c

A frequency distribution groups observations into classes, or intervals. An interval is a range of values.

Relative frequency is the percentage of total observations falling within an interval.

Cumulative relative frequency for an interval is the sum of the relative frequencies for all values less than or equal to that interval's maximum value.

LOS 2.d

A contingency table is a two-dimensional array with which we can analyze two variables at the same time. The rows represent some attributes of one of the variables and the columns represent those attributes for the other variable. The data in each cell show the joint frequency with which we observe a pair of attributes simultaneously. The total of frequencies for a row or a column is the marginal frequency for that attribute.

LOS 2.e

A histogram is a bar chart of data that has been grouped into a frequency distribution.

A frequency polygon plots the midpoint of each interval on the horizontal axis and the absolute frequency for that interval on the vertical axis, and it connects the midpoints with straight lines.

A cumulative frequency distribution chart is a line chart of the cumulative absolute frequency or the cumulative relative frequency.

Bar charts can be used to illustrate relative sizes, degrees, or magnitudes. A grouped or clustered bar chart can illustrate two categories at once. In a stacked bar chart, the height of each bar represents the cumulative frequency for a category, and the colors within each bar represent joint frequencies. A tree map is another method for visualizing the relative sizes of categories.

A word cloud is generated by counting the uses of specific words in a text file. It displays the words that appear most often, in type sizes that are scaled to the frequency of their use.

Line charts are particularly useful for exhibiting time series. Multiple time series can be displayed on a line chart if their scales are comparable. It is also possible to display two time series on a line chart if their scales are different by using left and right vertical axes. A technique for adding a dimension to a line chart is to create a bubble line chart.

A scatter plot is a way of displaying how two variables tend to change together. The vertical axis represents one variable and the horizontal axis represents a second variable. Each point in the scatter plot shows the values of both variables at one specific point in time.

A heat map uses color and shade to display data frequency.

LOS 2.f

Which chart types tend to be most effective depends on what they are intended to visualize:

- **Relationships.** Scatter plots, scatter plot matrices, and heat maps.
- **Comparisons.** Bar charts, tree maps, and heat maps for comparisons among categories; line charts and bubble line charts for comparisons over time.
- **Distributions.** Histograms, frequency polygons, and cumulative distribution charts for numerical data; bar charts, tree maps, and heat maps for categorical data; and word clouds for unstructured data.

LOS 2.g

The arithmetic mean is the average:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Population mean and sample mean are examples of arithmetic means.

The geometric mean is used to find a compound growth rate:

$$G = \sqrt[n]{X_1 \times X_2 \times \dots \times X_n}$$

The weighted mean weights each value according to its influence:

$$\bar{X}_w = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$$

The harmonic mean can be used to find an average purchase price, such as dollars per share for equal periodic investments:

$$\bar{X}_H = \frac{N}{\sum_{i=1}^N \frac{1}{X_i}}$$

The median is the midpoint of a data set when the data are arranged from largest to smallest.

The mode of a data set is the value that occurs most frequently.

LOS 2.h

Arithmetic mean is used to estimate expected value, value of a single outcome from a distribution.

Geometric mean is used calculate or estimate periodic compound returns over multiple periods.

Harmonic mean is used to calculate the average price paid with equal periodic investments.

A trimmed mean omits outliers and a winsorized mean replaces outliers with given values, reducing the effect of outliers on the mean in both cases.

LOS 2.i

Quantile is the general term for a value at or below which a stated proportion of the data in a distribution lies. Examples of quantiles include the following:

- Quartile. The distribution is divided into quarters.
- Quintile. The distribution is divided into fifths.
- Decile. The distribution is divided into tenths.
- Percentile. The distribution is divided into hundredths (percents).

LOS 2.j

The range is the difference between the largest and smallest values in a data set.

Mean absolute deviation (MAD) is the average of the absolute values of the deviations from the arithmetic mean:

$$MAD = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

Variance is defined as the mean of the squared deviations from the arithmetic mean or from the expected value of a distribution:

- Sample variance $= s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{n - 1}$, where \bar{X} = sample mean and n = sample size.

Standard deviation is the positive square root of the variance and is frequently used as a quantitative measure of risk.

The coefficient of variation for sample data, $CV = \frac{s}{\bar{X}}$, is the ratio of the standard deviation of the sample to its mean (expected value of the underlying distribution).

LOS 2.k

Target downside deviation or semideviation is a measure of downside risk. Calculating target downside deviation is similar to calculating standard deviation, but in this case, we choose a target against which to measure each outcome and only include outcomes below that target when calculating the numerator.

The formula for target downside deviation is:

$$s_{\text{target}} = \sqrt{\frac{\sum_{\text{all } X_i < B} (X_i - B)^2}{n - 1}}, \text{ where } B \text{ is the target value.}$$

LOS 2.1

Skewness describes the degree to which a distribution is not symmetric about its mean. A right-skewed distribution has positive skewness. A left-skewed distribution has negative skewness.

For a positively skewed, unimodal distribution, the mean is greater than the median, which is greater than the mode.

For a negatively skewed, unimodal distribution, the mean is less than the median, which is less than the mode.

LOS 2.m

Kurtosis measures the peakedness of a distribution and the probability of extreme outcomes (thickness of tails):

- Excess kurtosis is measured relative to a normal distribution, which has a kurtosis of 3.
- Positive values of excess kurtosis indicate a distribution that is leptokurtic (fat tails, more peaked), so the probability of extreme outcomes is greater than for a normal distribution.
- Negative values of excess kurtosis indicate a platykurtic distribution (thin tails, less peaked).

LOS 2.n

Correlation is a standardized measure of association between two random variables. It ranges in value from -1 to $+1$ and is equal to $\frac{\text{Cov}_{A,B}}{\sigma_A \sigma_B}$.

Scatterplots are useful for revealing nonlinear relationships that are not measured by correlation.

Correlation does not imply that changes in one variable cause changes in the other. Spurious correlation may result by chance or from the relationships of two variables to a third variable.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 2.1

- B** We can perform mathematical operations on numerical data but not on categorical data. Numerical data can be discrete or continuous. (LOS 2.a)
- A** Panel data combine time series data with cross-sectional data and are typically organized as data tables, which are two-dimensional arrays. (LOS 2.a,b)
- C** Intervals within a frequency distribution should always be non-overlapping and closed-ended so that each data value can be placed into only one interval. Interval widths should be defined so that data are adequately summarized without losing valuable characteristics. (LOS 2.c)
- A** The value 34% is the joint probability that a voter supports both Jones and Williams. Because it is stated as a percentage, this value is a relative frequency. The totals for each row and column are marginal frequencies. An absolute frequency is a number of occurrences, not a percentage of occurrences. (LOS 2.d)

Module Quiz 2.2

- A** In a histogram, the intervals are on the horizontal axis and the frequency is on the vertical axis. (LOS 2.e)
- A** In a stacked bar chart, the height or length of a bar represents the cumulative frequency of a category. In a grouped or clustered bar chart, each category is displayed with bars side by side that together represent the cumulative frequency. (LOS 2.e)
- B** With a scatter plot matrix, an analyst can visualize the relationships among three variables by organizing scatter plots of the relationships between each pair of variables. Bubble line charts are typically used to visualize two variables over time. Frequency polygons are best used to visualize distributions. (LOS 2.f)

Module Quiz 2.3

- A** $[22\% + 5\% + -7\% + 11\% + 2\% + 11\%] / 6 = 7.3\%$ (LOS 2.g)
- B** To find the median, rank the returns in order and take the middle value: -7%, 2%, 5%, 11%, 11%, 22%. In this case, because there is an even number of observations, the median is the average of the two middle values, or $(5\% + 11\%) / 2 = 8.0\%$. (LOS 2.g)
- A** The winsorized mean substitutes a value for some of the largest and smallest observations. The trimmed mean removes some of the largest and smallest observations. (LOS 2.g)
- B**
$$\bar{X}_H = \frac{3}{\frac{1}{3} + \frac{1}{4} + \frac{1}{5}} = 3.83$$
 (LOS 2.g)
- C** Because returns are compounded, the geometric mean is appropriate.
 $[(1.22)(1.05)(0.93)(1.11)(1.02)(1.11)]^{1/6} - 1 = 6.96\%$
 (LOS 2.h)

Module Quiz 2.4

- C** With eight observations, the location of the 65th percentile is:

$$(8 + 1) \times 65/100 = 5.85 \text{ observations}$$

The fifth observation is 7 and the sixth observation is 9, so the value at 5.85 observations is $7 + 0.85(9 - 7) = 8.7$. (LOS 2.i)

- A** The sample standard deviation is the square root of the sample variance:

$$s = \sqrt{\frac{(22 - 7.3)^2 + (5 - 7.3)^2 + (-7 - 7.3)^2 + (11 - 7.3)^2 + (2 - 7.3)^2 + (11 - 7.3)^2}{6 - 1}}$$

$$= \sqrt{96.3} = 9.8\%$$

(LOS 2.j)

3. **A** Deviations from the target return:

$$22\% - 11\% = 11\%$$

$$5\% - 11\% = -6\%$$

$$-7\% - 11\% = -18\%$$

$$11\% - 11\% = 0\%$$

$$2\% - 11\% = -9\%$$

$$11\% - 11\% = 0\%$$

$$\text{Target downside deviation} = \sqrt{\frac{(-6)^2 + (-18)^2 + (-9)^2}{6 - 1}} = \sqrt{88.2} = 9.39\%$$

(LOS 2.k)

Module Quiz 2.5

1. **A** A distribution with a mean greater than its median is positively skewed, or skewed to the right. The skew pulls the mean. Kurtosis deals with the overall shape of a distribution, not its skewness. (LOS 2.l)
2. **B** A distribution that has a greater percentage of small deviations from the mean and a greater percentage of extremely large deviations from the mean will be leptokurtic and will exhibit excess kurtosis (positive). The distribution will be more peaked and have fatter tails than a normal distribution. (LOS 2.m)
3. **B** Correlation of +0.25 indicates a positive linear relationship between the variables—one tends to be above its mean when the other is above its mean. The value 0.25 indicates that the linear relationship is not particularly strong. Correlation does not imply causation. (LOS 2.n)

¹ Tyler Vigen, "Spurious Correlations," www.tylervigen.com

READING 3

PROBABILITY CONCEPTS

EXAM FOCUS

This reading covers important terms and concepts associated with probability theory. We describe random variables, events, outcomes, conditional probability, and joint probability, and introduce probability rules such as the addition rule and multiplication rule. Finance practitioners use these rules frequently. We also discuss expected value, standard deviation, covariance, and correlation for individual asset and portfolio returns. A well-prepared candidate will be able to calculate and interpret these widely used measures. This review also discusses counting rules, which lay the foundation for the binomial probability distribution that is covered in the next reading.

MODULE 3.1: CONDITIONAL AND JOINT PROBABILITIES



Video covering this content is available online.

LOS 3.a: Define a random variable, an outcome, and an event.

LOS 3.b: Identify the two defining properties of probability, including mutually exclusive and exhaustive events, and compare and contrast empirical, subjective, and a priori probabilities.

- A **random variable** is an uncertain quantity/number.
- An **outcome** is an observed value of a random variable.
- An **event** is a single outcome or a set of outcomes.
- **Mutually exclusive events** are events that cannot both happen at the same time.
- **Exhaustive events** are those that include all possible outcomes.

Consider rolling a 6-sided die one time. The number that comes up is a *random variable*. If you roll a 4, that is an *outcome*. Rolling a 4 is also an *event*, as is rolling an even number. The possible outcomes from 1 to 6 are *mutually exclusive* (you cannot get a 3 and a 5 on the same roll) and *exhaustive* (you cannot roll a 7 or a 0).

There are **two defining properties of probability**:

- The probability of occurrence of any event (E_i) is between 0 and 1 (i.e., $0 \leq P(E_i) \leq 1$).
- If a set of events, E_1, E_2, \dots, E_n , is mutually exclusive and exhaustive, the probabilities of those events sum to 1 (i.e., $\sum P(E_i) = 1$).

The first of the defining properties introduces the term $P(E_i)$, which is shorthand for the “probability of event i .” If $P(E_i) = 0$, the event will never happen. If $P(E_i) = 1$, the event is certain to occur, and the outcome is not random. The probability of rolling any one of the numbers 1–6 with a fair die is $1/6 = 0.1667 = 16.7\%$. The set of events—rolling a number equal to 1, 2, 3, 4, 5, or 6—is an exhaustive set of outcomes (events). The six possible outcomes are mutually exclusive events, if a 2 is rolled, none of the other values can be the result of that roll. The probability of this set of events thus is 100% (equal to 1).

An **empirical probability** is established by analyzing past data (outcomes). An **a priori probability** is determined using a formal reasoning and inspection process (not data). Inspecting a coin and reasoning that the probability of each side coming up when the coin is flipped is an example of an a priori probability.

A **subjective probability** is the least formal method of developing probabilities and involves the use of personal judgment. An analyst may know many things about a firm’s performance and have expectations about the overall market that are all used to arrive at a subjective probability, such as “I believe there is a 70% probability that Acme Foods will outperform the market this year.” Empirical and a priori probabilities, by contrast, are **objective probabilities**.

LOS 3.c: Describe the probability of an event in terms of odds for and against the event.

Stating the **odds** that an event will or will not occur is an alternative way of expressing probabilities. Consider an event that has a probability of occurrence of 0.125, which is one-eighth. The *odds* that the event will occur are $\frac{0.125}{(1 - 0.125)} = \frac{1/8}{7/8} = \frac{1}{7}$, which we state as “the odds for the event occurring are one-to-seven.” The *odds against* the event occurring are the reciprocal of $1/7$, which is seven-to-one.

We can also get the probability of an event from the odds by reversing these calculations. If we know that the odds for an event are one-to-six, we can compute the probability of occurrence as $\frac{1}{1+6} = \frac{1}{7} = 0.1429 = 14.29\%$. Alternatively, the probability that the event will not occur is $\frac{6}{1+6} = \frac{6}{7} = 0.8571 = 85.71\%$.

LOS 3.d: Calculate and interpret conditional probabilities.

Unconditional probability (a.k.a. *marginal probability*) refers to the probability of an event regardless of the past or future occurrence of other events. If we are concerned with the probability of an economic recession, regardless of the occurrence of changes in interest rates or inflation, we are concerned with the unconditional probability of a recession.

A **conditional probability** is one where the occurrence of one event affects the probability of the occurrence of another event. In symbols we write “the probability of A occurring, given that B has occurred as $\text{Prob}(A|B)$ or $P(A|B)$. For example, we might be concerned with the probability of a recession *given* that the monetary authority has increased interest rates. This is a conditional probability. The key word to watch for here is “given.” Using probability notation, “the probability of A *given* the occurrence of B” is expressed as $P(A | B)$, where the vertical bar

(|) indicates “given,” or “conditional upon.” For our interest rate example above, the probability of a recession *given* an increase in interest rates is expressed as $P(\text{recession} \mid \text{increase in interest rates})$. A conditional probability of an occurrence is also called its **likelihood**.

Consider a numerical example. If the Fed increases the USD policy rate, there is a 70% probability that a recession will follow. If the Fed does not increase the USD policy rate, there is a 20% probability that a recession will follow. These are both conditional expectations, $P(\text{recession} \mid \text{rate increase}) = 70\%$, and $P(\text{recession} \mid \text{no rate increase}) = 20\%$.

LOS 3.e: Demonstrate the application of the multiplication and addition rules for probability.

The **addition rule of probability** is used to determine the probability that at least one of two events will occur:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

where B_1, B_2, \dots, B_N is a mutually exclusive and exhaustive set of outcomes.

The **joint probability** of two events is the probability that they will both occur. We can calculate this from the conditional probability that A will occur given B occurs (a conditional probability) and the probability that B will occur (the unconditional probability of B). This calculation is sometimes referred to as the **multiplication rule of probability**. Using the notation for conditional and unconditional probabilities, we can express this rule as:

$$P(AB) = P(A \mid B) \times P(B)$$

This expression is read as follows: “The joint probability of A and B, $P(AB)$, is equal to the conditional probability of A *given* B, $P(A \mid B)$, times the unconditional probability of B, $P(B)$.”

This relationship can be rearranged to define the conditional probability of A given B as follows:

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

EXAMPLE: Multiplication rule of probability

Consider the following information:

- $P(I) = 0.4$, the probability of the monetary authority increasing interest rates (I) is 40%.
- $P(R \mid I) = 0.7$, the probability of a recession (R) given an increase in interest rates is 70%.

What is $P(RI)$, the joint probability of a recession *and* an increase in interest rates?

Answer:

Applying the multiplication rule, we get the following result:

$$P(RI) = P(R \mid I) \times P(I)$$

$$P(RI) = 0.7 \times 0.4$$

$$P(RI) = 0.28$$

Don't let the cumbersome notation obscure the simple logic of this result. If an interest rate increase will occur 40% of the time and lead to a recession 70% of the time when it occurs, the joint probability of an interest rate increase and a resulting recession is $(0.4)(0.7) = (0.28) = 28\%$.

Calculating the Probability That at Least One of Two Events Will Occur

The *addition rule* for probabilities is used to determine the probability that at least one of two events will occur. For example, given two events, A and B, the addition rule can be used to determine the probability that either A or B will occur. If the events are *not* mutually exclusive, double counting must be avoided by subtracting the joint probability that *both* A and B will occur from the sum of the unconditional probabilities. This is reflected in the following general expression for the addition rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

For mutually exclusive events, where the joint probability, $P(AB)$, is zero, the probability that either A or B will occur is simply the sum of the unconditional probabilities for each event, $P(A \text{ or } B) = P(A) + P(B)$.

Figure 3.1: Venn Diagram for Events That Are Not Mutually Exclusive

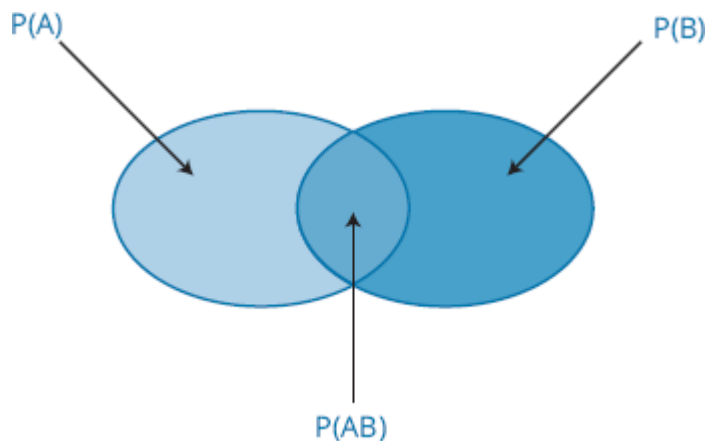


Figure 3.1 illustrates the addition rule with a Venn Diagram and highlights why the joint probability must be subtracted from the sum of the unconditional probabilities. Note that if the events are *mutually exclusive*, the sets do not intersect, $P(AB) = 0$, and the probability that one of the two events will occur is simply $P(A) + P(B)$.

EXAMPLE: Addition rule of probability

Using the information in our previous interest rate and recession example and the fact that the unconditional probability of a recession, $P(R)$, is 34%, determine the probability that either interest rates will increase *or* a recession will occur.

Answer:

Given that $P(R) = 0.34$, $P(I) = 0.40$, and $P(RI) = 0.28$, we can compute $P(R \text{ or } I)$ as follows:

$$P(R \text{ or } I) = P(R) + P(I) - P(RI)$$

$$P(R \text{ or } I) = 0.34 + 0.40 - 0.28$$

$$P(R \text{ or } I) = 0.46$$

Calculating a Joint Probability of Any Number of Independent Events

On the roll of two dice, the joint probability of getting two 4s is calculated as:

$$P(4 \text{ on first die and } 4 \text{ on second die}) = P(4 \text{ on first die}) \times P(4 \text{ on second die}) = 1/6 \times 1/6 = 1/36 = 0.0278$$

On the flip of two coins, the probability of getting two heads is:

$$P(\text{heads on first coin and heads on second coin}) = 1/2 \times 1/2 = 1/4 = 0.25$$

Hint: When dealing with *independent events*, the word *and* indicates multiplication, and the word *or* indicates addition. In probability notation:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB), \text{ and } P(A \text{ and } B) = P(A) \times P(B)$$

The multiplication rule we used to calculate the joint probability of two independent events may be applied to any number of independent events, as the following example illustrates.

EXAMPLE: Joint probability for more than two independent events

What is the probability of rolling three 4s in one simultaneous toss of three dice?

Answer:

Since the probability of rolling a 4 for each die is $1/6$, the probability of rolling three 4s is:

$$P(\text{three 4s on the roll of three dice}) = 1/6 \times 1/6 \times 1/6 = 1/216 = 0.00463$$



MODULE QUIZ 3.1

1. An event that includes all of the possible outcomes is said to be:
 - A. random.
 - B. exclusive.
 - C. exhaustive.
2. Which of the following values *cannot* be the probability of an event?
 - A. 0.00.
 - B. 1.00.
 - C. 1.25.
3. The probability that the DJIA will increase tomorrow is $2/3$. The probability of an increase in the DJIA stated as odds is:
 - A. two-to-one.
 - B. one-to-three.
 - C. two-to-three.
4. The multiplication rule of probability determines the joint probability of two events as the product of:

- A. two conditional probabilities.
 - B. two unconditional probabilities.
 - C. a conditional probability and an unconditional probability.
5. If events A and B are mutually exclusive, then:
- A. $P(A | B) = P(A)$.
 - B. $P(AB) = P(A) \times P(B)$.
 - C. $P(A \text{ or } B) = P(A) + P(B)$.
6. Two mutually exclusive events:
- A. will both occur.
 - B. cannot both occur.
 - C. may both occur.
7. At a charity ball, 800 names are put into a hat. Four of the names are identical. On a random draw, what is the probability that one of these four names will be drawn?
- A. 0.004.
 - B. 0.005.
 - C. 0.010.

MODULE 3.2: CONDITIONAL EXPECTATIONS AND EXPECTED VALUE



Video covering this content is available online.

LOS 3.f: Compare and contrast dependent and independent events.

Independent events refer to events for which the occurrence of one has no influence on the occurrence of the others. The definition of independent events can be expressed in terms of conditional probabilities. Events A and B are independent if and only if:

$$P(A | B) = P(A), \text{ or equivalently, } P(B | A) = P(B)$$

If this condition is not satisfied, the events are dependent events (i.e., the occurrence of one is dependent on the occurrence of the other).

In our interest rate and recession example, recall that events I and R are not independent; the occurrence of I affects the probability of the occurrence of R. In this example, the independence conditions for I and R are violated because:

$P(R) = 0.34$, but $P(R | I) = 0.7$; the probability of a recession is greater when there is an increase in interest rates.

The best examples of independent events are found with the a priori probabilities of dice tosses or coin flips. A die has “no memory.” Therefore, the event of rolling a 4 on the second toss is independent of rolling a 4 on the first toss. This idea may be expressed as:

$$P(4 \text{ on second toss} | 4 \text{ on first toss}) = P(4 \text{ on second toss}) = 1/6 \text{ or } 0.167$$

Because the two events are independent, the conditional probability of a 4 on the second toss is the same as its unconditional probability.

The idea of independent events also applies to flips of a coin: $P(\text{heads on second toss} | \text{heads on first toss}) = P(\text{heads on second toss}) = 1/2 \text{ or } 0.50$

LOS 3.g: Calculate and interpret an unconditional probability using the total probability rule.

The **total probability rule** is used to determine the unconditional probability of an event, given conditional probabilities:

$$P(A) = P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \dots + P(A | B_N)P(B_N)$$

In general, the unconditional probability of event R, $P(R) = P(R | S_1) \times P(S_1) + P(R | S_2) \times P(S_2) + \dots + P(R | S_N) \times P(S_N)$, where the set of events $\{S_1, S_2, \dots, S_N\}$ is mutually exclusive and exhaustive.

EXAMPLE: An investment application of unconditional probability

Building upon our ongoing example about interest rates and economic recession, we can assume that a recession can only occur with either of the two events—interest rates increase (I) or interest rates do not increase (I^C)—since these events are mutually exclusive and exhaustive. I^C is read “the complement of I,” which means “not I.” Therefore, the probability of I^C is $1 - P(I)$. It is logical, therefore, that the sum of the two joint probabilities must be the unconditional probability of a recession. This can be expressed as follows:

$$P(R) = P(RI) + P(RI^C)$$

Applying the multiplication rule, we may restate this expression as:

$$P(R) = P(R | I) \times P(I) + P(R | I^C) \times P(I^C)$$

Assume that $P(R | I) = 0.70$, $P(R | I^C)$, the probability of recession if interest rates do not rise, is 10% and that $P(I) = 0.40$ so that $P(I^C) = 0.60$. The unconditional probability of a recession can be calculated as follows:

$$\begin{aligned} P(R) &= P(R | I) \times P(I) + P(R | I^C) \times P(I^C) \\ &= (0.70)(0.40) + (0.10)(0.60) \\ &= 0.28 + 0.06 = 0.34 \end{aligned}$$

LOS 3.h: Calculate and interpret the expected value, variance, and standard deviation of random variables.

The **expected value** of a random variable is the weighted average of the possible outcomes for the variable. The mathematical representation for the expected value of random variable X , that can take on any of the values from x_1 to x_n , is:

$$E(X) = \sum P(x_i)x_i = P(x_1)x_1 + P(x_2)x_2 + \dots + P(x_n)x_n$$

EXAMPLE: Expected earnings per share

The probability distribution of EPS for Ron's Stores is given in the figure below. Calculate the expected earnings per share.

EPS Probability Distribution

Probability	Earnings Per Share
10%	£1.80
20%	£1.60
40%	£1.20
30%	£1.00
100%	

Answer:

The expected EPS is simply a weighted average of each possible EPS, where the weights are the probabilities of each possible outcome.

$$E[\text{EPS}] = 0.10(1.80) + 0.20(1.60) + 0.40(1.20) + 0.30(1.00) = \text{£}1.28$$

Variance and *standard deviation* measure the dispersion of a random variable around its expected value, sometimes referred to as the **volatility** of a random variable. Variance can be calculated as the probability-weighted sum of the squared deviations from the mean (or expected value). The standard deviation is the positive square root of the variance. The following example illustrates the calculations for a probability model of possible returns.

EXAMPLE: Expected Value, Variance, and Standard Deviation from a probability model

Using the probabilities given in the table below, calculate the expected return on Stock A, the variance of returns on Stock A, and the standard deviation of returns on Stock A.

Event	Probability	R_A	Probability $\times R_A$	R_A $- E(R_A)$	$[R_A$ $- E(R_A)]^2$	Probability \times $[R_A - E(R_A)]^2$
Boom	30%	20%	0.06	0.07	0.0049	0.00147
Normal	50%	12%	0.06	-0.01	0.0001	0.00005
Slow	20%	5%	0.01	-0.08	0.0064	0.00128
			$E(R_A) =$ 0.13	$\text{Var}(R_A) =$ 0.00280		

Answer:

$$E(R_A) = (0.30 \times 0.20) + (0.50 \times 0.12) + (0.20 \times 0.05) = 0.13 = 13\%$$

The expected return for Stock A is the probability-weighted sum of the returns under the three different economic scenarios.

In column 5, we have calculated the differences between the returns under each economic scenario and the expected return of 13%.

In column 6, we squared all the differences from column 5, and in the final column we have multiplied the probabilities of each economic scenario times the squared deviation of returns from the expected returns, and their sum, 0.00280, is the variance of R_A .

The standard deviation of $R_A = \sqrt{0.0028} = 0.0529$

Note that in a previous reading we estimated the standard deviation of a distribution from sample data, rather than from a probability model of returns. For the sample standard deviation, we divided the sum of the squared deviations from the mean by $n - 1$, where n was the size of the sample. Here we have no “ n ,” but we use the probability weights instead, as they describe the entire distribution of outcomes.

LOS 3.i: Explain the use of conditional expectation in investment applications.

Expected values or expected returns can be calculated using conditional probabilities. As the name implies, **conditional expected values** are contingent on the outcome of some other event. An analyst would use a conditional expected value to revise his expectations when new information arrives.

Consider the effect a tariff on steel imports might have on the returns of a domestic steel producer’s stock. The stock’s conditional expected return, given that the government imposes the tariff, will be higher than the conditional expected return if the tariff is not imposed.

Using the total probability rule, we can estimate the (unconditional) expected return on the stock as the sum of the expected return given no tariff, times the probability a tariff will not be enacted, and the expected return given a tariff, times the probability a tariff will be enacted.

LOS 3.j: Interpret a probability tree and demonstrate its application to investment problems.

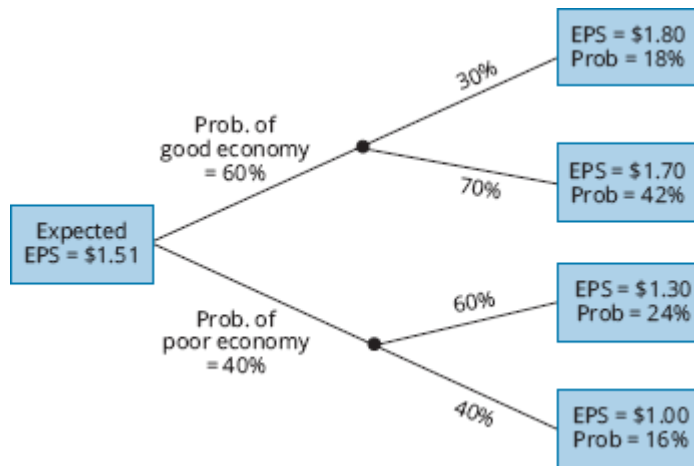
You might well wonder where the returns and probabilities used in calculating expected values come from. A general framework called a **probability tree** is used to show the probabilities of various outcomes. In Figure 3.2, we have shown estimates of EPS for four different events: (1) a good economy and relatively good results at the company, (2) a good economy and relatively poor results at the company, (3) a poor economy and relatively good results at the company, and (4) a poor economy and relatively poor results at the company. Using the rules of probability, we can calculate the probabilities of each of the four EPS outcomes shown in the boxes on the right-hand side of the “tree.”

The expected EPS of \$1.51 is simply calculated as:

$$0.18 \times 1.80 + 0.42 \times 1.70 + 0.24 \times 1.30 + 0.16 \times 1.00 = \$1.51$$

Note that the probabilities of the four possible outcomes sum to 1.

Figure 3.2: A Probability Tree



MODULE QUIZ 3.2

- Two events are said to be independent if the occurrence of one event:
 - means that the second event cannot occur.
 - means that the second event is certain to occur.
 - does not affect the probability of the occurrence of the other event.
- An analyst estimates that a share price has an 80% probability of increasing if economic growth exceeds 3%, a 40% probability of increasing if economic growth is between zero and 3%, and a 10% probability of increasing if economic growth is negative. If economic growth has a 25% probability of exceeding 3% and a 25% probability of being negative, what is the probability that the share price increases?
 - 22.5%.
 - 42.5%.
 - 62.5%.
- $P(A|B) = 40\%$ and $P(B) = 30\%$ and $P(A) = 40\%$. It is *most likely* that:
 - A and B are dependent.
 - A and B are independent.
 - A and B are mutually exclusive.

MODULE 3.3: PORTFOLIO VARIANCE, BAYES, AND COUNTING PROBLEMS



Video covering this content is available online.

LOS 3.k: Calculate and interpret the expected value, variance, standard deviation, covariances, and correlations of portfolio returns.

Portfolio expected return. The expected return of a portfolio composed of n assets with weights, w_i , and expected returns, R_i , can be determined using the following formula:

$$E(R_p) = \sum_{i=1}^n w_i E(R_i) = w_1 E(R_1) + w_2 E(R_2) + \dots + w_n E(R_n)$$

The expected return and variance for a portfolio of assets can be determined using the properties of the individual assets in the portfolio. To do this, it is necessary to establish the portfolio weight for each asset. As indicated in the formula, the weight, w , of portfolio asset i is simply the market value currently invested in the asset divided by the current market value of the entire portfolio.

$$w_i = \frac{\text{market value of investment in asset } i}{\text{market value of the portfolio}}$$

In many finance situations, we are interested in how two random variables move in relation to each other. For investment applications, one of the most frequently analyzed pairs of random variables is the returns of two assets. Investors and managers frequently ask questions such as “what is the relationship between the return for Stock A and Stock B?” or “what is the relationship between the performance of the S&P 500 and that of the automotive industry?”

Covariance is a measure of how two assets move together. It is the expected value of the product of the deviations of the two random variables from their respective expected values. A common symbol for the covariance between random variables X and Y is $\text{Cov}(X,Y)$. Since we will be mostly concerned with the covariance of asset returns, the following formula has been written in terms of the covariance of the return of asset i , R_i , and the return of asset j , R_j :

$$\text{Cov}(R_i, R_j) = E\{[R_i - E(R_i)][R_j - E(R_j)]\}$$

The following are *properties of covariance*:

- The covariance of a random variable with itself is its variance of R_A ; that is, $\text{Cov}(R_A, R_A) = \text{Var}(R_A)$.
- Covariance may range from negative infinity to positive infinity.
- A positive covariance indicates that when one random variable is above its mean, the other random variable tends to be above its mean as well.
- A negative covariance indicates that when one random variable is above its mean, the other random variable tends to be below its mean.

The sample covariance for a sample of returns data can be calculated as:

$$s_{x,y} = \frac{\sum_{i=1}^n \{ [R_{1,i} - \bar{R}_1] [R_{2,i} - \bar{R}_2] \}}{n - 1}$$

where:

$R_{1,i}$ = an observation of returns on asset 1

$R_{2,i}$ = an observation of returns on asset 2

\bar{R}_1 = mean return of asset 1

\bar{R}_2 = mean of asset 2

n = number of observations in the sample

A **covariance matrix** shows the covariances between returns on a group of assets.

Figure 3.3: Covariance matrix for three assets

Asset	A	B	C
A	$\text{Cov}(R_A, R_A)$	$\text{Cov}(R_A, R_B)$	$\text{Cov}(R_A, R_C)$
B	$\text{Cov}(R_B, R_A)$	$\text{Cov}(R_B, R_B)$	$\text{Cov}(R_B, R_C)$
C	$\text{Cov}(R_C, R_A)$	$\text{Cov}(R_C, R_B)$	$\text{Cov}(R_C, R_C)$

Note that the diagonal terms are the variances of each asset’s returns, i.e., $\text{Cov}(R_A, R_A) = \text{Var}(R_A)$.

The covariance between the returns on two assets does not depend on order, i.e., $\text{Cov}(R_A, R_B) = \text{Cov}(R_B, R_A)$, so in this covariance matrix only 3 of the (off-diagonal) covariance terms are unique. In general for n assets, there are n variance terms (on the diagonal) and $n(n - 1)/2$ unique covariance terms.

Portfolio variance. To calculate the variance of portfolio returns, we use the asset weights, returns variances, and returns covariances.

$$\text{Var}(R_p) = \sum_{i=1}^N \sum_{j=1}^N w_i w_j \text{Cov}(R_i, R_j)$$

The variance of a portfolio composed of two risky assets, A and B can be expressed as:

$$\begin{aligned} \text{Var}(R_p) &= w_A w_A \text{Cov}(R_A, R_A) + w_A w_B \text{Cov}(R_A, R_B) + w_B w_A \text{Cov}(R_B, R_A) \\ &+ w_B w_B \text{Cov}(R_B, R_B) \end{aligned}$$

which we can write more simply as:

$$\text{Var}(R_p) = w_A^2 \text{Var}(R_A) + w_B^2 \text{Var}(R_B) + 2w_A w_B \text{Cov}(R_A, R_B), \text{ or}$$

$$\sigma_p^2 = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + 2w_A w_B \text{Cov}_{AB}.$$

For a 3-asset portfolio, the portfolio variance is:

$$\sigma_p^2 = w_A^2 \sigma_A^2 + w_B^2 \sigma_B^2 + w_C^2 \sigma_C^2 + 2w_A w_B \text{Cov}_{AB} + 2w_A w_C \text{Cov}_{AC} + 2w_B w_C \text{Cov}_{BC}$$

Consider a portfolio with three assets: an index of domestic stocks (60%), an index of domestic bonds (30%), and an index of international equities (10%). A covariance matrix of the three assets is shown below.

Figure 3.4: Covariance matrix for the three assets

Asset	Domestic Stocks	Domestic Bonds	International Equities
Domestic Stocks	400	44	180
Domestic Bonds	44	70	35
International Equities	180	35	450

Portfolio returns variance =

$$(0.6^2)400 + (0.3^2)70 + (0.1^2)450 + 2(0.6)(0.3)44 + 2(0.6)(0.1)180 + 2(0.3)(0.1)35 = 194.34$$

Portfolio returns standard deviation $\sqrt{194.34} = 13.94\%$.

Note that the units of variance and covariance are $\%^2$, i.e., 0.001. When we put these values in as whole numbers (in $\%^2$), the portfolio variance is in $\%^2$, and the standard deviation is in whole percentages. We could also put variance and covariance in as decimals and get both the portfolio returns variance and standard deviation as decimals.

From the formula for portfolio returns variance, we can see that the lower the covariance terms, the lower the portfolio variance (and standard deviation). This is true for positive values of covariance, as well as negative values.

Correlation

Recall that the correlation coefficient for two variables is $\rho_{AB} = \frac{\text{Cov}_{AB}}{\sigma_A \sigma_B}$, so that

($\text{Cov}_{AB} = \rho_{AB} \times \sigma_A \sigma_B$). This can be substituted for Cov_{AB} in our formula for portfolio returns variance. With this substitution we can use a correlation matrix to calculate portfolio returns variance, rather than using covariances.

Figure 3.5: Correlation matrix for the three assets

Asset	Domestic Stocks	Domestic Bonds	International Equities
Domestic Stocks	1.000	0.263	0.424
Domestic Bonds	0.263	1.000	0.197
International Equities	0.424	0.197	1.000

Note that the correlations of asset returns with themselves (the diagonal terms) are all 1.

LOS 3.I: Calculate and interpret the covariances of portfolio returns using the joint probability function.

EXAMPLE: Covariance of returns from a joint probability function

Assume that the economy can be in three possible states (S) next year: boom, normal, or slow economic growth. An expert source has calculated that $P(\text{boom}) = 0.30$, $P(\text{normal}) = 0.50$, and $P(\text{slow}) = 0.20$. The returns for Asset A, R_A , and Asset B, R_B , under each of the economic states are provided in the probability model as follows. What is the covariance of the returns for Asset A and Asset B?

Joint Probability Function

	$R_B = 30\%$	$R_B = 10\%$	$R_B = 0\%$
$R_A = 20\%$	0.30	0	0
$R_A = 12\%$	0	0.50	0
$R_A = 5\%$	0	0	0.20

The table gives us the joint probability of returns on Assets A and B, e.g., the probability that the return on Asset A is 20% and the return on Asset B is 30% is 30% and the probability that the return on Asset A is 12% and the return on Asset B is 10% is 50%.

Answer:

First, we must calculate the expected returns for each of the assets.

$$E(R_A) = (0.3)(0.20) + (0.5)(0.12) + (0.2)(0.05) = 0.13$$

$$E(R_B) = (0.3)(0.30) + (0.5)(0.10) + (0.2)(0.00) = 0.14$$

The covariance can now be computed using the procedure described in the following table.

Covariance Calculation

Probability	R_A	R_B	Probability $\times [R_A - E(R_A)] \times [R_B - E(R_B)]$
0.3	0.20	0.30	$(0.3)(0.2 - 0.13)(0.3 - 0.14) = 0.00336$
0.5	0.12	0.10	$(0.5)(0.12 - 0.13)(0.1 - 0.14) = 0.00020$
0.2	0.05	0.00	$(0.2)(0.05 - 0.13)(0 - 0.14) = 0.00224$

Covariance of returns for Asset A and Asset B is

$$0.00336 + 0.00020 + 0.00224 = 0.005784$$

LOS 3.m: Calculate and interpret an updated probability using Bayes' formula.

Bayes' formula is used to update a given set of prior probabilities for a given event in response to the arrival of new information. The rule for updating prior probability of an event is:

$$\text{updated probability} = \frac{\text{probability of new information for a given event}}{\text{unconditional probability of new information}} \times \text{prior probability of event}$$

We can derive Bayes' formula using the multiplication rule and noting that $P(AB) = P(BA)$.

$$P(B|A) \times P(A) = P(BA), \text{ and } P(A|B) \times P(B) = P(AB).$$

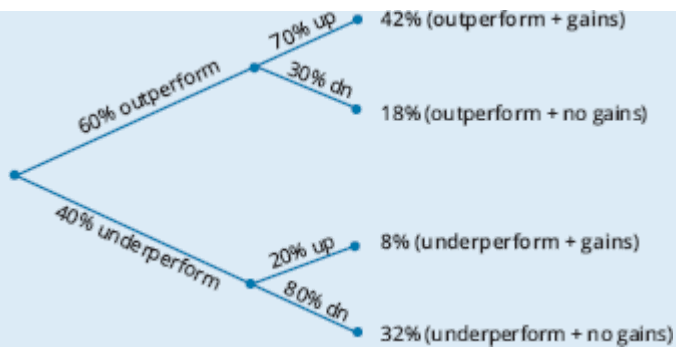
Because $P(BA) = P(AB)$, we can write $P(B|A) P(A) = P(A|B) P(B)$ and $\frac{P(B|A)P(A)}{P(B)}$ which equals $\frac{P(BA)}{P(B)}$ the joint probability of A and B, divided by the unconditional probability of B.

The following example illustrates the use of Bayes' formula. Note that A is outperform and A^C is underperform, $P(BA)$ is (outperform + gains), $P(A^C B)$ is (underperform + gains), and the unconditional probability $P(B)$ is $P(AB) + P(A^C B)$, by the total probability rule.

EXAMPLE: Bayes' formula

There is a 60% probability the economy will outperform, and if it does, there is a 70% probability a stock will go up and a 30% probability the stock will go down. There is a 40% probability the economy will underperform, and if it does, there is a 20% probability the stock in question will increase in value (have gains) and an 80% probability it will not. Given that the stock increased in value, calculate the probability that the economy outperformed.

Answer:



In the figure above, we have multiplied the probabilities to calculate the probabilities of each of the four outcome pairs. Note that these sum to 1. Given that the stock has gains, what is our updated probability of an outperforming economy? We sum the probability of stock gains in both states (outperform and underperform) to get $42\% + 8\% = 50\%$. Given that the stock has gains, the probability that the economy has outperformed is $\frac{42\%}{50\%} = 84\%$.

LOS 3.n: Identify the most appropriate method to solve a particular counting problem and analyze counting problems using factorial, combination, and permutation concepts.

Labeling refers to the situation where there are n items that can each receive one of k different labels. The number of items that receives label 1 is n_1 and the number that receive label 2 is n_2 , and so on, such that $n_1 + n_2 + n_3 + \dots + n_k = n$. The total number of ways that the labels can be assigned is:

$$\frac{n!}{(n_1!) \times (n_2!) \times \dots \times (n_k!)}$$

where:

the symbol “!” stands for **factorial**. For example, $4! = 4 \times 3 \times 2 \times 1 = 24$, and $2! = 2 \times 1 = 2$.

The general expression for n factorial is:

$$n! = n \times (n - 1) \times (n - 2) \times (n - 3) \times \dots \times 1, \text{ where by definition, } 0! = 1$$

Calculator help: On the TI, factorial is [2nd] [x!] (above the multiplication sign). To compute $4!$ on the TI, enter [4][2nd][x!] = 24.

EXAMPLE: Labeling

Consider a portfolio consisting of eight stocks. Your goal is to designate four of the stocks as “long-term holds,” three of the stocks as “short-term holds,” and one stock as “sell.” How many ways can these eight stocks be labeled?

Answer:

There are $8! = 40,320$ total possible sequences that can be followed to assign the three labels to the eight stocks. However, the order that each stock is assigned a label does not matter. For example, it does not matter which of the stocks labeled “long-term” is the first to be labeled. Thus, there are $4!$ ways to assign the long-term label. Continuing this reasoning to the other categories, there are $4! \times 3! \times 1!$ equivalent sequences for assigning the labels. To eliminate the

counting of these redundant sequences, the total number of possible sequences (8!) must be divided by the number of redundant sequences (4! × 3! × 1!).

Thus, the number of *different* ways to label the eight stocks is:

$$\frac{8!}{4! \times 3! \times 1!} = \frac{40,320}{24 \times 6 \times 1} = 280$$

If there are n labels ($k = n$), we have $\frac{n!}{1} = n!$. The number of ways to assign n different labels to n items is simply $n!$.

A special case of labeling arises when the number of labels equals 2 ($k = 2$). That is, the n items can only be in one of two groups, and $n_1 + n_2 = n$. In this case, we can let $r = n_1$ and $n_2 = n - r$. Since there are only two categories, we usually talk about choosing r items. Then $(n - r)$ items are not chosen. The general formula for labeling when $k = 2$ is called the **combination formula** (or *binomial formula*) and is expressed as:

$${}_n C_r = \frac{n!}{(n-r)!r!}$$

where ${}_n C_r$ is the number of possible ways (combinations) of selecting r items from a set of n items when the order of selection is not important. This is also written $\binom{n}{r}$ and read “ n choose r .”

Another useful formula is the **permutation formula**. A permutation is a specific ordering of a group of objects. The question of how many different groups of size r in specific order can be chosen from n objects is answered by the permutation formula. The number of permutations of r objects from n objects is:

$${}_n P_r = \frac{n!}{(n-r)!}$$

We will give an example using this formula shortly.



PROFESSOR'S NOTE

The combination formula ${}_n C_r$ and the permutation formula ${}_n P_r$ are both available on the TI calculator. To calculate the number of different groups of three stocks from a list of eight stocks (i.e., ${}_8 C_3$), the sequence is 8 [2nd] [${}_n C_r$] 3 [=], which yields 56. If we want to know the number of differently ordered groups of three that can be selected from a list of eight, we enter 8 [2nd] [${}_n P_r$] 3 [=] to get 336, which is the number of permutations, $\frac{8!}{(8-3)!}$. This function is not available on the HP calculator. Remember, current policy permits you to bring both calculators to the exam, if you choose.

EXAMPLE: Number of choices in any order

How many ways can three stocks be sold from an 8-stock portfolio?

Answer:

Here we use the combination formula (or function on your calculator) as order does not matter. Note that this is equivalent to using the labeling formula with two labels, sold and not sold. Thus, the answer is:

$$\frac{8!}{5! \times 3!} = 56$$

In the preceding two examples, ordering did not matter. The order of selection could, however, be important. For example, suppose we want to liquidate only one stock position per week over the next three weeks. Once we choose three particular stocks to sell, the order in which they are sold must be determined. In this case, the concept of permutation comes into play.

The permutation formula implies that there are $r!$ more ways to choose r items if the order of selection is important than if order is not important. That is, ${}_n P_r = r! \times {}_n C_r$.

EXAMPLE: Permutation

How many ways are there to sell three stocks out of eight if the order of the sales is important?

Answer:

$${}_n P_r = {}_8 P_3 = \frac{8!}{(8-3)!} = \frac{8!}{5!} = 336$$

This is $3!$ times the 56 possible combinations computed in the preceding example for selecting the three stocks when the order was not important.

Five guidelines may be used to determine which counting method to employ when dealing with counting problems:

- The *multiplication rule of counting* is used when there are *two or more groups*. The key is that only *one* item may be selected from each group. If there are k steps required to complete a task and each step can be done in n ways, the number of different ways to complete the task is $n_1 \times n_2 \times \dots \times n_k$.
- *Factorial* is used by itself when there are *no groups*—we are only arranging a given set of n items. Given n items, there are $n!$ ways of arranging them.
- The *labeling formula* applies to *three or more subgroups* of predetermined size. Each element of the entire group must be assigned a place, or label, in one of the three or more subgroups.
- The *combination formula* applies to *only two groups* of predetermined size. Look for the word “choose” or “combination.”
- The *permutation formula* applies to *only two groups* of predetermined size. Look for a specific reference to “order” being important.



MODULE QUIZ 3.3

1. Given the conditional probabilities in the table below and the unconditional probabilities $P(Y = 1) = 0.3$ and $P(Y = 2) = 0.7$, what is the expected value of X ?

x_i	$P(x_i Y = 1)$	$P(x_i Y = 2)$
0	0.2	0.1
5	0.4	0.8
10	0.4	0.1

A. 5.0.

- B. 5.3.
C. 5.7.
2. A discrete uniform distribution (each event has an equal probability of occurrence) has the following possible outcomes for X : [1, 2, 3, 4]. The variance of this distribution is *closest* to:
A. 1.00.
B. 1.25.
C. 2.00.
3. The correlation of returns between Stocks A and B is 0.50. The covariance between these two securities is 0.0043, and the standard deviation of the return of Stock B is 26%. The variance of returns for Stock A is:
A. 0.0011.
B. 0.0331.
C. 0.2656.
4. An analyst believes Davies Company has a 40% probability of earning more than \$2 per share. She estimates that the probability that Davies Company's credit rating will be upgraded is 70% if its earnings per share are greater than \$2 and 20% if its earnings per share are \$2 or less. Given the information that Davies Company's credit rating has been upgraded, what is the updated probability that its earnings per share are greater than \$2?
A. 50%.
B. 60%.
C. 70%.
5. Consider a universe of 10 bonds from which an investor will ultimately purchase six bonds for his portfolio. If the order in which he buys these bonds is not important, how many potential 6-bond combinations are there?
A. 7.
B. 210.
C. 5,040.
6. There are 10 sprinters in the finals of a race. How many different ways can the gold, silver, and bronze medals be awarded?
A. 120.
B. 720.
C. 1,440.

KEY CONCEPTS

LOS 3.a

A random variable is an uncertain value determined by chance.

An outcome is the realization of a random variable.

An event is a set of one or more outcomes. Two events that cannot both occur are termed "mutually exclusive," and a set of events that includes all possible outcomes is an "exhaustive" set of events.

LOS 3.b

The two properties of probability are as follows:

- The sum of the probabilities of all possible mutually exclusive events is 1.
- The probability of any event cannot be greater than 1 or less than 0.

A priori probability measures predetermined probabilities based on well-defined inputs; empirical probability measures probability from observations or experiments; and subjective

probability is an informed guess.

LOS 3.c

Probabilities can be stated as odds that an event will or will not occur. If the probability of an event is A out of B trials (A/B), the “odds for” are A to (B – A) and the “odds against” are (B – A) to A.

LOS 3.d

Unconditional probability (marginal probability) is the probability of an event occurring.

Conditional probability, $P(A | B)$, is the probability of an event A occurring given that event B has occurred.

LOS 3.e

The multiplication rule of probability is used to determine the joint probability of two events:

$$P(AB) = P(A | B) \times P(B)$$

The addition rule of probability is used to determine the probability that at least one of two events will occur:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

The total probability rule is used to determine the unconditional probability of an event, given conditional probabilities:

$$P(A) = P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \dots + P(A | B_N)P(B_N)$$

where B_1, B_2, \dots, B_N is a mutually exclusive and exhaustive set of outcomes.

LOS 3.f

The probability of an independent event is unaffected by the occurrence of other events, but the probability of a dependent event is changed by the occurrence of another event. Events A and B are independent if and only if:

$$P(A | B) = P(A), \text{ or equivalently, } P(B | A) = P(B)$$

LOS 3.g

Using the total probability rule, the unconditional probability of A is the probability-weighted sum of the conditional probabilities:

$$P(A) = \sum_{i=1}^n [P_i(B_i)] \times P(A | B_i)$$

where B_i is a set of mutually exclusive and exhaustive events.

LOS 3.h

The expected value of a random variable is the weighted average of its possible outcomes:

$$E(X) = \sum P(x_i)x_i = P(x_1)x_1 + P(x_2)x_2 + \dots + P(x_n)x_n$$

Variance can be calculated as the probability-weighted sum of the squared deviations from the mean or expected value. The standard deviation is the positive square root of the variance.

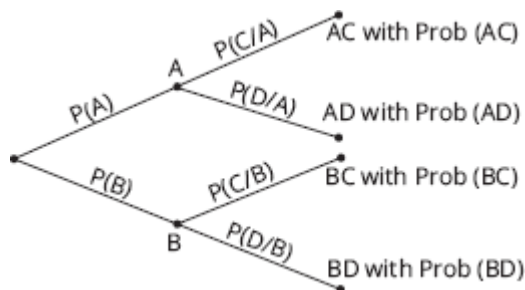
LOS 3.i

Conditional expected values depend on the outcome of some other event.

Forecasts of expected values for a stock's return, earnings, and dividends can be refined, using conditional expected values, when new information arrives that affects the expected outcome.

LOS 3.j

A probability tree shows the probabilities of two events and the conditional probabilities of two subsequent events.



LOS 3.k

The expected value of a random variable, $E(X)$, equals $\sum_{i=1}^n P_i(x_i) X_i$.

The variance of a random variable, $\text{Var}(X)$, equals

$$\sum_{i=1}^n P(X_i) [X_i - E(X)]^2 = \sigma_X^2$$

Standard deviation: $\sigma_X = \sqrt{\sigma_X^2}$.

The expected returns and variance of a 2-asset portfolio are given by:

$$E(R_P) = w_1 E(R_1) + w_2 E(R_2)$$

$$\begin{aligned} \text{Var}(R_P) &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \text{Cov}_{1,2} \\ &= w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \sigma_1 \sigma_2 \rho_{1,2} \end{aligned}$$

LOS 3.l

Given the joint probabilities for X_i and Y_j , i.e., $P(X_i Y_j)$, the covariance is calculated as:

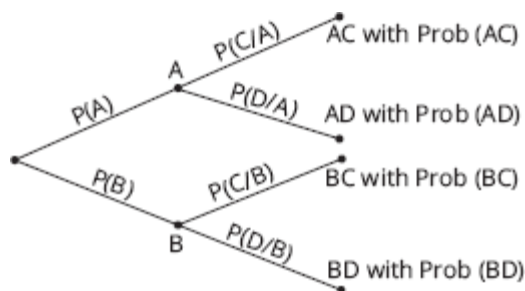
$$\sum_{i=1}^n P(X_i Y_i) [X_i - E(X)] [Y_i - E(Y)]$$

LOS 3.m

Bayes' formula for updating probabilities based on the occurrence of an event O is:

$$P(I|O) = \frac{P(O|I)}{P(O)} \times P(I)$$

Equivalently, based on the tree diagram below, $P(A|C) = \frac{P(AC)}{P(AC) + P(BC)}$



LOS 3.n

The number of ways to order n objects is n factorial, $n! = n \times (n - 1) \times (n - 2) \times \dots \times 1$.

There are $\frac{N!}{n_1! \times n_2! \times \dots \times n_k!}$ ways to assign k different labels to n items, where n_i is the number of items with the label i .

The number of ways to choose a subset of size r from a set of size n when order doesn't matter is $\frac{n!}{(n-r)!r!}$ combinations; when order matters, there are $\frac{n!}{(n-r)!}$ permutations.

ANSWERS TO MODULE QUIZ QUESTIONS

Module Quiz 3.1

1. **C** An event is said to be exhaustive if it includes all possible outcomes. (LOS 3.a)
2. **C** Probabilities may range from 0 (meaning no chance of occurrence) through 1 (which means a sure thing). (LOS 3.b)
3. **A** Odds for E = $P(E) / [1 - P(E)] = \frac{2/3}{1/3} = 2/1 = \text{two-to-one}$ (LOS 3.c)
4. **C** By the multiplication rule of probability, the joint probability of two events, $P(AB)$, is the product of a conditional probability, $P(A | B)$, and an unconditional probability, $P(B)$. (LOS 3.d, LOS 3.e)
5. **C** There is no intersection of events when events are mutually exclusive. $P(A | B) = P(A) \times P(B)$ is only true for independent events. Note that since A and B are mutually exclusive (cannot both happen), $P(A | B)$ and $P(AB)$ must both be equal to zero. (LOS 3.a, LOS 3.d)
6. **B** One or the other may occur, but not both. (LOS 3.a)
7. **B** $P(\text{name 1 or name 2 or name 3 or name 4}) = 1/800 + 1/800 + 1/800 + 1/800 = 4/800 = 0.005$. (LOS 3.e)

Module Quiz 3.2


1. **C** Two events are said to be independent if the occurrence of one event does not affect the probability of the occurrence of the other event. (LOS 3.f)
2. **B** The three outcomes given for economic growth are mutually exclusive and exhaustive. The probability that economic growth is positive but less than 3% is $100\% - 25\% - 25\% = 50\%$. Using the total probability rule, the probability that the share price increases is $(80\%)(25\%) + (40\%)(50\%) + (10\%)(25\%) = 42.5\%$. (LOS 3.g)
3. **B** From the values given, $P(A|B) = P(A)$, so A and B are independent. $P(A|B) \times P(B) = P(AB) = 12\%$, so A and B are not mutually exclusive (if they were $P(AB)$ would equal 0). (LOS 3.g)

Module Quiz 3.3

1. **B** $E(X | Y = 1) = (0.2)(0) + (0.4)(5) + (0.4)(10) = 6$
 $E(X | Y = 2) = (0.1)(0) + (0.8)(5) + (0.1)(10) = 5$
 $E(X) = (0.3)(6) + (0.7)(5) = 5.30$
(LOS 3.k)
2. **B** Expected value = $(1/4)(1 + 2 + 3 + 4) = 2.5$
Variance = $(1/4)[(1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2] = 1.25$

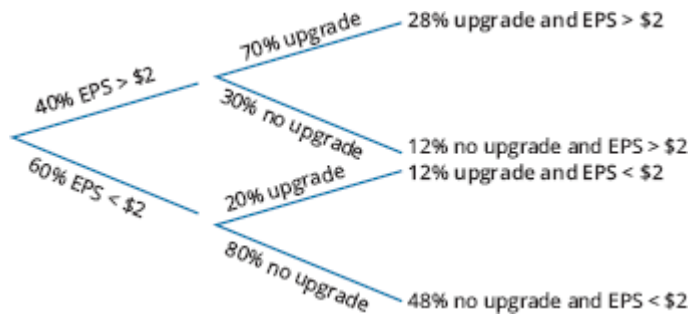
Note that since each observation is equally likely, each has 25% (1/4) chance of occurrence. (LOS 3.k)

3. A

 Numberfigure

(LOS 3.k)

4. C This is an application of Bayes' formula. As the tree diagram below shows, the updated probability that earnings per share are greater than \$2 is $\frac{28\%}{28\% + 12\%} = 70\%$



(LOS 3.m)

5. B

$${}^nC_r = \frac{n!}{(n-r)!r!} = {}^{10}C_6 = \frac{10!}{(10-6)!6!} = \frac{10!}{4!6!} = 210. \text{ (LOS 3.n)}$$

6. B Since the order of the top-three finishers matters, we need to use the permutation formula.

$${}^{10}P_3 = \frac{10!}{(10-3)!} = 720$$

(LOS 3.n)

READING 4

COMMON PROBABILITY DISTRIBUTIONS

EXAM FOCUS

This reading contains a lot of key material. Learn the difference between discrete and continuous probability distributions. The binomial and normal distributions are the most important here. You must learn the properties of both distributions and memorize the formula for the probability of a particular value when given a binomial probability distribution. Learn what shortfall risk is and how to calculate and use Roy's safety-first criterion. Know how to standardize a normally distributed random variable, use a z-table, determine probabilities using a cumulative distribution function, and construct confidence intervals. Learn the critical values for the often-used confidence intervals. You will use these skills repeatedly in the readings that follow. Additionally, understand the basic features of the lognormal distribution and Monte Carlo simulation. Candidates should know how to get continuously compounded rates of return from holding period returns. The chi-square and F-distribution are introduced here and we will use them in our reading on hypothesis testing.

MODULE 4.1: UNIFORM AND BINOMIAL DISTRIBUTIONS



Video covering this content is available online.

LOS 4.a: Define a probability distribution and compare and contrast discrete and continuous random variables and their probability functions.

A **probability distribution** describes the probabilities of all the possible outcomes for a random variable. The probabilities of all possible outcomes must sum to 1. A simple probability distribution is that for the roll of one fair die, there are six possible outcomes and each one has a probability of $1/6$, so they sum to 1. The probability distribution of all the possible returns on the S&P 500 Index for the next year is a more complex version of the same idea.

A **discrete random variable** is one for which the number of possible outcomes can be counted, and for each possible outcome, there is a measurable and positive probability. An example of a discrete random variable is the number of days it will rain in a given month, because there is a countable number of possible outcomes, ranging from zero to the number of days in the month.

A **probability function**, denoted $p(x)$, specifies the probability that a random variable is equal to a specific value. More formally, $p(x)$ is the probability that random variable X takes on the value x , or $p(x) = P(X = x)$.

The two key properties of a probability function are:

- $0 \leq p(x) \leq 1$.
- $\sum p(x) = 1$, the sum of the probabilities for *all* possible outcomes, x , for a random variable, X , equals 1.

EXAMPLE: Evaluating a probability function

Consider the following function: $X = \{1, 2, 3, 4\}$, $p(x) = \frac{x}{10}$, else $p(x) = 0$

Determine whether this function satisfies the conditions for a probability function.

Answer:

Note that all of the probabilities are between 0 and 1, and the sum of all probabilities equals 1:

$$\sum p(x) = \frac{1}{10} + \frac{2}{10} + \frac{3}{10} + \frac{4}{10} = 0.1 + 0.2 + 0.3 + 0.4 = 1$$

Both conditions for a probability function are satisfied.

A **continuous random variable** is one for which the number of possible outcomes is infinite, even if lower and upper bounds exist. The actual amount of daily rainfall between zero and 100 inches is an example of a continuous random variable because the actual amount of rainfall can take on an infinite number of values. Daily rainfall can be measured in inches, half inches, quarter inches, thousandths of inches, or even smaller increments. Thus, the number of possible daily rainfall amounts between zero and 100 inches is essentially infinite.

The assignment of probabilities to the possible outcomes for discrete and continuous random variables provides us with discrete probability distributions and continuous probability distributions. The difference between these types of distributions is most apparent for the following properties:

- For a *discrete distribution*, $p(x) = 0$ when x cannot occur, or $p(x) > 0$ if it can. Recall that $p(x)$ is read: “the probability that random variable $X = x$.” For example, the probability of it raining on 33 days in June is zero because this cannot occur, but the probability of it raining 25 days in June has some positive value.
- For a *continuous distribution*, $p(x) = 0$ even though x can occur. We can only consider $P(x_1 \leq X \leq x_2)$ where x_1 and x_2 are actual numbers. For example, the probability of receiving 2 inches of rain in June is zero because 2 inches is a single point in an infinite range of possible values. On the other hand, the probability of the amount of rain being between 1.99999999 and 2.00000001 inches has some positive value. In the case of continuous distributions, $P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2)$ because $p(x_1) = p(x_2) = 0$.

In finance, some discrete distributions are treated as though they are continuous because the number of possible outcomes is very large. For example, the increase or decrease in the price of a stock traded on an American exchange is recorded in dollars and cents. Yet, the probability of a change of exactly \$1.33 or \$1.34 or any other specific change is almost zero. It is customary, therefore, to speak in terms of the probability of a range of possible price change, say between \$1.00 and \$2.00. In other words $p(\text{price change} = 1.33)$ is essentially zero, but $p(\$1 < \text{price change} < \$2)$ is greater than zero.

LOS 4.b: Calculate and interpret probabilities for a random variable given its cumulative distribution function.

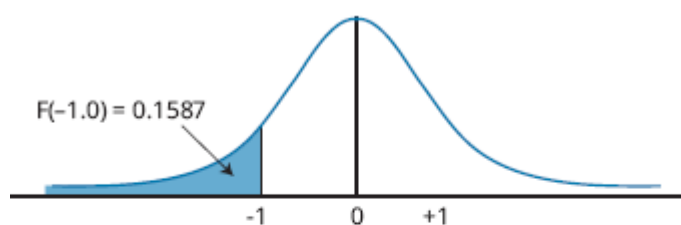
A **cumulative distribution function (cdf)**, or simply *distribution function*, defines the probability that a random variable, X , takes on a value equal to or less than a specific value, x . It represents the sum, or *cumulative value*, of the probabilities for the outcomes up to and including a specified outcome. The cumulative distribution function for a random variable, X , may be expressed as $F(x) = P(X \leq x)$.

Consider the probability function defined earlier for $X = \{1, 2, 3, 4\}$, $p(x) = x / 10$. For this distribution, $F(3) = 0.6 = 0.1 + 0.2 + 0.3$, and $F(4) = 1 = 0.1 + 0.2 + 0.3 + 0.4$. This means that $F(3)$ is the cumulative probability that outcomes 1, 2, or 3 occur, and $F(4)$ is the cumulative probability that one of the possible outcomes occurs.

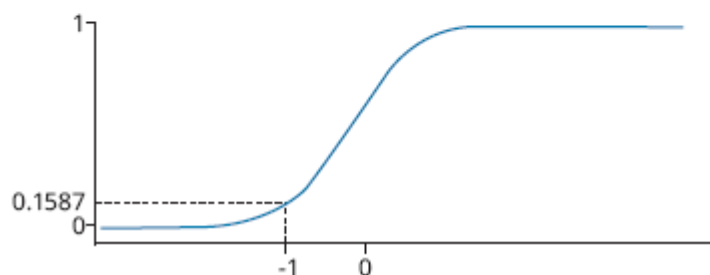
Figure 4.1 shows an example of a cumulative distribution function (for a standard normal distribution, described later in this topic). There is a 15.87% probability of a value less than -1 . This is the total area to the left of -1 in the pdf in Panel (a), and the y-axis value of the cdf for a value of -1 in Panel (b).

Figure 4.1: Standard Normal Probability Density and Cumulative Distribution Functions

(a) Probability density function



(b) Cumulative distribution function



EXAMPLE: Cumulative distribution function

Return on equity for a firm is defined as a continuous distribution over the range from -20% to $+30\%$ and has a cumulative distribution function of $F(x) = (x + 20) / 50$. **Calculate** the probability that ROE will be between 0% and 15% .

Answer:

To determine the probability that ROE will be between 0% and 15% , we can first calculate the probability that ROE will be less than or equal to 15% , or $F(15)$, and then subtract the probability that ROE will be less than zero, or $F(0)$.

$$P(0 \leq x \leq 15) = F(15) - F(0)$$

$$F(15) = (15 + 20) / 50 = 0.70$$

$$F(0) = (0 + 20) / 50 = 0.40$$

$$F(15) - F(0) = 0.70 - 0.40 = 0.30 = 30\%$$

LOS 4.c: Describe the properties of a discrete uniform random variable, and calculate and interpret probabilities given the discrete uniform distribution function.

A **discrete uniform random variable** is one for which the probabilities for all possible outcomes for a discrete random variable are equal. For example, consider the *discrete uniform probability distribution* defined as $X = \{1, 2, 3, 4, 5\}$, $p(x) = 0.2$. Here, the probability for each outcome is equal to 0.2 [i.e., $p(1) = p(2) = p(3) = p(4) = p(5) = 0.2$]. Also, the cumulative distribution function for the n th outcome, $F(x_n) = np(x)$, and the probability for a range of outcomes is $p(x)k$, where k is the number of possible outcomes in the range.

EXAMPLE: Discrete uniform distribution

Determine $p(6)$, $F(6)$, and $P(2 \leq X \leq 8)$ for the discrete uniform distribution function defined as:

$$X = \{2, 4, 6, 8, 10\}, p(x) = 0.2$$

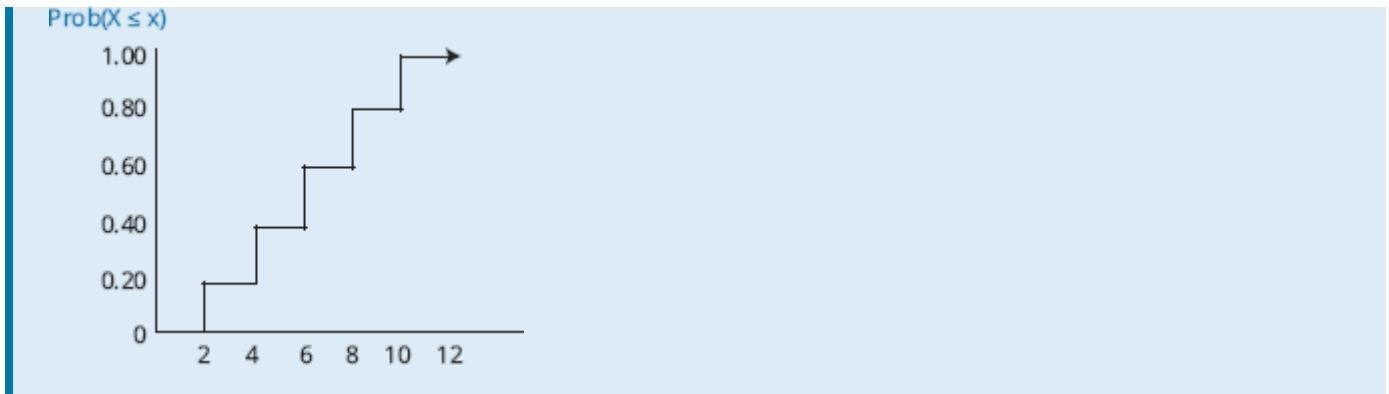
Answer:

$p(6) = 0.2$, since $p(x) = 0.2$ for all x . $F(6) = P(X \leq 6) = np(x) = 3(0.2) = 0.6$. Note that $n = 3$ since 6 is the third outcome in the range of possible outcomes. $P(2 \leq X \leq 8) = 4(0.2) = 0.8$. Note that $k = 4$, since there are four outcomes in the range $2 \leq X \leq 8$. The following figures illustrate the concepts of a probability function and cumulative distribution function for this distribution.

Probability and Cumulative Distribution Functions

$X = x$	Probability of x Prob ($X = x$)	Cumulative Distribution Function Prob ($X < x$)
2	0.20	0.20
4	0.20	0.40
6	0.20	0.60
8	0.20	0.80

Cumulative Distribution Function for $X \sim \text{Uniform } \{2, 4, 6, 8, 10\}$



LOS 4.d: Describe the properties of the continuous uniform distribution, and calculate and interpret probabilities given a continuous uniform distribution.

The **continuous uniform distribution** is defined over a range that spans between some lower limit, a , and some upper limit, b , which serve as the parameters of the distribution. Outcomes can only occur between a and b , and since we are dealing with a continuous distribution, even if $a < x < b$, $P(X = x) = 0$. Formally, the properties of a continuous uniform distribution may be described as follows:

- For all $a \leq x_1 < x_2 \leq b$ (i.e., for all x_1 and x_2 between the boundaries a and b).
- $P(X < a \text{ or } X > b) = 0$ (i.e., the probability of X outside the boundaries is zero).
- $P(x_1 \leq X \leq x_2) = (x_2 - x_1)/(b - a)$. This defines the probability of outcomes between x_1 and x_2 .

Don't miss how simple this is just because the notation is so mathematical. For a continuous uniform distribution, the probability of outcomes in a range that is one-half the whole range is 50%. The probability of outcomes in a range that is one-quarter as large as the whole possible range is 25%.

EXAMPLE: Continuous uniform distribution

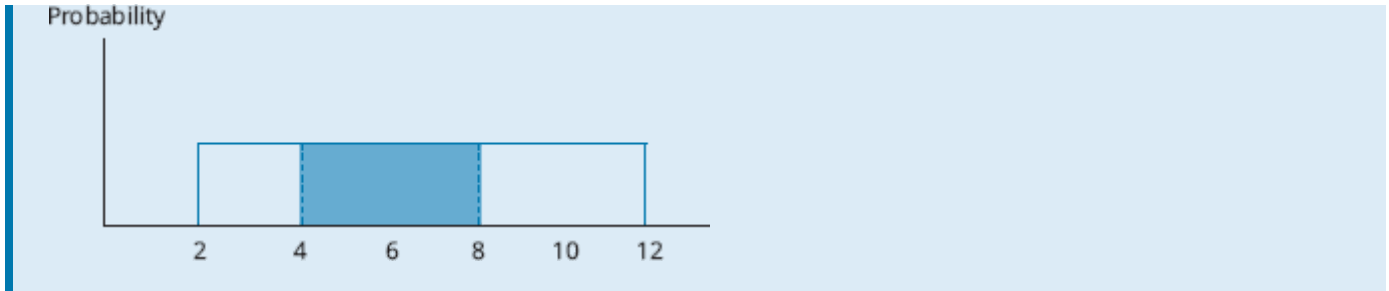
X is uniformly distributed between 2 and 12. **Calculate** the probability that X will be between 4 and 8.

Answer:

$$\frac{8 - 4}{12 - 2} = \frac{4}{10} = 40\%$$

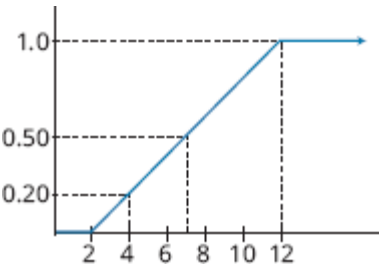
The figure below illustrates this continuous uniform distribution. Note that the area bounded by 4 and 8 is 40% of the total probability between 2 and 12 (which is 100%).

Continuous Uniform Distribution



Because outcomes are equal over equal-size possible intervals, the cdf is linear over the variable's range. The cdf for the distribution in the example, $\text{Prob}(X < x)$, is shown in Figure 4.2.

Figure 4.2: CDF for a Continuous Uniform Variable



LOS 4.e: Describe the properties of a Bernoulli random variable and a binomial random variable, and calculate and interpret probabilities given the binomial distribution function.

A **binomial random variable** may be defined as the number of “successes” in a given number of trials, whereby the outcome can be either “success” or “failure.” The probability of success, p , is constant for each trial, and the trials are independent. A binomial random variable for which the number of trials is 1 is called a **Bernoulli random variable**. Think of a trial as a mini-experiment (or “Bernoulli trial”). The final outcome is the number of successes in a series of n trials. Under these conditions, the binomial probability function defines the probability of x successes in n trials. It can be expressed using the following formula:

$$p(x) = P(X = x) = (\text{number of ways to choose } x \text{ from } n)p^x(1 - p)^{n-x}$$

where:

(number of ways to choose x from n) =

$$\frac{n!}{(n-x)!x!} \text{ which may also be denoted as } \binom{n}{x} \text{ or stated as “} n \text{ choose } x \text{”}$$

p = the probability of “success” on each trial [don’t confuse it with $p(x)$]

So the probability of exactly x successes in n trials is:

$$p(x) = \frac{n!}{(n-x)!x!} p^x(1 - p)^{n-x}$$

EXAMPLE: Binomial probability

Assuming a binomial distribution, compute the probability of drawing three black beans from a bowl of black and white beans if the probability of selecting a black bean in any given attempt is 0.6. You will draw five beans from the bowl.

Answer:

$$P(X = 3) = p(3) = \frac{5!}{2!3!}(0.6)^3(0.4)^2 = (120 / 12)(0.216)(0.160) = 0.3456$$

Some intuition about these results may help you remember the calculations. Consider that a (very large) bowl of black and white beans has 60% black beans and that each time you select a bean, you replace it in the bowl before drawing again. We want to know the probability of selecting exactly three black beans in five draws, as in the previous example.

One way this might happen is BBBWW. Since the draws are independent, the probability of this is easy to calculate. The probability of drawing a black bean is 60%, and the probability of drawing a white bean is $1 - 60\% = 40\%$. Therefore, the probability of selecting BBBWW, in order, is $0.6 \times 0.6 \times 0.6 \times 0.4 \times 0.4 = 3.456\%$. This is the $p^3(1 - p)^2$ from the formula and p is 60%, the probability of selecting a black bean on any single draw from the bowl. BBBWW is not, however, the only way to choose exactly three black beans in five trials. Another possibility is BBWWB, and a third is BWWBB. Each of these will have exactly the same probability of occurring as our initial outcome, BBBWW. That's why we need to answer the question of how many ways (different orders) there are for us to choose

three black beans in five draws. Using the formula, there are $\frac{5!}{3!(5 - 3)!} = 10$ ways; $10 \times 3.456\% = 34.56\%$, the answer we computed above.

Expected Value and Variance of a Binomial Random Variable

For a given series of n trials, the expected number of successes, or $E(X)$, is given by the following formula:

$$\text{expected value of } X = E(X) = np$$

The intuition is straightforward; if we perform n trials and the probability of success on each trial is p , we expect np successes.

The variance of a binomial random variable is given by:

$$\text{variance of } X = np(1 - p)$$

EXAMPLE: Expected value of a binomial random variable

Based on empirical data, the probability that the Dow Jones Industrial Average (DJIA) will increase on any given day has been determined to equal 0.67. Assuming that the only other outcome is that it decreases, we can state $p(\text{UP}) = 0.67$ and $p(\text{DOWN}) = 0.33$. Further, assume that movements in the DJIA are independent (i.e., an increase in one day is independent of what happened on another day).

Using the information provided, compute the expected value of the number of up days in a 5-day period.

Answer:

Using binomial terminology, we define success as UP, so $p = 0.67$. Note that the definition of success is critical to any binomial problem.

$$E(X | n = 5, p = 0.67) = (5)(0.67) = 3.35$$

Recall that the “|” symbol means *given*. Hence, the preceding statement is read as: the expected value of X given that $n = 5$, and the probability of success = 67% is 3.35.

We should note that since the binomial distribution is a discrete distribution, the result $X = 3.35$ is not possible. However, if we were to record the results of many 5-day periods, the average number of up days (successes) would converge to 3.35.



MODULE QUIZ 4.1

- Which of the following is *least likely* an example of a discrete random variable?
 - The number of stocks a person owns.
 - The time spent by a portfolio manager with a client.
 - The number of days it rains in a month in Iowa City.
- For a continuous random variable X , the probability of any single value of X is:
 - one.
 - zero.
 - determined by the cdf.
- Which of the following is *least likely* a probability distribution?
 - $X = [1,2,3,4]$; $\text{Prob}[X_i] = \frac{x_i^2}{30}$.
 - $X = [5,10]$; $\text{Prob}[X_i] = \frac{8^{-x_i}}{5}$.
 - $X = [5,10]$; $\text{Prob}[X_i] = \frac{x_i - 3}{9}$.

Use the following table to answer Questions 4 through 7.

Probability distribution of a discrete random variable X								
X	0	1	2	3	4	5	6	7
$P(X)$	0.04	0.11	0.18	0.24	0.14	0.17	0.09	0.03

- The cdf of 5, or $F(5)$ is:
 - 0.17.
 - 0.71.
 - 0.88.
- The probability that X is *greater* than 3 is:
 - 0.24.
 - 0.43.
 - 0.67.
- What is $P(2 \leq X \leq 5)$?
 - 0.17.
 - 0.38.
 - 0.73.
- The expected value of the random variable X is:
 - 3.35.
 - 3.70.
 - 5.47.
- A continuous uniform distribution has the parameters $a = 4$ and $b = 10$. The $F(20)$ is:
 - 0.25.

- B. 0.50.
C. 1.00.
9. Which of the following is *least likely* a condition of a binomial experiment?
A. There are only two trials.
B. The trials are independent.
C. If p is the probability of success, and q is the probability of failure, then $p + q = 1$.
10. Which of the following statements *least accurately* describes the binomial distribution?
A. It is a discrete distribution.
B. The probability of an outcome of zero is zero.
C. The combination formula is used in computing probabilities.
11. A recent study indicated that 60% of all businesses have a fax machine. From the binomial probability distribution table, the probability that exactly four businesses will have a fax machine in a random selection of six businesses is:
A. 0.138.
B. 0.276.
C. 0.311.
12. Ten percent of all college graduates hired stay with the same company for more than five years. In a random sample of six recently hired college graduates, the probability that exactly two will stay with the same company for more than five years is *closest to*:
A. 0.098.
B. 0.114.
C. 0.185.
13. Assume that 40% of candidates who sit for the CFA[®] examination pass it the first time. Of a random sample of 15 candidates who are sitting for the exam for the first time, what is the expected number of candidates that will pass?
A. 0.375.
B. 4.000.
C. 6.000.

MODULE 4.2: NORMAL DISTRIBUTIONS



LOS 4.f: Explain the key properties of the normal distribution.

Video covering this content is available online.

The normal distribution is important for many reasons. Many of the random variables that are relevant to finance and other professional disciplines follow a normal distribution. In the area of investment and portfolio management, the normal distribution plays a central role in portfolio theory.

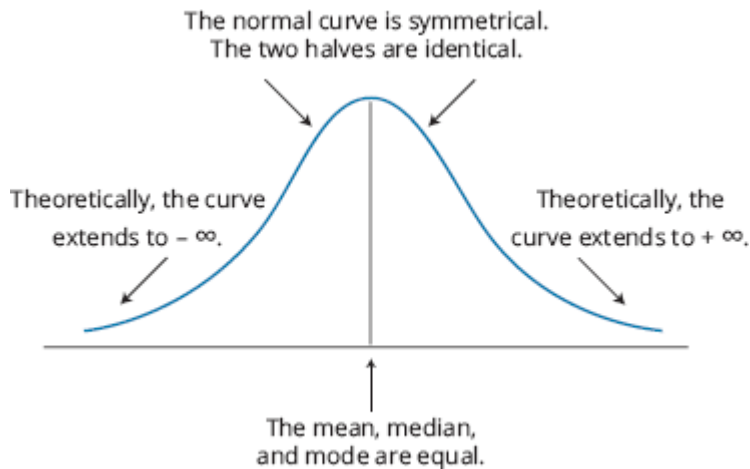
The **normal distribution** has the following key properties:

- It is completely described by its mean, μ , and variance, σ^2 , stated as $X \sim N(\mu, \sigma^2)$. In words, this says that “ X is normally distributed with mean μ and variance σ^2 .”
- Skewness = 0, meaning that the normal distribution is symmetric about its mean, so that $P(X \leq \mu) = P(\mu \leq X) = 0.5$, and mean = median = mode.
- Kurtosis = 3; this is a measure of how flat the distribution is. Recall that excess kurtosis is measured relative to 3, the kurtosis of the normal distribution.
- A linear combination of normally distributed random variables is also normally distributed.

- The probabilities of outcomes farther above and below the mean get smaller and smaller but do not go to zero (the tails get very thin but extend infinitely).

Many of these properties are evident from examining the graph of a normal distribution's probability density function as illustrated in Figure 4.3.

Figure 4.3: Normal Distribution Probability Density Function



LOS 4.g: Contrast a multivariate distribution and a univariate distribution, and explain the role of correlation in the multivariate normal distribution.

Up to this point, our discussion has been strictly focused on **univariate distributions** (i.e., the distribution of a single random variable). In practice, however, the relationships between two or more random variables are often relevant. For instance, investors and investment managers are frequently interested in the interrelationship among the returns of one or more assets. In fact, as you will see in your study of asset pricing models and modern portfolio theory, the return on a given stock and the return on the S&P 500 or some other market index will have special significance. Regardless of the specific variables, the simultaneous analysis of two or more random variables requires an understanding of multivariate distributions.

A **multivariate distribution** specifies the probabilities associated with a group of random variables and is meaningful only when the behavior of each random variable in the group is in some way dependent on the behavior of the others. Both discrete and continuous random variables can have multivariate distributions. Multivariate distributions between two discrete random variables are described using joint probability tables. For continuous random variables, a multivariate *normal* distribution may be used to describe them if all the individual variables follow a normal distribution. As previously mentioned, one of the characteristics of a normal distribution is that a linear combination of normally distributed random variables is normally distributed as well. For example, if the return of each stock in a portfolio is normally distributed, the return on the portfolio will also be normally distributed.

The Role of Correlation in the Multivariate Normal Distribution

Similar to a univariate normal distribution, a multivariate normal distribution can be described by the mean and variance of the individual random variables. Additionally, it is necessary to

specify the correlation between the individual pairs of variables when describing a multivariate distribution. Correlation is the feature that distinguishes a multivariate distribution from a univariate normal distribution. Correlation indicates the strength of the linear relationship between a pair of random variables.

Using asset returns as our random variables, the multivariate normal distribution for the returns on n assets can be completely defined by the following three sets of parameters:

- n means of the n series of returns $(\mu_1, \mu_2, \dots, \mu_n)$.
- n variances of the n series of returns $(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$.
- $0.5n(n - 1)$ pair-wise correlations.

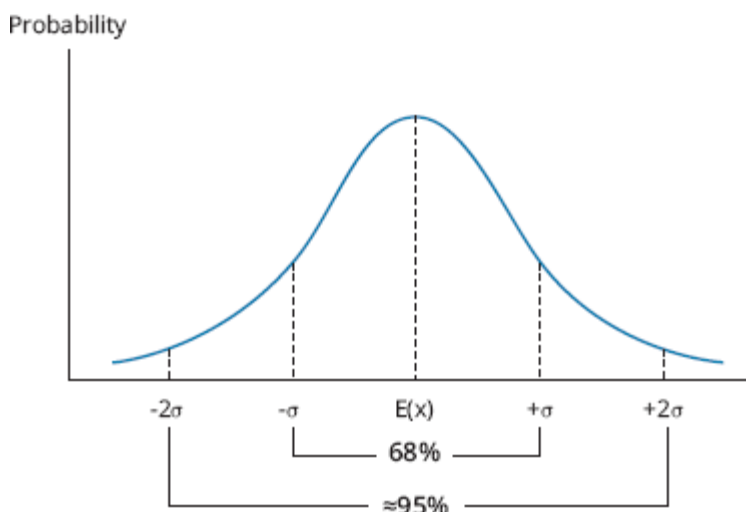
For example, if there are two assets, $n = 2$, then the multivariate returns distribution can be described with two means, two variances, and one correlation $[0.5(2)(2 - 1) = 1]$. If there are four assets, $n = 4$, the multivariate distribution can be described with four means, four variances, and six correlations $[0.5(4)(4 - 1) = 6]$. When building a portfolio of assets, all other things being equal, it is desirable to combine assets having low returns correlation because this will result in a portfolio with a lower variance than one composed of assets with higher correlations.

LOS 4.h: Calculate the probability that a normally distributed random variable lies inside a given interval.

A **confidence interval** is a range of values around the expected outcome within which we expect the actual outcome to be some specified percentage of the time. A 95% confidence interval is a range that we expect the random variable to be in 95% of the time. For a normal distribution, this interval is based on the expected value (sometimes called a point estimate) of the random variable and on its variability, which we measure with standard deviation.

Confidence intervals for a normal distribution are illustrated in Figure 4.4. For any normally distributed random variable, 68% of the outcomes are within one standard deviation of the expected value (mean), and approximately 95% of the outcomes are within two standard deviations of the expected value.

Figure 4.4: Confidence Intervals for a Normal Distribution



In practice, we will not know the actual values for the mean and standard deviation of the distribution, but will have estimated them as \bar{x} and s . The three confidence intervals of most interest are given by the following:

- The 90% confidence interval for \bar{x} is $\bar{x} - 1.65s$ to $\bar{x} + 1.65s$.
- The 95% confidence interval for \bar{x} is $\bar{x} - 1.96s$ to $\bar{x} + 1.96s$.
- The 99% confidence interval for \bar{x} is $\bar{x} - 2.58s$ to $\bar{x} + 2.58s$.

EXAMPLE: Confidence intervals

The average return of a mutual fund is 10.5% per year and the standard deviation of annual returns is 18%. If returns are approximately normal, what is the 95% confidence interval for the mutual fund return next year?

Answer:

Here μ and σ are 10.5% and 18%, respectively. Thus, the 95% confidence interval for the return, R , is:

$$10.5 \pm 1.96(18) = -24.78\% \text{ to } 45.78\%$$

Symbolically, this result can be expressed as:

$$P(-24.78 < R < 45.78) = 0.95 \text{ or } 95\%$$

The interpretation is that the annual return is expected to be within this interval 95% of the time, or 95 out of 100 years.

LOS 4.i: Explain how to standardize a random variable.

The **standard normal distribution** is a normal distribution that has been standardized so that it has a mean of zero and a standard deviation of 1 [i.e., $N(0,1)$]. To standardize an observation from a given normal distribution, the *z-value* of the observation must be calculated. The *z-value* represents the number of standard deviations a given observation is from the population mean. *Standardization* is the process of converting an observed value for a random variable to its *z-value*. The following formula is used to standardize a random variable:

$$z = \frac{\text{observation} - \text{population mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$



PROFESSOR'S NOTE

The term *z-value* will be used for a standardized observation in this document. The terms *z-score* and *z-statistic* are also commonly used.

EXAMPLE: Standardizing a random variable (calculating z-values)

Assume that the annual earnings per share (EPS) for a population of firms are normally distributed with a mean of \$6 and a standard deviation of \$2.

What are the z-values for EPS of \$2 and \$8?

Answer:

If EPS = $x = \$8$, then $z = (x - \mu) / \sigma = (\$8 - \$6) / \$2 = +1$

If EPS = $x = \$2$, then $z = (x - \mu) / \sigma = (\$2 - \$6) / \$2 = -2$

Here, $z = +1$ indicates that an EPS of \$8 is one standard deviation above the mean, and $z = -2$ means that an EPS of \$2 is two standard deviations below the mean.

LOS 4.j: Calculate and interpret probabilities using the standard normal distribution.

Now we will show how to use standardized values (z-values) and a table of probabilities for Z to determine probabilities. A portion of a table of the cumulative distribution function for a standard normal distribution is shown in Figure 4.5. We will refer to this table as the z-table, as it contains values generated using the cumulative distribution function for a standard normal distribution, denoted by $F(Z)$. Thus, the values in the z-table are the probabilities of observing a z-value that is less than a given value, z [i.e., $P(Z < z)$]. The numbers in the first column are z-values that have only one decimal place. The columns to the right supply probabilities for z-values with two decimal places.

Note that the z-table in Figure 4.5 only provides probabilities for positive z-values. This is not a problem because we know from the symmetry of the standard normal distribution that $F(-Z) = 1 - F(Z)$. The tables in the back of many texts actually provide probabilities for negative z-values, but we will work with only the positive portion of the table because this may be all you get on the exam. In Figure 4.5, we can find the probability that a standard normal random variable will be less than 1.66, for example. The table value is 95.15%. The probability that the random variable will be less than -1.66 is simply $1 - 0.9515 = 0.0485 = 4.85\%$, which is also the probability that the variable will be greater than $+1.66$.

Figure 4.5: Cumulative Probabilities for a Standard Normal Distribution

Cdf Values for the Standard Normal Distribution: The z-Table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.5	.6915	Please note that several of the rows have been deleted to save space.*								
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

*A complete cumulative standard normal table is included in Appendix A.

PROFESSOR'S NOTE



When you use the standard normal probabilities, you have formulated the problem in terms of standard deviations from the mean. Consider a security with returns that are approximately normal, an expected return of 10%, and standard deviation of returns of 12%. The probability of returns greater than 30% is calculated based on the number of standard deviations that 30% is above the expected return of 10%. 30% is 20% above the expected return of 10%, which is $20 / 12 = 1.67$ standard deviations above the mean. We look up the probability of returns less than 1.67 standard deviations above the mean (0.9525 or 95.25% from Figure 4.5) and calculate the probability of returns more than 1.67 standard deviations above the mean as $1 - 0.9525 = 4.75\%$.

EXAMPLE: Using the z-table (1)

Considering again EPS distributed with $\mu = \$6$ and $\sigma = \$2$, what is the probability that EPS will be \$9.70 or more?

Answer:

Here we want to know $P(\text{EPS} > \$9.70)$, which is the area under the curve to the right of the z-value corresponding to $\text{EPS} = \$9.70$ (see the following figure).

The z-value for $\text{EPS} = \$9.70$ is:

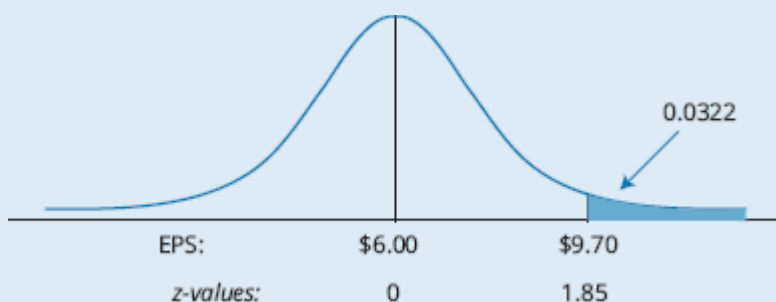
$$z = \frac{x - \mu}{\sigma} = \frac{9.70 - 6}{2} = 1.85$$

That is, \$9.70 is 1.85 standard deviations above the mean EPS value of \$6.

From the z-table we have $F(1.85) = 0.9678$, but this is $P(\text{EPS} \leq 9.70)$. We want $P(\text{EPS} > 9.70)$, which is $1 - P(\text{EPS} \leq 9.70)$.

$$P(\text{EPS} > 9.70) = 1 - 0.9678 = 0.0322, \text{ or } 3.2\%$$

$P(\text{EPS} > \$9.70)$



EXAMPLE: Using the z-table (2)

Using the distribution of EPS with $\mu = \$6$ and $\sigma = \$2$ again, what percentage of the observed EPS values are likely to be less than \$4.10?

Answer:

As shown graphically in the following figure, we want to know $P(\text{EPS} < \$4.10)$. This requires a 2-step approach like the one taken in the preceding example.

First, the corresponding z-value must be determined as follows:

$$z = \frac{\$4.10 - \$6}{\$2} = -0.95,$$

so \$4.10 is 0.95 standard deviations below the mean of \$6.00.

Now, from the z-table for negative values in the back of this book, we find that $F(-0.95) = 0.1711$, or 17.11%.

Finding a Left-Tail Probability



The z-table gives us the probability that the outcome will be more than 0.95 standard deviations below the mean.

LOS 4.k: Define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy's safety-first criterion.

Shortfall risk is the probability that a portfolio value or return will fall below a particular (target) value or return over a given time period.

Roy's safety-first criterion states that the optimal portfolio minimizes the probability that the return of the portfolio falls below some minimum acceptable level. This minimum acceptable level is called the **threshold level**. Symbolically, Roy's safety-first criterion can be stated as:

$$\text{minimize } P(R_p < R_L)$$

where:

R_p = portfolio return

R_L = threshold level return

If portfolio returns are normally distributed, then Roy's safety-first criterion can be stated as:

$$\text{maximize the SFRatio, where } \text{SFRatio} = \frac{E(R_p) - R_L}{\sigma_p}$$

The reasoning behind the safety-first criterion is illustrated in Figure 4.6. Assume an investor is choosing between two portfolios: Portfolio A with expected return of 12% and standard deviation of returns of 18%, and Portfolio B with expected return of 10% and standard deviation of returns of 12%. The investor has stated that he wants to minimize the probability of losing money (negative returns). Assuming that returns are normally distributed, the portfolio with the larger SFR using 0% as the threshold return (R_L) will be the one with the lower probability of negative returns.

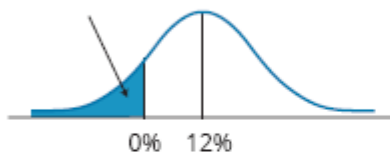
Figure 4.6: The Safety-First Criterion and Shortfall Risk

A. Normally Distributed Returns

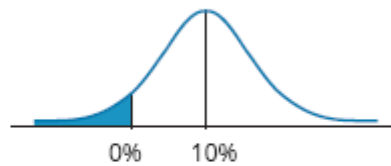
Portfolio A: $E(R) = 12\%$ $\sigma_A = 18\%$

Portfolio B: $E(R) = 10\%$ $\sigma_B = 12\%$

Probability of returns $< 0\%$
- i.e. short fall risk

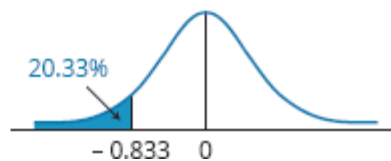
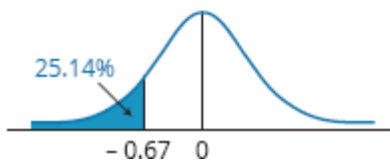


$$SFR_A = \frac{12 - 0}{18} = 0.667$$



$$SFR_B = \frac{10 - 0}{12} = 0.833$$

B. Standard Normal



Panel B of Figure 4.6 relates the SFRatio to the standard normal distribution. Note that the SFR is the number of standard deviations *below* the mean. Thus, the portfolio with the larger SFR has the lower probability of returns below the threshold return, which is a return of 0% in our example. Using a z-table for negative values, we can find the probabilities in the left-hand tails as indicated. These probabilities (25% for Portfolio A and 20% for Portfolio B) are also the shortfall risk for a target return of 0%, that is, the probability of negative returns. Portfolio B has the higher SFR which means it has the lower probability of negative returns.

In summary, when choosing among portfolios with normally distributed returns using Roy's safety-first criterion, there are two steps:

Step 1: Calculate the SFRatio $= \frac{E(R_P) - R_L}{\sigma_P}$

Step 2: Choose the portfolio that has the *largest* SFRatio.

EXAMPLE: Roy's safety-first criterion

For the next year, the managers of a \$120 million college endowment plan have set a minimum acceptable end-of-year portfolio value of \$123.6 million. Three portfolios are being considered which have the expected returns and standard deviation shown in the first two rows of the following table. Determine which of these portfolios is the most desirable using Roy's safety-first criterion and the probability that the portfolio value will fall short of the target amount.

Answer:

The threshold return is $R_L = (123.6 - 120) / 120 = 0.030 = 3\%$. The SFRs are shown in the table below. As indicated, the best choice is Portfolio A because it has the largest SFR.

Roy's Safety-First Ratios

Portfolio	Portfolio A	Portfolio B	Portfolio C
$E(R_p)$	9%	11%	6.6%
σ_p	12%	20%	8.2%
SFRatio	$0.5 = \frac{9-3}{12}$	$0.4 = \frac{11-3}{20}$	$0.44 = \frac{6.6-3}{8.2}$

The probability of an ending value for Portfolio A less than \$123.6 million (a return less than 3%) is simply $F(-0.5)$, which we can find on the z-table for negative values. The probability is $0.3085 = 30.85\%$.



MODULE QUIZ 4.2

- A key property of a normal distribution is that it:
 - has zero skewness.
 - is asymmetrical.
 - has zero kurtosis.
- Which of the following parameters is necessary to describe a multivariate normal distribution?
 - Beta.
 - Correlation.
 - Degrees of freedom.

Use the following table to answer Question 3.

z	0.00	0.01	0.02	0.03	0.04
1.0	0.8413	0.8438	0.8461	0.8485	0.8508
1.1	0.8643	0.8665	0.8686	0.8708	0.8729
1.2	0.8849	0.8869	0.8888	0.8907	0.8925

- A study of hedge fund investors found that their annual household incomes are normally distributed with a mean of \$175,000 and a standard deviation of \$25,000. The percentage of hedge fund investors that have incomes greater than \$150,000 is *closest* to:
 - 34.13%.
 - 68.26%.
 - 84.13%.
- For the standard normal distribution, the z -value gives the distance between the mean and a point in terms of:
 - the variance.
 - the standard deviation.
 - the center of the curve.
- For a standard normal distribution, $F(0)$ is:
 - 0.0.
 - 0.1.
 - 0.5.

Use the following table to answer Questions 6 and 7.

Portfolio	Portfolio A	Portfolio B	Portfolio C
$E(R_p)$	5%	11%	18%
σ_p	8%	21%	40%

6. Given a threshold level of return of 4%, use Roy's safety-first criterion to choose the optimal portfolio.
 - A. Portfolio A.
 - B. Portfolio B.
 - C. Portfolio C.
7. Given a threshold level of return of 0%, use Roy's safety-first criterion to choose the optimal portfolio.
 - A. Portfolio A.
 - B. Portfolio B.
 - C. Portfolio C.

MODULE 4.3: LOGNORMAL, T, CHI-SQUARE, AND F DISTRIBUTIONS



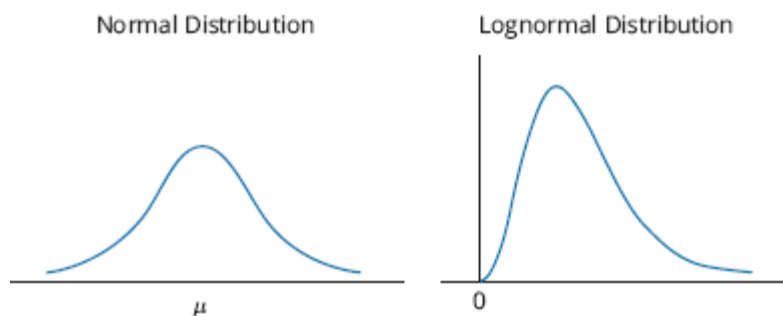
Video covering this content is available online.

LOS 4.1: Explain the relationship between normal and lognormal distributions and why the lognormal distribution is used to model asset prices.

The **lognormal distribution** is generated by the function e^x , where x is normally distributed. Since the natural logarithm, \ln , of e^x is x , the logarithms of lognormally distributed random variables are normally distributed, thus the name.

Figure 4.7 illustrates the differences between a normal distribution and a lognormal distribution.

Figure 4.7: Normal vs. Lognormal Distributions



In Figure 4.7, we can see that:

- The lognormal distribution is skewed to the right.
- The lognormal distribution is bounded from below by zero so that it is useful for modeling asset prices, which never take negative values.

If we used a normal distribution of returns to model asset prices over time, we would admit the possibility of returns less than -100% , which would admit the possibility of asset prices less than zero. Using a lognormal distribution to model *price relatives* avoids this problem. A price relative is just the end-of-period price of the asset divided by the beginning price (S_1/S_0) and is equal to $(1 + \text{the holding period return})$. To get the end-of-period asset price, we can simply multiply the price relative times the beginning-of-period asset price. Since a lognormal distribution takes a minimum value of zero, end-of-period asset prices cannot be less than zero. A price relative of zero corresponds to a holding period return of -100% (i.e., the asset price has gone to zero). Recall that we used price relatives as the up-move and down-move (multiplier) terms in constructing a binomial tree for stock price changes over a number of periods.

LOS 4.m: Calculate and interpret a continuously compounded rate of return, given a specific holding period return.

Discretely compounded returns are just the compound returns we are familiar with, given some discrete compounding period, such as semiannual or quarterly. Recall that the more frequent the compounding, the greater the effective annual return. For a stated rate of 10% , semiannual compounding results in an effective yield of $\left(1 + \frac{0.10}{2}\right)^2 - 1 = 10.25\%$ and monthly compounding results in an effective yield of $\left(1 + \frac{0.10}{12}\right)^{12} - 1 = 10.47\%$. Daily or even hourly compounding will produce still larger effective yields. The limit of this exercise, as the compounding periods get shorter and shorter, is called **continuous compounding**. The effective annual rate, based on continuous compounding for a stated annual rate of R_{cc} , can be calculated from the formula:

$$\text{effective annual rate} = e^{R_{cc}} - 1$$

Based on a stated rate of 10% , the effective rate with continuous compounding is $e^{0.10} - 1 = 10.5171\%$. Please verify this by entering 0.1 in your calculator and finding the e^x function.

Since the natural log, \ln , of e^x is x , we can get the continuously compounded rate from an effective annual rate by using the \ln calculator function. Using our previous example, $\ln(1 + 10.517\%) = \ln 1.105171 = 10\%$. Verify this by entering 1.105171 in your calculator and then entering the \ln key.

We can use this method to find the continuously compounded rate that will generate a particular holding period return. If we are given a holding period return of 12.5% for the year, the equivalent continuously compounded rate is $\ln 1.125 = 11.778\%$. Since the calculation is based on 1 plus the holding period return, we can also do the calculation directly from the *price relative*. The price relative is just the end-of-period value divided by the beginning-of-period value. The continuously compounded rate of return is:

$$\ln\left(\frac{S_1}{S_0}\right) = \ln(1 + \text{HPR}) = R_{cc}$$

EXAMPLE: Calculating continuously compounded returns

A stock was purchased for \$100 and sold one year later for \$120. Calculate the investor's annual rate of return on a continuously compounded basis.

Answer:

$$\ln\left(\frac{120}{100}\right) = 18.232\%$$

If we had been given the return (20%) instead, the calculation is:

$$\ln(1 + 0.20) = 18.232\%$$

One property of continuously compounded rates of return is that they are additive for multiple periods. Note that the (effective) holding period return over two years is calculated by doubling the continuously compounded annual rate. If $R_{cc} = 10\%$, the (effective) holding period return over two years is $e^{(0.10)2} - 1 = 22.14\%$. In general, the holding period return after T years, when the annual continuously compounded rate is R_{cc} , is given by:

$$HPR_T = e^{R_{cc} \times T} - 1$$

Given investment results over a 2-year period, we can calculate the 2-year continuously compounded return and divide by 2 to get the annual rate. Consider an investment that appreciated from \$1,000 to \$1,221.40 over a 2-year period. The 2-year continuously compounded rate is $\ln(1,221.40 / 1,000) = 20\%$, and the annual continuously compounded rate (R_{cc}) is $20\% / 2 = 10\%$.

LOS 4.n: Describe the properties of the Student's t -distribution, and calculate and interpret its degrees of freedom.

Student's t -distribution, or simply the t -distribution, is a bell-shaped probability distribution that is symmetrical about its mean. It is the appropriate distribution to use when constructing confidence intervals based on *small samples* ($n < 30$) from populations with *unknown variance* and a normal, or approximately normal, distribution. It may also be appropriate to use the t -distribution when the population variance is unknown and the sample size is large enough that the central limit theorem will assure that the sampling distribution is approximately normal.

Student's t -distribution has the following properties:

- It is symmetrical.
- It is defined by a single parameter, the **degrees of freedom (df)**, where the degrees of freedom are equal to the number of sample observations minus 1, $n - 1$, for sample means.
- It has more probability in the tails ("fatter tails") than the normal distribution.
- As the degrees of freedom (the sample size) gets larger, the shape of the t -distribution more closely approaches a standard normal distribution.

When *compared to the normal distribution*, the t -distribution is flatter with more area under the tails (i.e., it has fatter tails). As the degrees of freedom for the t -distribution increase, however, its shape approaches that of the normal distribution.

The degrees of freedom for tests based on sample means are $n - 1$ because, given the mean, only $n - 1$ observations can be unique.

The t -distribution is a symmetrical distribution that is centered about zero. The shape of the t -distribution is dependent on the number of degrees of freedom, and degrees of freedom are based on the number of sample observations. The t -distribution is flatter and has thicker tails than the standard normal distribution. As the number of observations increases (i.e., the degrees of freedom increase), the t -distribution becomes more spiked and its tails become thinner. As the number of degrees of freedom increases without bound, the t -distribution converges to the standard normal distribution (z -distribution). The thickness of the tails relative to those of the z -distribution is important in hypothesis testing because thicker tails mean more observations away from the center of the distribution (more outliers). Hence, hypothesis testing using the t -distribution makes it more difficult to reject the null relative to hypothesis testing using the z -distribution.

The table in Figure 4.8 contains one-tailed critical values for the t -distribution at the 0.05 and 0.025 levels of significance with various degrees of freedom (df). Note that, unlike the z -table, the t -values are contained within the table, and the probabilities are located at the column headings.

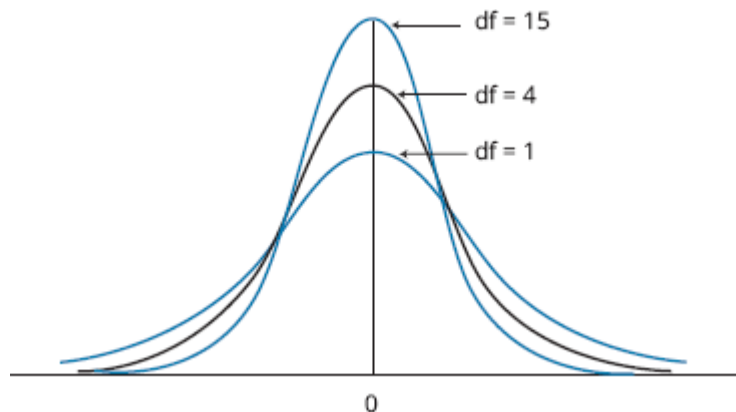
Note that the significance level for a two-tailed test is two times the one-tail probabilities. As degrees of freedom increase, the values under the column for one-tail probabilities of 2.5% approach 1.96, our critical value for a two-tailed test for a normally distributed variable at the 5% significance level. We can see that the critical values increase as the degrees of freedom decrease. Confidence intervals for a t -statistic are wider than those for a z -statistic. Because the tails of the t -distribution are fatter than those for a normally distributed test statistic, we must increase the width of the confidence interval to leave a given percentage of the outcomes in each tail.

Figure 4.8: Table of Critical t -Values

df	One-Tailed Probabilities, p	
	$p = 0.05$	$p = 0.025$
5	2.015	2.571
10	1.812	2.228
15	1.753	2.131
20	1.725	2.086
25	1.708	2.060
30	1.697	2.042
40	1.684	2.021
50	1.676	2.009
60	1.671	2.000
70	1.667	1.994
80	1.664	1.990
90	1.662	1.987
100	1.660	1.984
120	1.658	1.980
∞	1.645	1.960

Figure 4.9 illustrates the different shapes of the t -distribution associated with different degrees of freedom. The tendency is for the t -distribution to look more and more like the normal distribution as the degrees of freedom increase. Practically speaking, the greater the degrees of freedom, the greater the percentage of observations near the center of the distribution and the lower the percentage of observations in the tails, which are thinner as degrees of freedom increase. This means that confidence intervals for a random variable that follows a t -distribution must be wider (narrower) when the degrees of freedom are less (more) for a given significance level.

Figure 4.9: t -Distributions for Different Degrees of Freedom (df)



LOS 4.o: Describe the properties of the chi-square distribution and the F -distribution, and calculate and interpret their degrees of freedom.

Like the t -distribution, a **chi-square distribution** (χ^2) is a family of distributions, each based on degrees of freedom. The chi-square distribution is the distribution of the sum of the squared values of n random variables, and k , the degrees of freedom, is equal to $n - 1$.

Because it is the sum of squared values, the chi-square distribution is bounded from below by zero. It is typically asymmetric, but its symmetry increases with the degrees of freedom. As degrees of freedom get larger, the chi-square distribution approaches the normal distribution in shape. The chi-square distribution is often used in tests of the value of the variance of a normally distributed population.

The **F -distribution** is the distribution of the quotient of two (appropriately scaled)

independent chi-square variables with degrees of freedom m and n :
$$F = \frac{\chi^2/m}{\chi^2/n}$$

where the numerator is a χ^2 variable with m degrees of freedom and the denominator is a χ^2 variable with n degrees of freedom. A common use of the F -distribution is to determine the probability that the variances of two independent normal distributions are equal.

The table of values for the F -distribution is constructed with the degrees of freedom for the numerator in the top margin, the degrees of freedom for the denominator in the side margin, and

the F -distribution values at the intersections of the degrees of freedom. Each F -distribution table is given for a specific level of significance.

The F -distribution, because it is the ratio of two chi-square values, cannot take on negative values. Therefore, like the chi-square distribution, it is bounded from below by zero. The F -distribution is also asymmetric. As the numerator and denominator degrees of freedom increase, the F -distribution becomes more symmetric and its shape becomes more like the bell curve of a normal distribution.



PROFESSOR'S NOTE

Tables for chi-square and F distributions appear in the Appendix to this book. We will use these distributions in our reading on Hypothesis Testing.

LOS 4.p: Describe Monte Carlo simulation.

Monte Carlo simulation is a technique based on the repeated generation of one or more risk factors that affect security values, in order to generate a distribution of security values. For each of the risk factors, the analyst must specify the parameters of the probability distribution that the risk factor is assumed to follow. A computer is then used to generate random values for each risk factor based on its assumed probability distributions. Each set of randomly generated risk factors is used with a pricing model to value the security. This procedure is repeated many times (100s, 1,000s, or 10,000s), and the distribution of simulated asset values is used to draw inferences about the expected (mean) value of the security and possibly the variance of security values about the mean as well.

As an example, consider the valuation of stock options that can only be exercised on a particular date. The main risk factor is the value of the stock itself, but interest rates could affect the valuation as well. The simulation procedure would be to:

1. Specify the probability distributions of stock prices and of the relevant interest rate, as well as the parameters (mean, variance, possibly skewness) of the distributions.
2. Randomly generate values for both stock prices and interest rates.
3. Value the options for each pair of risk factor values.
4. After many iterations, calculate the mean option value and use that as your estimate of the option's value.

Monte Carlo simulation is used to:

- Value complex securities.
- Simulate the profits/losses from a trading strategy.
- Calculate estimates of value at risk (VaR) to determine the riskiness of a portfolio of assets and liabilities.
- Simulate pension fund assets and liabilities over time to examine the variability of the difference between the two.
- Value portfolios of assets that have nonnormal returns distributions.

The limitations of Monte Carlo simulation are that it is fairly complex and will provide answers that are no better than the assumptions about the distributions of the risk factors and the pricing/valuation model that is used. Also, simulation is not an analytic method, but a statistical one, and cannot provide the insights that analytic methods can.



MODULE QUIZ 4.3

1. For a lognormal distribution:
 - A. the mean equals the median.
 - B. the probability of a negative outcome is zero.
 - C. the probability of a positive outcome is 50%.
2. If a stock's initial price is \$20 and its year-end price is \$23, then its continuously compounded annual (stated) rate of return is:
 - A. 13.64%.
 - B. 13.98%.
 - C. 15.00%.
3. A stock doubled in value last year. Its continuously compounded return over the period was *closest to*:
 - A. 18.2%.
 - B. 69.3%.
 - C. 100.0%.
4. Which of the following is *least likely* a property of Student's t -distribution?
 - A. As the degrees of freedom get larger, the variance approaches zero.
 - B. It is defined by a single parameter, the degrees of freedom, which is equal to $n - 1$.
 - C. It has more probability in the tails and less at the peak than a standard normal distribution.
5. Which of the following statements about the F -distribution and chi-square distribution is *least accurate*? Both distributions:
 - A. are typically asymmetrical.
 - B. are bounded from below by zero.
 - C. have means that are less than their standard deviations.

KEY CONCEPTS

LOS 4.a

A probability distribution lists all the possible outcomes of an experiment, along with their associated probabilities.

A discrete random variable has positive probabilities associated with a finite number of outcomes.

A continuous random variable has positive probabilities associated with a range of outcome values—the probability of any single value is zero.

LOS 4.b

Given the cumulative distribution function for a random variable, the probability that an outcome will be less than or equal to a specific value is represented by the area under the probability distribution to the left of that value.

LOS 4.c

A discrete uniform distribution is one where there are n discrete, equally likely outcomes.

For a discrete uniform distribution with n possible outcomes, the probability for each outcome equals $1/n$.

LOS 4.d

A continuous uniform distribution is one where the probability of X occurring in a possible range is the length of the range relative to the total of all possible values. Letting a and b be the lower and upper limit of the uniform distribution, respectively, then for:

$$a \leq x_1 \leq x_2 \leq b, P(x_1 \leq X \leq x_2) = \frac{x_2 - x_1}{b - a}$$

LOS 4.e

The binomial distribution is a probability distribution for a binomial (discrete) random variable that has two possible outcomes.

For a binomial distribution, if the probability of success is p , the probability of x successes in n trials is:

$$p(x) = P(X = x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} = {}_n C_x \times p^x (1-p)^{n-x}$$

LOS 4.f

The normal probability distribution and normal curve have the following characteristics:



- The normal curve is symmetrical and bell-shaped with a single peak at the exact center of the distribution.
- Mean = median = mode, and all are in the exact center of the distribution.
- The normal distribution can be completely defined by its mean and standard deviation because the skew is always zero and kurtosis is always 3.

LOS 4.g

Multivariate distributions describe the probabilities for more than one random variable, whereas a univariate distribution is for a single random variable.

The correlation(s) of a multivariate distribution describes the relation between the outcomes of its variables relative to their expected values.

LOS 4.h

A confidence interval is a range within which we have a given level of confidence of finding a point estimate (e.g., the 90% confidence interval for X is  Numberfigure - 1.65s to  Numberfigure + 1.65s).

Confidence intervals for any normally distributed random variable are:

- 90%: $\mu \pm 1.65$ standard deviations.
- 95%: $\mu \pm 1.96$ standard deviations.
- 99%: $\mu \pm 2.58$ standard deviations.

The probability that a normally distributed random variable X will be within A standard deviations of its mean, μ , [i.e., $P(\mu - A\sigma \leq X \leq \mu + A\sigma)$], may be calculated as $F(A) - F(-A)$, where $F(A)$ is the cumulative standard normal probability of A , or as $1 - 2[F(-A)]$.

LOS 4.i

The standard normal probability distribution has a mean of 0 and a standard deviation of 1.

A normally distributed random variable X can be standardized as $Z = \frac{x - \mu}{\sigma}$ and Z will be normally distributed with mean = 0 and standard deviation 1.

LOS 4.j

The z -table is used to find the probability that X will be less than or equal to a given value.

- $P(X < x) = F(x) = F\left[\frac{x - \mu}{\sigma}\right] = F(z)$, which is found in the standard normal probability table.
- $P(X > x) = 1 - P(X < x) = 1 - F(z)$.

LOS 4.k

Shortfall risk is the probability that a portfolio's value (or return) will fall below a specific value over a given period of time.

The safety-first ratio for portfolio P , based on a target return R_T , is:

$$\text{SFRatio} = \frac{E(R_P) - R_T}{\sigma_P}$$

Greater safety-first ratios are preferred and indicate a smaller shortfall probability. Roy's safety-first criterion states that the optimal portfolio minimizes shortfall risk.

LOS 4.l

If x is normally distributed, e^x follows a lognormal distribution. A lognormal distribution is often used to model asset prices, since a lognormal random variable cannot be negative and can take on any positive value.

LOS 4.m

As we decrease the length of discrete compounding periods (e.g., from quarterly to monthly) the effective annual rate increases. As the length of the compounding period in discrete compounding gets shorter and shorter, the compounding becomes continuous, where the effective annual rate = $e^i - 1$.

For a holding period return (HPR) over any period, the equivalent continuously compounded rate over the period is $\ln(1 + \text{HPR})$.

LOS 4.n

The t -distribution is similar, but not identical, to the normal distribution in shape—it is defined by the degrees of freedom and has fatter tails compared to the normal distribution.

Degrees of freedom for the t -distribution are equal to $n - 1$. Student's t -distribution is closer to the normal distribution when degrees of freedom are greater, and confidence intervals are narrower when degrees of freedom are greater.

LOS 4.o

A chi-square distribution has $n - 1$ degrees of freedom, is asymmetric, is bounded from below by zero, and becomes more symmetric and approaches a bell-curve shape as its degrees of freedom increase.

A ratio of two chi-square random variables with degrees of freedom m and n follows an F -distribution with degrees of freedom m and n . An F -distribution is asymmetric, bounded from

below by zero, and approaches a bell-curve shape as its degrees of freedom increase.

LOS 4.p

Monte Carlo simulation uses randomly generated values for risk factors, based on their assumed distributions, to produce a distribution of possible security values. Its limitations are that it is fairly complex and will provide answers that are no better than the assumptions used.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 4.1

1. **B** Time is usually a continuous random variable; the others are discrete. (LOS 4.a)
2. **B** For a continuous distribution $p(x) = 0$ for all X ; only ranges of value of X have positive probabilities. (LOS 4.a)
3. **B** $\frac{8-5}{5} + \frac{8-10}{5} = \frac{1}{5}$, and $\frac{8-10}{5}$ is negative, so this satisfies neither of the requirements for a probability distribution. The others have $P[X_i]$ between zero and 1 and $\sum P[X_i] = 1$, and thus satisfy both requirements for a probability distribution. (LOS 4.a)
4. **C** $(0.04 + 0.11 + 0.18 + 0.24 + 0.14 + 0.17) = 0.88$ (LOS 4.b)
5. **B** $(0.14 + 0.17 + 0.09 + 0.03) = 0.43$ (LOS 4.b)
6. **C** $(0.18 + 0.24 + 0.14 + 0.17) = 0.73$ (LOS 4.b)
7. **A** $0 + 1(0.11) + 2(0.18) + 3(0.24) + 4(0.14) + 5(0.17) + 6(0.09) + 7(0.03) = 3.35$ (LOS 4.b)
8. **C** $F(x)$ is the cumulative probability, $P(x < 20)$ here. Because all the observations in this distribution are between 4 and 10, the probability of an outcome less than 20 is 100%. (LOS 4.d)
9. **A** There may be any number of independent trials, each with only two possible outcomes. (LOS 4.e)
10. **B** With only two possible outcomes, there must be some positive probability for each. If this were not the case, the variable in question would not be a random variable, and a probability distribution would be meaningless. It does not matter if one of the possible outcomes happens to be zero. (LOS 4.e)
11. **C** Success = having a fax machine. $[6! / 4!(6-4)!](0.6)^4(0.4)^{6-4} = 15(0.1296)(0.16) = 0.311$. (LOS 4.e)
12. **A** Success = staying for five years. $[6! / 2!(6-2)!](0.10)^2(0.90)^{6-2} = 15(0.01)(0.656) = 0.0984$. (LOS 4.e)
13. **C** Success = passing the exam. Then, $E(\text{success}) = np = 15 \times 0.4 = 6$. (LOS 4.e)

Module Quiz 4.2

1. **A** Normal distributions are symmetrical (i.e., have zero skewness) and their kurtosis is equal to 3. (LOS 4.f)
2. **B** To describe a multivariate normal distribution, we must consider the correlations among the variables, as well as the means and variances of the variables. (LOS 4.g)
3. **C** $1 - F(-1) = F(1) = 0.8413$. There is an 84.13% probability that a randomly chosen income is not more than one standard deviation below the mean. (LOS 4.h)
4. **B** This is true by the formula for z . (LOS 4.i)
5. **C** By the symmetry of the z -distribution and $F(0) = 0.5$. Half the distribution lies on each side of the mean. (LOS 4.j)

6. **C** $SFR = (18 - 4) / 40 = 0.35$ is the largest value. (LOS 4.k)

7. **A** $SFR = (5 - 0) / 8 = 0.625$ is the largest value. (LOS 4.k)

Module Quiz 4.3

1. **B** A lognormally distributed variable is never negative. (LOS 4.l)

2. **B** $\ln(23 / 20) = 0.1398$ (LOS 4.m)

3. **B** $\ln(2) = 0.6931$ (LOS 4.m)

4. **A** As the degrees of freedom get larger, the t -distribution approaches the normal distribution. As the degrees of freedom fall, the peak of the t -distribution flattens and its tails get fatter (more probability in the tails—that's why, all else the same, the critical t increases as the df decreases). (LOS 4.n)

5. **C** There is no consistent relationship between the mean and standard deviation of the chi-square distribution or F -distribution. (LOS 4.o)

READING 5

SAMPLING AND ESTIMATION

EXAM FOCUS

This reading covers sampling and making inferences about population parameters (and other statistics) from sample data. It is essential that you know the central limit theorem, for it allows us to use sampling statistics to construct confidence intervals for point estimates of population means. Make sure you can calculate confidence intervals for population means given sample parameter estimates and a level of significance, and know when it is appropriate to use the z -statistic versus the t -statistic. You should also understand the various procedures for selecting samples, and recognize the sources of bias in selecting sample data.

MODULE 5.1: SAMPLING METHODS, CENTRAL LIMIT THEOREM, AND STANDARD ERROR



Video covering this content is available online.

LOS 5.a: Compare and contrast probability samples with non-probability samples and discuss applications of each to an investment problem.

Probability sampling refers to selecting a sample when we know the probability of each sample member in the overall population. With **random sampling**, each item is assumed to have the same probability of being selected. If we have a population of data and select our sample by using a computer to randomly select a number of observations from the population, each data point has an equal probability of being selected and we call this **simple random sampling**. If we want to estimate the mean profitability for a population of firms, this may be an appropriate method.

Non-probability sampling is based on either low cost and easy access to some data items, or on using the judgment of the researcher in selecting specific data items. Less randomness in selection may lead to greater sampling error.

LOS 5.b: Explain sampling error.

Sampling error is the difference between a sample statistic (such as the mean, variance, or standard deviation of the sample) and its corresponding population parameter (the true mean, variance, or standard deviation of the population). For example, the sampling error for the mean is as follows:

$$\text{sampling error of the mean} = \text{sample mean} - \text{population mean} = \bar{x} - \mu$$

It is important to recognize that the sample statistic itself is a random variable and therefore has a probability distribution. The **sampling distribution** of the sample statistic is a probability distribution of all possible sample statistics computed from a set of equal-size samples that were randomly drawn from the same population. Think of it as the probability distribution of a statistic from many samples.

For example, suppose a random sample of 100 bonds is selected from a population of a major municipal bond index consisting of 1,000 bonds, and then the mean return of the 100-bond sample is calculated. Repeating this process many times will result in many different estimates of the population mean return (i.e., one for each sample). The distribution of these estimates of the mean is the *sampling distribution of the mean*.

It is important to note that this sampling distribution is distinct from the distribution of the actual prices of the 1,000 bonds in the underlying population and will have different parameters.

LOS 5.c: Compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling.

Probability Sampling Methods

Simple random sampling is a method of selecting a sample in such a way that each item or person in the population being studied has the same likelihood of being included in the sample. As an example of simple random sampling, assume that you want to draw a sample of five items out of a group of 50 items. This can be accomplished by numbering each of the 50 items, placing them in a hat, and shaking the hat. Next, one number can be drawn randomly from the hat. Repeating this process (experiment) four more times results in a set of five numbers. The five drawn numbers (items) comprise a simple random sample from the population. In applications like this one, a random-number table or a computer random-number generator is often used to create the sample. Another way to form an approximately random sample is **systematic sampling**, selecting every n th member from a population.

Stratified random sampling uses a classification system to separate the population into smaller groups based on one or more distinguishing characteristics. From each subgroup, or stratum, a random sample is taken and the results are pooled. The size of the samples from each stratum is based on the size of the stratum relative to the population.

Stratified sampling is often used in bond indexing because of the difficulty and cost of completely replicating the entire population of bonds. In this case, bonds in a population are categorized (stratified) according to major bond risk factors such as duration, maturity, coupon rate, and the like. Then, samples are drawn from each separate category and combined to form a final sample.

To see how this works, suppose you want to construct a portfolio of 100 bonds that is indexed to a major municipal bond index of 1,000 bonds, using a stratified random sampling approach. First, the entire population of 1,000 municipal bonds in the index can be classified on the basis of maturity and coupon rate. Then, cells (stratum) can be created for different maturity/coupon combinations, and random samples can be drawn from each of the maturity/coupon cells. To sample from a cell containing 50 bonds with 2- to 4-year maturities and coupon rates less than

5%, we would select five bonds. The number of bonds drawn from a given cell corresponds to the cell's weight relative to the population (index), or $(50 / 1000) \times 100 = 5$ bonds. This process is repeated for all the maturity/coupon cells, and the individual samples are combined to form the portfolio.

By using stratified sampling, we guarantee that we sample five bonds from this cell. If we had used simple random sampling, there would be no guarantee that we would sample any of the bonds in the cell. Or, we may have selected more than five bonds from this cell.

Cluster sampling is also based on subsets of a population, but in this case we are assuming that each subset (cluster) is representative of the overall population with respect to the item we are sampling. For example, we may have data on personal incomes for a state's residents by county. The data for each county is a cluster.

In **one-stage cluster sampling**, a random sample of clusters is selected and all the data in those clusters comprise the sample. In **two-stage cluster sampling**, random samples from each of the selected clusters comprise the sample. Contrast this with stratified random sampling, in which random samples are selected from every subgroup.

To the extent that the subgroups do not have the same distribution as the entire population of the characteristic we are interested in, cluster sampling will have greater sampling error than simple random sampling. Two-stage cluster sampling can be expected to have greater sampling error than one-stage cluster sampling. Lower cost and less time required to assemble the sample are the primary advantages of cluster sampling, and it may be most appropriate for a smaller pilot study.

Non-Probability Sampling Methods

Convenience sampling refers to selecting sample data based on its ease of access, using data that are readily available. Because such a sample is typically not random, sampling error will be greater. It is most appropriate for an initial look at the data prior to adopting a sampling method with less sampling error.

Judgmental sampling refers to samples for which each observation is selected from a larger data set by the researcher, based on her experience and judgment. As an example, a researcher interested in assessing company compliance with accounting standards may have experience suggesting that evidence of noncompliance is typically found in certain ratios derived from the financial statements. The researcher may select only data on these items. Researcher bias (or simply poor judgment) may lead to samples that have excessive sampling error. In the absence of bias or poor judgment, judgmental sampling may produce a more representative sample or allow the researcher to focus on a sample that offers good data on the characteristic or statistic of interest.

An important consideration when sampling is ensuring that the distribution of data of interest is constant for the whole population being sampled. For example, judging a characteristic U.S. banks using data from 2005 to 2015 may not be appropriate. It may well be that regulatory reform of the banking industry after the financial crisis of 2007–2008 resulted in significant changes in banking practices, so that the mean of a statistic precrisis and its mean value across the population of banks post-crisis are quite different. Pooling the data over the whole period

from 2005 to 2015 would not be appropriate if this is the case, and the sample mean calculated from it would not be a good estimate of either precrisis or post-crisis mean values.

LOS 5.d: Explain the central limit theorem and its importance.

The **central limit theorem** states that for simple random samples of size n from a *population* with a mean μ and a finite variance σ^2 , the sampling distribution of the sample mean \bar{x} approaches a normal probability distribution with mean μ and a variance equal to $\frac{\sigma^2}{n}$ as the sample size becomes large.

The central limit theorem is extremely useful because the normal distribution is relatively easy to apply to hypothesis testing and to the construction of confidence intervals. Specific inferences about the population mean can be made from the sample mean, *regardless of the population's distribution*, as long as the sample size is “sufficiently large,” which usually means $n \geq 30$.

Important properties of the central limit theorem include the following:

- If the sample size n is sufficiently large ($n \geq 30$), the sampling distribution of the sample means will be approximately normal. Remember what's going on here, random samples of size n are repeatedly being taken from an overall larger population. Each of these random samples has its own mean, which is itself a random variable, and this set of sample means has a distribution that is approximately normal.
 - The mean of the population, μ , and the mean of the distribution of all possible sample means are equal.
 - The variance of the distribution of sample means is $\frac{\sigma^2}{n}$, the population variance divided by the sample size.
-

LOS 5.e: Calculate and interpret the standard error of the sample mean.

The **standard error of the sample mean** is the standard deviation of the distribution of the sample means.

When the standard deviation of the population, σ , is *known*, the standard error of the sample mean is calculated as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where:

$\sigma_{\bar{x}}$ = standard error of the sample mean

σ = standard deviation of the population

n = size of the sample

EXAMPLE: Standard error of sample mean (known population variance)

The mean hourly wage for Iowa farm workers is \$13.50 with a *population standard deviation* of \$2.90. Calculate and interpret the standard error of the sample mean for a sample size of 30.

Answer:

Because the population standard deviation, σ , is *known*, the standard error of the sample mean is expressed as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$2.90}{\sqrt{30}} = \$0.53$$

This means that if we were to take many samples of size 30 from the Iowa farm worker population and prepare a sampling distribution of the sample means, we would get a distribution with an expected mean of \$13.50 and standard error (standard deviation of the sample means) of \$0.53.

Practically speaking, the *population's standard deviation is almost never known*. Instead, the standard error of the sample mean must be estimated by dividing the standard deviation of *the sample* by \sqrt{n} :

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

EXAMPLE: Standard error of sample mean (unknown population variance)

Suppose a sample contains the past 30 monthly returns for McCreary, Inc. The mean return is 2% and the *sample* standard deviation is 20%. Calculate and interpret the standard error of the sample mean.

Answer:

Since σ is unknown, the standard error of the sample mean is:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{20\%}{\sqrt{30}} = 3.6\%$$

This implies that if we took all possible samples of size 30 from McCreary's monthly returns and prepared a sampling distribution of the sample means, the mean would be 2% with a standard error of 3.6%.

EXAMPLE: Standard error of sample mean (unknown population variance)

Continuing with our example, suppose that instead of a sample size of 30, we take a sample of the past 200 monthly returns for McCreary, Inc. In order to highlight the effect of sample size on the sample standard error, let's assume that the mean return and standard deviation of this larger sample remain at 2% and 20%, respectively. Now, calculate the standard error of the sample mean for the 200-return sample.

Answer:

The standard error of the sample mean is computed as:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{20\%}{\sqrt{200}} = 1.4\%$$

The result of the preceding two examples illustrates an important property of sampling distributions. Notice that the value of the standard error of the sample mean decreased from 3.6% to 1.4% as the sample size increased from 30 to 200. This is because as the sample size increases, the sample mean gets closer, on average, to the true mean of the population. In other words, the distribution of the sample means about the population mean gets smaller and smaller, so the standard error of the sample mean decreases.



PROFESSOR'S NOTE

I get a lot of questions about when to use σ and $\frac{\sigma}{\sqrt{n}}$. Just remember that the standard deviation of the means of multiple samples is less than the standard deviation of single observations. If the standard deviation of monthly stock returns is 2%, the standard error (deviation) of the average monthly return over the next six months is $\frac{2\%}{\sqrt{6}} = 0.82\%$. The average of several observations of a random variable will be less widely dispersed (have lower standard deviation) around the expected value than will a single observation of the random variable.

LOS 5.f: Identify and describe desirable properties of an estimator.

Regardless of whether we are concerned with point estimates or confidence intervals, there are certain statistical properties that make some estimates more desirable than others. These desirable properties of an estimator are **unbiasedness**, **efficiency**, and **consistency**.

- An *unbiased* estimator is one for which the expected value of the estimator is equal to the parameter you are trying to estimate. For example, because the expected value of the sample mean is equal to the population mean [$E(\bar{X}) = \mu$], the sample mean is an unbiased estimator of the population mean.
- An unbiased estimator is also *efficient* if the variance of its sampling distribution is smaller than all the other unbiased estimators of the parameter you are trying to estimate. The sample mean, for example, is an unbiased and efficient estimator of the population mean.
- A *consistent* estimator is one for which the accuracy of the parameter estimate increases as the sample size increases. As the sample size increases, the standard error of the sample mean falls, and the sampling distribution bunches more closely around the population mean. In fact, as the sample size approaches infinity, the standard error approaches zero.



MODULE QUIZ 5.1

1. An important difference between two-stage cluster sampling and stratified random sampling is that compared to stratified random sampling, two-stage cluster sampling:
 - A. uses all members of each sub-group (strata).
 - B. takes random samples all sub-groups (strata).
 - C. will not preserve differences in a characteristic across sub-groups.
2. Sampling error is defined as:
 - A. an error that occurs when a sample of less than 30 elements is drawn.
 - B. an error that occurs during collection, recording, and tabulation of data.

- C. the difference between the value of a sample statistic and the value of the corresponding population parameter.
- The mean age of all CFA candidates is 28 years. The mean age of a random sample of 100 candidates is found to be 26.5 years. The difference of 1.5 years is called:
 - the random error.
 - the sampling error.
 - the population error.
 - A simple random sample is a sample drawn in such a way that each member of the population has:
 - some chance of being selected in the sample.
 - an equal chance of being included in the sample.
 - a 1% chance of being included in the sample.
 - To apply the central limit theorem to the sampling distribution of the sample mean, the sample is usually considered to be large if n is *greater* than:
 - 20.
 - 25.
 - 30.
 - If n is large and the population standard deviation is unknown, the standard error of the sampling distribution of the sample mean is *equal* to:
 - the sample standard deviation divided by the sample size.
 - the population standard deviation multiplied by the sample size.
 - the sample standard deviation divided by the square root of the sample size.
 - The standard error of the sampling distribution of the sample mean for a sample size of n drawn from a population with a mean of μ and a standard deviation of σ is:
 - sample standard deviation divided by the sample size.
 - sample standard deviation divided by the square root of the sample size.
 - population standard deviation divided by the square root of the sample size.
 - Assume that a population has a mean of 14 with a standard deviation of 2. If a random sample of 49 observations is drawn from this population, the standard error of the sample mean is *closest* to:
 - 0.04.
 - 0.29.
 - 2.00.
 - The population's mean is 30 and the mean of a sample of size 100 is 28.5. The variance of the sample is 25. The standard error of the sample mean is *closest* to:
 - 0.05.
 - 0.25.
 - 0.50.
 - Which of the following is *least likely* a desirable property of an estimator?
 - Reliability.
 - Efficiency.
 - Consistency.

MODULE 5.2: CONFIDENCE INTERVALS, RESAMPLING, AND SAMPLING BIASES



Video covering this content is available online.

LOS 5.g: Contrast a point estimate and a confidence interval estimate of a population parameter.

LOS 5.h: Calculate and interpret a confidence interval for a population mean, given a normal distribution with 1) a known population variance, 2) an unknown population variance, or 3) an unknown population variance and a large sample size.

Point estimates are single (sample) values used to estimate population parameters. The formula used to compute the point estimate is called the estimator. For example, the sample mean, \bar{x} , is an estimator of the population mean μ and is computed using the familiar formula:

$$\bar{x} = \frac{\sum x}{n}$$

The value generated with this calculation for a given sample is called the *point estimate* of the mean.

A **confidence interval** is a range of values in which the population parameter is expected to lie. Confidence interval estimates result in a range of values within which the actual value of a parameter will lie, given the probability of $1 - \alpha$. Here, alpha, α , is called the *level of significance* for the confidence interval, and the probability $1 - \alpha$ is referred to as the *degree of confidence*. For example, we might estimate that the population mean of random variables will range from 15 to 25 with a 95% degree of confidence, or at the 5% level of significance.

Confidence intervals are usually constructed by adding or subtracting an appropriate value from the point estimate. In general, confidence intervals take on the following form:

point estimate \pm (reliability factor \times standard error)

where:

point estimate = value of a sample statistic of the population parameter

reliability factor = number that depends on the sampling distribution of the point estimate and the probability that the point estimate falls in the confidence interval, $(1 - \alpha)$

standard error = standard error of the point estimate

If the population has a *normal distribution with a known variance*, a **confidence interval for the population mean** can be calculated as:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where:

\bar{x} = *point estimate* of the population mean (sample mean).

$z_{\alpha/2}$ = *reliability factor*, a standard normal random variable for which the probability in the right-hand tail of the distribution is $\alpha/2$. In other words, this is the z-score that leaves $\alpha/2$ of probability in the upper tail.

$\frac{\sigma}{\sqrt{n}}$ = the *standard error* of the sample mean where σ is the known standard deviation of the population, and n is the sample size.

The most commonly used standard normal distribution reliability factors are:

$z_{\alpha/2} = 1.645$ for 90% confidence intervals (the significance level is 10%, 5% in each tail).

$z_{\alpha/2} = 1.960$ for 95% confidence intervals (the significance level is 5%, 2.5% in each tail).

$z_{\alpha/2} = 2.575$ for 99% confidence intervals (the significance level is 1%, 0.5% in each tail).

Do these numbers look familiar? They should! In our review of common probability distributions, we found the probability under the standard normal curve between $z = -1.96$ and

$z = +1.96$ to be 0.95, or 95%. Owing to symmetry, this leaves a probability of 0.025 under each tail of the curve beyond $z = -1.96$ or $z = +1.96$, for a total of 0.05, or 5%—just what we need for a significance level of 0.05, or 5%.

EXAMPLE: Confidence interval

Consider a practice exam that was administered to 36 Level I candidates. Their mean score on this practice exam was 80. Assuming a population standard deviation equal to 15, construct and interpret a 99% confidence interval for the mean score on the practice exam for all candidates. *Note that, in this example, the population standard deviation is known, so we don't have to estimate it.*

Answer:

At a confidence level of 99%, $z_{\alpha/2} = z_{0.005} = 2.58$. So, the 99% confidence interval is calculated as follows:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 80 \pm 2.58 \frac{15}{\sqrt{36}} = 80 \pm 6.45$$

Thus, the 99% confidence interval ranges from 73.55 to 86.45.

EXAMPLE: Confidence intervals for a population mean and for a single observation

Annual returns on energy stocks are approximately normally distributed with a mean of 9% and standard deviation of 6%. Construct a 90% confidence interval for the annual returns of a randomly selected energy stock and a 90% confidence interval for the mean of the annual returns for a sample of 12 energy stocks.

Answer:

A 90% confidence interval for a single observation is 1.645 *standard deviations* from the sample mean.

$$9\% \pm 1.645(6\%) = -0.87\% \text{ to } 18.87\%$$

A 90% confidence interval for the population mean is 1.645 *standard errors* from the sample mean.

$$9\% \pm 1.645 \frac{6\%}{\sqrt{12}} = 6.15\% \text{ to } 11.85\%$$

Confidence intervals can be interpreted from a probabilistic perspective or a practical perspective. With regard to the outcome of the practice exam example, these two perspectives can be described as follows:

- *Probabilistic interpretation.* After repeatedly taking samples of CFA candidates, administering the practice exam, and constructing confidence intervals for each sample's mean, 99% of the resulting confidence intervals will, in the long run, include the population mean.
- *Practical interpretation.* We are 99% confident that the population mean score is between 73.55 and 86.45 for candidates from this population.

Confidence Intervals for the Population Mean: Normal With Unknown Variance

If the distribution of the *population is normal with unknown variance*, we can use the *t*-distribution to construct a confidence interval:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where:

\bar{x} = the point estimate of the population mean

$t_{\alpha/2}$ = the *t*-reliability factor (a.k.a. *t*-statistic or critical *t*-value) corresponding to a *t*-distributed random variable with $n - 1$ degrees of freedom, where n is the sample size. The area under the tail of the *t*-distribution to the right of $t_{\alpha/2}$ is $\alpha/2$.

$\frac{s}{\sqrt{n}}$ = standard error of the sample mean

s = sample standard deviation

Unlike the standard normal distribution, the reliability factors for the *t*-distribution depend on the sample size, so we cannot rely on a commonly used set of reliability factors. Instead, reliability factors for the *t*-distribution have to be looked up in a table of Student's *t*-distribution, like the one at the back of this book.

Owing to the relatively fatter tails of the *t*-distribution, confidence intervals constructed using *t*-reliability factors ($t_{\alpha/2}$) will be more conservative (wider) than those constructed using *z*-reliability factors.

EXAMPLE: Confidence intervals

Let's return to the McCreary, Inc., example. Recall that we took a sample of the past 30 monthly stock returns for McCreary, Inc., and determined that the mean return was 2% and the sample standard deviation was 20%. Since the population variance is unknown, the standard error of the sample was estimated to be:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{20\%}{\sqrt{30}} = 3.6\%$$

Now, let's construct a 95% confidence interval for the mean monthly return.

Answer:

Here, we will use the *t*-reliability factor because the population variance is unknown. Since there are 30 observations, the degrees of freedom are $29 = 30 - 1$. Remember, because this is a two-tailed test at the 95% confidence level, the probability under each tail must be $\alpha/2 = 2.5\%$, for a total of 5%. So, referencing the one-tailed probabilities for Student's *t*-distribution at the back of this book, we find the critical *t*-value (reliability factor) for $\alpha/2 = 0.025$ and $df = 29$ to be $t_{29, 2.5} = 2.045$. Thus, the 95% confidence interval for the population mean is:

$$2\% \pm 2.045 \left(\frac{20\%}{\sqrt{30}} \right) = 2\% \pm 2.045(3.6\%) = 2\% \pm 7.4\%$$

Thus, the 95% confidence has a lower limit of -5.4% and an upper limit of $+9.4\%$.

We can interpret this confidence interval by saying that we are 95% confident that the population mean monthly return for McCreary stock is between -5.4% and $+9.4\%$.



PROFESSOR'S NOTE

You should practice looking up reliability factors (a.k.a. critical t -values or t -statistics) in a t -table. The first step is always to compute the degrees of freedom, which is $n - 1$. The second step is to find the appropriate level of alpha or significance. This depends on whether the test you're concerned with is one-tailed (use α) or two-tailed (use $\alpha/2$). In this review, our tests will always be two-tailed because confidence intervals are designed to compute an upper and lower limit. Thus, we will use $\alpha/2$. To look up $t_{29, 2.5}$, find the 29 df row and match it with the 0.025 column; $t = 2.045$ is the result. We'll do more of this in our study of hypothesis testing.

Confidence Interval for a Population Mean When the Population Variance Is Unknown, Given a Large Sample From Any Type of Distribution

We now know that the z -statistic should be used to construct confidence intervals when the population distribution is normal and the variance is known, and the t -statistic should be used when the distribution is normal but the variance is unknown. But what do we do when the distribution is *nonnormal*?

As it turns out, the size of the sample influences whether or not we can construct the appropriate confidence interval for the sample mean.

- If the *distribution is nonnormal* but the *population variance is known*, the z -statistic can be used as long as the sample size is large ($n \geq 30$). We can do this because the central limit theorem assures us that the distribution of the sample mean is approximately normal when the sample is large.
- If the *distribution is nonnormal* and the *population variance is unknown*, the t -statistic can be used as long as the sample size is large ($n \geq 30$). It is also acceptable to use the z -statistic, although use of the t -statistic is more conservative.

This means that if we are sampling from a nonnormal distribution (which is sometimes the case in finance), *we cannot create a confidence interval if the sample size is less than 30*. So, all else equal, make sure you have a sample of at least 30, and the larger, the better.

Figure 5.1 summarizes this discussion.



PROFESSOR'S NOTE

You should commit the criteria in the following table to memory.

Figure 5.1: Criteria for Selecting the Appropriate Test Statistic

When sampling from a:	Test Statistic	
	Small Sample ($n < 30$)	Large Sample ($n \geq 30$)
Normal distribution with <i>known</i> variance	z-statistic	z-statistic
Normal distribution with <i>unknown</i> variance	t-statistic	t-statistic*
Nonnormal distribution with <i>known</i> variance	not available	z-statistic
Nonnormal distribution with <i>unknown</i> variance	not available	t-statistic*

*The z-statistic is theoretically acceptable here, but use of the t-statistic is more conservative.

All of the preceding analysis depends on the sample we draw from the population being random. If the sample isn't random, the central limit theorem doesn't apply, our estimates won't have the desirable properties, and we can't form unbiased confidence intervals. Surprisingly, creating a *random sample* is not as easy as one might believe. There are a number of potential mistakes in sampling methods that can bias the results. These biases are particularly problematic in financial research, where available historical data are plentiful, but the creation of new sample data by experimentation is restricted.

LOS 5.i: Describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic.

Previously, we used the sample variance to calculate the standard error of our estimate of the mean. The standard error provides better estimates of the distribution of sample means when the sample is unbiased and the distribution of sample means is approximately normal.

Two alternative methods of estimating the standard error of the sample mean involve resampling of the data. The first of these, termed the **jackknife**, calculates multiple sample means, each with one of the observations removed from the sample. The standard deviation of these sample means can then be used as an estimate of the standard error of sample means. The jackknife is a computationally simple tool and can be used when the number of observations available is relatively small. It can remove bias from statistical estimates.

The jackknife (so named because it is a handy and readily available tool) was developed when computational power was not as available and low-cost as it has become. A **bootstrap** method is more computationally demanding but has some advantages. To estimate the standard error of the sample mean, we draw repeated samples of size n from the full data set (replacing the sampled observations each time). We can then directly calculate the standard deviation of these sample means as our estimate of the standard error of the sample mean.

The bootstrap method can improve accuracy compared to using only the data in a single sample, and can be used to construct confidence intervals for a variety of statistics in addition to the mean, such as the median. It can be used to estimate the distributions of complex statistics, including those that do not have an analytic form.

LOS 5.j: Describe the issues regarding selection of the appropriate sample size, data snooping bias, sample selection bias, survivorship bias, look-ahead bias, and time-period bias.

We have seen so far that a larger sample reduces the sampling error and the standard deviation of the sample statistic around its true (population) value. Confidence intervals are narrower when samples are larger and the standard errors of the point estimates of population parameters are less.

There are two limitations on this idea of “larger is better” when it comes to selecting an appropriate sample size. One is that larger samples may contain observations from a different population (distribution). If we include observations that come from a different population (one with a different population parameter), we will not necessarily improve, and may even reduce, the precision of our population parameter estimates. The other consideration is cost. The costs of using a larger sample must be weighed against the value of the increase in precision from the increase in sample size. Both of these factors suggest that the largest possible sample size is not always the most appropriate choice.

Data snooping occurs when analysts repeatedly use the same database to search for patterns or trading rules until one that “works” is discovered. For example, empirical research has provided evidence that value stocks appear to outperform growth stocks. Some researchers argue that this anomaly is actually the product of data snooping. Because the data set of historical stock returns is quite limited, it is difficult to know for sure whether the difference between value and growth stock returns is a true economic phenomenon or simply a chance pattern that was stumbled upon after repeatedly looking for any identifiable pattern in the data.

Data-snooping bias refers to results where the statistical significance of the pattern is overestimated because the results were found through data snooping.

When reading research findings that suggest a profitable trading strategy, make sure you heed the following warning signs of data snooping:

- Evidence that many different variables were tested, most of which are unreported, until significant ones were found.
- The lack of any economic theory that is consistent with the empirical results.

The best way to avoid data snooping is to test a potentially profitable trading rule on a data set different from the one you used to develop the rule (i.e., use out-of-sample data).

Sample selection bias occurs when some data is systematically excluded from the analysis, usually because of the lack of availability. This practice renders the observed sample to be nonrandom, and any conclusions drawn from this sample can’t be applied to the population because the observed sample and the portion of the population that was not observed are different.

Survivorship bias is the most common form of sample selection bias. A good example of the existence of survivorship bias in investments is the study of mutual fund performance. Most mutual fund databases, like Morningstar[®]’s, only include funds currently in existence—the “survivors.” They do not include funds that have ceased to exist due to closure or merger.

This would not be a problem if the characteristics of the surviving funds and the missing funds were the same; then the sample of survivor funds would still be a random sample drawn from the population of mutual funds. As one would expect, however, and as evidence has shown, the funds that are dropped from the sample have lower returns relative to the surviving funds. Thus,

the surviving sample is biased toward the better funds (i.e., it is not random). The analysis of a mutual fund sample with survivorship bias will yield results that overestimate the average mutual fund return because the database only includes the better-performing funds. The solution to survivorship bias is to use a sample of funds that all started at the same time and not drop funds that have been dropped from the sample.

Look-ahead bias occurs when a study tests a relationship using sample data that was not available on the test date. For example, consider the test of a trading rule that is based on the price-to-book ratio at the end of the fiscal year. Stock prices are available for all companies at the same point in time, while end-of-year book values may not be available until 30 to 60 days after the fiscal year ends. In order to account for this bias, a study that uses price-to-book value ratios to test trading strategies might estimate the book value as reported at fiscal year end and the market value two months later.

Time-period bias can result if the time period over which the data is gathered is either too short or too long. If the time period is too short, research results may reflect phenomena specific to that time period, or perhaps even data mining. If the time period is too long, the fundamental economic relationships that underlie the results may have changed.

For example, research findings may indicate that small stocks outperformed large stocks during 1980–1985. This may well be the result of time-period bias—in this case, using too short a time period. It's not clear whether this relationship will continue in the future or if it is just an isolated occurrence.

On the other hand, a study that quantifies the relationship between inflation and unemployment during the period from 1940 to 2000 will also result in time-period bias—because this period is too long, and it covers a fundamental change in the relationship between inflation and unemployment that occurred in the 1980s. In this case, the data should be divided into two subsamples that span the period before and after the change.



MODULE QUIZ 5.2

1. A random sample of 100 computer store customers spent an average of \$75 at the store. Assuming the distribution is normal and the population standard deviation is \$20, the 95% confidence interval for the population mean is *closest* to:
 - A. \$71.08 to \$78.92.
 - B. \$73.89 to \$80.11.
 - C. \$74.56 to \$79.44.
2. Best Computers, Inc., sells computers and computer parts by mail. A sample of 25 recent orders showed the mean time taken to ship these orders was 70 hours with a sample standard deviation of 14 hours. Assuming the population is normally distributed, the 99% confidence interval for the population mean is:
 - A. 70 ± 2.80 hours.
 - B. 70 ± 6.98 hours.
 - C. 70 ± 7.83 hours.
3. What is the *most appropriate* test statistic for constructing confidence intervals for the population mean when the population is normally distributed, but the variance is unknown?
 - A. The z -statistic at α with n degrees of freedom.
 - B. The t -statistic at $\alpha/2$ with n degrees of freedom.
 - C. The t -statistic at $\alpha/2$ with $n - 1$ degrees of freedom.

4. When constructing a confidence interval for the population mean of a nonnormal distribution when the population variance is unknown and the sample size is large ($n > 30$), an analyst may acceptably use:
 - A. either a z -statistic or a t -statistic.
 - B. only a z -statistic at α with n degrees of freedom.
 - C. only a t -statistic at $\alpha/2$ with n degrees of freedom.
5. Jenny Fox evaluates managers who have a cross-sectional population standard deviation of returns of 8%. If returns are independent across managers, how large of a sample does Fox need so the standard error of sample means is 1.265%?
 - A. 7.
 - B. 30.
 - C. 40.
6. Annual returns on small stocks have a population mean of 12% and a population standard deviation of 20%. If the returns are normally distributed, a 90% confidence interval on mean returns over a 5-year period is:
 - A. 5.40% to 18.60%.
 - B. -2.75% to 26.75%.
 - C. -5.52% to 29.52%.
7. Which of the following techniques to improve the accuracy of confidence intervals on a statistic is *most* computationally demanding?
 - A. The jackknife.
 - B. Systematic resampling.
 - C. Bootstrapping.
8. An analyst who uses historical data that was not publicly available at the time period being studied will have a sample with:
 - A. look-ahead bias.
 - B. time-period bias.
 - C. sample selection bias.
9. Which of the following is *most closely* associated with survivorship bias?
 - A. Price-to-book studies.
 - B. Stratified bond sampling studies.
 - C. Mutual fund performance studies.

KEY CONCEPTS

LOS 5.a

Probability sampling refers to sampling methods based on randomly chosen samples and assuming that members of a population are equally likely to be chosen for the samples.

Non-probability sampling refers to choosing sample data that are not random but based on low cost and availability of the sample data, or specifically chosen based on the experience and judgment of the researcher.

LOS 5.b

Sampling error is the difference between a sample statistic and its corresponding population parameter (e.g., the sample mean minus the population mean).

LOS 5.c

Simple random sampling is a method of selecting a sample in such a way that each item or person in the population being studied has the same probability of being included in the sample.

Stratified random sampling involves randomly selecting samples proportionally from subgroups that are formed based on one or more distinguishing characteristics of the data, so that random samples from the subgroups will have the same distribution of these characteristics as the overall population.

Cluster sampling is also based on subgroups (not necessarily based on data characteristics) of a larger data set. In one-stage cluster sampling, the sample is formed from randomly chosen clusters (subsets) of the overall data set. In two-stage cluster sampling, random samples are taken from each of the randomly chosen clusters (subgroups).

Convenience sampling refers to selecting sample data based on its ease of access, using data that are readily available. Judgmental sampling refers to samples for which each observation is selected from a larger data set by the researcher, based on her experience and judgment. Both are examples of non-probability sampling and are non-random.

LOS 5.d

The central limit theorem states that for a population with a mean μ and a finite variance σ^2 , the sampling distribution of the sample mean of all possible samples of size n (for $n \geq 30$) will be approximately normally distributed with a mean equal to μ and a variance equal to σ^2 / n .

LOS 5.e

The standard error of the sample mean is the standard deviation of the distribution of the sample means and is calculated as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, where σ , the population standard deviation, is known, and as $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, where s , the sample standard deviation, is used because the population standard deviation is unknown.

LOS 5.f

Desirable statistical properties of an estimator include unbiasedness (sign of estimation error is random), efficiency (lower sampling error than any other unbiased estimator), and consistency (variance of sampling error decreases with larger sample size).

LOS 5.g

Point estimates are single-value estimates of population parameters. An estimator is a formula used to compute a point estimate.

Confidence intervals are ranges of values, within which the actual value of the parameter will lie with a given probability.

$$\text{confidence interval} = \text{point estimate} \pm (\text{reliability factor} \times \text{standard error})$$

The reliability factor is a number that depends on the sampling distribution of the point estimate and the probability that the point estimate falls in the confidence interval.

LOS 5.h

For a normally distributed population, a confidence interval for its mean can be constructed using a z -statistic when variance is known, and a t -statistic when the variance is unknown. The z -statistic is acceptable in the case of a normal population with an unknown variance if the sample size is large (30+).

In general, we have:

- $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ when the variance is known, and
- $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ when the variance is unknown and the sample standard deviation must be used.

LOS 5.i

Two resampling techniques to improve our estimates of the distribution of sample statistics are the jackknife and bootstrapping. With the jackknife, we calculate n sample means, one with each observation in a sample of size n removed, and base our estimate on the standard error of sample means of size n . It can remove bias from our estimates based on the sample standard deviation without resampling.

With bootstrapping, we use the distribution of sample means (or other statistics) from a large number of samples of size n , drawn from a large data set. Bootstrapping can improve our estimates of the distribution of various sample statistics and provide such estimates when analytical methods will not.

LOS 5.j

Increasing the sample size will generally improve parameter estimates and narrow confidence intervals. The cost of more data must be weighed against these benefits, and adding data that is not generated by the same distribution will not necessarily improve accuracy or narrow confidence intervals.

Potential mistakes in the sampling method can bias results. These biases include data snooping (significant relationships that have occurred by chance), sample selection bias (selection is nonrandom), look-ahead bias (basing the test at a point in time on data not available at that time), survivorship bias (using only surviving mutual funds, hedge funds, etc.), and time-period bias (the relation does not hold over other time periods).

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 5.1

1. **C** With cluster sampling, the randomly selected subgroups may have different distributions of the relevant characteristic relative to the entire population. Cluster sampling uses only randomly selected subgroups, whereas stratified random sampling samples all subgroups to match the distribution of characteristics across the entire population. (LOS 5.a)
2. **C** An example might be the difference between a particular sample mean and the average value of the overall population. (LOS 5.b)
3. **B** The sampling error is the difference between the population parameter and the sample statistic. (LOS 5.b)
4. **B** In a simple random sample, each element of the population has an equal probability of being selected. Choice C allows for an equal chance, but only if there are 100 elements in the population from which the random sample is drawn. (LOS 5.c)
5. **C** Sample sizes of 30 or greater are typically considered large. (LOS 5.d)
6. **C** The formula for the standard error when the population standard deviation is unknown is $s_{\bar{x}} = \frac{s}{\sqrt{n}}$. (LOS 5.e)
7. **C** The formula for the standard error when the population standard deviation is known is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. (LOS 5.e)

8. **B** $s_{\bar{x}} = \frac{s}{\sqrt{n}}$. Given $s = 2$, $s_{\bar{x}} = \frac{2}{\sqrt{49}} = \frac{2}{7} = 0.2857$. (LOS 5.e)
9. **C** $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Given $\sigma^2 = 25$, $\sigma_{\bar{x}} = \frac{5}{\sqrt{100}} = \frac{5}{10} = 0.5$. (LOS 5.e)
10. **A** Efficiency, consistency, and unbiasedness are desirable properties of an estimator. (LOS 5.f)

Module Quiz 5.2

1. **A** Since the population variance is known and $n \geq 30$, the confidence interval is determined as $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. $z_{\alpha/2} = z_{0.025} = 1.96$. So, the confidence interval is $75 \pm 1.96(20/10) = 75 \pm 3.92 = 71.08$ to 78.92. (LOS 5.h)
2. **C** Since the population variance is unknown and $n < 30$, the confidence interval is determined as $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$. Look up $t_{\alpha/2}$ and $df = n - 1$ to get critical t -value. $t_{0.01/2}$ and $df = 24$ is 2.797. So, the confidence interval is $70 \pm 2.797(14 / 5) = 70 \pm 7.83$. (LOS 5.h)
3. **C** Use the t -statistic at $\alpha/2$ and $n - 1$ degrees of freedom when the population variance is unknown. While the z -statistic is acceptable when the sample size is large, sample size is not given here, and the t -statistic is always appropriate under these conditions. (LOS 5.h)
4. **A** When the sample size is large, and the central limit theorem can be relied on to assure a sampling distribution that is normal, either the t -statistic or the z -statistic is acceptable for constructing confidence intervals for the population mean. The t -statistic, however, will provide a more conservative range (wider) at a given level of significance. (LOS 5.h)
5. **C** $1.265 = \frac{8}{\sqrt{n}}$, $n = \left(\frac{8}{1.265}\right)^2 \approx 40$. (LOS 5.h)
6. **B** With a known population standard deviation of returns and a normally distributed population, we can use the z -distribution. The sample mean for a sample of five years will have a standard deviation of $\frac{20}{\sqrt{5}} = 8.94\%$. A 90% confidence interval around the mean return of 12% is $12\% \pm 1.65(8.94\%) = -2.75\%$ to 26.75%. (LOS 5.h)
7. **C** Bootstrapping, repeatedly drawing samples of equal size from a large data set, is more computationally demanding than the jackknife. We have not defined "systematic resampling" as a specific technique. (LOS 5.i)
8. **A** The primary example of look-ahead bias is using year-end financial information in conjunction with market pricing data to compute ratios like the price/earnings (P/E). The E in the denominator is typically not available for 30–60 days after the end of the period. Hence, data that was available on the test date (P) is mixed with information that was not available (E). That is, the P is "ahead" of the E. (LOS 5.j)
9. **C** Mutual fund performance studies are most closely associated with survivorship bias because only the better-performing funds remain in the sample over time. (LOS 5.j)

READING 6

HYPOTHESIS TESTING

EXAM FOCUS

This review addresses common hypothesis tests. Included are tests about population means, population variances, differences in means, mean differences, differences in variances, correlation, and independence. Make sure you understand what a p -value is and how to interpret one. The various hypothesis tests have test statistics with different distributions. These distributions include standard normal (z -test), Student's t -distribution (t -test), chi-square, and F -distributions. Candidates should understand the basic characteristics of each of these distributions and how to apply them. Memorize the standard hypothesis testing procedure presented in this review. Finally, there are some non-parametric tests that you should understand and be able to interpret.

MODULE 6.1: HYPOTHESIS TESTS AND TYPES OF ERRORS



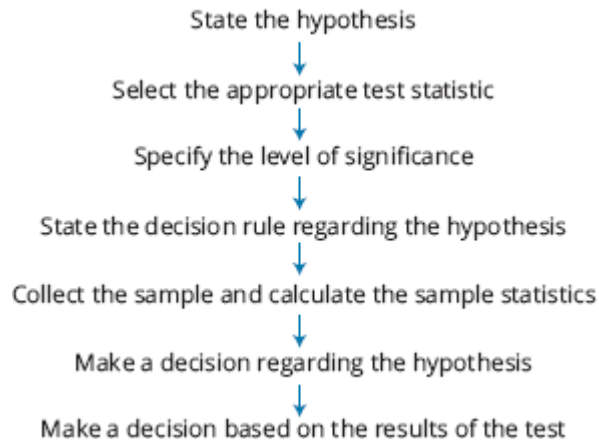
Video covering this content is available online.

LOS 6.a: Define a hypothesis, describe the steps of hypothesis testing, and describe and interpret the choice of the null and alternative hypotheses.

A hypothesis is a statement about the value of a population parameter developed for the purpose of testing a theory or belief. Hypotheses are stated in terms of the population parameter to be tested, like the population mean, μ . For example, a researcher may be interested in the mean daily return on stock options. Hence, the hypothesis may be that the mean daily return on a portfolio of stock options is positive.

Hypothesis testing procedures, based on sample statistics and probability theory, are used to determine whether a hypothesis is a reasonable statement and should not be rejected or if it is an unreasonable statement and should be rejected. The process of hypothesis testing consists of a series of steps shown in Figure 6.1.

Figure 6.1: Hypothesis Testing Procedure*



*Source: Wayne W. Daniel and James C. Terrell, Business Statistics, Basic Concepts and Methodology, Houghton Mifflin, Boston, 1997.

The Null Hypothesis and Alternative Hypothesis

The **null hypothesis**, designated H_0 , is the hypothesis that the researcher wants to reject. It is the hypothesis that is actually tested and is the basis for the selection of the test statistics. The null is generally stated as a simple statement about a population parameter. Typical statements of the null hypothesis for the population mean include $H_0: \mu = \mu_0$, $H_0: \mu \leq \mu_0$, and $H_0: \mu \geq \mu_0$, where μ is the population mean and μ_0 is the hypothesized value of the population mean.



PROFESSOR'S NOTE

The null hypothesis always includes the “equal to” condition.

The **alternative hypothesis**, designated H_a , is what is concluded if there is sufficient evidence to reject the null hypothesis. It is usually the alternative hypothesis that you are really trying to assess. Why? Because you can never really prove anything with statistics, when the null hypothesis is discredited, the implication is that the alternative hypothesis is valid.

LOS 6.b: Compare and contrast one-tailed and two-tailed tests of hypotheses.

The alternative hypothesis can be one-sided or two-sided. A one-sided test is referred to as a **one-tailed test**, and a two-sided test is referred to as a **two-tailed test**. Whether the test is one- or two-sided depends on the proposition being tested. If a researcher wants to test whether the return on stock options is greater than zero, a one-tailed test should be used. However, a two-tailed test should be used if the research question is whether the return on options is simply different from zero. Two-sided tests allow for deviation on both sides of the hypothesized value (zero). In practice, most hypothesis tests are constructed as two-tailed tests.

A **two-tailed test** for the population mean may be structured as:

$$H_0: \mu = \mu_0 \text{ versus } H_a: \mu \neq \mu_0$$

Since the alternative hypothesis allows for values above and below the hypothesized parameter, a two-tailed test uses two **critical values** (or **rejection points**).

The *general decision rule for a two-tailed test* is:

Reject H_0 if:

test statistic > upper critical value or

test statistic < lower critical value

Let's look at the development of the decision rule for a two-tailed test using a z -distributed test statistic (a z -test) at a 5% level of significance, $\alpha = 0.05$.

- At $\alpha = 0.05$, the computed test statistic is compared with the critical z -values of ± 1.96 . The values of ± 1.96 correspond to $\pm z_{\alpha/2} = \pm z_{0.025}$, which is the range of z -values within which 95% of the probability lies. These values are obtained from the cumulative probability table for the standard normal distribution (z -table), which is included at the back of this book.
- If the computed test statistic falls outside the range of critical z -values (i.e., test statistic > 1.96, or test statistic < -1.96), we reject the null and conclude that the sample statistic is sufficiently different from the hypothesized value.
- If the computed test statistic falls within the range ± 1.96 , we conclude that the sample statistic is not sufficiently different from the hypothesized value ($\mu = \mu_0$ in this case), and we fail to reject the null hypothesis.

The *decision rule* (rejection rule) for a two-tailed z -test at $\alpha = 0.05$ can be stated as:

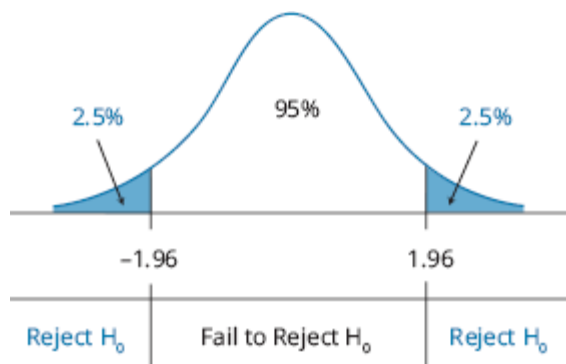
Reject H_0 if:

test statistic < -1.96 or

test statistic > 1.96

Figure 6.2 shows the standard normal distribution for a two-tailed hypothesis test using the z -distribution. Notice that the significance level of 0.05 means that there is $0.05 / 2 = 0.025$ probability (area) under each tail of the distribution beyond ± 1.96 .

Figure 6.2: Two-Tailed Hypothesis Test Using the Standard Normal (z) Distribution



For a **one-tailed hypothesis test** of the population mean, the null and alternative hypotheses are either:

Upper tail: $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$, or

Lower tail: $H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$

The appropriate set of hypotheses depends on whether we believe the population mean, μ , to be greater than (upper tail) or less than (lower tail) the hypothesized value, μ_0 . Using a z -test at the 5% level of significance, the computed test statistic is compared with the critical values of 1.645 for the upper tail tests (i.e., $H_a: \mu > \mu_0$) or -1.645 for lower tail tests (i.e., $H_a: \mu < \mu_0$). These

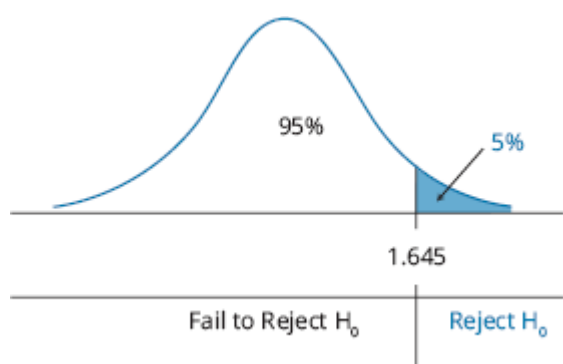
critical values are obtained from a z-table, where $-z_{0.05} = -1.645$ corresponds to a cumulative probability equal to 5%, and the $z_{0.05} = 1.645$ corresponds to a cumulative probability of 95% ($1 - 0.05$).

Let's use the upper tail test structure where $H_0: \mu \leq \mu_0$ and $H_a: \mu > \mu_0$.

- If the calculated test statistic is greater than 1.645, we conclude that the sample statistic is sufficiently greater than the hypothesized value. In other words, we reject the null hypothesis.
- If the calculated test statistic is less than 1.645, we conclude that the sample statistic is not sufficiently different from the hypothesized value, and we fail to reject the null hypothesis.

Figure 6.3 shows the standard normal distribution and the rejection region for a one-tailed test (upper tail) at the 5% level of significance.

Figure 6.3: One-Tailed Hypothesis Test Using the Standard Normal (z) Distribution



The Choice of the Null and Alternative Hypotheses

The most common null hypothesis will be an “equal to” hypothesis. Combined with a “not equal to” alternative, this will require a two-tailed test. The alternative is often the hoped-for hypothesis. The null will include the “equal to” sign and the alternative will include the “not equal to” sign. When the null is that a coefficient is equal to zero, we hope to reject it and show the significance of the relationship.

When the null is less than or equal to, the (mutually exclusive) alternative is framed as greater than, and a one-tail test is appropriate. If we are trying to demonstrate that a return is greater than the risk-free rate, this would be the correct formulation. We will have set up the null and alternative hypothesis so that rejection of the null will lead to acceptance of the alternative, our goal in performing the test. As with a two-tailed test, the null for a one-tailed test will include the “equal to” sign (i.e., either “greater than or equal to” or “less than or equal to”). The alternative will include the opposite sign to the null—either “less than” or “greater than.”

LOS 6.c: Explain a test statistic, Type I and Type II errors, a significance level, how significance levels are used in hypothesis testing, and the power of a test.

Hypothesis testing involves two statistics: the test statistic calculated from the sample data and the *critical value* of the test statistic. The value of the computed test statistic relative to the critical value is a key step in assessing the validity of a hypothesis.

A test statistic is calculated by comparing the point estimate of the population parameter with the hypothesized value of the parameter (i.e., the value specified in the null hypothesis). With reference to our option return example, this means we are concerned with the difference between the mean return of the sample (i.e., $\bar{x} = 0.001$) and the hypothesized mean return (i.e., $\mu_0 = 0$). As indicated in the following expression, the **test statistic** is the difference between the sample statistic and the hypothesized value, scaled by the standard error of the sample statistic.

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothesized value}}{\text{standard error of the sample statistic}}$$

The standard error of the sample statistic is the adjusted standard deviation of the sample. When the sample statistic is the sample mean, \bar{x} , the standard error of the sample statistic for sample size n , is calculated as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

when the population standard deviation, σ , is known, or

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

when the population standard deviation, σ , is not known. In this case, it is estimated using the standard deviation of the sample, s .



PROFESSOR'S NOTE

Don't be confused by the notation here. A lot of the literature you will encounter in your studies simply uses the term $\sigma_{\bar{x}}$ for the standard error of the test statistic, regardless of whether the population standard deviation or sample standard deviation was used in its computation.

As you will soon see, a test statistic is a random variable that may follow one of several distributions, depending on the characteristics of the sample and the population. We will look at four distributions for test statistics: the t -distribution, the z -distribution (standard normal distribution), the chi-square distribution, and the F -distribution. The critical value for the appropriate test statistic—the value against which the computed test statistic is compared—depends on its distribution.

Type I and Type II Errors

Keep in mind that hypothesis testing is used to make inferences about the parameters of a given population on the basis of statistics computed for a sample that is drawn from that population. We must be aware that there is some probability that the sample, in some way, does not represent the population, and any conclusion based on the sample about the population may be made in error.

When drawing inferences from a hypothesis test, there are two types of errors:

- **Type I error:** the rejection of the null hypothesis when it is actually true.
- **Type II error:** the failure to reject the null hypothesis when it is actually false.

The **significance level** is the probability of making a Type I error (rejecting the null when it is true) and is designated by the Greek letter alpha (α). For instance, a significance level of 5% ($\alpha = 0.05$) means there is a 5% chance of rejecting a true null hypothesis. When conducting

hypothesis tests, a significance level must be specified in order to identify the critical values needed to evaluate the test statistic.

The Power of a Test

While the significance level of a test is the probability of rejecting the null hypothesis when it is true, the **power of a test** is the probability of correctly rejecting the null hypothesis when it is false. The power of a test is actually one minus the probability of making a Type II error, or $1 - P(\text{Type II error})$. In other words, the probability of rejecting the null when it is false (power of the test) equals one minus the probability of *not* rejecting the null when it is false (Type II error). When more than one test statistic may be used, the power of the test for the competing test statistics may be useful in deciding which test statistic to use. Ordinarily, we wish to use the test statistic that provides the most powerful test among all possible tests.

Figure 6.4 shows the relationship between the level of significance, the power of a test, and the two types of errors.

Figure 6.4: Type I and Type II Errors in Hypothesis Testing

Decision	True Condition	
	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Incorrect decision Type II error
Reject H_0	Incorrect decision Type I error Significance level, α , $= P(\text{Type I error})$	Correct decision Power of the test $= 1 - P(\text{Type II error})$

Sample size and the choice of significance level (Type I error probability) will together determine the probability of a Type II error. The relation is not simple, however, and calculating the probability of a Type II error in practice is quite difficult. Decreasing the significance level (probability of a Type I error) from 5% to 1%, for example, will increase the probability of failing to reject a false null (Type II error) and therefore reduce the power of the test. Conversely, for a given sample size, we can increase the power of a test only with the cost that the probability of rejecting a true null (Type I error) increases. For a given significance level, we can decrease the probability of a Type II error and increase the power of a test, only by increasing the sample size.

LOS 6.d: Explain a decision rule and the relation between confidence intervals and hypothesis tests, and determine whether a statistically significant result is also economically meaningful.

The decision for a hypothesis test is to either reject the null hypothesis or fail to reject the null hypothesis. Note that it is statistically incorrect to say “accept” the null hypothesis; it can only be supported or rejected. The **decision rule** for rejecting or failing to reject the null hypothesis is based on the distribution of the test statistic. For example, if the test statistic follows a normal distribution, the decision rule is based on critical values determined from the standard normal

distribution (z-distribution). Regardless of the appropriate distribution, it must be determined if a one-tailed or two-tailed hypothesis test is appropriate before a decision rule (rejection rule) can be determined.

A decision rule is specific and quantitative. Once we have determined whether a one- or two-tailed test is appropriate, the significance level we require, and the distribution of the test statistic, we can calculate the exact critical value for the test statistic. Then we have a decision rule of the following form: if the test statistic is (greater, less than) the value X, reject the null.

The Relation Between Confidence Intervals and Hypothesis Tests

A confidence interval is a range of values within which the researcher believes the true population parameter may lie.

A confidence interval is determined as:

$$\left\{ \left[\text{sample statistic} - \left(\text{critical value} \right) \left(\text{standard error} \right) \right] \leq \text{population parameter} \leq \left[\text{sample statistic} + \left(\text{critical value} \right) \left(\text{standard error} \right) \right] \right\}$$

The interpretation of a confidence interval is that for a level of confidence of 95%; for example, there is a 95% probability that the true population parameter is contained in the interval.

From the previous expression, we see that a confidence interval and a hypothesis test are linked by the critical value. For example, a 95% confidence interval uses a critical value associated with a given distribution at the 5% level of significance. Similarly, a hypothesis test would compare a test statistic to a critical value at the 5% level of significance. To see this relationship more clearly, the expression for the confidence interval can be manipulated and restated as:

$$-\text{critical value} \leq \text{test statistic} \leq +\text{critical value}$$

This is the range within which we fail to reject the null for a two-tailed hypothesis test at a given level of significance.

EXAMPLE: Confidence intervals and two-tailed hypothesis tests

A researcher has gathered data on the daily returns on a portfolio of call options over a recent 250-day period. The mean daily return has been 0.1%, and the sample standard deviation of daily portfolio returns is 0.25%. The researcher believes that the mean daily portfolio return is not equal to zero.

1. Construct a 95% confidence interval for the population mean daily return over the 250-day sample period.
2. Construct a hypothesis test of the researcher's belief.

Answer:

1. Given a sample size of 250 with a standard deviation of 0.25%, the standard error can be computed as $s_x = \frac{s}{\sqrt{n}} = \frac{0.25\%}{\sqrt{250}} = 0.0158\%$.

At the 5% level of significance, the critical z-values for the confidence interval are $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$. Thus, given a sample mean equal to 0.1%, the 95% confidence interval for the population mean is:

$$0.1 - 1.96(0.0158) \leq \mu \leq 0.1 + 1.96(0.0158), \text{ or}$$

$$0.069\% \leq \mu \leq 0.131\%$$

2. First we need to specify the null and alternative hypotheses. Note that the null hypothesis must contain "equal to." That is, it can be $=$, \leq , or \geq .

$$H_0: \mu_0 = 0 \text{ versus } H_a: \mu_0 \neq 0$$

Since the null hypothesis is an equality, this is a two-tailed test. At a 5% level of significance, the critical z-values for a two-tailed test are ± 1.96 , so the decision rule can be stated as:

$$\text{Reject } H_0 \text{ if test statistic} < -1.96 \text{ or test statistic} > +1.96$$

Using the standard error of the sample mean we calculated above, our test statistic is:

$$\frac{0.001}{\left(\frac{0.0025}{\sqrt{250}}\right)} = \frac{0.001}{0.000158} = 6.33$$

Since $6.33 > 1.96$, we reject the null hypothesis that the mean daily option return is equal to zero.

Notice the similarity of this analysis with our confidence interval. We rejected the hypothesis $\mu = 0$ because the sample mean of 0.1% is more than 1.96 standard errors from zero. Based on the 95% confidence interval, we reject $\mu = 0$ because zero is more than 1.96 standard errors from the sample mean of 0.1%.

Whether a Statistically Significant Result is Also Economically Meaningful

Statistical significance does not necessarily imply **economic significance**. For example, we may have tested a null hypothesis that a strategy of going long all the stocks that satisfy some criteria and shorting all the stocks that do not satisfy the criteria resulted in returns that were less than or equal to zero over a 20-year period. Assume we have rejected the null in favor of the alternative hypothesis that the returns to the strategy are greater than zero (positive). This does not necessarily mean that investing in that strategy will result in economically meaningful positive returns. Several factors must be considered.

One important consideration is transactions costs. Once we consider the costs of buying and selling the securities, we may find that the mean positive returns to the strategy are not enough to generate positive returns. Taxes are another factor that may make a seemingly attractive strategy a poor one in practice. A third reason that statistically significant results may not be economically significant is risk. In the above strategy, we have additional risk from short sales (they may have to be closed out earlier than in the test strategy). Since the statistically significant results were for a period of 20 years, it may be the case that there is significant variation from year to year in the returns from the strategy, even though the mean strategy

return is greater than zero. This variation in returns from period to period is an additional risk to the strategy that is not accounted for in our test of statistical significance.

Any of these factors could make committing funds to a strategy unattractive, even though the statistical evidence of positive returns is highly significant. By the nature of statistical tests, a very large sample size can result in highly (statistically) significant results that are quite small in absolute terms.



MODULE QUIZ 6.1

1. To test whether the mean of a population is greater than 20, the appropriate null hypothesis is that the population mean is:
 - A. less than 20.
 - B. greater than 20.
 - C. less than or equal to 20.
2. Which of the following statements about hypothesis testing is *most accurate*?
 - A. A Type II error is rejecting the null when it is actually true.
 - B. The significance level equals one minus the probability of a Type I error.
 - C. A two-tailed test with a significance level of 5% has z -critical values of ± 1.96 .
3. For a hypothesis test with a probability of a Type II error of 60% and a probability of a Type I error of 5%, which of the following statements is *most accurate*?
 - A. The power of the test is 40%, and there is a 5% probability that the test statistic will exceed the critical value(s).
 - B. There is a 95% probability that the test statistic will be between the critical values if this is a two-tailed test.
 - C. There is a 5% probability that the null hypothesis will be rejected when actually true, and the probability of rejecting the null when it is false is 40%.
4. If the significance level of a test is 0.05 and the probability of a Type II error is 0.15, what is the power of the test?
 - A. 0.850.
 - B. 0.950.
 - C. 0.975.

MODULE 6.2: P -VALUES AND TESTS OF MEANS



LOS 6.e: Explain and interpret the p -value as it relates to hypothesis testing.

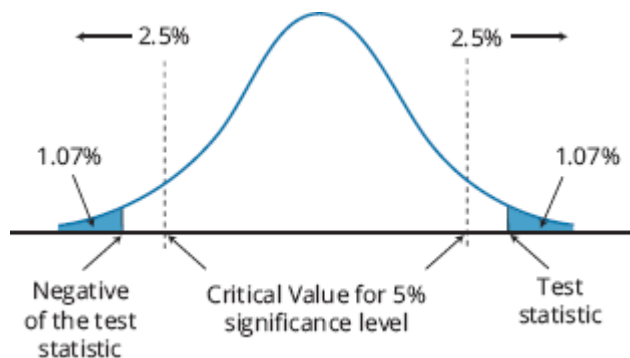
Video covering this content is available online.

The **p -value** is the probability of obtaining a test statistic that would lead to a rejection of the null hypothesis, assuming the null hypothesis is true. It is the smallest level of significance for which the null hypothesis can be rejected. For one-tailed tests, the p -value is the probability that lies above the computed test statistic for upper tail tests or below the computed test statistic for lower tail tests. For two-tailed tests, the p -value is the probability that lies above the positive value of the computed test statistic *plus* the probability that lies below the negative value of the computed test statistic.

Consider a two-tailed hypothesis test about the mean value of a random variable at the 5% significance level where the test statistic is 2.3, greater than the upper critical value of 1.96. If we consult the Z -table, we find the probability of getting a value greater than 2.3 is $(1 - 0.9893)$

= 1.07%. Since it's a two-tailed test, our p -value is $2 \times 1.07 = 2.14\%$, as illustrated in Figure 6.5. At a 3%, 4%, or 5% significance level, we would reject the null hypothesis, but at a 2% or 1% significance level, we would not. Many researchers report p -values without selecting a significance level and allow the reader to judge how strong the evidence for rejection is.

Figure 6.5: Two-Tailed Hypothesis Test With p -Value = 2.14%



LOS 6.f: Describe how to interpret the significance of a test in the context of multiple tests.

Recall that the probability of a Type I error is the probability that a true null hypothesis will be rejected (and is also the significance level of a test). Statisticians refer to these incorrect rejections of the null hypothesis as **false positives**. For a test of the hypothesis that the mean return to an investment strategy is equal to zero, with a significance level of 5%, we will get a false positive 5% of the time, on average. That is, our test statistic will be outside the critical values (in the tails of the distribution) and the p -value of our test statistic will be less than 0.05. If we do a single test, this conclusion is correct.

While we might think that if we get more than 5 false positives with 100 tests, we should reject the null hypothesis, with multiple tests this may be misleading. There is an adjustment to the p -values of our tests, however, that will improve the accuracy of our conclusions.

The procedure is illustrated in the following example. Consider 20 tests at the 10% significance level, for which we get 4 test statistics with p -values less than 10%. This is more than the 10% of 20 that we would expect to get if the null hypothesis is true, and we might believe we should reject the null based on these results. The accepted method to use in this case is to rank the p -values in ascending order and calculate an adjusted significance number for each test with a p -value less than 10%. We then compare these adjusted significance numbers to the reported p -values. We will only count tests as actual rejections if their adjusted significance based on p -value rank is greater than or equal to their reported p -values.

Figure 6.6 illustrates an example of this procedure. We list the p -values for the tests that are less than 10% in ascending order. The adjusted significance levels are calculated with the following formula:

$$\text{Adjusted significance} = \alpha \times \frac{\text{Rank of } p\text{-value}}{\text{Number of tests}}$$

Figure 6.6: Adjusted significance

Rank (Test #)	p -value	Adjusted Significance	p -value \leq adjusted significance
1 (Test 12)	0.004	0.005	Yes
2 (Test 4)	0.010	0.010	Yes
3 (Test 17)	0.053	0.015	No
4 (Test 9)	0.076	0.020	No

From the results reported in Figure 6.6, we see that only two of the tests should actually be counted as rejections. Because only 2 of the 20 tests (tests 12 and 4) qualify as actual rejections based on comparison of their p -values with the adjusted significance values for their rank, our rejection rate is 10%. When the null hypothesis is true, two rejections from 20 tests is just what we would expect with a significance level of 10%. In this case, we will not reject the null hypothesis.



PROFESSOR'S NOTE

The LOS here says, “Describe how to interpret the significance of a test ...” It does not indicate that calculations will be required. Perhaps if you just remember that we compare the reported p -values (ranked from lowest to highest) to the adjusted significance levels (significance level times rank / number of tests), and then count only those tests for which the p -values are less than their adjusted significance levels as rejections, you’ll be able to handle any questions based on this LOS.

LOS 6.g: Identify the appropriate test statistic and interpret the results for a hypothesis test concerning the population mean of both large and small samples when the population is normally or approximately normally distributed and the variance is (1) known or (2) unknown.

When hypothesis testing, the choice between using a critical value based on the t -distribution or the z -distribution depends on sample size, the distribution of the population, and whether or not the variance of the population is known.

The t -Test

The t -test is a widely used hypothesis test that employs a test statistic that is distributed according to a t -distribution. Following are the rules for when it is appropriate to use the t -test for hypothesis tests of the population mean.

Use the t -test if the population variance is unknown and either of the following conditions exist:

- The sample is large ($n \geq 30$).
- The sample is small (less than 30), but the distribution of the population is normal or approximately normal.

If the sample is small and the distribution is nonnormal, we have no reliable statistical test.

The computed value for the test statistic based on the t -distribution is referred to as the t -statistic. For hypothesis tests of a population mean, a t -statistic with $n - 1$ degrees of freedom is computed as:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where:

\bar{x} = sample mean

μ_0 = hypothesized population mean (i.e., the null)

s = standard deviation of the sample

n = sample size



PROFESSOR'S NOTE

This computation is not new. It is the same test statistic computation that we have been performing all along. Note the use of the sample standard deviation, s , in the standard error term in the denominator.

To conduct a t -test, the t -statistic is compared to a critical t -value at the desired level of significance with the appropriate degrees of freedom.

In the real world, the underlying variance of the population is rarely known, so the t -test enjoys widespread application.

The z -Test

The z -test is the appropriate hypothesis test of the population mean when the *population is normally distributed with known variance*. The computed test statistic used with the z -test is referred to as the z -statistic. The z -statistic for a hypothesis test for a population mean is computed as follows:

$$z\text{-statistic} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

where:

\bar{x} = sample mean

μ_0 = hypothesized population mean

σ = standard deviation of the *population*

n = sample size

To test a hypothesis, the z -statistic is compared to the critical z -value corresponding to the significance of the test. Critical z -values for the most common levels of significance are displayed in Figure 6.7. You should have these memorized by now.

Figure 6.7: Critical z -Values

Level of Significance	Two-Tailed Test	One-Tailed Test
0.10 = 10%	± 1.65	+1.28 or -1.28
0.05 = 5%	± 1.96	+1.65 or -1.65
0.01 = 1%	± 2.58	+2.33 or -2.33

When the *sample size is large* and the *population variance is unknown*, the z -statistic is:

$$z\text{-statistic} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where:

\bar{x} = sample mean

μ_0 = hypothesized population mean

s = standard deviation of the *sample*

n = sample size

Note the use of the sample standard deviation, s , versus the population standard deviation, σ . Remember, this is acceptable if the sample size is large, although the t -statistic is the more conservative measure when the population variance is unknown.

EXAMPLE: **z-test or t-test?**

Referring to our previous option portfolio mean return example, determine which test statistic (z or t) should be used.

Answer:

The population variance for our sample of returns is unknown. Hence, the t -distribution is appropriate. With 250 observations, however, the sample is considered to be large, so the z -distribution would also be acceptable. Because our sample is so large, the critical values for the t and z are almost identical. Hence, there is almost no difference in the likelihood of rejecting a true null.

EXAMPLE: **The z-test**

When your company's gizmo machine is working properly, the mean length of gizmos is 2.5 inches. However, from time to time the machine gets out of alignment and produces gizmos that are either too long or too short. When this happens, production is stopped and the machine is adjusted. To check the machine, the quality control department takes a gizmo sample each day. Today, a random sample of 49 gizmos showed a mean length of 2.49 inches. The population standard deviation is known to be 0.021 inches. Using a 5% significance level, determine if the machine should be shut down and adjusted.

Answer:

Let μ be the mean length of all gizmos made by this machine, and let x be the corresponding mean for the sample.

Let's follow the hypothesis testing procedure presented earlier in Figure 6.1. Again, you should know this process!

Statement of hypothesis. For the information provided, the null and alternative hypotheses are appropriately structured as:

$H_0: \mu = 2.5$ (The machine does not need an adjustment.)

$H_a: \mu \neq 2.5$ (The machine needs an adjustment.)

Note that since this is a two-tailed test, H_a allows for values above and below 2.5.

Select the appropriate test statistic. Since the population variance is known and the sample size is > 30 , the z-statistic is the appropriate test statistic. The z-statistic is computed as:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Specify the level of significance. The level of significance is given at 5%, implying that we are willing to accept a 5% probability of rejecting a true null hypothesis.

State the decision rule regarding the hypothesis. The \neq sign in the alternative hypothesis indicates that the test is two-tailed with two rejection regions, one in each tail of the standard normal distribution curve. Because the total area of both rejection regions combined is 0.05 (the significance level), the area of the rejection region in each tail is 0.025. You should know that the critical z-values for $\pm z_{0.025}$ are ± 1.96 . This means that the null hypothesis should not be rejected if the computed z-statistic lies between -1.96 and $+1.96$ and should be rejected if it lies outside of these critical values. The decision rule can be stated as:

Reject H_0 if $-z_{0.025} > z\text{-statistic} > z_{0.025}$, or equivalently,

Reject H_0 if: $-1.96 > z\text{-statistic} > +1.96$

Collect the sample and calculate the test statistic. The value of x from the sample is 2.49. Since σ is given as 0.021, we calculate the z-statistic using σ as follows:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{2.49 - 2.5}{0.021 / \sqrt{49}} = \frac{-0.01}{0.003} = -3.33$$

Make a decision regarding the hypothesis. The calculated value of the z-statistic is -3.33 . Since this value is less than the critical value, $-z_{0.025} = -1.96$, it falls in the rejection region in the left tail of the z-distribution. Hence, there is sufficient evidence to reject H_0 .

Make a decision based on the results of the test. Based on the sample information and the results of the test, it is concluded that the machine is out of adjustment and should be shut down for repair.

EXAMPLE: One-tailed test

Using the data from the previous example and a 5% significance level, test the hypothesis that the mean length of gizmos is less than 2.5 inches.

Answer:

In this case, we use a one-tailed test with the following structure:

$$H_0: \mu \geq 2.5 \text{ versus } H_a: \mu < 2.5$$

The appropriate decision rule for this one-tailed test at a significance level of 5% is:

Reject H_0 if test statistic < -1.645

The test statistic is computed in the same way, regardless of whether we are using a one-tailed or a two-tailed test. From the previous example, we know that the test statistic for the gizmo

sample is -3.33 . Because $-3.33 < -1.645$, we can reject the null hypothesis and conclude that the mean length is statistically less than 2.5 at a 5% level of significance.



MODULE QUIZ 6.2

1. A researcher has 28 quarterly excess returns to an investment strategy and believes these returns are approximately normally distributed. The mean return on this sample is 1.645% and the standard deviation is 5.29%. For a test with a 5% significance level of the hypothesis that excess returns are less than or equal to zero, the researcher should:
 - A. reject the null hypothesis because the critical value for the test is 1.645.
 - B. not draw any conclusion because the sample size is less than 30.
 - C. fail to reject the null because the critical value is greater than 1.645.
2. An analyst wants to test a hypothesis concerning the population mean of monthly returns for a composite that has existed for 24 months. The analyst may appropriately use:
 - A. a t -test but not a z -test if returns for the composite are normally distributed.
 - B. either a t -test or a z -test if returns for the composite are normally distributed.
 - C. a t -test but not a z -test, regardless of the distribution of returns for the composite.

Use the following segment of Student's t -distribution for Question 3.

Level of Significance for One-Tailed Test				
df	0.100	0.050	0.025	0.01
Level of Significance for Two-Tailed Test				
df	0.20	0.10	0.05	0.02
11	1.363	1.796	2.201	2.718
12	1.356	1.782	2.179	2.681
13	1.350	1.771	2.160	2.650
14	1.345	1.761	2.145	2.624
15	1.341	1.753	2.131	2.602

3. From a sample of 14 observations, an analyst calculates a t -statistic to test a hypothesis that the population mean is equal to zero. If the analyst chooses a 5% significance level, the appropriate critical value is:
 - A. less than 1.80.
 - B. greater than 2.15.
 - C. between 1.80 and 2.15.

MODULE 6.3: MEAN DIFFERENCES AND DIFFERENCE IN MEANS



Video covering this content is available online.

LOS 6.h: Identify the appropriate test statistic and interpret the results for a hypothesis test concerning the equality of the population means of two at least approximately normally distributed populations based on independent random samples with equal assumed variances.

Up to this point, we have been concerned with tests of a single population mean. In practice, we frequently want to know if there is a difference between the means of two populations. The t -

test for differences between means requires that we are reasonably certain that our samples are independent and that they are taken from two populations that are normally distributed.



PROFESSOR'S NOTE

Please note the language of the LOS here. Candidates must “Identify the appropriate test statistic and interpret the results of a hypothesis test....” Certainly you should know that this is a t -test, and that we reject the hypothesis of equality when the test statistic is outside the critical t -values. Don't worry about memorizing the following formulas.

A pooled variance is used with the t -test for testing the hypothesis that the means of two normally distributed populations are equal, when the variances of the populations are unknown but assumed to be equal.

Assuming independent samples, the t -statistic is computed as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

where:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

s_1^2 = variance of the first sample

s_2^2 = variance of the second sample

n_1 = number of observations in the first sample

n_2 = number of observations in the second sample

Note: The degrees of freedom, df , is $(n_1 + n_2 - 2)$.

Since we assume that the variances are equal, we just add the variances of the two sample means in order to calculate the standard error in the denominator.

The intuition here is straightforward. If the sample means are very close together, the numerator of the t -statistic (and the t -statistic itself) are small, and we do not reject equality. If the sample means are far apart, the numerator of the t -statistic (and the t -statistic itself) are large, and we reject equality. Perhaps not as easy to remember is the fact that this test is only valid for two populations that are independent and normally distributed.

EXAMPLE: Difference between means – equal variances

Sue Smith is investigating whether the abnormal returns for acquiring firms during merger announcement periods differ for horizontal and vertical mergers. She estimates the abnormal returns for a sample of acquiring firms associated with horizontal mergers and a sample of acquiring firms involved in vertical mergers. Smith finds that abnormal returns from horizontal mergers have a mean of 1.0% and a standard deviation of 1.0%, while abnormal returns from vertical mergers have a mean of 2.5% and a standard deviation of 2.0%.

Smith assumes that the samples are independent, the population means are normally distributed, and the population variances are equal.

Smith calculates the t -statistic as -5.474 and the degrees of freedom as 120 . Using a 5% significance level, should Smith reject or fail to reject the null hypothesis that the abnormal returns to acquiring firms during the announcement period are the same for horizontal and vertical mergers?

Answer:

Since this is a two-tailed test, the structure of the hypotheses takes the following form:

$$H_0: \mu_1 - \mu_2 = 0 \text{ versus } H_a: \mu_1 - \mu_2 \neq 0$$

where:

μ_1 = the mean of the abnormal returns for the horizontal mergers

μ_2 = the mean of the abnormal returns for the vertical mergers

From the following t -table segment, the critical t -value for a 5% level of significance at $\alpha / 2 = p = 0.025$ with $df = 120$, is 1.980 .

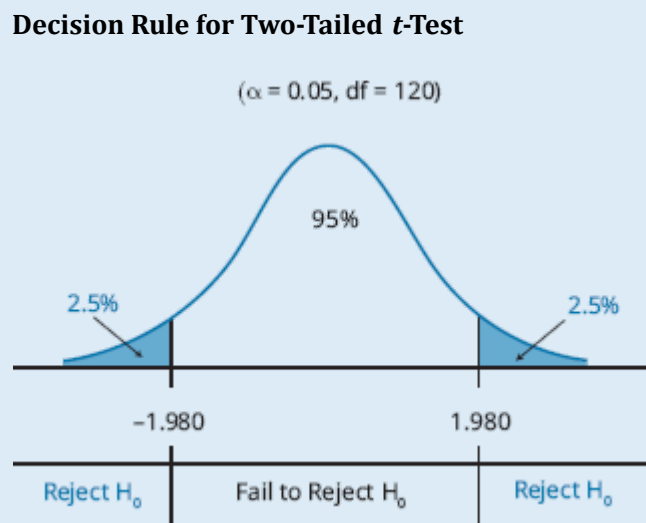
Partial t -Table

df	One-Tailed Probabilities (p)		
	$p = 0.10$	$p = 0.05$	$p = 0.025$
110	1.289	1.659	1.982
120	1.289	1.658	1.980
200	1.286	1.653	1.972

Thus, the decision rule can be stated as:

Reject H_0 if t -statistic < -1.980 or t -statistic > 1.980

The rejection region for this test is illustrated in the following figure.



Since the test statistic, -5.474 , falls to the left of the lower critical t -value, Smith can reject the null hypothesis and conclude that mean abnormal returns are different for horizontal and vertical mergers.

LOS 6.i: Identify the appropriate test statistic and interpret the results for a hypothesis test concerning the mean difference of two normally distributed populations.

While the test in the previous section was of the difference between the means of two independent samples, sometimes our samples may be dependent. If the observations in the two samples both depend on some other factor, we can construct a “paired comparisons” test of whether the means of the differences between observations for the two samples are different. Dependence may result from an event that affects both sets of observations for a number of companies or because observations for two firms over time are both influenced by market returns or economic conditions.

For an example of a paired comparisons test, consider a test of whether the returns on two steel firms were equal over a 5-year period. We can't use the difference in means test because we have reason to believe that the samples are not independent. To some extent, both will depend on the returns on the overall market (market risk) and the conditions in the steel industry (industry-specific risk). In this case, our pairs will be the returns on each firm over the same time periods, so we use the differences in monthly returns for the two companies. The paired comparisons test is just a test of whether the average difference between monthly returns is significantly different from zero, based on the standard error of the differences in monthly returns.

Remember, the paired comparisons test also requires that the sample data be normally distributed. Although we frequently just want to test the hypothesis that the mean of the differences in the pairs is zero ($\mu_{dz} = 0$), the general form of the test for any hypothesized mean difference, μ_{dz} , is as follows:

$$H_0: \mu_d = \mu_{dz} \text{ versus } H_a: \mu_d \neq \mu_{dz}$$

where:

μ_d = mean of the population of paired differences

μ_{dz} = hypothesized mean of paired differences, which is commonly zero

For one-tail tests, the hypotheses are structured as either:

$$H_0: \mu_d \leq \mu_{dz} \text{ versus } H_a: \mu_d > \mu_{dz}, \text{ or } H_0: \mu_d \geq \mu_{dz} \text{ versus } H_a: \mu_d < \mu_{dz}$$

For the paired comparisons test, the t -statistic with $n - 1$ degrees of freedom is computed as:

$$t = \frac{\bar{d} - \mu_{dz}}{s_{\bar{d}}}$$

where:

$$\bar{d} = \text{sample mean difference} = \frac{1}{n} \sum_{i=1}^n d_i$$

d_i = difference between the i th pair of observations

$$s_{\bar{d}} = \text{standard error of the mean difference} = \frac{s_d}{\sqrt{n}}$$

$$s_d = \text{sample standard deviation} = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

n = the number of paired observations

EXAMPLE: Paired comparisons test

Joe Andrews is examining changes in estimated betas for the common stock of companies in the telecommunications industry before and after deregulation. Andrews believes that the betas may decline because of deregulation since companies are no longer subject to the uncertainties of rate regulation or that they may increase because there is more uncertainty regarding competition in the industry. Andrews calculates a t -statistic of 10.26 for this hypothesis test, based on a sample size of 39. Using a 5% significance level, determine whether there is a change in betas.

Answer:

Because the mean difference may be positive or negative, a two-tailed test is in order here. Thus, the hypotheses are structured as:

$$H_0: \mu_d = 0 \text{ versus } H_a: \mu_d \neq 0$$

There are $39 - 1 = 38$ degrees of freedom. Using the t -distribution, the two-tailed critical t -values for a 5% level of significance with $df = 38$ is ± 2.024 . As indicated in the following table, the critical t -value of 2.024 is located at the intersection of the $p = 0.025$ column and the $df = 38$ row. The one-tailed probability of 0.025 is used because we need 2.5% in each tail for 5% significance with a two-tailed test.

Partial t -Table

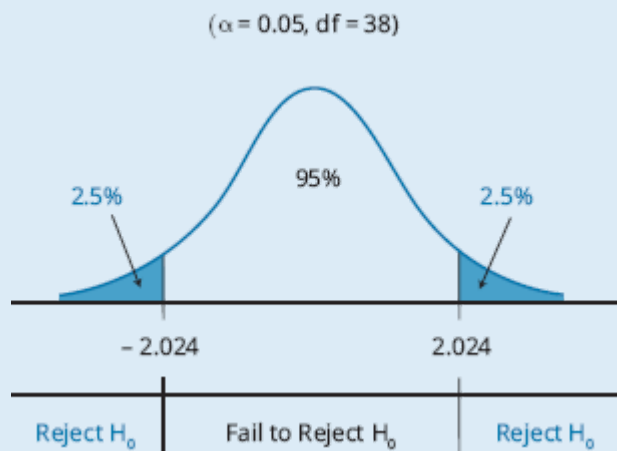
df	One-Tailed Probabilities (p)		
	$p = 0.10$	$p = 0.05$	$p = 0.025$
38	1.304	1.686	2.024
39	1.304	1.685	2.023
40	1.303	1.684	2.021

Thus, the decision rule becomes:

$$\text{Reject } H_0 \text{ if } t\text{-statistic} < -2.024, \text{ or } t\text{-statistic} > 2.024$$

This decision rule is illustrated in the following figure.

Decision Rule for a Two-Tailed Paired Comparisons Test



The test statistic, 10.26, is greater than the critical t -value, 2.024—it falls in the rejection region to the right of 2.024 in the previous figure. Thus, we reject the null hypothesis of no

difference, concluding that there *is* a statistically significant difference between mean firm betas before and after deregulation.

Keep in mind that we have been describing two distinct hypothesis tests, one about the significance of the difference between the means of two populations and one about the significance of the mean of the differences between pairs of observations. Here are rules for when these tests may be applied:

- The test of the differences in means is used when there are two *independent samples*.
- A test of the significance of the mean of the differences between paired observations is used when the samples are *not independent*.



PROFESSOR'S NOTE

The LOS here say “Identify the appropriate test statistic and interpret the results ...” I can’t believe candidates are expected to memorize these formulas (or that you would be a better analyst if you did). You should instead focus on the fact that both of these tests involve *t*-statistics and depend on the degrees of freedom. Also note that when samples are independent, you can use the difference in means test, and when they are dependent, we must use the paired comparison (mean differences) test. In that case, with a null hypothesis that there is no difference in means, the test statistic is simply the mean of the differences between each pair of observations, divided by the standard error of those differences. This is just a straightforward *t*-test of whether the mean of a sample is zero, which might be considered “fair game” for the exam.



MODULE QUIZ 6.3

1. Which of the following assumptions is *least likely* required for the difference in means test based on two samples?
 - A. The two samples are independent.
 - B. The two populations are normally distributed.
 - C. The two populations have equal variances.
2. William Adams wants to test whether the mean monthly returns over the last five years are the same for two stocks. If he assumes that the returns distributions are normal and have equal variances, the type of test and test statistic are *best* described as:
 - A. paired comparisons test, *t*-statistic.
 - B. paired comparisons test, *F*²-statistic.
 - C. difference in means test, *t*-statistic.

MODULE 6.4: TESTS OF VARIANCE, CORRELATION, AND INDEPENDENCE



Video covering this content is available online.

LOS 6.j: Identify the appropriate test statistic and interpret the results for a hypothesis test concerning (1) the variance of a normally distributed population and (2) the equality of the variances of two normally distributed populations based on two independent random samples.

The *chi-square test* is used for hypothesis tests concerning the variance of a normally distributed population. Letting σ^2 represent the true population variance and σ_0^2 represent the hypothesized variance, the hypotheses for a two-tailed test of a single population variance are structured as:

$$H_0: \sigma^2 = \sigma_0^2 \text{ versus } H_a: \sigma^2 \neq \sigma_0^2$$

The hypotheses for one-tailed tests are structured as:

$$H_0: \sigma^2 \leq \sigma_0^2 \text{ versus } H_a: \sigma^2 > \sigma_0^2 \text{ or}$$

$$H_0: \sigma^2 \geq \sigma_0^2 \text{ versus } H_a: \sigma^2 < \sigma_0^2$$

Hypothesis testing of the population variance requires the use of a chi-square distributed test statistic, denoted χ^2 . The **chi-square distribution** is asymmetrical and approaches the normal distribution in shape as the degrees of freedom increase.

To illustrate the chi-square distribution, consider a two-tailed test with a 5% level of significance and 30 degrees of freedom. As displayed in Figure 6.8, the critical chi-square values are 16.791 and 46.979 for the lower and upper bounds, respectively. These values are obtained from a chi-square table, which is used in the same manner as a *t*-table. A portion of a chi-square table is presented in Figure 6.9.

Note that the chi-square values in Figure 6.9 correspond to the probabilities in the right tail of the distribution. As such, the 16.791 in Figure 6.8 is from the column headed 0.975 because 95% + 2.5% of the probability is to the right of it. The 46.979 is from the column headed 0.025 because only 2.5% probability is to the right of it. Similarly, at a 5% level of significance with 10 degrees of freedom, Figure 6.9 shows that the critical chi-square values for a two-tailed test are 3.247 and 20.483.

Figure 6.8: Decision Rule for a Two-Tailed Chi-Square Test

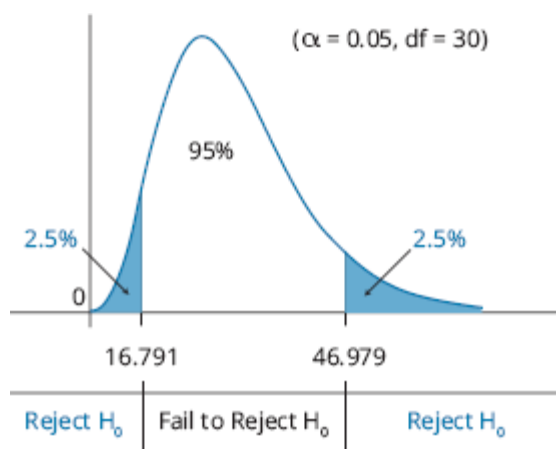


Figure 6.9: Chi-Square Table

Degrees of Freedom	Probability in Right Tail					
	0.975	0.95	0.90	0.1	0.05	0.025
9	2.700	3.325	4.168	14.684	16.919	19.023
10	3.247	3.940	4.865	15.987	18.307	20.483
11	3.816	4.575	5.578	17.275	19.675	21.920
30	16.791	18.493	20.599	40.256	43.773	46.979

The chi-square test statistic, χ^2 , with $n - 1$ degrees of freedom, is computed as:

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

where:

n = sample size

s^2 = sample variance

σ_0^2 = hypothesized value for the population variance.

Similar to other hypothesis tests, the chi-square test compares the test statistic, χ_{n-1}^2 , to a critical chi-square value at a given level of significance and $n - 1$ degrees of freedom. Note that since the chi-square distribution is bounded below by zero, chi-square values cannot be negative.

EXAMPLE: Chi-square test for a single population variance

Historically, High-Return Equity Fund has advertised that its monthly returns have a standard deviation equal to 4%. This was based on estimates from the 2005–2013 period. High-Return wants to verify whether this claim still adequately describes the standard deviation of the fund's returns. High-Return collected monthly returns for the 24-month period between 2013 and 2015 and measured a standard deviation of monthly returns of 3.8%. High-Return calculates a test statistic of 20.76. Using a 5% significance level, determine if the more recent standard deviation is different from the advertised standard deviation.

Answer:

The null hypothesis is that the standard deviation is equal to 4% and, therefore, the variance of monthly returns for the population is $(0.04)^2 = 0.0016$. Since High-Return simply wants to test whether the standard deviation has changed, up or down, a two-sided test should be used. The hypothesis test structure takes the form:

$$H_0: \sigma_0^2 = 0.0016 \text{ versus } H_a: \sigma^2 \neq 0.0016$$

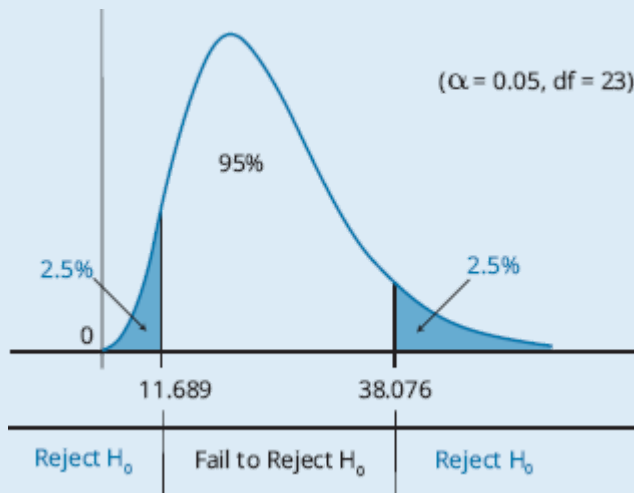
The appropriate test statistic for tests of variance is a chi-square statistic.

With a 24-month sample, there are 23 degrees of freedom. Using the table of chi-square values in Appendix E at the back of this book, for 23 degrees of freedom and probabilities of 0.975 and 0.025, we find two critical values, 11.689 and 38.076. Thus, the decision rule is:

$$\text{Reject } H_0 \text{ if } \chi^2 < 11.689, \text{ or } \chi^2 > 38.076$$

This decision rule is illustrated in the following figure.

Decision Rule for a Two-Tailed Chi-Square Test of a Single Population Variance



Since the computed test statistic, χ^2 , falls between the two critical values, we cannot reject the null hypothesis that the variance is equal to 0.0016. The recently measured standard deviation is close enough to the advertised standard deviation that we cannot say that it is different from 4%, at a 5% level of significance.

Testing the Equality of the Variances of Two Normally Distributed Populations, Based on Two Independent Random Samples

The hypotheses concerned with the equality of the variances of two populations are tested with an F -distributed test statistic. Hypothesis testing using a test statistic that follows an F -distribution is referred to as the F -test. The F -test is used under the assumption that the populations from which samples are drawn are normally distributed and that the samples are independent.

If we let σ_1^2 and σ_2^2 represent the variances of normal Population 1 and Population 2, respectively, the hypotheses for the two-tailed F -test of differences in the variances can be structured as:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ versus } H_a: \sigma_1^2 \neq \sigma_2^2$$

and the one-sided test structures can be specified as:

$$H_0: \sigma_1^2 \leq \sigma_2^2 \text{ versus } H_a: \sigma_1^2 > \sigma_2^2, \text{ or } H_0: \sigma_1^2 \geq \sigma_2^2 \text{ versus } H_a: \sigma_1^2 < \sigma_2^2$$

The test statistic for the F -test is the ratio of the sample variances. The F -statistic is computed as:

$$F = \frac{s_1^2}{s_2^2}$$

where:

s_1^2 = variance of the sample of n_1 observations drawn from Population 1

s_2^2 = variance of the sample of n_2 observations drawn from Population 2

Note that $n_1 - 1$ and $n_2 - 1$ are the degrees of freedom used to identify the appropriate critical value from the F -table (provided in the Appendix).

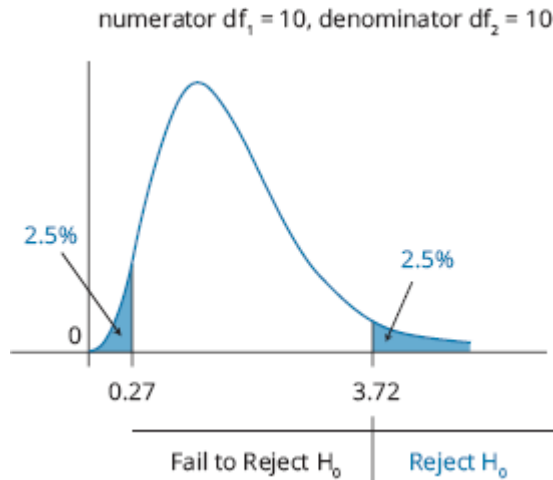


PROFESSOR'S NOTE

Always put the larger variance in the numerator (s_1^2). Following this convention means we only have to consider the critical value for the right-hand tail.

An F -distribution is presented in Figure 6.10. As indicated, the F -distribution is right-skewed and is bounded by zero on the left-hand side. The shape of the F -distribution is determined by *two separate degrees of freedom*, the numerator degrees of freedom, df_1 , and the denominator degrees of freedom, df_2 .

Figure 6.10: F -Distribution



Note that when the sample variances are equal, the value of the test statistic is 1. The upper critical value is always greater than one (the numerator is significantly greater than the denominator), and the lower critical value is always less than one (the numerator is significantly smaller than the denominator). In fact, the lower critical value is the reciprocal of the upper critical value. For this reason, in practice we put the larger sample variance in the numerator and consider only the upper critical value.

EXAMPLE: F -test for equal variances

Annie Cower is examining the earnings for two different industries. Cower suspects that the variance of earnings in the textile industry is different from the variance of earnings in the paper industry. To confirm this suspicion, Cower has looked at a sample of 31 textile manufacturers and a sample of 41 paper companies. She measured the sample standard deviation of earnings across the textile industry to be \$4.30 and that of the paper industry companies to be \$3.80. Cower calculates a test statistic of 1.2805. Using a 5% significance level, determine if the earnings of the textile industry have a different standard deviation than those of the paper industry.

Answer:

In this example, we are concerned with whether the variance of earnings for companies in the textile industry is equal to the variance of earnings for companies in the paper industry. As such, the test hypotheses can be appropriately structured as:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ versus } H_a: \sigma_1^2 \neq \sigma_2^2$$

For tests of difference between variances, the appropriate test statistic is:

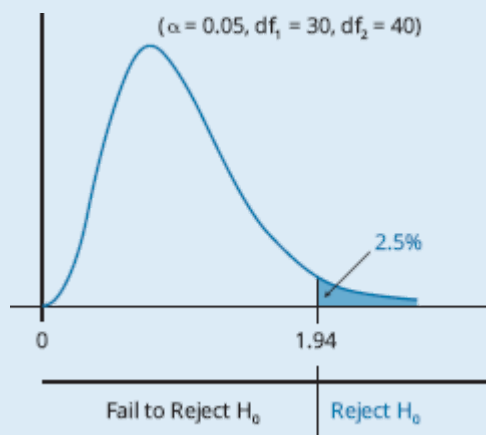
$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 is the larger sample variance.

Using the sample sizes for the two industries, the critical F -value for our test is found to be 1.94. This value is obtained from the table of the F -distribution for 2.5% in the upper tail, with $df_1 = 30$ and $df_2 = 40$. Thus, if the computed F -statistic is greater than the critical value of 1.94, the null hypothesis is rejected. The decision rule, illustrated in the following figure, can be stated as:

Reject H_0 if $F > 1.94$

Decision Rule for F -Test



Since the calculated F -statistic of 1.2805 is less than the critical F -statistic of 1.94, Cower cannot reject the null hypothesis. Cower should conclude that the earnings variances of the industries are not significantly different from one another at a 5% level of significance.

LOS 6.k: Compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test.

Parametric tests rely on assumptions regarding the distribution of the population and are specific to population parameters. For example, the z -test relies upon a mean and a standard deviation to define the normal distribution. The z -test also requires that either the sample is large, relying on the central limit theorem to assure a normal sampling distribution, or that the population is normally distributed.

Nonparametric tests either do not consider a particular population parameter or have few assumptions about the population that is sampled. Nonparametric tests are used when there is concern about quantities other than the parameters of a distribution or when the assumptions of parametric tests can't be supported. They are also used when the data are not suitable for parametric tests (e.g., ranked observations).

Situations where a nonparametric test is called for are the following:

1. The assumptions about the distribution of the random variable that support a parametric test are not met. An example would be a hypothesis test of the mean value for a variable that comes from a distribution that is not normal and is of small size so that neither the t -test nor the z -test is appropriate.
2. When data are ranks (an ordinal measurement scale) rather than values.
3. The hypothesis does not involve the parameters of the distribution, such as testing whether a variable is normally distributed. We can use a nonparametric test, called a runs test, to determine whether data are random. A runs test provides an estimate of the probability that a series of changes (e.g., +, +, -, -, +, -,....) are random.

LOS 6.I: 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.

Correlation measures the strength of the relationship between two variables. If the correlation between two variables is zero, there is no linear relationship between them. When the sample correlation coefficient for two variables is different from zero, we must address the question of whether the true population correlation coefficient (ρ) is equal to zero. The appropriate test statistic for the hypothesis that the population correlation equals zero, when the two variables are normally distributed, is:

$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where r = sample correlation and n = sample size. This test statistic follows a t -distribution with $n - 2$ degrees of freedom. It is worth noting that the test statistic increases, not only with the sample correlation coefficient, but also with sample size.

EXAMPLE: Test of the hypothesis that the population correlation coefficient equals zero

A researcher computes the sample correlation coefficient for two normally distributed random variables as 0.35, based on a sample size of 42. Determine whether to reject the hypothesis that the population correlation coefficient is equal to zero at a 5% significance level.

Answer:

Our test statistic is $\frac{0.35\sqrt{42-2}}{\sqrt{1-0.35^2}} = 2.363$. Using the t -table with $42 - 2 = 40$ degrees of freedom for a two-tailed test and a significance level of 5%, we can find the critical value of 2.021. Because our computed test statistic of 2.363 is greater than 2.021, we reject the hypothesis that the population mean is zero and conclude that it is not equal to zero. That is, the two populations are correlated, in this case positively.



PROFESSOR'S NOTE

The correlation coefficient we refer to here is the Pearson correlation coefficient, which is a measure of the linear relationship between two variables. There are other

correlation coefficients that better measure the strength of any non-linear relationship between two variables.

The **Spearman rank correlation test**, a non-parametric test, can be used to test whether two sets of ranks are correlated. Ranks are simply ordered values. If there is a tie (equal values), the ranks are shared, so if 2nd and 3rd rank is the same, the ranks are shared and each gets a rank if $(2 + 3) / 2 = 2.5$.

The Spearman rank correlation, r_s , (when all ranks are integer values) is calculated as:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where:

r_s = rank correlation

n = sample size

d_i = difference between two ranks

We can test the significance of the Spearman rank correlation calculated with the above formula using the same test statistic we used for estimating the significance of a parametric (Pearson) correlation coefficient:

$$\frac{r_s \sqrt{n - 2}}{\sqrt{1 - r_s^2}}$$

When the sample size is greater than 30, the test statistic follows a t -distribution with $n - 2$ degrees of freedom.

LOS 6.m: Explain tests of independence based on contingency table data.

A contingency or two-way table shows the number of observations from a sample that have a combination of two characteristics. Figure 6.11 is a contingency table where the characteristics are earnings growth (low, medium, or high) and dividend yield (low, medium, or high). We can use the data in the table to test the hypothesis that the two characteristics, earnings growth and dividend yield, are independent of each other.

Figure 6.11: Contingency Table for Categorical Data

Earnings Growth	Dividend Yield			Total
	Low	Medium	High	
Low	28	53	42	123
Medium	42	32	39	113
High	49	25	14	88
Total	119	110	95	324

We index our three categories of earnings growth from low to high with $i = 1, 2, \text{ or } 3$, and our three categories of dividend yield from low to high with $j = 1, 2, \text{ or } 3$. From the table, we see in cell 1,1 that 28 firms have both low earnings growth and low dividend yield. We see in cell 3,2 that 25 firms have high earnings growth and medium dividends.

For our test, we are going to compare the actual table values to what the values would be if the two characteristics were independent. The test statistic is a chi-square test statistic calculated as:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

O_{ij} = the number of observations in cell i, j , row i and column j (i.e., observed frequency)

E_{ij} = the expected number of observations for cell i, j

r = the number of row categories

c = the number of column categories

The degrees of freedom are $[(r - 1)(c - 1)]$, which is 4 in our example for dividend yield and earnings growth.

$E_{i,j}$, the expected number of observations in cell ij , is: $\frac{\text{total for row } i \times \text{total for column } j}{\text{total for all columns and rows}}$.

The expected number of observations for cell 2,2 is $\frac{110 \times 113}{324} = 38.4$.

In calculating our test statistic, the term for cell 2,2 is then $\frac{(32 - 38.4)^2}{38.4} = 1.067$.

Figure 6.12 shows the expected frequencies for each pair of categories in our earnings growth and dividend yield contingency table.

Figure 6.12: Contingency Table for Expected Frequencies

Earnings Growth	Dividend Yield		
	Low	Medium	High
Low	45.2	41.8	36.1
Medium	41.5	38.4	33.1
High	32.3	29.9	25.8

For our test statistic, we sum, for all nine cells, the squared difference between the expected frequency and observed frequency, divided by the expected frequency. The resulting sum is 27.43.

Our degrees of freedom are $(3 - 1) \times (3 - 1) = 4$. The critical value for a significance level of 5% (from the chi-square table in the Appendix) with 4 degrees of freedom is 9.488. Based on our sample data, we can reject the hypothesis that the earnings growth and dividend yield categories are independent.



MODULE QUIZ 6.4

- The appropriate test statistic for a test of the equality of variances for two normally distributed random variables, based on two independent random samples, is:
 - the t -test.
 - the F -test.
 - the χ^2 test.
- The appropriate test statistic to test the hypothesis that the variance of a normally distributed population is equal to 13 is:

- A. the t -test.
 - B. the F -test.
 - C. the χ^2 test.
3. For a parametric test of whether a correlation coefficient is equal to zero, it is *least likely* that:
 - A. degrees of freedom are $n - 1$.
 - B. the test statistic follows a t -distribution.
 - C. the test statistic increases with a greater sample size.
 4. The test statistic for a Spearman rank correlation test for a sample size greater than 30 follows:
 - A. a t -distribution.
 - B. a normal distribution.
 - C. a chi-square distribution.
 5. A contingency table can be used to test:
 - A. a null hypothesis that rank correlations are equal to zero.
 - B. whether multiple characteristics of a population are independent.
 - C. the number of p -values from multiple tests that are less than adjusted critical values.

KEY CONCEPTS

LOS 6.a

The hypothesis testing process requires a statement of a null and an alternative hypothesis, the selection of the appropriate test statistic, specification of the significance level, a decision rule, the calculation of a sample statistic, a decision regarding the hypotheses based on the test, and a decision based on the test results.

The null hypothesis is what the researcher wants to reject. The alternative hypothesis is what the researcher wants to support, and it is accepted when the null hypothesis is rejected.

LOS 6.b

A two-tailed test results from a two-sided alternative hypothesis (e.g., $H_a: \mu \neq \mu_0$). A one-tailed test results from a one-sided alternative hypothesis (e.g., $H_a: \mu > \mu_0$, or $H_a: \mu < \mu_0$).

LOS 6.c

The test statistic is the value that a decision about a hypothesis will be based on. For a test about the value of the mean of a distribution:

$$\text{test statistic} = \frac{\text{sample mean} - \text{hypothesized mean}}{\text{standard error of the sample mean}}$$

A Type I error is the rejection of the null hypothesis when it is actually true, while a Type II error is the failure to reject the null hypothesis when it is actually false.

The significance level can be interpreted as the probability that a test statistic will reject the null hypothesis by chance when it is actually true (i.e., the probability of a Type I error). A significance level must be specified to select the critical values for the test.

The power of a test is the probability of rejecting the null when it is false. The power of a test = $1 - P(\text{Type II error})$.

LOS 6.d

Hypothesis testing compares a computed test statistic to a critical value at a stated level of significance, which is the decision rule for the test.

A hypothesis about a population parameter is rejected when the sample statistic lies outside a confidence interval around the hypothesized value for the chosen level of significance.

Statistical significance does not necessarily imply economic significance. Even though a test statistic is significant statistically, the size of the gains to a strategy to exploit a statistically significant result may be absolutely small or simply not great enough to outweigh transactions costs.

LOS 6.e

The p -value for a hypothesis test is the smallest significance level for which the hypothesis would be rejected. For example, a p -value of 7% means the hypothesis can be rejected at the 10% significance level but cannot be rejected at the 5% significance level.

LOS 6.f

When multiple tests are performed on different samples from a population, the p -values of each test are ranked, from lowest to highest, and compared to the adjusted critical values for each rank. When the proportion of the total number of ranked tests for which reported p -values are less than their adjusted critical values is greater than the significance level, the null hypothesis is rejected.

LOS 6.g

With unknown population variance, the t -statistic is used for tests about the mean of a normally distributed population: $t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$. If the population variance is known, the appropriate test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ for tests about the mean of a population.

LOS 6.h

For two independent samples from two normally distributed populations, the difference in means can be tested with a t -statistic. When the two population variances are assumed to be equal, the denominator is based on the variance of the pooled samples.

LOS 6.i

A paired comparisons test is concerned with the mean of the differences between the paired observations of two dependent, normally distributed samples. A t -statistic, $t = \frac{\bar{d}}{s_d}$, where $s_d = \frac{s_d}{\sqrt{n}}$, and \bar{d} is the average difference of the n paired observations, is used to test whether the means of two dependent normal variables are equal. Values outside the critical t -values lead us to reject equality.

LOS 6.j

The test of a hypothesis about the population variance for a normally distributed population uses a chi-square test statistic: $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma_0^2}$, where n is the sample size, s^2 is the sample variance, and σ_0^2 is the hypothesized value for the population variance. Degrees of freedom are $n - 1$.

The test comparing two variances based on independent samples from two normally distributed populations uses an F -distributed test statistic: $F = \frac{s_1^2}{s_2^2}$, where s_1^2 is the variance of the first sample and s_2^2 is the (smaller) variance of the second sample.

LOS 6.k

Parametric tests, like the t -test, F -test, and chi-square tests, make assumptions regarding the distribution of the population from which samples are drawn. Nonparametric tests either do not consider a particular population parameter or have few assumptions about the sampled population. Nonparametric tests are used when the assumptions of parametric tests can't be supported or when the data are not suitable for parametric tests.

LOS 6.l

To test a hypothesis that a population correlation coefficient equals zero, the appropriate test statistic is a t -statistic with $n - 2$ degrees of freedom, calculated as $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where r is the sample correlation coefficient.

A non-parametric test of correlation can be performed when we have only ranks (e.g., deciles of investment performance). The Spearman rank correlation test tests whether the ranks for

multiple periods are correlated. The rank correlation is $r = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$, where d_i^2 is the sum of the squared difference in pairs of ranks and n is the number of sample periods. The test statistic follows a t -distribution for samples sizes greater than 30.

LOS 6.m

A contingency table can be used to test the hypothesis that two characteristics (categories) of a sample of items are independent. The test statistic follows a chi-square distribution and is calculated as:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

O_{ij} = the number of observations in cell i, j , row i and column j (i.e., observed frequency)

E_{ij} = the expected number of observations for cell i, j of the contingency table with independence

r = the number of row categories and c = the number of column categories

The degrees of freedom are $[(r - 1)(c - 1)]$. If the test statistic is greater than the critical chi-square value for a given level of significance, we reject the hypothesis that the two characteristics are independent.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 6.1

- C** To test whether the population mean is greater than 20, the test would attempt to reject the null hypothesis that the mean is less than or equal to 20. The null hypothesis must always include the "equal to" condition. (LOS 6.a)
- C** Rejecting the null when it is actually true is a Type I error. A Type II error is failing to reject the null hypothesis when it is false. The significance level equals the probability of a Type I error. (LOS 6.c)

- C** A Type I error is rejecting the null hypothesis when it's true. The probability of rejecting a false null is $[1 - \text{Prob Type II}] = [1 - 0.60] = 40\%$, which is called the power of the test. A and B are not necessarily true, since the null may be false and the probability of rejection unknown. (LOS 6.c)
- A** The power of a test is $1 - P(\text{Type II error}) = 1 - 0.15 = 0.85$. (LOS 6.c)

Module Quiz 6.2

- C** The standard error is $\frac{5.29}{\sqrt{28}} = 1.0$. Test statistic = $1.645/1.0 = 1.645$. The critical value for t -test is greater than the critical value for a z -test at a 5% significance level (which is 1.645 for a one-tailed test), so the calculated test statistic of 1.645 must be less than the critical value for a t -test (which is 1.703 for a one-tailed test with 27 degrees of freedom) and we cannot reject the null hypothesis that mean excess return is greater than zero. (LOS 6.f)
- A** With a small sample size, a t -test may be used if the population is approximately normally distributed. If the population has a nonnormal distribution, no test statistic is available unless the sample size is large. (LOS 6.g)
- B** This is a two-tailed test with $14 - 1 = 13$ degrees of freedom. From the t -table, 2.160 is the critical value to which the analyst should compare the calculated t -statistic. (LOS 6.g)

Module Quiz 6.3

- C** When the variances are assumed to be unequal, we just calculate the denominator (standard error) differently and use both sample variances to calculate the t -statistic. (LOS 6.h)
- A** Since the observations are likely dependent (both related to market returns), a paired comparisons (mean differences) test is appropriate and is based on a t -statistic. (LOS 6.h, LOS 6.i)

Module Quiz 6.4

- B** The F -test is the appropriate test. (LOS 6.j)
- C** A test of the population variance is a chi-square test. (LOS 6.j)
- A** Degrees of freedom are $n - 2$ for a test of the hypothesis that correlation is equal to zero. The test statistic increases with sample size (degrees of freedom increase) and follows a t -distribution. (LOS 6.l)
- A** The test statistic for the Spearman rank correlation test follows a t -distribution. (LOS 6.l)
- B** A contingency table is used to determine whether two characteristics of a group are independent. (LOS 6.m)

READING 7

INTRODUCTION TO LINEAR REGRESSION

EXAM FOCUS

This introduction covers simple linear regression, which involves two variables: an independent and a dependent variable. Candidates should be able to construct a simple regression model and state the assumptions under which a linear model is valid. Given the estimated model parameters (coefficients), you should be able to use the model to predict the dependent variable. Finally, you may be required to interpret an ANOVA table and test the significance of estimated regression coefficients. Note that an F -test, in the context of a simple regression, is equivalent to a t -test of the significance of the estimated slope coefficient.

MODULE 7.1: LINEAR REGRESSION: INTRODUCTION



Video covering this content is available online.

LOS 7.a: Describe a simple linear regression model and the roles of the dependent and independent variables in the model.

The purpose of **simple linear regression** is to explain the variation in a dependent variable in terms of the variation in a single independent variable. Here, the term “variation” is interpreted as the degree to which a variable differs from its mean value. Don’t confuse variation with variance—they are related but are not the same.

$$\text{variation in } Y = \text{variation in } Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- The **dependent variable** is the variable whose variation is explained by the independent variable. We are interested in answering the question, “What explains fluctuations in the dependent variable?” The dependent variable is also referred to as the *explained variable*, the *endogenous variable*, or the *predicted variable*.
- The **independent variable** is the variable used to explain the variation of the dependent variable. The independent variable is also referred to as the *explanatory variable*, the *exogenous variable*, or the *predicting variable*.

EXAMPLE: Dependent vs. independent variables

Suppose that we want to predict stock returns based on GDP growth. Which variable is the independent variable?

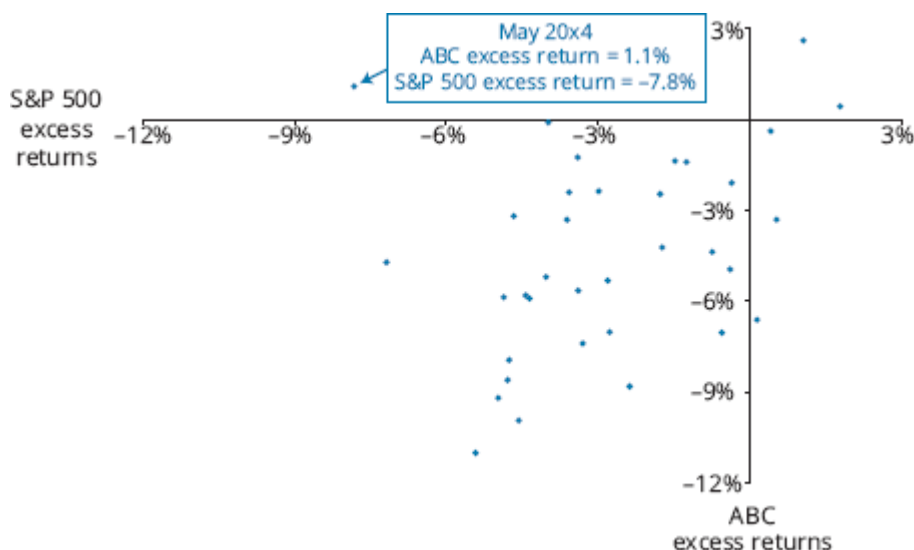
Answer:

Because GDP is going to be used as a *predictor* of stock returns, stock returns are being *explained* by GDP. Hence, stock returns are the dependent (explained) variable, and GDP is the independent (explanatory) variable.

Suppose we want to use excess returns on the S&P 500 (the independent variable) to explain the variation in excess returns on ABC common stock (the dependent variable). For this model, we define excess return as the difference between the actual return and the return on 1-month Treasury bills.

We would start by creating a scatter plot with ABC excess returns on the vertical axis and S&P 500 excess returns on the horizontal axis. Monthly excess returns for both variables from June 20x2 to May 20x5 are plotted in Figure 7.1. For example, look at the point labeled May 20x4. In that month, the excess return on the S&P 500 was -7.8% and the excess return on ABC was 1.1% .

Figure 7.1: Scatter Plot of ABC Excess Returns vs. S&P 500 Index Excess Returns



The two variables in Figure 7.1 appear to be positively correlated: excess ABC returns tended to be positive (negative) in the same month that S&P 500 excess returns were positive (negative). This is not the case for all the observations, however (for example, May 20x4). In fact, the correlation between these variables is approximately 0.40.

LOS 7.b: Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation.

Simple Linear Regression Model

The following linear regression model is used to describe the relationship between two variables, X and Y :

$$Y_i = b_0 + b_1 X_i + \epsilon_i, i = 1, \dots, n$$

where:

Y_i = i th observation of the dependent variable, Y

X_i = i th observation of the independent variable, X

b_0 = regression intercept term

b_1 = regression slope coefficient

ϵ_i = **residual** for the i th observation (also referred to as the disturbance term or error term)

Based on this regression model, the regression process estimates an equation for a line through a scatter plot of the data that “best” explains the observed values for Y in terms of the observed values for X .

The linear equation, often called the line of best fit or **regression line**, takes the following form:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i, i = 1, 2, 3 \dots, n$$

where:

\hat{Y}_i = estimated value of Y_i given X_i

\hat{b}_0 = estimated intercept term

\hat{b}_1 = estimated slope coefficient



PROFESSOR'S NOTE

The hat “^” above a variable or parameter indicates a predicted value.

The regression line is just one of the many possible lines that can be drawn through the scatter plot of X and Y . The criteria used to estimate this line is the essence of linear regression. The regression line is the line that minimizes the sum of the squared differences (vertical distances) between the Y -values predicted by the regression equation ($\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$) and the *actual* Y -values, Y_i . The sum of the squared vertical distances between the estimated and actual Y -values is referred to as the **sum of squared errors (SSE)**.

Thus, the regression line is the line that minimizes the SSE. This explains why simple linear regression is frequently referred to as **ordinary least squares (OLS)** regression, and the values determined by the estimated regression equation, \hat{Y}_i , are called least squares estimates.

The estimated **slope coefficient** (\hat{b}_1) for the regression line describes the change in Y for a one-unit change in X . It can be positive, negative, or zero, depending on the relationship between the regression variables. The slope term is calculated as:

$$\hat{b}_1 = \frac{\text{Cov}_{XY}}{\sigma_X^2}$$

The intercept term (\hat{b}_0) is the line's intersection with the Y -axis at $X = 0$. It can be positive, negative, or zero. A property of the least squares method is that the intercept term may be expressed as:

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

where:

\bar{Y} = mean of Y

\bar{X} = mean of X

The intercept equation highlights the fact that the regression line passes through a point with coordinates equal to the mean of the independent and dependent variables (i.e., the point \bar{X}, \bar{Y}).

EXAMPLE: Computing the slope coefficient and intercept term

Compute the slope coefficient and intercept term for the ABC regression example using the following information:

$$\begin{array}{ll} \text{Cov}(\text{S\&P 500}, \text{ABC}) = 0.000336 & \text{Mean return, S\&P 500} = -2.70\% \\ \text{Var}(\text{S\&P 500}) = 0.000522 & \text{Mean return, ABC} = -4.05\% \end{array}$$

Answer:

The slope coefficient is calculated as $\hat{\beta}_1 = 0.000336 / 0.000522 = 0.64$.

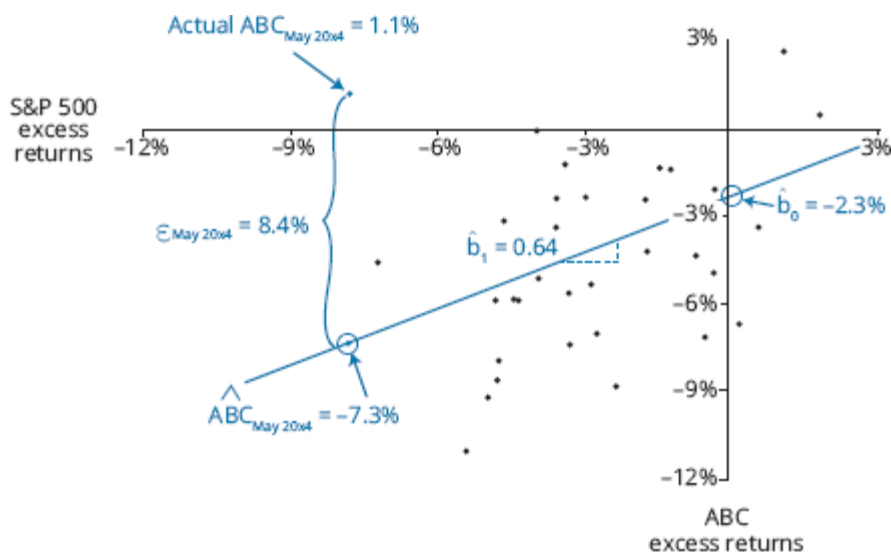
The intercept term is:

$$\hat{\beta}_0 = \overline{\text{ABC}} - \hat{\beta}_1 \overline{\text{S\&P 500}} = -4.05\% - 0.64(-2.70\%) = -2.3\%$$

The estimated regression line that minimizes the SSE in our ABC stock return example is shown in Figure 7.2.

This regression line has an intercept of -2.3% and a slope of 0.64 . The model predicts that if the S&P 500 excess return is -7.8% (May 20x4 value), then the ABC excess return would be $-2.3\% + (0.64)(-7.8\%) = -7.3\%$. The residual (error) for the May 20x4 ABC prediction is 8.4% , the difference between the actual ABC excess return of 1.1% and the predicted return of -7.3% .

Figure 7.2: Estimated Regression Equation for ABC vs. S&P 500 Excess Returns



Interpreting a Regression Coefficient

The estimated intercept represents the value of the dependent variable at the point of intersection of the regression line and the axis of the dependent variable (usually the vertical axis). In other words, the intercept is an estimate of the dependent variable when the independent variable is zero.

We also mentioned earlier that the estimated slope coefficient is interpreted as the expected change in the dependent variable for a one-unit change in the independent variable. For example, an estimated slope coefficient of 2 would indicate that the dependent variable is expected to change by two units for every one-unit change in the independent variable.

EXAMPLE: Interpreting regression coefficients

In the ABC regression example, the estimated slope coefficient was 0.64 and the estimated intercept term was -2.3% . Interpret each coefficient estimate.

Answer:

The slope coefficient of 0.64 can be interpreted to mean that when excess S&P 500 returns increase (decrease) by 1%, ABC excess return is expected to increase (decrease) by 0.64%.

The intercept term of -2.3% can be interpreted to mean that when the excess return on the S&P 500 is zero, the expected return on ABC stock is -2.3% .



PROFESSOR'S NOTE

The slope coefficient in a linear regression of the excess return of an individual security (the Y -variable) on the excess return on the market (the X -variable) is called the stock's beta, which is an estimate of systematic risk of ABC's stock. Notice that ABC is less risky than the average because its returns tend to increase or decrease by less than the change in the market returns. A stock with a beta (regression line slope) of one would have an average level of systematic risk and a stock with a beta greater than one would have more-than-average systematic risk. We will apply this concept in the Portfolio Management topic area.

Keep in mind, however, that any conclusions regarding the importance of an independent variable in explaining a dependent variable are based on the statistical significance of the slope coefficient. The magnitude of the slope coefficient tells us nothing about the strength of the linear relationship between the dependent and independent variables. A hypothesis test must be conducted, or a confidence interval must be formed, to assess the explanatory power of the independent variable. Later in this reading we will perform these hypothesis tests.

LOS 7.c: Explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated.

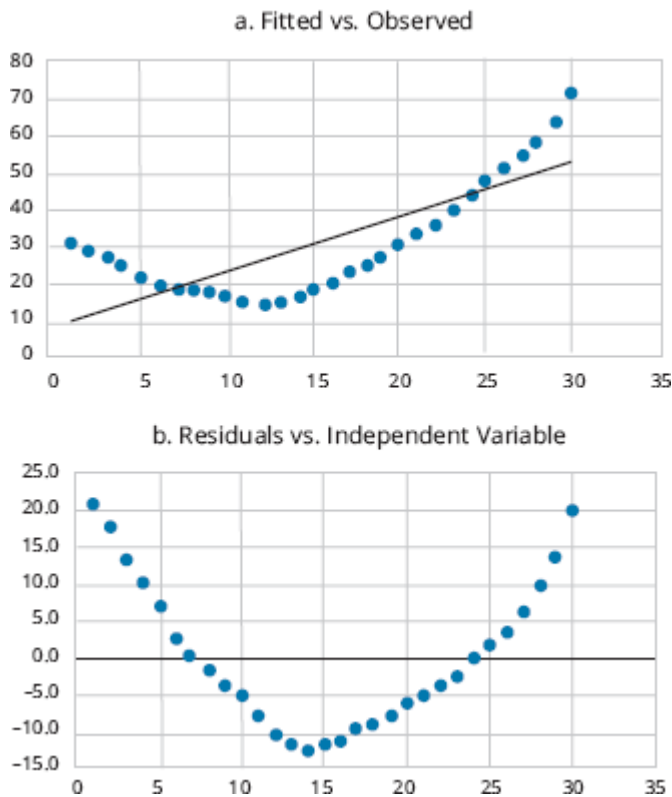
Linear regression is based on a number of assumptions. Most of the assumptions pertain to the regression model's residual term (ϵ). Linear regression assumes the following:

1. A linear relationship exists between the dependent and the independent variables.
2. The variance of the residual term is constant for all observations (homoskedasticity).
3. The residual term is independently distributed; that is, the residual for one observation is not correlated with that of another observation.
4. The residual term is normally distributed.

Linear Relationship

A linear regression model is not appropriate when the underlying relationship between X and Y is nonlinear. In Panel a of Figure 7.3, we illustrate a regression line fitted to a nonlinear relationship. Note that the prediction errors (vertical distances from the dots to the line) are positive for low values of X, then increasingly negative for higher values of X, and then turning positive for still greater values of X. One way of checking for linearity is to examine the model residuals (prediction errors) in relation to the independent regression variable. In Panel b, we show the pattern of residuals over the range of the independent variable: positive, negative, then positive.

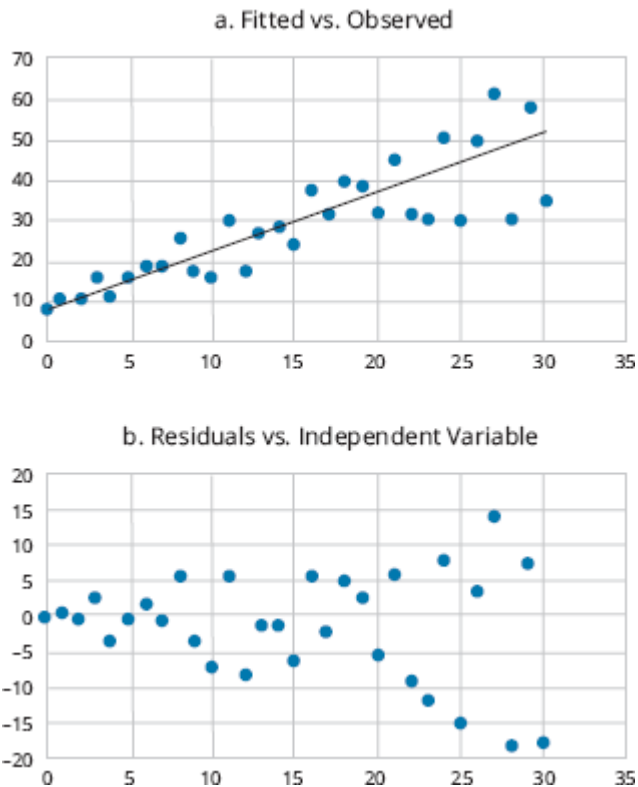
Figure 7.3: Nonlinear Relationship



Homoskedasticity

Homoskedasticity refers to the case where prediction errors all have the same variance. **Heteroskedasticity** refers to the situation when the assumption of homoskedasticity is violated. Figure 7.4 Panel a shows a scatter plot of observations around a fitted regression line where the residuals (prediction errors) increase in magnitude with larger values of the independent variable X. Panel b shows the residuals plotted versus the value of the independent variable, and also illustrates that the variance of the error terms is not likely constant for all observations.

Figure 7.4: Heteroskedasticity

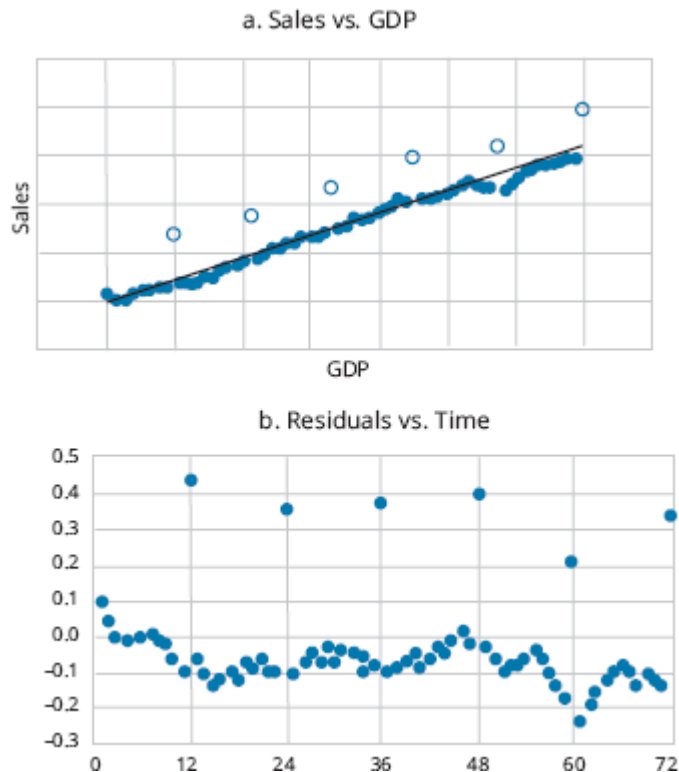


Another type of heteroskedasticity results if the variance of the error term changes over time (rather than with the magnitude of the independent variable). We could observe this by plotting the residuals from a linear regression model versus the dates of each observation and finding that the magnitude of the errors exhibits a pattern of changing over time. To illustrate this, we could plot the residuals versus a time index (as the X variable). Residuals would exhibit a pattern of increasing over time.

Independence

Suppose we collect a company's monthly sales and plot them against monthly GDP as in Figure 7.5 Panel a and observe that some prediction errors (the unfilled dots) are noticeably larger than others. To investigate this, we plot the residuals versus time, as in Panel b. The residuals plot illustrates that there are large negative prediction errors every 12 months (in December). This suggests that there is seasonality in sales such that December sales (the unfilled dots in Figure 7.5) are noticeably further from their predicted values than sales for the other months. If the relationship between X and Y is not independent, the residuals are not independent, and our estimates of variance, as well as our estimates of the model parameters, will not be correct.

Figure 7.5: Independence



Normality

When the residuals (prediction errors) are normally distributed, we can conduct hypothesis testing for evaluating the goodness of fit of the model (discussed later). With a large sample size, based on the central limit theorem, our parameter estimates may be valid, even when the residuals are not normally distributed.

Outliers are observations (one or a few) that are far from our regression line (have large prediction errors or X values that are far from the others). Outliers will influence our parameter estimates so that the OLS model will not fit the other observations well.



MODULE QUIZ 7.1

- Which of the following is *least likely* a necessary assumption of simple linear regression analysis?
 - The residuals are normally distributed.
 - There is a constant variance of the error term.
 - The dependent variable is uncorrelated with the residuals.
- What is the *most appropriate* interpretation of a slope coefficient estimate of 10.0?
 - The predicted value of the dependent variable when the independent variable is zero is 10.0.
 - For every one unit change in the independent variable, the model predicts that the dependent variable will change by 10 units.
 - For every 1-unit change in the independent variable, the model predicts that the dependent variable will change by 0.1 units.

MODULE 7.2: GOODNESS OF FIT AND HYPOTHESIS TESTS



Video covering this content is available online.

LOS 7.d: Calculate and interpret the coefficient of determination and the F -statistic in a simple linear regression.

LOS 7.e: 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.

Analysis of variance (ANOVA) is a statistical procedure for analyzing the total variability of the dependent variable. Let's define some terms before we move on to ANOVA tables:

- **Total sum of squares (SST)** measures the total variation in the dependent variable. SST is equal to the sum of the squared differences between the actual Y -values and the mean of Y .

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- **Sum of squares regression (SSR)** measures the variation in the dependent variable that is explained by the independent variable. SSR is the sum of the squared distances between the predicted Y -values and the mean of Y .

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

- **Sum of squared errors (SSE)** measures the unexplained variation in the dependent variable. It's also known as the sum of squared residuals or the residual sum of squares. SSE is the sum of the squared vertical distances between the actual Y -values and the predicted Y -values on the regression line.

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

You probably will not be surprised to learn that:

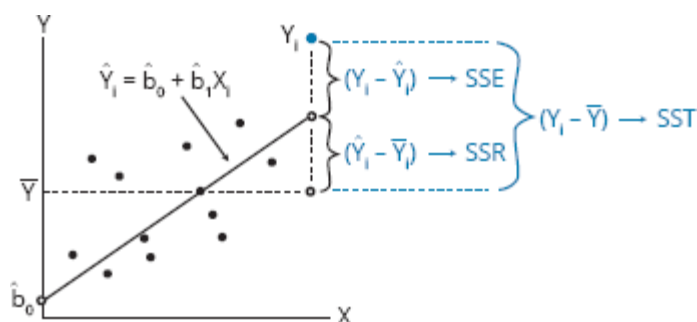
total variation = explained variation + unexplained variation

or:

$$SST = SSR + SSE$$

Figure 7.6 illustrates how the total variation in the dependent variable (SST) is composed of SSR and SSE.

Figure 7.6: Components of Total Variation



The output of the ANOVA procedure is an ANOVA table, which is a summary of the variation in the dependent variable. ANOVA tables are included in the regression output of many statistical

software packages. You can think of the ANOVA table as the source of the data for the computation of many of the regression concepts discussed in this reading. A generic ANOVA table for a simple linear regression (one independent variable) is presented in Figure 7.7.

Figure 7.7: ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	SSR	$MSR = \frac{SSR}{k} = \frac{SSR}{1} = SSR$
Error (unexplained)	$n - 2$	SSE	$MSE = \frac{SSE}{n - 2}$
Total	$n - 1$	SST	



PROFESSOR'S NOTE

In Figure 7.7, k is the number of slope parameters estimated and n is the number of observations. In general (including regressions with more than one independent variable), the regression $df = k$ and the error $df = (n - k - 1)$. Because we are limited to simple linear regressions in this reading (one independent variable), we use $k = 1$ for the regression degrees of freedom and $n - 1 - 1 = n - 2$ for the error degrees of freedom.

Standard Error of Estimate (SEE)

SEE for a regression is the standard deviation of its residuals. The lower the SEE, the better the model fit.

$$SEE = \sqrt{MSE}$$

Coefficient of Determination (R^2)

The **coefficient of determination** (R^2) is defined as the percentage of the total variation in the dependent variable explained by the independent variable. For example, an R^2 of 0.63 indicates that the variation of the independent variable explains 63% of the variation in the dependent variable.

$$R^2 = SSR / SST$$



PROFESSOR'S NOTE

For simple linear regression (i.e., with one independent variable), the coefficient of determination, R^2 , may be computed by simply squaring the correlation coefficient, r . In other words, $R^2 = r^2$ for a regression with one independent variable.

EXAMPLE: Using the ANOVA table

Complete the ANOVA table for the ABC regression example and calculate the R^2 and the standard error of estimate (SEE).

Partial ANOVA Table for ABC Regression Example

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	?	0.00756	?
Error (unexplained)	?	0.04064	?
Total	?	?	

Answer:

Recall that the data included three years of monthly return observations, so the total number of observations (n) is 36.

Completed ANOVA Table for ABC Regression Example

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	0.0076	0.0076
Error (unexplained)	34	0.0406	0.0012
Total	35	0.0482	

$$R^2 = \frac{\text{explained variation (SSR)}}{\text{total variation (SST)}} = \frac{0.0076}{0.0482} = 0.158 \text{ or } 15.8\%$$

$$SEE = \sqrt{MSE} = \sqrt{0.0012} = 0.035$$

The F-Statistic

An F -test assesses how well a set of independent variables, as a group, explains the variation in the dependent variable.

The F -statistic is calculated as:

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

where:

MSR = mean regression sum of squares

MSE = mean squared error

Important: This is always a one-tailed test!

For simple linear regression, there is only one independent variable, so the F -test is equivalent to a t -test for statistical significance of the slope coefficient:

$$H_0: b_1 = 0 \text{ versus } H_a: b_1 \neq 0$$

To determine whether b_1 is statistically significant using the F -test, the calculated F -statistic is compared with the critical F -value, F_c , at the appropriate level of significance. The degrees of freedom for the numerator and denominator with one independent variable are:

$$df_{\text{numerator}} = k = 1$$

$$df_{\text{denominator}} = n - k - 1 = n - 2$$

where:

n = number of observations

The decision rule for the F -test is: reject H_0 if $F > F_c$.

Rejecting the null hypothesis that the value of the slope coefficient equals zero at a stated level of significance indicates that the independent variable and the dependent variable have a significant linear relationship.

EXAMPLE: Calculating and interpreting the F -statistic

Use the completed ANOVA table from the previous example to calculate and interpret the F -statistic. Test the null hypothesis at the 5% significance level that the slope coefficient is equal to 0.

Answer:

$$F = \frac{MSR}{MSE} = \frac{0.0076}{0.0012} = 6.33$$

$$df_{\text{numerator}} = k = 1$$

$$df_{\text{denominator}} = n - k - 1 = 36 - 1 - 1 = 34$$

The null and alternative hypotheses are: $H_0: b_1 = 0$ versus $H_a: b_1 \neq 0$. The critical F -value for 1 and 34 degrees of freedom at a 5% significance level is approximately 4.1. (Remember, it's a one-tail test, so we use the 5% F -table!) Therefore, we can reject the null hypothesis and conclude that the slope coefficient is significantly different than zero.

LOS 7.f: Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance.

A t -test may also be used to test the hypothesis that the true slope coefficient, b_1 , is equal to a hypothesized value. Letting \hat{b}_1 be the point estimate for b_1 , the appropriate test statistic with $n - 2$ degrees of freedom is:

$$t_{b_1} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

The decision rule for tests of significance for regression coefficients is:

$$\text{Reject } H_0 \text{ if } t > +t_{\text{critical}} \text{ or } t < -t_{\text{critical}}$$

Rejection of the null supports the alternative hypothesis that the slope coefficient is *different* from the hypothesized value of b_1 . To test whether an independent variable explains the variation in the dependent variable (i.e., it is statistically significant), the null hypothesis is that the true slope is zero ($b_1 = 0$). The appropriate test structure for the null and alternative hypotheses is:

$H_0: b_1 = 0$ versus $H_a: b_1 \neq 0$

EXAMPLE: Hypothesis test for significance of regression coefficients

The estimated slope coefficient from the ABC example is 0.64 with a standard error equal to 0.26. Assuming that the sample has 36 observations, determine if the estimated slope coefficient is significantly different than zero at a 5% level of significance.

Answer:

The calculated test statistic is:

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.64 - 0}{0.26} = 2.46$$

The critical two-tailed t -values are ± 2.03 (from the t -table with $df = 36 - 2 = 34$). Because $t > t_{\text{critical}}$ (i.e., $2.46 > 2.03$), we reject the null hypothesis and conclude that the slope is different from zero.

Note that the t -test for a simple linear regression is equivalent to a t -test for the correlation coefficient between x and y :

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$



MODULE QUIZ 7.2

Use the following data to answer Questions 1 and 2.

An analyst is interested in predicting annual sales for XYZ Company, a maker of paper products. The following table reports a regression of the annual sales for XYZ against paper product industry sales.

Regression Output

Parameters	Coefficient	Standard Error of the Coefficient
Intercept	-94.88	32.97
Slope (industry sales)	0.2796	0.0363

The correlation between company and industry sales is 0.9757. The regression was based on five observations.

- Which of the following is *closest* to the value and reports the *most likely* interpretation of the R^2 for this regression?
 - The R^2 is 0.048, indicating that the variability of industry sales explains about 4.8% of the variability of company sales.
 - The R^2 is 0.952, indicating that the variability of industry sales explains about 95.2% of the variability of company sales.
 - The R^2 is 0.952, indicating that the variability of company sales explains about 95.2% of the variability of industry sales.

2. Based on the regression results, XYZ Company's market share of any increase in industry sales is expected to be *closest* to:
- 4%.
 - 28%.
 - 45%.

Use the following information to answer Questions 3 and 4.

A study was conducted by the British Department of Transportation to estimate urban travel time between locations in London, England. Data was collected for motorcycles and passenger cars. Simple linear regression was conducted using data sets for both types of vehicles, where Y = urban travel time in minutes and X = distance between locations in kilometers. The following results were obtained:

Regression Results for Travel Times Between Distances in London		
Passenger cars:	$\hat{Y} = 1.85 + 3.86X$	$R^2 = 0.758$
Motorcycles:	$\hat{Y} = 2.50 + 1.93X$	$R^2 = 0.676$

3. The estimated increase in travel time for a motorcycle commuter planning to move 8 km farther from his workplace in London is *closest* to:
- 31 minutes.
 - 15 minutes.
 - 0.154 hours.
4. Based on the regression results, which model is more reliable?
- The passenger car model because $3.86 > 1.93$.
 - The motorcycle model because $1.93 < 3.86$.
 - The passenger car model because $0.758 > 0.676$.
5. Consider the following statement: In a simple linear regression, the appropriate degrees of freedom for the critical t -value used to calculate a confidence interval around both a parameter estimate and a predicted Y -value is the same as the number of observations minus two. The statement is:
- justified.
 - not justified, because the appropriate of degrees of freedom used to calculate a confidence interval around a parameter estimate is the number of observations.
 - not justified, because the appropriate of degrees of freedom used to calculate a confidence interval around a predicted Y -value is the number of observations.
6. What is the appropriate alternative hypothesis to test the statistical significance of the intercept term in the following regression?
- $$Y = a_1 + a_2(X) + \varepsilon$$
- $H_A: a_1 \neq 0$.
 - $H_A: a_1 > 0$.
 - $H_A: a_2 \neq 0$.

MODULE 7.3: PREDICTING DEPENDENT VARIABLES AND FUNCTIONAL FORMS



Video covering this content is available online.

LOS 7.g: 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.

Predicted values are values of the dependent variable based on the estimated regression coefficients and a prediction about the value of the independent variable. They are the values that are *predicted* by the regression equation, given an estimate of the independent variable.

For a simple regression, the predicted (or forecast) value of Y is:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_p$$

where:

\hat{Y} = predicted value of the dependent variable

X_p = forecasted value of the independent variable

EXAMPLE: Predicting the dependent variable

Given the ABC regression equation:

$$\widehat{ABC} = -2.3\% + (0.64)(\widehat{S\&P\ 500})$$

Calculate the predicted value of ABC excess returns if forecasted S&P 500 excess returns are 10%.

Answer:

The predicted value for ABC excess returns is determined as follows:

$$\widehat{ABC} = -2.3\% + (0.64)(10\%) = 4.1\%$$

Confidence Intervals for Predicted Values

The equation for the confidence interval for a predicted value of Y is:

$$\hat{Y} \pm (t_c \times s_f) \Rightarrow [\hat{Y} - (t_c \times s_f) < Y < \hat{Y} + (t_c \times s_f)]$$

where:

t_c = two-tailed critical t -value at the desired level of significance with $df = n - 2$

s_f = standard error of the forecast

The challenge with computing a confidence interval for a predicted value is calculating s_f . On the Level I exam it's highly unlikely that you will have to calculate the standard error of the forecast (it will probably be provided if you need to compute a confidence interval for the dependent variable). However, if you do need to calculate s_f it can be done with the following formula for the variance of the forecast:

$$s_f^2 = SEE^2 \left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_x^2} \right]$$

where:

SEE^2 = variance of the residuals = the square of the standard error of estimate

s_x^2 = variance of the independent variable

\bar{X} = value of the independent variable for which the forecast was made

EXAMPLE: Confidence interval for a predicted value

Calculate a 95% prediction interval on the predicted value of ABC excess returns from the previous example. Suppose the standard error of the forecast is 3.67, and the forecast value of S&P 500 excess returns is 10%.

Answer:

The predicted value for ABC excess returns is:

$$\widehat{ABC} = -2.3\% + (0.64)(10\%) = 4.1\%$$

The 5% two-tailed critical t -value with 34 degrees of freedom is 2.03. The prediction interval at the 95% confidence level is:

$$\widehat{ABC} \pm (t_c \times s_f) \Rightarrow [4.1\% \pm (2.03 \times 3.67\%)] = 4.1\% \pm 7.5\%$$

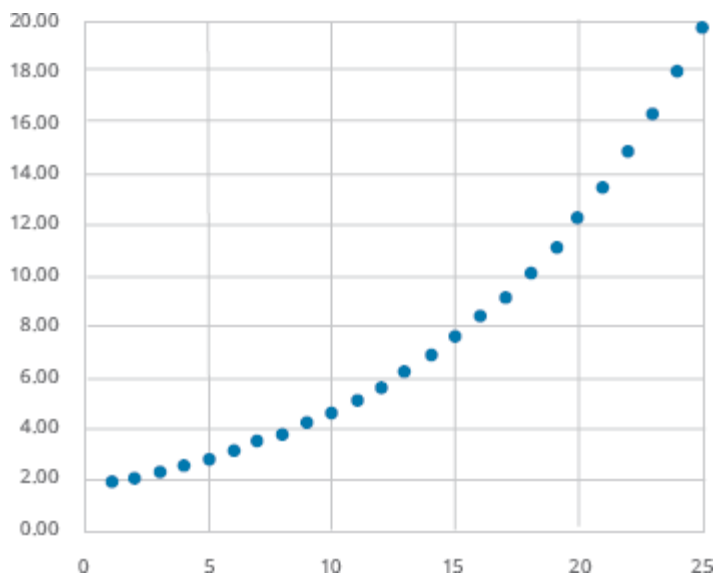
or -3.4% to 11.6%

We can interpret this to mean that, given a forecast value for S&P 500 excess returns of 10%, we can be 95% confident that the ABC excess returns will be between -3.4% and 11.6%.

LOS 7.h: Describe different functional forms of simple linear regressions.

One of the assumptions of linear regression is that the relationship between X and Y is linear. What if that assumption is violated? Consider $Y = \text{EPS}$ for a company and $X = \text{time index}$. Suppose that EPS is growing at approximately 10% annually. Figure 7.8 shows the plot of actual EPS versus time.

Figure 7.8: Nonlinear Relationship



In such a situation, transforming one of both of the variables can produce a linear relationship. The appropriate transformation depends on the relationship between the two variables. One

often-used transformation is to take the natural log of one or both of the variables. Some examples are:

- **Log-lin model.** If the dependent variable is logarithmic while the independent variable is linear.
- **Lin-log model.** If the dependent variable is linear while the independent variable is logarithmic.
- **Log-log model.** Both the dependent variable and the independent variable are logarithmic.

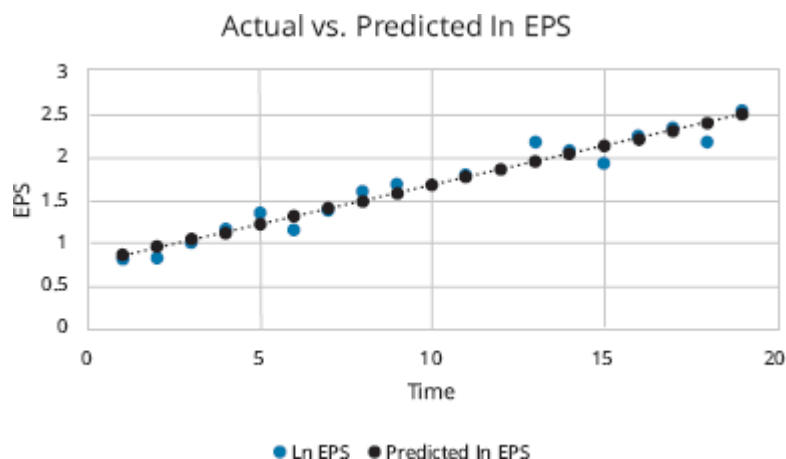
Log-Lin Model

Taking the natural logarithm of the dependent variable, our model now becomes:

$$\ln Y_i = b_0 + b_1 X_i + \varepsilon_i$$

In this model, the slope coefficient is interpreted as the *relative* change in the dependent variable for an absolute change in the independent variable. Figure 7.9 shows the results after taking the natural log of EPS, and fitting that data using a log-lin model.

Figure 7.9: Log-Lin Model, EPS Data



Lin-Log Model

Taking the natural logarithm of the independent variable, our model now becomes:

$$Y_i = b_0 + b_1 \ln(X)_i + \varepsilon_i$$

In this model, the slope coefficient is interpreted as the *absolute* change in the dependent variable for a *relative* change in the independent variable.

Log-Log Model

Taking the natural logarithm of both variables, our model now becomes:

$$\ln Y_i = b_0 + b_1 \ln(X)_i + \varepsilon_i$$

In this model, the slope coefficient is interpreted as the relative change in the dependent variable for a relative change in the independent variable.

Selection of Functional Form

Selecting the correct functional form involves determining the nature of the variables and evaluation of the goodness of fit measures (e.g., R^2 , SEE, F -stat).



MODULE QUIZ 7.3

- The variation in the dependent variable explained by the independent variable is measured by:
 - the mean squared error.
 - the sum of squared errors.
 - the regression sum of squares.
- Results from a regression analysis are presented in the following figures.

Estimated Coefficients

Coefficient	Coefficient Estimate	Standard Error
b_0	0.0023	0.0022
b_1	1.1163	0.0624

Partial ANOVA Table

Source of Variation	Sum of Squares
Regression (explained)	0.0228
Error (unexplained)	0.0024

Are the intercept term and the slope coefficient statistically significantly different from zero at the 5% significance level?

<u>Intercept term significant?</u>	<u>Slope coefficient significant?</u>
A. Yes	Yes
B. Yes	No
C. No	Yes

- Partial ANOVA Table

Source of Variation	Sum of Squares
Regression (explained)	0.0228
Error (unexplained)	0.0024

To test the following hypothesis: $H_0: b_1 \leq 1$ versus $H_1: b_1 > 1$, at the 1% significance level, the calculated t -statistic and the appropriate conclusion are:

<u>Calculated t-statistic</u>	<u>Appropriate conclusion</u>
A. 1.86	Reject H_0
B. 1.86	Fail to reject H_0
C. 2.44	Reject H_0

- The appropriate regression model for a linear relationship between the relative change in an independent variable and the absolute change in the dependent variable is a:
 - log-lin model.
 - lin-log model.
 - lin-lin model.

5. For a regression model of $Y = 5 + 3.5X$, the analysis (based on a large data sample) provides the standard error of the forecast as 2.5 and the standard error of the slope coefficient as 0.8. A 90% confidence interval for the estimate of Y when the value of the independent variable is 10 is *closest to*:
- A. 35.1 to 44.9.
 - B. 35.6 to 44.4.
 - C. 35.9 to 44.1.

KEY CONCEPTS

LOS 7.a

Linear regression provides an estimate of the linear relationship between an independent variable (the explanatory variable) and a dependent variable (the predicted variable).

LOS 7.b

The general form of a simple linear regression model is $Y_i = b_0 + b_1X_i + \varepsilon_i$.

The least-squares model minimizes the sum of squared errors.

- $\hat{b}_0 = Y - \hat{b}_1X$ = fitted intercept
- \hat{b}_1 fitted slope coefficient = cov (X,Y) / variance of X

The estimated intercept, \hat{b}_0 , represents the value of the dependent variable at the point of intersection of the regression line and the axis of the dependent variable (usually the vertical axis). The estimated slope coefficient, \hat{b}_1 , is interpreted as the change in the dependent variable for a one-unit change in the independent variable.

LOS 7.c

Assumptions made regarding simple linear regression include the following:

1. A linear relationship exists between the dependent and the independent variable.
2. The variance of the residual term is constant (homoskedasticity).
3. The residual term is independently distributed (residuals are uncorrelated).
4. The residual term is normally distributed.

LOS 7.d, e

ANOVA Table for Simple Linear Regression ($k = 1$)

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	SSR	$MSR = \frac{SSR}{k} = \frac{SSR}{1} = SSR$
Error (unexplained)	$n - 2$	SSE	$MSE = \frac{SSE}{n - 2}$
Total	$n - 1$	SST	

The standard error of the estimate in a simple linear regression is calculated as:

$$SEE = \sqrt{\frac{SSE}{n - 2}}$$

The standard error of the estimate in a simple linear regression is calculated as:

$$SEE = \sqrt{\frac{SSE}{n-2}}$$

The coefficient of determination, R^2 , is the proportion of the total variation of the dependent variable explained by the regression:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

In simple linear regression, because there is only one independent variable ($k = 1$), the F -test tests the same null hypothesis as testing the statistical significance of b_1 using the t -test: $H_0: b_1 = 0$ versus $H_a: b_1 \neq 0$. With only one independent variable, F is calculated as:

$$F\text{-stat} = \frac{MSR}{MSE} \text{ with } 1 \text{ and } n - 2 \text{ degrees of freedom}$$

LOS 7.f

A t -test with $n - 2$ degrees of freedom is used to conduct hypothesis tests of the estimated regression parameters:

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

LOS 7.g

A predicted value of the dependent variable, \hat{Y} , is determined by inserting the predicted value of the independent variable, X_p , in the regression equation and calculating $\hat{Y}_p = \hat{b}_0 + \hat{b}_1 X_p$

The confidence interval for a predicted Y -value is $[\hat{Y} - (t_c \times s_f) < Y < \hat{Y} + (t_c \times s_f)]$, where s_f is the standard error of the forecast.

LOS 7.h

Dependent Variable	Independent Variable	Model	Slope Interpretation
Logarithmic	Linear	Log-lin	<i>Relative</i> change in dependent variable for an absolute change in the independent variable
Linear	Logarithmic	Lin-log	<i>Absolute</i> change in dependent variable for a relative change in the independent variable
Logarithmic	Logarithmic	Log-log	<i>Relative</i> change in dependent variable for a <i>relative</i> change in the independent variable

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 7.1

- C** The model does not assume that the dependent variable is uncorrelated with the residuals. It does assume that the independent variable is uncorrelated with the residuals. (LOS 7.c)
- B** The slope coefficient is best interpreted as the predicted change in the dependent variable for a 1-unit change in the independent variable. If the slope coefficient estimate is 10.0 and the independent variable changes by 1

unit, the dependent variable is expected to change by 10 units. The intercept term is best interpreted as the value of the dependent variable when the independent variable is equal to zero. (LOS 7.b)

Module Quiz 7.2

1. **B** The R^2 is computed as the correlation squared: $(0.9757)^2 = 0.952$.

The interpretation of this R^2 is that 95.2% of the variation in Company XYZ's sales is explained by the variation in industry sales. The independent variable (industry sales) explains the variation in the dependent variable (company sales). This interpretation is based on the economic reasoning used in constructing the regression model. (LOS 7.d)

2. **B** The slope coefficient of 0.2796 indicates that a \$1 million increase in industry sales will result in an increase in firm sales of approximately 28% of that amount (\$279,600). (LOS 7.b)

3. **B** The slope coefficient is 1.93, indicating that each additional kilometer increases travel time by 1.93 minutes:

$$1.93 \times 8 = 15.44$$

(LOS 7.b)

4. **C** The higher R^2 for the passenger car model indicates that regression results are more reliable. Distance is a better predictor of travel time for cars. Perhaps the aggressiveness of the driver is a bigger factor in travel time for motorcycles than it is for autos. (LOS 7.d)

5. **A** In simple linear regression, the appropriate degrees of freedom for both confidence intervals is the number of observations in the sample (n) minus two. (LOS 7.d)

6. **A** In this regression, a_1 is the intercept term. To test the statistical significance means to test the null hypothesis that a_1 is equal to zero, versus the alternative that a_1 is not equal to zero. (LOS 7.d)

Module Quiz 7.3

1. **C** The regression sum of squares measures the amount of variation in the dependent variable explained by the independent variable (i.e., the explained variation). The sum of squared errors measures the variation in the dependent variable not explained by the independent variable. The mean squared error is equal to the sum of squared errors divided by its degrees of freedom. (Module 7.2, LOS 7.e)

2. **C** The critical two-tailed 5% t -value with 34 degrees of freedom is approximately 2.03. The calculated t -statistics for the intercept term and slope coefficient are, respectively, $0.0023 / 0.0022 = 1.05$ and $1.1163 / 0.0624 = 17.9$. Therefore, the intercept term is not statistically different from zero at the 5% significance level, while the slope coefficient is. (LOS 7.g)

3. **B** Note that this is a one-tailed test. The critical one-tailed 1% t -value with 34 degrees of freedom is approximately 2.44. The calculated t -statistic for the slope coefficient is $(1.1163 - 1) / 0.0624 = 1.86$. Therefore, the slope coefficient is not statistically different from one at the 1% significance level and the analyst should fail to reject the null hypothesis. (LOS 7.g)

4. **B** The appropriate model would be a lin-log model, in which the values of the dependent variable (Y) are regressed on the natural logarithms of the independent variable (X), $Y = b_0 + b_1 \ln X$. (LOS 7.h)

5. **C** The estimate of Y , given $X = 10$ is: $Y = 5 + 3.5(10) = 40$. The critical value for a 90% confidence interval with a large sample size (z -statistic) is approximately 1.65. Given the standard error of the forecast of 2.5, the confidence interval for the estimated value of Y is $40 \pm 1.65(2.5) = 35.875$ to 44.125. (LOS 7.g)

TOPIC QUIZ: QUANTITATIVE METHODS

You have now finished the Quantitative Methods topic section. Please log into your Schweser online dashboard and take the Topic Quiz on Quantitative Methods. The Topic Quiz provides immediate feedback on how effective your study has been for this material. The number of questions on this quiz is approximately the number of questions for the topic on one-half of the actual Level I CFA exam. Questions are more exam-like than typical Module Quiz or QBank questions; a score of less than 70% indicates that your study likely needs improvement. These tests are best taken timed; allow 1.5 minutes per question.

After you've completed this Topic Quiz, select "Performance Tracker" to view a breakdown of your score. Select "Compare with Others" to display how your score on the Topic Quiz compares to the scores of others who entered their answers.

READING 8

TOPICS IN DEMAND AND SUPPLY ANALYSIS

EXAM FOCUS

The Level I Economics curriculum assumes candidates are familiar with concepts such as supply and demand, utility-maximizing consumers, and the product and cost curves of firms. CFA Institute has posted three assigned readings to its website as prerequisites for Level I Economics. If you have not studied economics before (or if it has been a while), you should review these readings, along with the video instruction, study notes, and review questions for each of them in your online Schweser Resource Library to get up to speed.

MODULE 8.1: ELASTICITY



LOS 8.a: Calculate and interpret price, income, and cross-price elasticities of demand and describe factors that affect each measure.

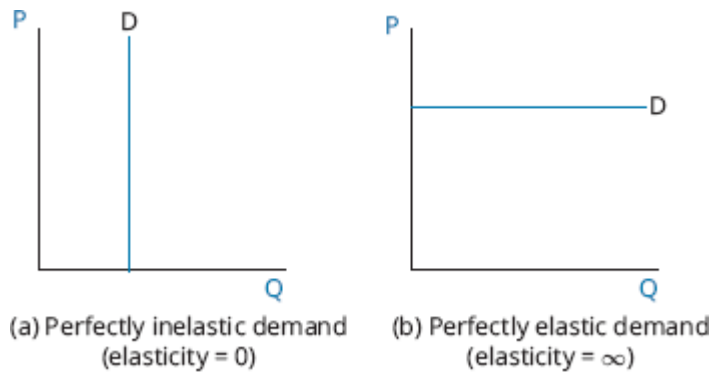
Video covering this content is available online.

Own-Price Elasticity of Demand

Own-price elasticity is a measure of the responsiveness of the quantity demanded to a change in price. It is calculated as the ratio of the percentage change in quantity demanded to a percentage change in price. With downward-sloping demand (i.e., an increase in price decreases quantity demanded), own-price elasticity is negative.

When the quantity demanded is very responsive to a change in price (absolute value of elasticity > 1), we say demand is elastic; when the quantity demanded is not very responsive to a change in price (absolute value of elasticity < 1), we say that demand is inelastic. In Figure 8.1, we illustrate the most extreme cases: perfectly elastic demand (at any higher price, quantity demanded decreases to zero) and perfectly inelastic demand (a change in price has no effect on quantity demanded).

Figure 8.1: Perfectly Inelastic and Perfectly Elastic Demand



When there are few or no good substitutes for a good, demand tends to be relatively inelastic. Consider a drug that keeps you alive by regulating your heart. If two pills per day keep you alive, you are unlikely to decrease your purchases if the price goes up and also quite unlikely to increase your purchases if price goes down.

When one or more goods are very good substitutes for the good in question, demand will tend to be very elastic. Consider two gas stations along your regular commute that offer gasoline of equal quality. A decrease in the posted price at one station may cause you to purchase all your gasoline there, while a price increase may lead you to purchase all your gasoline at the other station. Remember, we calculate demand and elasticity while holding the prices of related goods (in this case, the price of gas at the other station) constant.

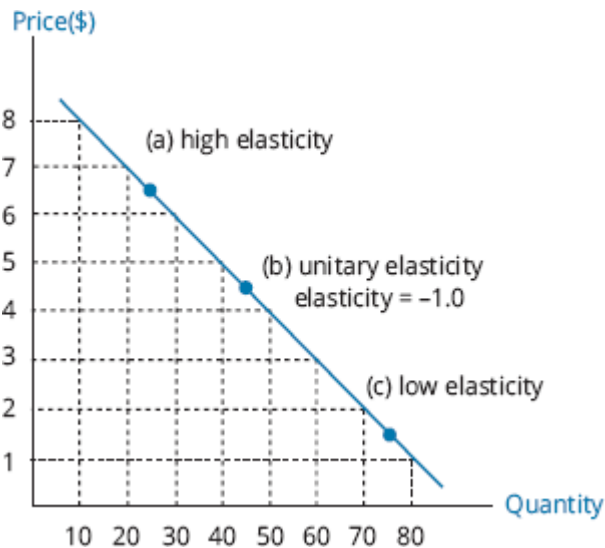
Other factors affect demand elasticity in addition to the quality and availability of substitutes:

- **Portion of income spent on a good.** The larger the proportion of income spent on a good, the more elastic an individual's demand for that good. If the price of a preferred brand of toothpaste increases, a consumer may not change brands or adjust the amount used if the customer prefers to simply pay the extra cost. When housing costs increase, however, a consumer will be much more likely to adjust consumption, because rent is a fairly large proportion of income.
- **Time.** Elasticity of demand tends to be greater the longer the time period since the price change. For example, when energy prices initially rise, some adjustments to consumption are likely made quickly. Consumers can lower the thermostat temperature. Over time, adjustments such as smaller living quarters, better insulation, more efficient windows, and installation of alternative heat sources are more easily made, and the effect of the price change on consumption of energy is greater.

It is important to understand that elasticity is not equal to the slope of a demand curve (except for the extreme examples of perfectly elastic or perfectly inelastic demand). Slope is dependent on the units that price and quantity are measured in. Elasticity is not dependent on units of measurement because it is based on percentage changes.

Figure 8.2 shows how elasticity changes along a linear demand curve. In the upper part of the demand curve, elasticity is greater (in absolute value) than 1; in other words, the percentage change in quantity demanded is greater than the percentage change in price. In the lower part of the curve, the percentage change in quantity demanded is smaller than the percentage change in price.

Figure 8.2: Price Elasticity Along a Linear Demand Curve



- At point (a), in a higher price range, the price elasticity of demand is greater than at point (c) in a lower price range.
- The elasticity at point (b) is -1.0 ; a 1% increase in price leads to a 1% decrease in quantity demanded. This is the point of greatest total revenue ($P \times Q$), which is $4.50 \times 45 = \$202.50$.
- At prices less than \$4.50 (inelastic range), total revenue will increase when price increases. The percentage decrease in quantity demanded will be less than the percentage increase in price.
- At prices above \$4.50 (elastic range), a price increase will decrease total revenue since the percentage decrease in quantity demanded will be greater than the percentage increase in price.

An important point to consider about the price and quantity combination for which price elasticity equals -1.0 (**unit** or **unitary elasticity**) is that total revenue (price \times quantity) is maximized at that price. An increase in price moves us to the elastic region of the curve so that the percentage decrease in quantity demanded is greater than the percentage increase in price, resulting in a decrease in total revenue. A decrease in price from the point of unitary elasticity moves us into the inelastic region of the curve so that the percentage decrease in price is more than the percentage increase in quantity demanded, resulting, again, in a decrease in total revenue.

Income Elasticity of Demand

Recall that one of the independent variables in our example of a demand function for gasoline was income. The sensitivity of quantity demanded to a change in income is termed **income elasticity**. Holding other independent variables constant, we can measure income elasticity as the ratio of the percentage change in quantity demanded to the percentage change in income.

For most goods, the sign of income elasticity is positive—an increase in income leads to an increase in quantity demanded. Goods for which this is the case are termed **normal goods**. For other goods, it may be the case that an increase in income leads to a decrease in quantity demanded. Goods for which this is true are termed **inferior goods**.

Cross-Price Elasticity of Demand

Recall that some of the independent variables in a demand function are the prices of related goods (related in the sense that their prices affect the demand for the good in question). The ratio of the percentage change in the quantity demanded of a good to the percentage change in the price of a related good is termed the **cross-price elasticity of demand**.

When an increase in the price of a related good increases demand for a good, the two goods are substitutes. If Bread A and Bread B are two brands of bread, considered good substitutes by many consumers, an increase in the price of one will lead consumers to purchase more of the other (substitute the other). When the cross-price elasticity of demand is positive (price of one is up and quantity demanded for the other is up), we say those goods are substitutes.

When an increase in the price of a related good decreases demand for a good, the two goods are **complements**. If an increase in the price of automobiles (less automobiles purchased) leads to a decrease in the demand for gasoline, they are complements. Right shoes and left shoes are perfect complements for most of us and, as a result, shoes are priced by the pair. If they were priced separately, there is little doubt that an increase in the price of left shoes would decrease the quantity demanded of right shoes. Overall, the cross-price elasticity of demand is more positive the better substitutes two goods are and more negative the better complements the two goods are.

Calculating Elasticities

The price elasticity of demand is defined as:

$$\frac{\% \Delta Q}{\% \Delta P} = \frac{\Delta Q / Q_0}{\Delta P / P_0} = \left(\frac{P_0}{Q_0} \right) \times \left(\frac{\Delta Q}{\Delta P} \right)$$

The term $\frac{\Delta Q}{\Delta P}$ is the slope of a demand *function* that (for a linear demand function) takes the form:

$$\text{quantity demanded} = A + B \times \text{price}$$

In such a function, B is the slope of the line. A demand *curve* is the inverse of the demand function, in which price is given as a function of quantity demanded.

As an example, consider a demand function with $A = 100$ and $B = -2$, so that $Q = 100 - 2P$. The slope, $\frac{\Delta Q}{\Delta P}$ of this line is -2 . The corresponding demand curve for this demand function is: $P = 100 / 2 - Q / 2 = 50 - 1/2 Q$. Therefore, given a demand curve, we can calculate the slope of the demand function as the reciprocal of slope term, $-1/2$, of the demand curve (i.e., the reciprocal of $-1/2$ is -2 , the slope of the demand function).

EXAMPLE: Calculating price elasticity of demand

A demand function for gasoline is as follows:

$$Q_{\text{Dgas}} = 138,500 - 12,500P_{\text{gas}}$$

Calculate the price elasticity at a gasoline price of \$3 per gallon.

Answer:

We can calculate the quantity demanded at a price of \$3 per gallon as $138,500 - 12,500(3) = 101,000$. Substituting 3 for P_0 , 101,000 for Q_0 , and $-12,500$ for $\left(\frac{\Delta Q}{\Delta P}\right)$, we can calculate the price elasticity of demand as:

$$E_{\text{Demand}} = \frac{\% \Delta Q}{\% \Delta P} = \left(\frac{3}{101,000} \right) \times (-12,500) = -0.37$$

For this demand function, at a price and quantity of \$3 per gallon and 101,000 gallons, demand is inelastic.

The techniques for calculating the income elasticity of demand and the cross-price elasticity of demand are the same, as illustrated in the following example. We assume values for all the independent variables, except the one of interest, then calculate elasticity for a given value of the variable of interest.

EXAMPLE: Calculating income elasticity and cross-price elasticity

An individual has the following demand function for gasoline:

$$Q_{D \text{ gas}} = 15 - 3P_{\text{gas}} + 0.02I + 0.11P_{\text{BT}} - 0.008P_{\text{auto}}$$

where income and car price are measured in thousands, and the price of bus travel is measured in average dollars per 100 miles traveled.

Assuming the average automobile price is \$22,000, income is \$40,000, the price of bus travel is \$25, and the price of gasoline is \$3, calculate and interpret the income elasticity of gasoline demand and the cross-price elasticity of gasoline demand with respect to the price of bus travel.

Answer:

Inserting the prices of gasoline, bus travel, and automobiles into our demand equation, we get:

$$Q_{D \text{ gas}} = 15 - 3(3) + 0.02(\text{income in thousands}) + 0.11(25) - 0.008(22)$$

and

$$Q_{D \text{ gas}} = 8.574 + 0.02(\text{income in thousands})$$

Our slope term on income is 0.02, and for an income of 40,000, $Q_{D \text{ gas}} = 9.374$ gallons.

The formula for the income elasticity of demand is:

$$\frac{\% \Delta Q}{\% \Delta I} = \frac{\Delta Q / Q_0}{\Delta I / I_0} = \left(\frac{I_0}{Q_0} \right) \times \left(\frac{\Delta Q}{\Delta I} \right)$$

Substituting our calculated values, we have:

$$\left(\frac{40}{9.374} \right) \times (0.02) = 0.085$$

This tells us that for these assumed values (at a single point on the demand curve), a 1% increase (decrease) in income will lead to an increase (decrease) of 0.085% in the quantity of gasoline demanded.

In order to calculate the cross-price elasticity of demand for bus travel and gasoline, we construct a demand function with only the price of bus travel as an independent variable:

$$Q_{D \text{ gas}} = 15 - 3P_{\text{gas}} + 0.02I + 0.11P_{\text{BT}} - 0.008P_{\text{auto}}$$

$$Q_{D \text{ gas}} = 15 - 3(3) + 0.02(40) + 0.11P_{\text{BT}} - 0.008(22)$$

$$Q_{D \text{ gas}} = 6.624 + 0.11P_{\text{BT}}$$

For a price of bus travel of \$25, the quantity of gasoline demanded is:

$$Q_{D \text{ gas}} = 6.624 + 0.11P_{\text{BT}}$$

$$Q_{D \text{ gas}} = 6.624 + 0.11(25) = 9.374 \text{ gallons}$$

The cross-price elasticity of the demand for gasoline with respect to the price of bus travel is:

$$\begin{aligned} \frac{\% \Delta Q}{\% \Delta P_{\text{BT}}} &= \frac{\Delta Q / Q_0}{\Delta P_{\text{BT}} / P_{0 \text{ BT}}} = \left(\frac{P_{0 \text{ BT}}}{Q_0} \right) \times \left(\frac{\Delta Q}{\Delta P_{\text{BT}}} \right) = \frac{25}{9.374} \times 0.11 \\ &= 0.293 \end{aligned}$$

As noted, gasoline and bus travel are substitutes, so the cross-price elasticity of demand is positive. We can interpret this value to mean that, for our assumed values, a 1% change in the price of bus travel will lead to a 0.293% change in the quantity of gasoline demanded in the same direction, other things equal.

MODULE 8.2: DEMAND AND SUPPLY



LOS 8.b: Compare substitution and income effects.

Video covering this content is available online.

When the price of Good X decreases, there is a **substitution effect** that shifts consumption towards more of Good X. Because the total expenditure on the consumer's original bundle of goods falls when the price of Good X falls, there is also an **income effect**. The income effect can be toward more or less consumption of Good X. This is the key point here: the substitution effect always acts to increase the consumption of a good that has fallen in price, while the income effect can either increase or decrease consumption of a good that has fallen in price.

Based on this analysis, we can describe three possible outcomes of a decrease in the price of Good X:

1. The substitution effect is positive, and the income effect is also positive—consumption of Good X will *increase*.
2. The substitution effect is positive, and the income effect is negative but smaller than the substitution effect—consumption of Good X will *increase*.
3. The substitution effect is positive, and the income effect is negative and larger than the substitution effect—consumption of Good X will *decrease*.

LOS 8.c: Contrast normal goods with inferior goods.



PROFESSOR'S NOTE

Candidates who are not already familiar with profit maximization based on a firm's cost curves (e.g., average cost and marginal cost) and firm revenue (e.g., average revenue, total revenue, and marginal revenue) should study the material in the CFA curriculum prerequisite reading "Demand and Supply Analysis: The Firm" prior to their study of the following material.

Earlier, we defined normal goods and inferior goods in terms of their income elasticity of demand. A normal good is one for which the income effect is positive. An inferior good is one for which the income effect is negative.

A specific good may be an inferior good for some ranges of income and a normal good for other ranges of income. For a really poor person or population (e.g., underdeveloped country), an increase in income may lead to greater consumption of noodles or rice. Now, if incomes rise a bit (e.g., college student or developing country), more meat or seafood may become part of the diet. Over this range of incomes, noodles can be an inferior good and ground meat a normal good. If incomes rise to a higher range (e.g., graduated from college and got a job), the consumption of ground meat may fall (inferior) in favor of preferred cuts of meat (normal).

For many of us, commercial airline travel is a normal good. When our incomes rise, vacations are more likely to involve airline travel, be more frequent, and extend over longer distances so that airline travel is a normal good. For wealthy people (e.g., hedge fund manager), an increase in income may lead to travel by private jet and a decrease in the quantity of commercial airline travel demanded.

A **Giffen good** is an inferior good for which the negative income effect outweighs the positive substitution effect when price falls. A Giffen good is theoretical and would have an upward-sloping demand curve. At lower prices, a smaller quantity would be demanded as a result of the dominance of the income effect over the substitution effect. Note that the existence of a Giffen good is not ruled out by the axioms of the theory of consumer choice.

A **Veblen good** is one for which a higher price makes the good more desirable. The idea is that the consumer gets utility from being seen to consume a good that has high status (e.g., Gucci bag), and that a higher price for the good conveys more status and increases its utility. Such a good could conceivably have a positively sloped demand curve for some individuals over some range of prices. If such a good exists, there must be a limit to this process, or the price would rise without limit. Note that the existence of a Veblen good does violate the theory of consumer choice. If a Veblen good exists, it is not an inferior good, so both the substitution and income effects of a price increase are to decrease consumption of the good.

LOS 8.d: Describe the phenomenon of diminishing marginal returns.

Factors of production are the resources a firm uses to generate output. Factors of production include:

- *Land*—where the business facilities are located.
- *Labor*—includes all workers from unskilled laborers to top management.
- *Capital*—sometimes called physical capital or plant and equipment to distinguish it from financial capital. Refers to manufacturing facilities, equipment, and machinery.
- *Materials*—refers to inputs into the productive process, including raw materials, such as iron ore or water, or manufactured inputs, such as wire or microprocessors.

For economic analysis, we often consider only two inputs, capital and labor. The quantity of output that a firm can produce can be thought of as a function of the amounts of capital and labor employed. Such a function is called a **production function**.

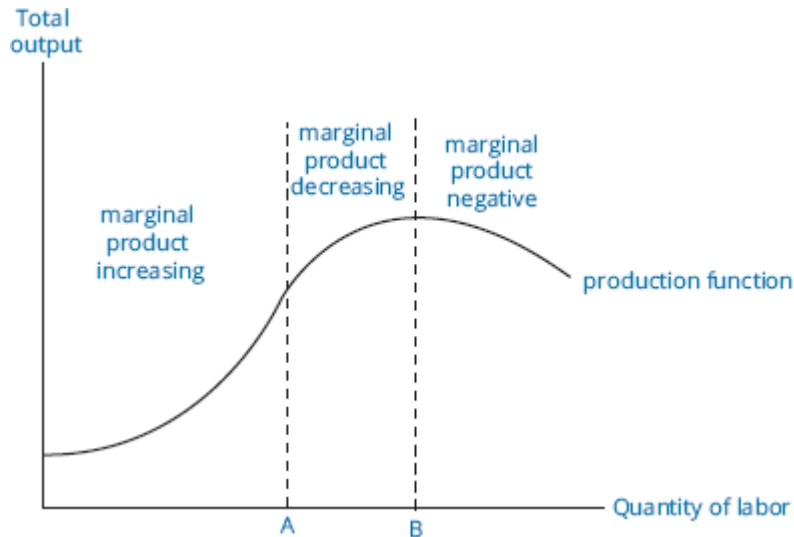
If we consider a given amount of capital (a firm's plant and equipment), we can examine the increase in production (increase in total product) that will result as we increase the amount of labor employed. The output with only one worker is considered the marginal product of the first unit of labor. The addition of a second worker will increase total product by the marginal product of the second worker. The marginal product of (additional output from) the second worker is likely greater than the marginal product of the first. This is true if we assume that two workers can produce more than twice as much output as one because of the benefits of teamwork or specialization of tasks. At this low range of labor input (remember, we are holding capital constant), we can say that the marginal product of labor is increasing.

As we continue to add additional workers to a fixed amount of capital, at some point, adding one more worker will increase total product by less than the addition of the previous worker, although total product continues to increase. When we reach the quantity of labor for which the additional output for each additional worker begins to decline, we have reached the point of **diminishing marginal productivity** of labor, or that labor has reached the point of **diminishing marginal returns**. Beyond this quantity of labor, the additional output from each additional worker continues to decline.

There is, theoretically, some quantity for labor for which the marginal product of labor is actually negative (i.e., the addition of one more worker actually decreases total output).

In Figure 8.3, we illustrate all three cases. For quantities of labor between zero and A, the marginal product of labor is increasing (slope is increasing). Beyond the inflection point in the production at quantity of labor A up to quantity B, the marginal product of labor is still positive but decreasing. The slope of the production function is positive but decreasing, and we are in a range of diminishing marginal productivity of labor. Beyond the quantity of labor B, adding additional workers decreases total output. The marginal product of labor in this range is negative, and the production function slopes downward.

Figure 8.3: Production Function—Capital Fixed, Labor Variable



LOS 8.e: Determine and interpret breakeven and shutdown points of production.

In economics, we define the **short run** for a firm as the time period over which some factors of production are fixed. Typically, we assume that capital is fixed in the short run so that a firm cannot change its scale of operations (plant and equipment) over the short run. All factors of production (costs) are variable in the **long run**. The firm can let its leases expire and sell its equipment, thereby avoiding costs that are fixed in the short run.

Shutdown and Breakeven Under Perfect Competition

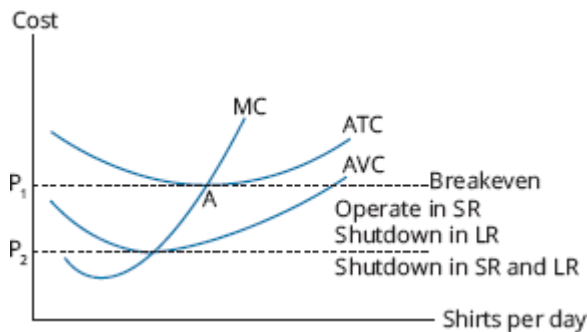
As a simple example of shutdown and breakeven analysis, consider a retail store with a 1-year lease (fixed cost) and one employee (quasi-fixed cost), so that variable costs are simply the store's cost of merchandise. If the total sales (total revenue) just covers both fixed and variable costs, price equals both average revenue and average total cost, so we are at the breakeven output quantity, and economic profit equals zero.

During the period of the lease (the short run), as long as items are being sold for more than their variable cost, the store should continue to operate to minimize losses. If items are being sold for less than their average variable cost, losses would be reduced by shutting down the business in the short run.

In the long run, a firm should shut down if the price is less than average total cost, regardless of the relation between price and average variable cost.

In the case of a firm under perfect competition, price = marginal revenue = average revenue, as we have noted. For a firm under perfect competition (a price taker), we can use a graph of cost functions to examine the profitability of the firm at different output prices. In Figure 8.4, at price P_1 , price and average revenue equal average total cost. At the output level of Point A, the firm is making an economic profit of zero. At a price above P_1 , economic profit is positive, and at prices less than P_1 , economic profit is negative (the firm has economic losses).

Figure 8.4: Shutdown and Breakeven



Because some costs are fixed in the short run, it will be better for the firm to continue production in the short run as long as average revenue is greater than average variable costs. At prices between P_1 and P_2 in Figure 8.4, the firm has losses, but the loss is less than the losses that would occur if all production were stopped. As long as total revenue is greater than total variable cost, at least some of the firm's fixed costs are covered by continuing to produce and sell its product. If the firm were to shut down, losses would be equal to the fixed costs that still must be paid. As long as price is greater than average variable costs, the firm will minimize its losses in the short run by continuing in business.

If average revenue is less average variable cost, the firm's losses are greater than its fixed costs, and it will minimize its losses by shutting down production in the short run. In this case (a price less than P_2 in Figure 8.4), the loss from continuing to operate is greater than the loss (total fixed costs) if the firm is shut down.

In the long run, all costs are variable, so a firm can avoid its (short-run) fixed costs by shutting down. For this reason, if price is expected to remain below minimum average total cost (Point A in Figure 8.4) in the long run, the firm will shut down rather than continue to generate losses.

To sum up, if average revenue is less than average variable cost in the short run, the firm should shut down. This is its **short-run shutdown point**. If average revenue is greater than average variable cost in the short run, the firm should continue to operate, even if it has losses. In the long run, the firm should shut down if average revenue is less than average total cost. This is the **long-run shutdown point**. If average revenue is just equal to average total cost, total revenue is just equal to total (economic) cost, and this is the firm's **breakeven point**.

- If $AR \geq ATC$, the firm should stay in the market in both the short and long run.
- If $AR \geq AVC$, but $AR < ATC$, the firm should stay in the market in the short run but will exit the market in the long run.
- If $AR < AVC$, the firm should shut down in the short run and exit the market in the long run.

Shutdown and Breakeven Under Imperfect Competition

For price-searcher firms (those that face downward-sloping demand curves), we could compare average revenue to ATC and AVC, just as we did for price-taker firms, to identify shutdown and breakeven points. However, marginal revenue is no longer equal to price.

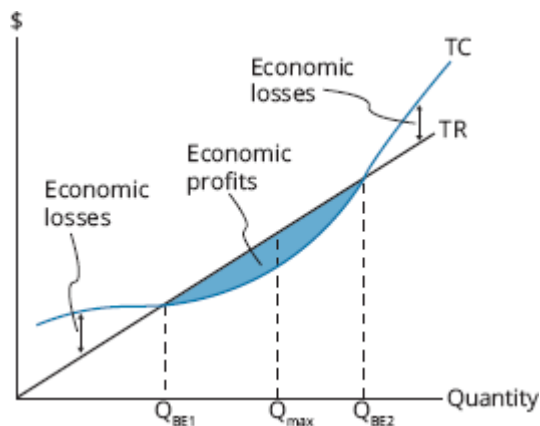
We can, however, still identify the conditions under which a firm is breaking even, should shut down in the short run, and should shut down in the long run in terms of total costs and total revenue. These conditions are:

- $TR = TC$: break even.
- $TC > TR > TVC$: firm should continue to operate in the short run but shut down in the long run.
- $TR < TVC$: firm should shut down in the short run and the long run.

Because price does not equal marginal revenue for a firm in imperfect competition, analysis based on total costs and revenues is better suited for examining breakeven and shutdown points.

The previously described relations hold for both price-taker and price-searcher firms. We illustrate these relations in Figure 8.5 for a price-taker firm (TR increases at a constant rate with quantity). Total cost equals total revenue at the breakeven quantities Q_{BE1} and Q_{BE2} . The quantity for which economic profit is maximized is shown as Q_{max} .

Figure 8.5: Breakeven Point Using the Total Revenue/Total Cost Approach



If the entire TC curve exceeds TR (i.e., no breakeven point), the firm will want to minimize the economic loss in the short run by operating at the quantity corresponding to the smallest (negative) value of $TR - TC$.

EXAMPLE: Short-run shutdown decision

For the last fiscal year, Legion Gaming reported total revenue of \$700,000, total variable costs of \$800,000, and total fixed costs of \$400,000. Should the firm continue to operate in the short run?

Answer:

The firm should shut down. Total revenue of \$700,000 is less than total costs of \$1,200,000 and also less than total variable costs of \$800,000. By shutting down, the firm will lose an amount equal to fixed costs of \$400,000. This is less than the loss of operating, which is $TR - TC = \$500,000$.

EXAMPLE: Long-run shutdown decision

Suppose instead that Legion reported total revenue of \$850,000. Should the firm continue to operate in the short run? Should it continue to operate in the long run?

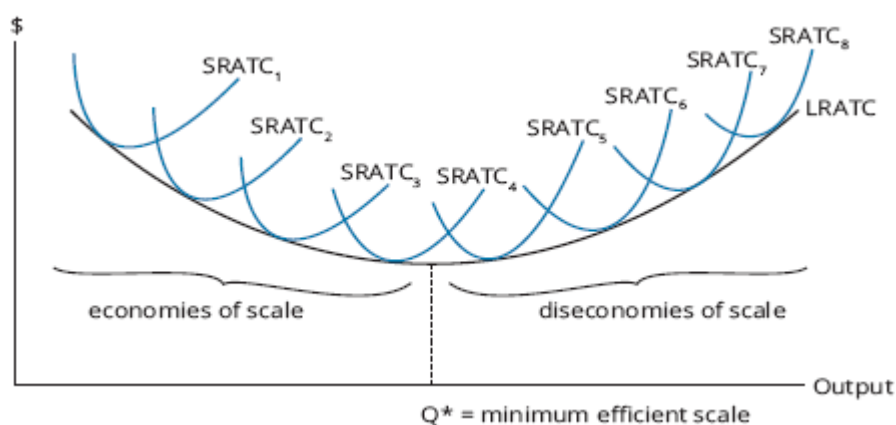
Answer:

In the short run, $TR > TVC$, and the firm should continue operating. The firm should consider exiting the market in the long run, as TR is not sufficient to cover all of the fixed costs and variable costs.

LOS 8.f: Describe how economies of scale and diseconomies of scale affect costs.

While plant size is fixed in the short run, in the long run, firms can choose their most profitable scale of operations. Because the long-run average total cost (LRATC) curve is drawn for many different plant sizes or scales of operation, each point along the curve represents the minimum ATC for a given plant size or scale of operations. In Figure 8.6, we show a firm's LRATC curve along with short-run average total cost (SRATC) curves for many different plant sizes, with $SRATC_{n+1}$ representing a larger scale of operations than $SRATC_n$.

Figure 8.6: Economies and Diseconomies of Scale



We draw the LRATC curve as U-shaped. Average total costs first decrease with larger scale and eventually increase. The lowest point on the LRATC corresponds to the scale or plant size at which the average total cost of production is at a minimum. This scale is sometimes called the **minimum efficient scale**. Under perfect competition, firms must operate at minimum efficient scale in long-run equilibrium, and LRATC will equal the market price. Recall that under perfect competition, firms earn zero economic profit in long-run equilibrium. Firms that have chosen a different scale of operations with higher average total costs will have economic losses and must either leave the industry or change to minimum efficient scale.

The downward-sloping segment of the long-run average total cost curve presented in Figure 8.6 indicates that **economies of scale** (or *increasing returns to scale*) are present. Economies of scale result from factors such as labor specialization, mass production, and investment in more efficient equipment and technology. In addition, the firm may be able to negotiate lower input prices with suppliers as firm size increases and more resources are purchased. A firm operating with economies of scale can increase its competitiveness by expanding production and reducing costs.

The upward-sloping segment of the LRATC curve indicates that **diseconomies of scale** are present. Diseconomies of scale may result as the increasing bureaucracy of larger firms leads to inefficiency, problems with motivating a larger workforce, and greater barriers to innovation

and entrepreneurial activity. A firm operating under diseconomies of scale will want to decrease output and move back toward the minimum efficient scale. The U.S. auto industry is an example of an industry that has exhibited diseconomies of scale.

There may be a relatively flat portion at the bottom of the LRATC curve that exhibits *constant returns to scale*. Over a range of constant returns to scale, costs are constant for the various plant sizes.



MODULE QUIZ 8.1, 8.2

- Total revenue is greatest in the part of a demand curve that is:
 - elastic
 - inelastic
 - unit elastic.
- A demand function for air conditioners is given by:
$$Q_{\text{air conditioner}} = 10,000 - 2 P_{\text{air conditioner}} + 0.0004 \text{ income} + 30 P_{\text{electric fan}} - 4 P_{\text{electricity}}$$
At current average prices, an air conditioner costs 5,000 yen, a fan costs 200 yen, and electricity costs 1,000 yen. Average income is 4,000,000 yen. The income elasticity of demand for air conditioners is *closest* to:
 - 0.0004.
 - 0.444.
 - 40,000.
- When the price of a good decreases, and an individual's consumption of that good also decreases, it is *most likely* that:
 - the income effect and substitution effect are both negative.
 - the substitution effect is negative and the income effect is positive.
 - the income effect is negative and the substitution effect is positive.
- A good is classified as an inferior good if its:
 - income elasticity is negative.
 - own-price elasticity is negative.
 - cross-price elasticity is negative.
- Increasing the amount of one productive input while keeping the amounts of other inputs constant results in diminishing marginal returns:
 - in all cases.
 - when it causes total output to decrease.
 - when the increase in total output becomes smaller.
- A firm's average revenue is greater than its average variable cost and less than its average total cost. If this situation is expected to persist, the firm should:
 - shut down in the short run and in the long run.
 - shut down in the short run but operate in the long run.
 - operate in the short run but shut down in the long run.
- If a firm's long-run average total cost increases by 6% when output is increased by 6%, the firm is experiencing:
 - economies of scale.
 - diseconomies of scale.
 - constant returns to scale.

KEY CONCEPTS

Elasticity is measured as the ratio of the percentage change in one variable to a percentage change in another. Three elasticities related to a demand function are of interest:

$$\text{own-price elasticity} = \frac{\% \text{ change in quantity demanded}}{\% \text{ change in own price}}$$

$$\text{cross-price elasticity} = \frac{\% \text{ change in quantity demanded}}{\% \text{ change in price of related good}}$$

$$\text{income elasticity} = \frac{\% \text{ change in quantity demanded}}{\% \text{ change in income}}$$

|own-price elasticity| > 1: demand is elastic

|own-price elasticity| < 1: demand is inelastic

cross-price elasticity > 0: related good is a substitute

cross-price elasticity < 0: related good is a complement

income elasticity < 0: good is an inferior good

income elasticity > 0: good is a normal good

LOS 8.b

When the price of a good decreases, the substitution effect leads a consumer to consume more of that good and less of goods for which prices have remained the same.

A decrease in the price of a good that a consumer purchases leaves her with unspent income (for the same combination of goods). The effect of this additional income on consumption of the good for which the price has decreased is termed the income effect.

LOS 8.c

For a normal good, the income effect of a price decrease is positive—income elasticity of demand is positive.

For an inferior good, the income effect of a price decrease is negative—income elasticity of demand is negative. An increase in income reduces demand for an inferior good.

A Giffen good is an inferior good for which the negative income effect of a price decrease outweighs the positive substitution effect, so that a decrease (increase) in the good's price has a net result of decreasing (increasing) the quantity consumed.

A Veblen good is also one for which an increase (decrease) in price results in an increase (decrease) in the quantity consumed. However, a Veblen good is not an inferior good and is not supported by the axioms of the theory of demand.

LOS 8.d

Marginal returns refer to the additional output that can be produced by using one more unit of a productive input while holding the quantities of other inputs constant. Marginal returns may increase as the first units of an input are added, but as input quantities increase, they reach a point at which marginal returns begin to decrease. Inputs beyond this quantity are said to produce diminishing marginal returns.

LOS 8.e

Under perfect competition:

- The breakeven quantity of production is the quantity for which price (P) = average total cost (ATC) and total revenue (TR) = total cost (TC).
- The firm should shut down in the long run if $P < ATC$ so that $TR < TC$.
- The firm should shut down in the short run (and the long run) if $P < \text{average variable cost (AVC)}$ so that $TR < \text{total variable cost (TVC)}$.

Under imperfect competition (firm faces downward sloping demand):

- Breakeven quantity is the quantity for which $TR = TC$.
- The firm should shut down in the long run if $TR < TC$.
- The firm should shut down in the short run (and the long run) if $TR < TVC$.

LOS 8.f

The long-run average total cost (LRATC) curve shows the minimum average total cost for each level of output assuming that the plant size (scale of the firm) can be adjusted. A downward-sloping segment of an LRATC curve indicates economies of scale (increasing returns to scale). Over such a segment, increasing the scale of the firm reduces ATC. An upward-sloping segment of an LRATC curve indicates diseconomies of scale, where average unit costs will rise as the scale of the business (and long-run output) increases.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 8.1, 8.2

1. **C** Total revenue is maximized at the quantity at which own-price elasticity equals -1 . (Module 8.1, LOS 8.a)
2. **B** Substituting current values for the independent variables other than income, the demand function becomes:

$$\begin{aligned}
 QD_{\text{air conditioner}} &= 10,000 - 2(5,000) + 0.0004 \text{ income} + 30(200) - 4(1,000) \\
 &= 0.0004 \text{ income} + 2,000.
 \end{aligned}$$

The slope of income is 0.0004, and for an income of 4,000,000 yen, $QD = 3,600$.
 Income elasticity = $I_0 / Q_0 \times \Delta Q / \Delta I = 4,000,000 / 3,600 \times 0.0004 = 0.444$. (Module 8.1, LOS 8.a)
3. **C** The substitution effect of a price decrease is always positive, but the income effect can be either positive or negative. Consumption of a good will decrease when the price of that good decreases only if the income effect is both negative and greater than the substitution effect. (Module 8.2, LOS 8.b)
4. **A** An inferior good is one that has a negative income elasticity of demand. (Module 8.2, LOS 8.c)
5. **C** Productive inputs exhibit diminishing marginal returns at the level where an additional unit of input results in a smaller increase in output than the previous unit of input. (Module 8.2, LOS 8.d)
6. **C** If a firm is generating sufficient revenue to cover its variable costs and part of its fixed costs, it should continue to operate in the short run. If average revenue is likely to remain below average total costs in the long run, the firm should shut down. (Module 8.2, LOS 8.e)
7. **B** Increasing long-run average total cost as a result of increasing output demonstrates diseconomies of scale. (Module 8.2, LOS 8.f)

READING 9

THE FIRM AND MARKET STRUCTURES

EXAM FOCUS

This reading covers four market structures: perfect competition, monopolistic competition, oligopoly, and monopoly. You need to be able to compare and contrast these structures in terms of numbers of firms, firm demand elasticity and pricing power, long-run economic profits, barriers to entry, and the amount of product differentiation and advertising. Finally, know the two quantitative concentration measures, their implications for market structure and pricing power, and their limitations in this regard. We will apply all of these concepts when we analyze industry competition and pricing power of companies in the Equity Investments topic area.

MODULE 9.1: PERFECT COMPETITION



LOS 9.a: Describe characteristics of perfect competition, monopolistic competition, oligopoly, and pure monopoly.

Video covering this content is available online.

In this reading, we examine four types of market structure: perfect competition, monopolistic competition, oligopoly, and monopoly. We can analyze where an industry falls along this spectrum by examining the following five factors:

1. Number of firms and their relative sizes.
2. Degree to which firms differentiate their products.
3. Bargaining power of firms with respect to pricing.
4. Barriers to entry into or exit from the industry.
5. Degree to which firms compete on factors other than price.

At one end of the spectrum is **perfect competition**, in which many firms produce identical products, and competition forces them all to sell at the market price. At the other extreme, we have **monopoly**, where only one firm is producing the product. In between are **monopolistic competition** (many sellers and differentiated products) and **oligopoly** (few firms that compete in a variety of ways). Each market structure has its own characteristics and implications for firm strategy, and we will examine each in turn.

Perfect competition refers to a market in which many firms produce identical products, barriers to entry into the market are very low, and firms compete for sales only on the basis of price. Firms face perfectly elastic (horizontal) demand curves at the price determined in the market because no firm is large enough to affect the market price. The market for wheat in a region is a

good approximation of such a market. Overall market supply and demand determine the price of wheat.

Monopolistic competition differs from perfect competition in that products are not identical. Each firm differentiates its product(s) from those of other firms through some combination of differences in product quality, product features, and marketing. The demand curve faced by each firm is downward sloping; while demand is elastic, it is not perfectly elastic. Prices are not identical because of perceived differences among competing products, and barriers to entry are low. The market for toothpaste is a good example of monopolistic competition. Firms differentiate their products through features and marketing with claims of more attractiveness, whiter teeth, fresher breath, and even of actually cleaning your teeth and preventing decay. If the price of your personal favorite increases, you are not likely to immediately switch to another brand as under perfect competition. Some customers would switch in response to a 10% increase in price and some would not. This is why firm demand is downward sloping.

The most important characteristic of an *oligopoly* market is that there are only a few firms competing. In such a market, each firm must consider the actions and responses of other firms in setting price and business strategy. We say that such firms are interdependent. While products are typically good substitutes for each other, they may be either quite similar or differentiated through features, branding, marketing, and quality. Barriers to entry are high, often because economies of scale in production or marketing lead to very large firms. Demand can be more or less elastic than for firms in monopolistic competition. The automobile market is dominated by a few very large firms and can be characterized as an oligopoly. The product and pricing decisions of Toyota certainly affect those of Ford and vice versa. Automobile makers compete based on price, but also through marketing, product features, and quality, which is often signaled strongly through brand name. The oil industry also has a few dominant firms but their products are very good substitutes for each other.

A *monopoly* market is characterized by a single seller of a product with no close substitutes. This fact alone means that the firm faces a downward-sloping demand curve (the market demand curve) and has the power to choose the price at which it sells its product. High barriers to entry protect a monopoly producer from competition. One source of monopoly power is the protection offered by copyrights and patents. Another possible source of monopoly power is control over a resource specifically needed to produce the product. Most frequently, monopoly power is supported by government. A **natural monopoly** refers to a situation where the average cost of production is falling over the relevant range of consumer demand. In this case, having two (or more) producers would result in a significantly higher cost of production and be detrimental to consumers. Examples of natural monopolies include the electric power and distribution business and other public utilities. When privately owned companies are granted such monopoly power, the price they charge is often regulated by government as well.

Sometimes market power is the result of *network effects* or *synergies* that make it very difficult to compete with a company once it has reached a critical level of market penetration. EBay gained such a large share of the online auction market that its information on buyers and sellers and the number of buyers who visit eBay essentially precluded others from establishing competing businesses. While it may have competition to some degree, its market share is such that it has negatively sloped demand and a good deal of pricing power. Sometimes we refer to such companies as having a moat around them that protects them from competition. It is best to

remember, however, that changes in technology and consumer tastes can, and usually do, reduce market power over time. Polaroid had a monopoly on instant photos for years, but the introduction of digital photography forced the firm into bankruptcy in 2001.

The table in Figure 9.1 shows the key features of each market structure.

Figure 9.1: Characteristics of Market Structures

	Perfect Competition	Monopolistic Competition	Oligopoly	Monopoly
Number of sellers	Many firms	Many firms	Few firms	Single firm
Barriers to entry	Very low	Low	High	Very high
Nature of substitute products	Very good substitutes	Good substitutes but differentiated	Very good substitutes or differentiated	No good substitutes
Nature of competition	Price only	Price, marketing, features	Price, marketing, features	Advertising
Pricing power	None	Some	Some to significant	Significant

LOS 9.b: Explain relationships between price, marginal revenue, marginal cost, economic profit, and the elasticity of demand under each market structure.

LOS 9.d: Describe and determine the optimal price and output for firms under each market structure.

LOS 9.f: Explain factors affecting long-run equilibrium under each market structure.

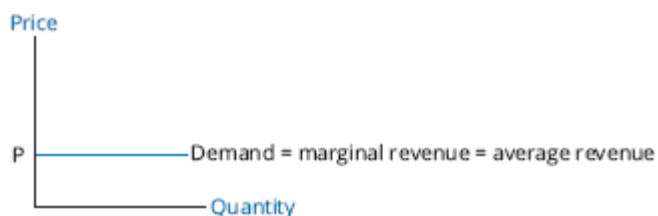


PROFESSOR'S NOTE

We cover these LOS together and slightly out of curriculum order so that we can present the complete analysis of each market structure to better help candidates understand the economics of each type of market structure.

Producer firms in perfect competition have no influence over market price. Market supply and demand determine price. As illustrated in Figure 9.2, the *individual firm's* demand schedule is *perfectly elastic* (horizontal).

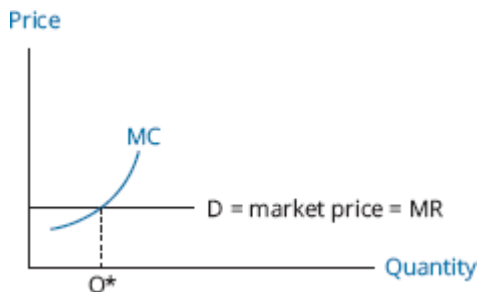
Figure 9.2: Price-Taker Demand



In a perfectly competitive market, a firm will continue to expand production until marginal revenue (MR) equals marginal cost (MC). Marginal revenue is the increase in total revenue from selling one more unit of a good or service. For a price taker, marginal revenue is simply the price

because all additional units are assumed to be sold at the same (market) price. In *pure competition*, a firm's marginal revenue is equal to the market price, and a firm's MR curve, presented in Figure 9.3, is identical to its demand curve. A profit maximizing firm will produce the quantity, Q^* , when $MC = MR$.

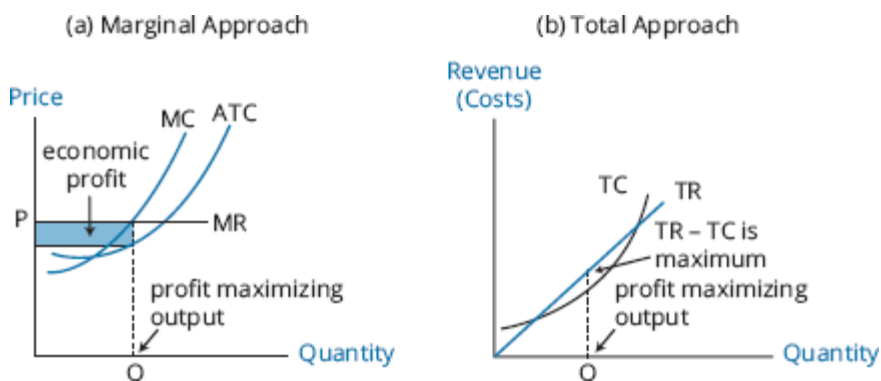
Figure 9.3: Profit Maximizing Output For A Price Taker



All firms maximize (economic) profit by producing and selling the quantity for which marginal revenue equals marginal cost. For a firm in a perfectly competitive market, this is the same as producing and selling the quantity for which marginal cost equals (market) price. Economic profit equals total revenues less the opportunity cost of production, which includes the cost of a normal return to all factors of production, including invested capital.

Panel (a) of Figure 9.4 illustrates that in the *short run*, economic profit is maximized at the quantity for which marginal revenue = marginal cost. As shown in Panel (b), profit maximization also occurs when total revenue exceeds total cost by the maximum amount.

Figure 9.4: Short-Run Profit Maximization



An *economic loss* occurs on any units for which marginal revenue is less than marginal cost. At any output above the quantity where $MR = MC$, the firm will be generating losses on its marginal production and will maximize profits by reducing output to where $MR = MC$.

In a perfectly competitive market, firms will not earn economic profits for any significant period of time. The assumption is that new firms (with average and marginal cost curves identical to those of existing firms) will enter the industry to earn economic profits, increasing market supply and eventually reducing market price so that it just equals firms' average total cost (ATC). In equilibrium, each firm is producing the quantity for which $P = MR = MC = ATC$, so that no firm earns economic profits and each firm is producing the quantity for which ATC is a

minimum (the quantity for which $ATC = MC$). This equilibrium situation is illustrated in Figure 9.5.

Figure 9.5: Equilibrium in a Perfectly Competitive Market

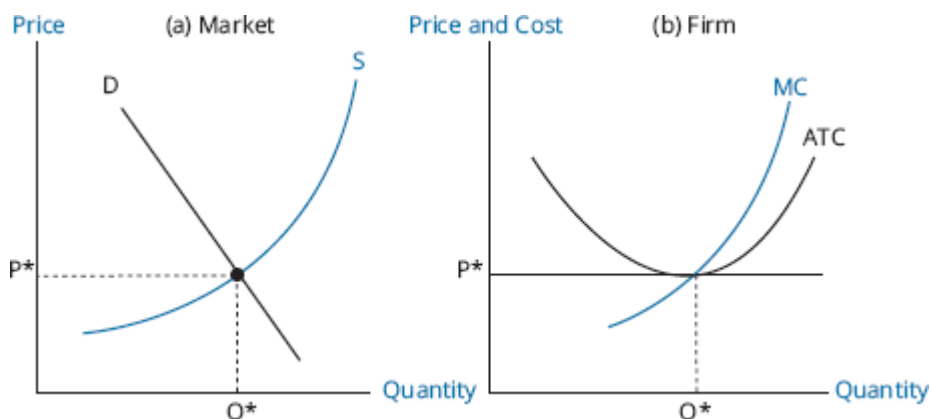
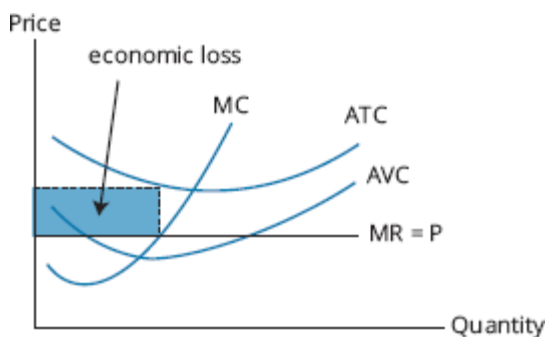


Figure 9.6 illustrates that firms will experience economic losses when price is below average total cost ($P < ATC$). In this case, the firm must decide whether to continue operating. A firm will minimize its losses in the short run by continuing to operate when price is less than ATC but greater than AVC . As long as the firm is covering its variable costs and some of its fixed costs, its loss will be less than its fixed (in the short run) costs. If the firm is only just covering its variable costs ($P = AVC$), the firm is operating at its **shutdown point**. If the firm is not covering its variable costs ($P < AVC$) by continuing to operate, its losses will be greater than its fixed costs. In this case, the firm will shut down (zero output) and lay off its workers. This will limit its losses to its fixed costs (e.g., its building lease and debt payments). If the firm does not believe price will ever exceed ATC in the future, going out of business is the only way to eliminate fixed costs.

Figure 9.6: Short-Run Loss

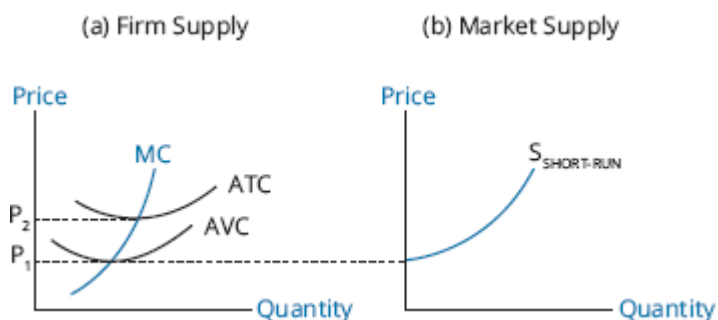


The *long-run equilibrium output* level for perfectly competitive firms is where $MR = MC = ATC$, which is where ATC is at a minimum. At this output, economic profit is zero and only a normal return is realized.

Recall that price takers should produce where $P = MC$. Referring to Panel (a) in Figure 9.7, a firm will shut down at a price below P_1 . Between P_1 and P_2 , a firm will continue to operate in the short run. At P_2 , the firm is earning a normal profit—economic profit equals zero. At prices above P_2 , a firm is making economic profits and will expand its production along the MC line.

Thus, the **short-run supply curve for a firm** is its MC line above the average variable cost curve, AVC. The supply curve shown in Panel (b) is the **short-run market supply curve**, which is the horizontal sum (add up the quantities from all firms at each price) of the MC curves for all firms in a given industry. Because firms will supply more units at higher prices, the short-run market supply curve slopes upward to the right.

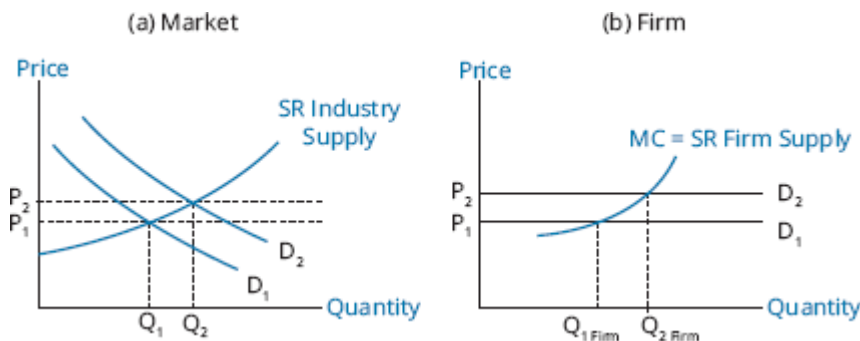
Figure 9.7: Short-Run Supply Curves



Changes in Demand, Entry and Exit, and Changes in Plant Size

In the short run, an increase in market demand (a shift of the market demand curve to the right) will increase both equilibrium price and quantity, while a decrease in market demand will reduce both equilibrium price and quantity. The change in equilibrium price will change the (horizontal) demand curve faced by each individual firm and the profit-maximizing output of a firm. These effects for an increase in demand are illustrated in Figure 9.8. An increase in market demand from D_1 to D_2 increases the short-run equilibrium price from P_1 to P_2 and equilibrium output from Q_1 to Q_2 . In Panel (b) of Figure 9.8, we see the short-run effect of the increased market price on the output of an individual firm. The higher price leads to a greater profit-maximizing output, $Q_{2 \text{ Firm}}$. At the higher output level, a firm will earn an economic profit in the short run. In the long run, some firms will increase their scale of operations in response to the increase in demand, and new firms will likely enter the industry. In response to a decrease in demand, the short-run equilibrium price and quantity will fall, and in the long run, firms will decrease their scale of operations or exit the market.

Figure 9.8: Short-Run Adjustment to an Increase in Demand Under Perfect Competition



A firm's long-run adjustment to a shift in industry demand and the resulting change in price may be either to alter the size of its plant or leave the market entirely. The marketplace abounds with examples of firms that have increased their plant sizes (or added additional production facilities) to increase output in response to increasing market demand. Other firms, such as Ford

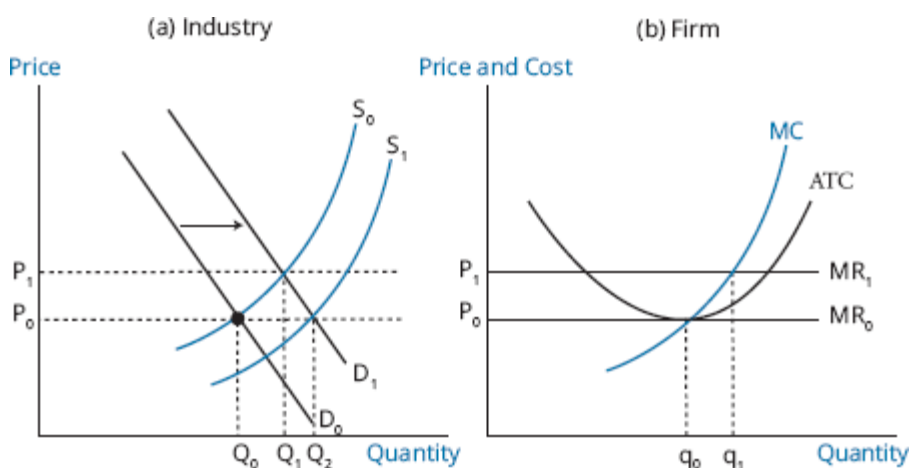
and GM, have decreased plant size to reduce economic losses. This strategy is commonly referred to as *downsizing*.

If an industry is characterized by firms earning economic profits, new firms will enter the market. This will cause industry supply to increase (the industry supply curve shifts downward and to the right), increasing equilibrium output and decreasing equilibrium price. Even though industry output increases, however, individual firms will produce less because as price falls, each individual firm will move down its own supply curve. The end result is that a firm's total revenue and economic profit will decrease.

If firms in an industry are experiencing economic losses, some of these firms will exit the market. This will decrease industry supply and increase equilibrium price. Each remaining firm in the industry will move up its individual supply curve and increase production at the higher market price. This will cause total revenues to increase, reducing any economic losses the remaining firms had been experiencing.

A *permanent change in demand* leads to the entry of firms to, or exit of firms from, an industry. Let's consider the permanent increase in demand illustrated in Figure 9.9. The initial long-run industry equilibrium condition shown in Panel (a) is at the intersection of demand curve D_0 and supply curve S_0 , at price P_0 and quantity Q_0 . As indicated in Panel (b) of Figure 9.9, at the market price of P_0 each firm will produce q_0 . At this price and output, each firm earns a normal profit, and economic profit is zero. That is, $MC = MR = P$, and ATC is at its minimum. Now, suppose industry demand permanently increases such that the industry demand curve in Panel (a) shifts to D_1 . The new market price will be P_1 and industry output will increase to Q_1 . At the new price P_1 , existing firms will produce q_1 and realize an economic profit because $P_1 > ATC$. Positive economic profits will cause new firms to enter the market. As these new firms increase total industry supply, the industry supply curve will gradually shift to S_1 , and the market price will decline back to P_0 . At the market price of P_0 , the industry will now produce Q_2 , with an increased number of firms in the industry, each producing at the original quantity, q_0 . The individual firms will no longer enjoy an economic profit because $ATC = P_0$ at q_0 .

Figure 9.9: Effects of a Permanent Increase in Demand



1. When a firm operates under conditions of pure competition, marginal revenue always equals:
 - A. price.
 - B. average cost.
 - C. marginal cost.
2. In which market structure(s) can a firm's supply function be described as its marginal cost curve above its average variable cost curve?
 - A. Oligopoly or monopoly.
 - B. Perfect competition only.
 - C. Perfect competition or monopolistic competition.
3. In a purely competitive market, economic losses indicate that:
 - A. price is below average total costs.
 - B. collusion is occurring in the market place.
 - C. firms need to expand output to reduce costs.
4. A purely competitive firm will tend to expand its output so long as:
 - A. marginal revenue is positive.
 - B. marginal revenue is greater than price.
 - C. market price is greater than marginal cost.
5. A firm is likely to operate in the short run as long as price is at least as great as:
 - A. marginal cost.
 - B. average total cost.
 - C. average variable cost.

MODULE 9.2: MONOPOLISTIC COMPETITION



Video covering this content is available online.

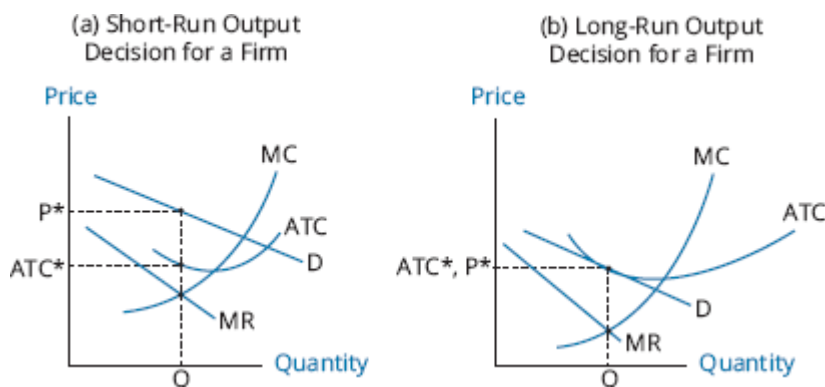
Monopolistic competition has the following market characteristics:

- *A large number of independent sellers:* (1) Each firm has a relatively small market share, so no individual firm has any significant power over price. (2) Firms need only pay attention to average market price, not the price of individual competitors. (3) There are too many firms in the industry for collusion (price fixing) to be possible.
- *Differentiated products:* Each producer has a product that is slightly different from its competitors (at least in the minds of consumers). The competing products are close substitutes for one another.
- *Firms compete on price, quality, and marketing* as a result of product differentiation. *Quality* is a significant product-differentiating characteristic. *Price* and output can be set by firms because they face downward-sloping demand curves, but there is usually a strong correlation between quality and the price that firms can charge. *Marketing* is a must to inform the market about a product's differentiating characteristics.
- *Low barriers to entry* so that firms are free to enter and exit the market. If firms in the industry are earning economic profits, new firms can be expected to enter the industry.

Firms in monopolistic competition face *downward-sloping demand* curves (they are price searchers). Their demand curves are highly *elastic* because competing products are perceived by consumers as close substitutes. Think about the market for toothpaste. All toothpaste is quite similar, but differentiation occurs due to taste preferences, influential advertising, and the reputation of the seller.

The price/output decision for monopolistic competition is illustrated in Figure 9.10. Panel (a) of Figure 9.10 illustrates the short-run price/output characteristics of monopolistic competition for a single firm. As indicated, firms in monopolistic competition maximize economic profits by producing where marginal revenue (MR) equals marginal cost (MC), and by charging the price for that quantity from the demand curve, D . Here the firm earns positive economic profits because price, P^* , exceeds average total cost, ATC^* . Due to low barriers to entry, competitors will enter the market in pursuit of these economic profits.

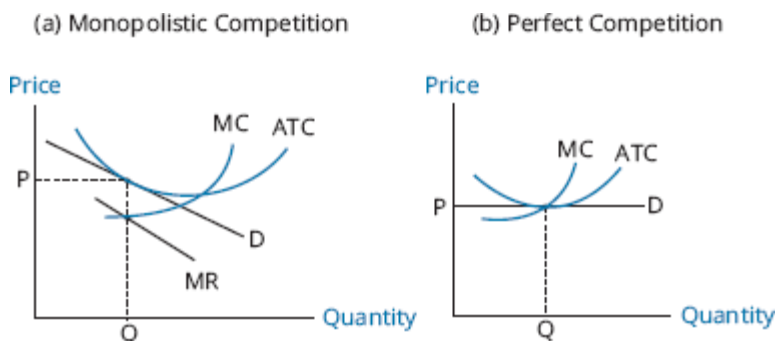
Figure 9.10: Short-Run and Long-Run Output Under Monopolistic Competition



Panel (b) of Figure 9.10 illustrates long-run equilibrium for a *representative* firm after new firms have entered the market. As indicated, the entry of new firms shifts the demand curve faced by each individual firm down to the point where price equals average total cost ($P^* = ATC^*$), such that economic profit is zero. At this point, there is no longer an incentive for new firms to enter the market, and long-run equilibrium is established. The firm in monopolistic competition continues to produce at the quantity where $MR = MC$ but no longer earns positive economic profits.

Figure 9.11 illustrates the differences between long-run equilibrium in markets with monopolistic competition and markets with perfect competition. Note that with monopolistic competition, price is greater than marginal cost (i.e., producers can realize a **markup**), average total cost is not at a minimum for the quantity produced (suggesting **excess capacity**, or an inefficient scale of production), and the price is slightly higher than under perfect competition. The point to consider here, however, is that perfect competition is characterized by no product differentiation. The question of the efficiency of monopolistic competition becomes, "Is there an economically efficient amount of product differentiation?"

Figure 9.11: Firm Output Under Monopolistic and Perfect Competition



In a world with only one brand of toothpaste, clearly average production costs would be lower. That fact alone probably does not mean that a world with only one brand/type of toothpaste would be a better world. While product differentiation has costs, it also has benefits to consumers.

Consumers definitely benefit from brand name promotion and advertising because they receive information about the nature of a product. This often enables consumers to make better purchasing decisions. Convincing consumers that a particular brand of deodorant will actually increase their confidence in a business meeting or make them more attractive to the opposite sex is not easy or inexpensive. Whether the perception of increased confidence or attractiveness from using a particular product is worth the additional cost of advertising is a question probably better left to consumers of the products. Some would argue that the increased cost of advertising and sales is not justified by the benefits of these activities.

Product innovation is a necessary activity as firms in monopolistic competition pursue economic profits. Firms that bring new and innovative products to the market are confronted with less-elastic demand curves, enabling them to increase price and earn economic profits. However, close substitutes and imitations will eventually erode the initial economic profit from an innovative product. Thus, firms in monopolistic competition must continually look for innovative product features that will make their products relatively more desirable to some consumers than those of the competition.

Innovation does not come without costs. The costs of product innovation must be weighed against the extra revenue that it produces. A firm is considered to be spending the optimal amount on innovation when the marginal cost of (additional) innovation just equals the marginal revenue (marginal benefit) of additional innovation.

Advertising expenses are high for firms in monopolistic competition. This is to inform consumers about the unique features of their products and to create or increase a perception of differences between products that are actually quite similar. We just note here that advertising costs for firms in monopolistic competition are greater than those for firms in perfect competition and those that are monopolies.

As you might expect, advertising costs increase the average total cost curve for a firm in monopolistic competition. The increase to average total cost attributable to advertising decreases as output increases, because more fixed advertising dollars are being averaged over a larger quantity. In fact, if advertising leads to enough of an increase in output (sales), it can actually decrease a firm's average total cost.

Brand names provide information to consumers by providing them with signals about the quality of the branded product. Many firms spend a significant portion of their advertising budget on brand name promotion. Seeing the brand name BMW likely tells a consumer more about the quality of a newly introduced automobile than an inspection of the vehicle itself would reveal. At the same time, the reputation BMW has for high quality is so valuable that the firm has an added incentive not to damage it by producing vehicles of low quality.



MODULE QUIZ 9.2

1. The demand for products from monopolistic competitors is relatively elastic due to:
A. high barriers to entry.

- B. the availability of many close substitutes.
 - C. the availability of many complementary goods.
2. Compared to a perfectly competitive industry, in an industry characterized by monopolistic competition:
 - A. both price and quantity are likely to be lower.
 - B. price is likely to be higher and quantity is likely to be lower.
 - C. quantity is likely to be higher and price is likely to be lower.
 3. A firm will *most likely* maximize profits at the quantity of output for which:
 - A. price equals marginal cost.
 - B. price equals marginal revenue.
 - C. marginal cost equals marginal revenue.

MODULE 9.3: OLIGOPOLY



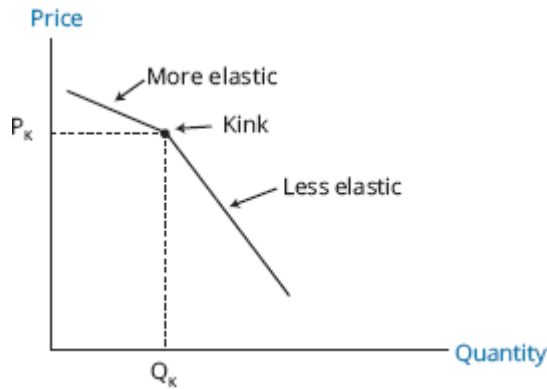
Video covering this content is available online.

Compared to monopolistic competition, an oligopoly market has higher barriers to entry and fewer firms. The other key difference is that the firms are interdependent, so a price change by one firm can be expected to be met by a price change by its competitors. This means that the actions of another firm will directly affect a given firm's demand curve for the product. Given this complicating fact, models of oligopoly pricing and profits must make a number of important assumptions. In the following, we describe four of these models and their implications for price and quantity:

1. Kinked demand curve model.
2. Cournot duopoly model.
3. Nash equilibrium model (prisoner's dilemma).
4. Stackelberg dominant firm model.

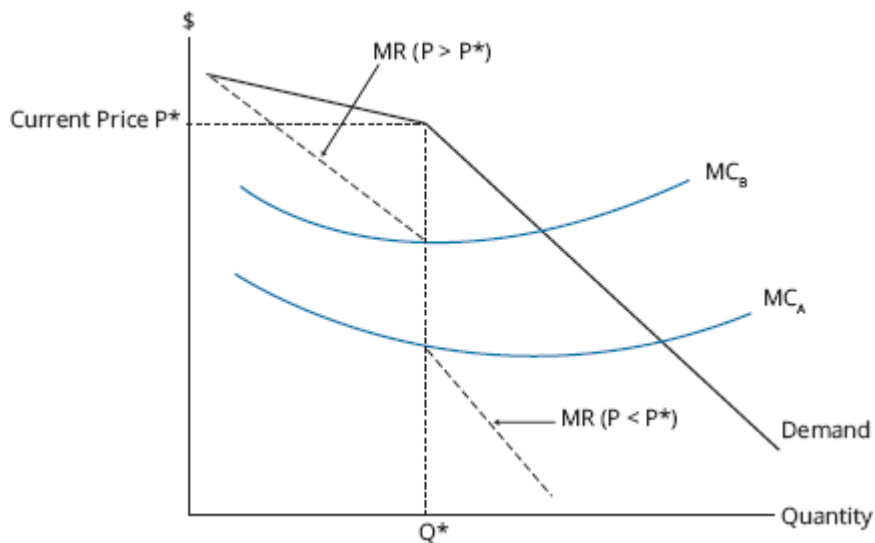
One traditional model of oligopoly, the **kinked demand curve model**, is based on the assumption that an increase in a firm's product price will not be followed by its competitors, but a decrease in price will. According to the kinked demand curve model, each firm believes that it faces a demand curve that is more elastic (flatter) above a given price (the kink in the demand curve) than it is below the given price. The kinked demand curve model is illustrated in Figure 9.12. The kink price is at price P_K , where a firm produces Q_K . A firm believes that if it raises its price above P_K , its competitors will remain at P_K , and it will lose market share because it has the highest price. Above P_K , the demand curve is considered to be relatively elastic, where a small price increase will result in a large decrease in demand. On the other hand, if a firm decreases its price below P_K , other firms will match the price cut, and all firms will experience a relatively small increase in sales relative to any price reduction. Therefore, Q_K is the profit-maximizing level of output.

Figure 9.12: Kinked Demand Curve Model



It is worth noting that with a kink in the market demand curve, we also get a gap in the associated marginal revenue curve, as shown in Figure 9.13. For any firm with a marginal cost curve passing through this gap, the price at which the kink is located is the firm's profit maximizing price.

Figure 9.13: Gap in Marginal Revenue Curve



We say that the decisions of firms in an oligopoly are interdependent; that is, the pricing decision of one firm depends on the pricing decision of another firm. Some models of market price equilibrium have a set of rules for the actions of oligopolists. These rules assume they choose prices based on the choices of the other firms. By specifying the decision rules that each firm follows, we can design a model that allows us to determine the equilibrium prices and quantities for firms operating in an oligopoly market. An early model of oligopoly pricing decisions is the **Cournot model**. In Cournot's model, two firms with identical marginal cost curves each choose their preferred selling price based on the price the other firm chose in the previous period. The equilibrium for an oligopoly with two firms (duopoly), in the Cournot model, is for both firms to sell the same amounts and same quantities, splitting the market equally at the equilibrium price. The equilibrium price is less than the price a single monopolist would charge, but greater than the equilibrium price that would result under perfect competition.

Another model, the **Stackelberg model**, uses a different set of rules. One firm is the "leader" and chooses its price first, and the other firm chooses a price based on the leader's price. In

equilibrium, under these rules, the leader charges a higher price and receives a greater proportion of the firms' total profits.

Firms determine their quantities simultaneously each period and, under the assumptions of the Cournot model, these quantities will change each period until they are equal. When each firm selects the same quantity, there is no longer any additional profit to be gained by changing quantity, and we have a stable equilibrium. The resulting market price is less than the profit maximizing price that a monopolist would charge, but higher than marginal cost, the price that would result from perfect competition. Additional analysis shows that as more firms are added to the model, the equilibrium market price falls towards marginal cost, which is the equilibrium price in the limit as the number of firms gets large.

These rules-based models are early versions of what are called *strategic games*, decision models in which the best choice for a firm depends on the actions (reactions) of other firms. A more general model of this strategic game was developed by Nobel Prize winner John Nash, who developed the concept of a **Nash equilibrium**. A Nash equilibrium is reached when the choices of all firms are such that there is no other choice that makes any firm better off (increases profits or decreases losses).

The concept of a Nash equilibrium can be applied to the situation presented in Figure 9.14, which shows the choices and resulting profits for two firms. Each firm can charge either a high price or a low price. If both firms charge a high price, Firm A earns 1,000 and Firm B earns 600. While Firm A would not charge the low price (it would earn less regardless of Firm B's decision), Firm B can increase profits to 700 by charging a low price. With Firm A charging a high price and Firm B charging a low price, neither firm can increase profits by changing its price strategy. Thus, we can identify the Nash equilibrium in this scenario as Firm B charging a low price and Firm A charging a high price.

Figure 9.14: Nash Equilibrium

	Firm B High Price	Firm B Low Price
Firm A High Price	A earns 1,000 B earns 600	A earns 600 B earns 700
Firm A Low Price	A earns 160 B earns 0	A earns 100 B earns 140

The firms could, however, collude. The greatest joint profits (1,600) are earned when both firms charge a high price. If Firm A offers to pay Firm B 200 for charging a high price, Firm A's profits increase from 600 to 1,000. After paying 200 to Firm B, Firm A still gains 200. Firm B's profits (including the payment of 200) increase from 700 to 800. Collusion, in this case, increases the profits of both firms, compared to the Nash equilibrium.

If firms can enter into and enforce an agreement regarding pricing and output, often they can all benefit. Such agreements among producers are illegal in many countries because they reduce competition.

An example of a collusive agreement is the OPEC **cartel**. Cartel member countries agree to restrict their oil production in order to increase the world price of oil. Members sometimes choose to "cheat" on the cartel agreement by producing more than the amount of oil they have

agreed to produce. If members of a cartel do not adhere to the agreement, taking advantage of the higher market price but failing to restrict output to the agreed-upon amount, the agreement can quickly break down.

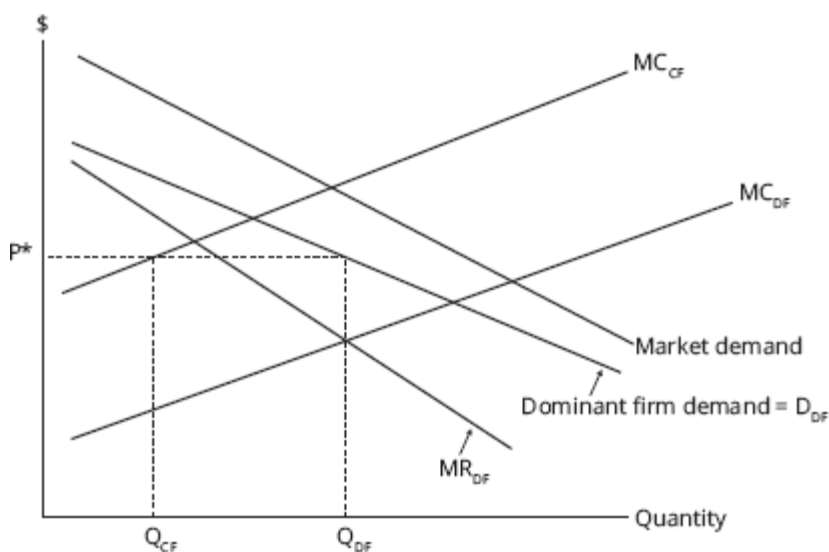
In general, collusive agreements to increase price in an oligopoly market will be more successful (have less cheating) when:

- There are fewer firms.
- Products are more similar (less differentiated).
- Cost structures are more similar.
- Purchases are relatively small and frequent.
- Retaliation by other firms for cheating is more certain and more severe.
- There is less actual or potential competition from firms outside the cartel.

A final model of oligopoly behavior to consider is the **dominant firm model**. In this model, there is a single firm that has a significantly large market share because of its greater scale and lower cost structure—the dominant firm (DF). In such a model, the market price is essentially determined by the dominant firm, and the other competitive firms (CF) take this market price as given.

The dominant firm believes that the quantity supplied by the other firms decreases at lower prices, so that the dominant firm's demand curve is related to the market demand curve as shown in Figure 9.15. Based on this demand curve (D_{DF}) and its associated marginal revenue (MR_{DF}) curve, the firm will maximize profits at a price of P^* . The competitive firms maximize profits by producing the quantity for which their marginal cost (MC_{CF}) equals P^* , quantity Q_{CF} .

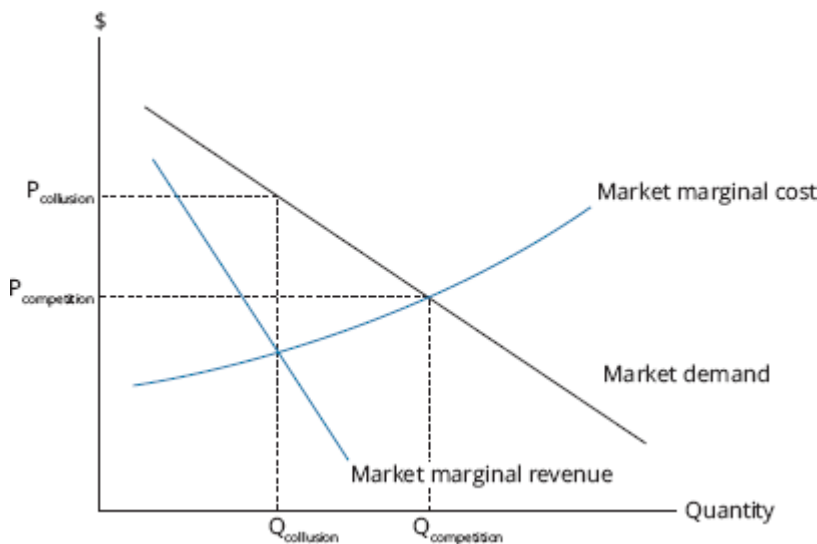
Figure 9.15: Dominant Firm Oligopoly



A price decrease by one of the competitive firms, which increases Q_{CF} in the short run, will lead to a decrease in price by the dominant firm, and competitive firms will decrease output and/or exit the industry in the long run. The long-run result of such a price decrease by competitors below P^* would then be to decrease the overall market share of competitor firms and increase the market share of the dominant firm.

Clearly, there are many possible outcomes in oligopoly markets that depend on the characteristics of the firms and the market itself. The important point is that the firms' decisions are interdependent so that the expected reaction of other firms is an important consideration. Overall, the resulting price will be somewhere between the price based on perfect collusion that would maximize total profits to all firms in the market (actually the monopoly price, which is addressed next) and the price that would result from perfect competition and generate zero economic profits in the long run. These two limiting outcomes are illustrated in Figure 9.16 as $P_{\text{collusion}}$ with $Q_{\text{collusion}}$ for perfect collusion and $P_{\text{competition}}$ and $Q_{\text{competition}}$ for perfect competition.

Figure 9.16: Collusion vs. Perfect Competition



MODULE QUIZ 9.3

- An oligopolistic industry has:
 - few barriers to entry.
 - few economies of scale.
 - a great deal of interdependence among firms.
- Consider a firm in an oligopoly market that believes the demand curve for its product is more elastic above a certain price than below this price. This belief fits *most closely* to which of the following models?
 - Cournot model.
 - Dominant firm model.
 - Kinked demand model.
- Consider an agreement between France and Germany that will restrict wine production so that maximum economic profit can be realized. The possible outcomes of the agreement are presented in the table below.

	Germany complies	Germany defaults
France complies	France gets €8 billion Germany gets €8 billion	France gets €2 billion Germany gets €10 billion
France defaults	France gets €10 billion Germany gets €2 billion	France gets €4 billion Germany gets €4 billion

Based on the concept of a Nash equilibrium, the *most likely* strategy followed by the two countries with respect to whether they comply with or default on the agreement will be:

- A. both countries will default.
- B. both countries will comply.
- C. one country will default and the other will comply.

MODULE 9.4: MONOPOLY AND CONCENTRATION



Video covering this content is available online.

A monopoly faces a downward-sloping demand curve for its product, so profit maximization involves a trade-off between price and quantity sold if the firm sells at the same price to all buyers. Assuming a single selling price, a monopoly firm must lower its price in order to sell a greater quantity. Unlike a firm in perfect competition, a firm facing a downward-sloping demand curve must determine what price to charge, hoping to find the price and output combination that will bring the maximum profit to the firm.

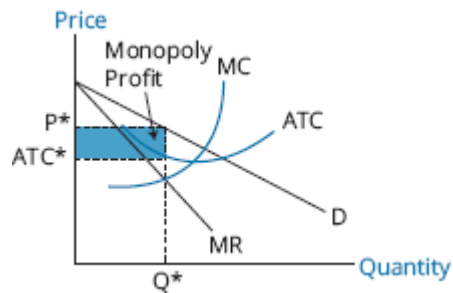
Two pricing strategies that are possible for a monopoly firm are *single-price* and *price discrimination*. If the monopoly's customers cannot resell the product to each other, the monopoly can maximize profits by charging different prices to different groups of customers. When price discrimination isn't possible, the monopoly will charge a single price. Price discrimination is described in more detail after we address single-price profit maximization.

To maximize profit, monopolists will expand output until marginal revenue (MR) equals marginal cost (MC). Due to high entry barriers, monopolist profits do not attract new market entrants. Therefore, long-run positive economic profits can exist. Do monopolists charge the highest possible price? The answer is no, because monopolists want to maximize profits, not price.

One way to calculate marginal revenue for a firm that faces a downward-sloping demand curve and sells all units for the same price is $MR = P\left(1 - \frac{1}{E_p}\right)$, where MR is marginal revenue, P is the current price, and E_p is the absolute value of the price elasticity of demand at price = P. Therefore, we can also express the single-price profit-maximizing output as that output for which $MC = P\left(1 - \frac{1}{E_p}\right)$.

Figure 9.17 shows the revenue-cost structure facing the monopolist. Note that production will expand until $MR = MC$ at optimal output Q^* . To find the price at which it will sell Q^* units, you must go to the demand curve. The demand curve itself does not determine the optimal behavior of the monopolist. Just like the perfect competition model, the profit maximizing output for a monopolist is where $MR = MC$. To ensure a profit, the demand curve must lie above the firm's average total cost (ATC) curve at the optimal quantity so that price > ATC. The optimal quantity will be in the elastic range of the demand curve.

Figure 9.17: Monopoly Short-Run Costs and Revenues



Once again, the *profit maximizing* output for a monopolistic firm is the one for which $MR = MC$. As shown in Figure 9.17, the profit maximizing output is Q^* , with a price of P^* , and an economic profit equal to $(P^* - ATC^*) \times Q^*$.

Monopolists are *price searchers* and have *imperfect information* regarding market demand. They must experiment with different prices to find the one that maximizes profit.

Price discrimination is the practice of charging different consumers different prices for the same product or service. Examples are different prices for airline tickets based on whether a Saturday-night stay is involved (separates business travelers and leisure travelers) and different prices for movie tickets based on age.

The motivation for a monopolist is to capture more consumer surplus as economic profit than is possible by charging a single price.

For price discrimination to work, the seller must:

- Face a downward-sloping demand curve.
- Have at least two identifiable groups of customers with *different price elasticities of demand* for the product.
- Be able to prevent the customers paying the lower price from reselling the product to the customers paying the higher price.

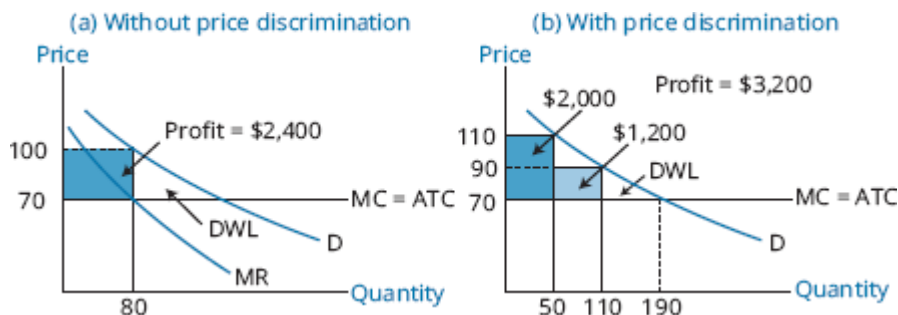
As long as these conditions are met, firm profits can be increased through price discrimination.

Figure 9.18 illustrates how price discrimination can increase the total quantity supplied and increase economic profits compared to a single-price pricing strategy. For simplicity, we have assumed no fixed costs and constant variable costs so that $MC = ATC$. In Panel (a), the single profit-maximizing price is \$100 at a quantity of 80 (where $MC = MR$), which generates a profit of \$2,400. In Panel (b), the firm is able to separate consumers, charges one group \$110 and sells them 50 units, and sells an additional 60 units to another group (with more elastic demand) at a price of \$90. Total profit is increased to \$3,200, and total output is increased from 80 units to 110 units.

Compared to the quantity produced under perfect competition, the quantity produced by a monopolist reduces the sum of consumer and producer surplus by an amount represented by the triangle labeled *deadweight loss* (DWL) in Panel (a) of Figure 9.18. Consumer surplus is reduced not only by the decrease in quantity but also by the increase in price relative to perfect competition. Monopoly is considered inefficient because the reduction in output compared to perfect competition reduces the sum of consumer and producer surplus. Because marginal benefit is greater than marginal cost, less than the efficient quantity of resources are allocated to the production of the good. Price discrimination reduces this inefficiency by increasing

output toward the quantity where marginal benefit equals marginal cost. Note that the deadweight loss is smaller in Panel (b). The firm gains from those customers with inelastic demand while still providing goods to customers with more elastic demand. This may even cause production to take place when it would not otherwise.

Figure 9.18: Effect of Price Discrimination on Output and Operating Profit



An extreme (and largely theoretical) case of price discrimination is perfect price discrimination. If it were possible for the monopolist to charge each consumer the maximum they are willing to pay for each unit, there would be no deadweight loss because a monopolist would produce the same quantity as under perfect competition. With perfect price discrimination, there would be no consumer surplus. It would all be captured by the monopolist.

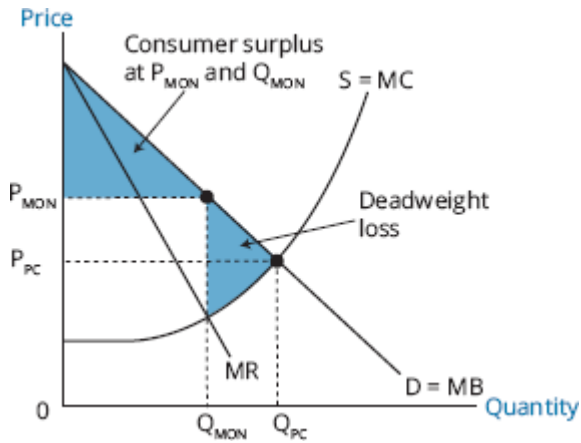
Figure 9.19 illustrates the difference in allocative efficiency between monopoly and perfect competition. Under perfect competition, the industry supply curve, S , is the sum of the supply curves of the many competing firms in the industry. The perfect competition equilibrium price and quantity are at the intersection of the industry supply curve and the market demand curve, D . The quantity produced is Q_{PC} at an equilibrium price P_{PC} . Because each firm is small relative to the industry, there is nothing to be gained by attempting to decrease output in an effort to increase price.

A monopolist facing the same demand curve, and with the same marginal cost curve, MC , will maximize profit by producing Q_{MON} (where $MC = MR$) and charging a price of P_{MON} .

The important thing to note here is that when compared to a perfectly competitive industry, the monopoly firm will produce less total output and charge a higher price.

Recall from our review of perfect competition that the efficient quantity is the one for which the sum of consumer surplus and producer surplus is maximized. In Figure 9.19, this quantity is where $S = D$, or equivalently, where marginal cost (MC) = marginal benefit (MB). *Monopoly creates a deadweight loss* relative to perfect competition because monopolies produce a quantity that does not maximize the sum of consumer surplus and producer surplus. A further loss of efficiency results from **rent seeking** when producers spend time and resources to try to acquire or establish a monopoly.

Figure 9.19: Perfect Competition vs. Monopoly

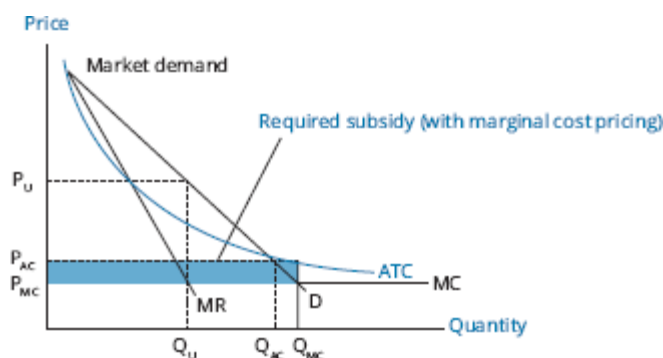


Natural Monopoly

In some industries, the economics of production lead to a single firm supplying the entire market demand for the product. When there are large economies of scale, it means that the average cost of production decreases as a single firm produces greater and greater output. An example is an electric utility. The fixed costs of producing electricity and building the power lines and related equipment to deliver it to homes are quite high. The marginal cost of providing electricity to an additional home or of providing more electricity to a home is, however, quite low. The more electricity provided, the lower the average cost per kilowatt hour. When the average cost of production for a single firm is falling throughout the relevant range of consumer demand, we say that the industry is a **natural monopoly**. The entry of another firm into the industry would divide the production between two firms and result in a higher average cost of production than for a single producer. Thus, large economies of scale in an industry present significant barriers to entry.

We illustrate the case of a natural monopoly in Figure 9.20. Left unregulated, a single-price monopolist will maximize profits by producing where $MR = MC$, producing quantity Q_U and charging P_U . Given the economies of scale, having another firm in the market would increase the ATC significantly. Note in Figure 9.20 that if two firms each produced approximately one-half of output Q_{AC} , average cost for each firm would be much higher than for a single producer producing Q_{AC} . Thus, there is a potential gain from monopoly because of lower average cost production when LRAC is decreasing so that economies of scale lead to a single supplier.

Figure 9.20: Natural Monopoly—Average Cost and Marginal Cost Pricing



Regulators often attempt to increase competition and efficiency through efforts to reduce artificial barriers to trade, such as licensing requirements, quotas, and tariffs.

Because monopolists produce less than the optimal quantity (do not achieve efficient resource allocation), government regulation may be aimed at improving resource allocation by regulating the prices monopolies may charge. This may be done through average cost pricing or marginal cost pricing.

Average cost pricing is the most common form of regulation. This would result in a price of P_{AC} and an output of Q_{AC} as illustrated in Figure 9.20. It forces monopolists to reduce price to where the firm's ATC intersects the market demand curve. This will:

- Increase output and decrease price.
- Increase social welfare (allocative efficiency).
- Ensure the monopolist a *normal* profit because price = ATC.

Marginal cost pricing, which is also referred to as *efficient regulation*, forces the monopolist to reduce price to the point where the firm's MC curve intersects the market demand curve. This increases output and reduces price, but causes the monopolist to incur a loss because price is below ATC, as illustrated in Figure 9.20. Such a solution requires a government subsidy in order to provide the firm with a normal profit and prevent it from leaving the market entirely.

Another way of regulating a monopoly is for the government to sell the monopoly right to the highest bidder. The right to build a gasoline station and food court on a tollway is one example. In theory, the winning bidder will be an efficient supplier that bids an amount equal to the value of expected economic profit and sets prices equal to long-run average cost.

LOS 9.c: Describe a firm's supply function under each market structure.

The short-run supply function for a firm under perfect competition is its marginal cost curve above its average variable cost curve, as described earlier. The short-run market supply curve is constructed simply by summing the quantities supplied at each price across all firms in the market.

In markets characterized as monopolistic competition, oligopoly, and monopoly, there is no well-defined supply function. This is because under all three of these market structures, firms face downward-sloping demand curves. In each case, the quantity supplied is determined by the intersection of marginal cost and marginal revenue, and the price charged is then determined by the demand curve the firm faces. We cannot construct a function of quantity supplied as a function of price as we can under perfect competition, where price equals marginal revenue. The quantity supplied depends not only on a firm's marginal cost, but on demand and marginal revenue (which change with quantity) as well.

LOS 9.e: Describe pricing strategy under each market structure.

We have covered each market structure separately in detail, so we will simply summarize optimal pricing strategies.

Perfect competition: Profits are maximized by producing the quantity for which marginal cost equals marginal revenue. Note that marginal revenue and price are equal so price also equals marginal cost at the profit-maximizing quantity.

Monopoly: Profits are also maximized by producing the quantity for which marginal revenue equals marginal cost. Because the firm's demand curve is downward sloping, price is greater than marginal revenue and greater than marginal cost.

Monopolistic competition: Profits are maximized when a firm produces the quantity for which marginal revenue equals marginal cost. Similar to a monopoly structure, the firm faces a downward sloping demand curve and price will be greater than marginal cost and marginal revenue.

Oligopoly: Because one of the key characteristics of oligopoly is the interdependence of firms' pricing and output decisions, the optimal pricing strategy depends on our assumptions about the reactions of other firms to each firm's actions. Here we note different possible assumptions and the strategy that is implied by each.

1. Kinked demand curve: This assumes competitors will match a price decrease but not a price increase. Firms produce the quantity for which marginal revenue equals marginal cost. However, the marginal revenue curve is discontinuous (there's a gap in it), so for many cost structures the optimal quantity is the same, given they face the same kinked demand curve.
2. Collusion: If all producers agree to share the market to maximize total industry profits, they will produce a total quantity for which marginal cost equals marginal revenue and charge the price from the industry demand curve at which that quantity can be sold. This is the same overall price and quantity as for a profit maximizing monopoly firm, but the oligopoly firms must agree to share this total output among themselves and share the economic profits as a result.
3. Dominant firm model: In this case, we assume one firm has the lowest cost structure and a large market share as a result. The dominant firm will maximize profits by producing the quantity for which its marginal cost equals its marginal revenue and charge the price on its firm demand curve for that quantity. Other firms in the market will essentially take that price as given and produce the quantity for which their marginal cost equals that price.
4. Game theory: Because of the interdependence of oligopoly firms' decisions, assumptions about how a competitor will react to a particular price and output decision by a competitor can determine the optimal output and pricing strategy. Given the variety of models and assumptions about competitor reactions, the long-run outcome is indeterminate. We can only say that the price will be between the monopoly price (if firms successfully collude) and the perfect competition price which equals marginal cost (if potential competition rules out prices above that level).

LOS 9.g: Describe the use and limitations of concentration measures in identifying market structure.

When examining the pricing power of firms in an industry, we would like to be able to measure elasticity of demand directly, but that is very difficult. Regulators often use percentage of

market sales (market share) to measure the degree of monopoly or market power of a firm. Often, mergers or acquisitions of companies in the same industry or market are not permitted by government authorities when they determine the market share of the combined firms will be too high and, therefore, detrimental to the economy.

Rather than estimate elasticity of demand, **concentration measures** for a market or industry are very often used as an indicator of market power. One concentration measure is the **N-firm concentration ratio**, which is calculated as the sum or the percentage market shares of the largest N firms in a market. While this measure is simple to calculate and understand, it does not directly measure market power or elasticity of demand.

One limitation of the N-firm concentration ratio is that it may be relatively insensitive to mergers of two firms with large market shares. This problem is reduced by using an alternative measure of market concentration, the **Herfindahl-Hirschman Index (HHI)**. The HHI is calculated as the sum of the squares of the market shares of the largest firms in the market. The following example illustrates this difference between the two measures and their calculation.

EXAMPLE: 4-firm concentration ratios

Given the market shares of the following firms, calculate the 4-firm concentration ratio and the 4-firm HHI, both before and after a merger of Acme and Blake.

Firm	Sales/Total Market Sales
Acme	25%
Blake	15%
Curtis	15%
Dent	10%
Erie	5%
Federal	5%

Answer:

Prior to the merger, the 4-firm concentration ratio for the market is $25 + 15 + 15 + 10 = 65\%$. After the merger, the Acme + Blake firm has 40% of the market, and the 4-firm concentration ratio is $40 + 15 + 10 + 5 = 70\%$. Although the 4-firm concentration ratio has only increased slightly, the market power of the largest firm in the industry has increased significantly from 25% to 40%.

Prior to the merger, the 4-firm HHI is $0.25^2 + 0.15^2 + 0.15^2 + 0.10^2 = 0.1175$.

After the merger, the 4-firm HHI is $0.40^2 + 0.15^2 + 0.10^2 + 0.05^2 = 0.1950$, a significant increase.

A second limitation that applies to both of our simple concentration measures is that barriers to entry are not considered in either case. Even a firm with high market share may not have much pricing power if barriers to entry are low and there is *potential competition*. With low barriers to entry, it may be the case that other firms stand ready to enter the market if firms currently in the market attempt to increase prices significantly. In this case, the elasticity of demand for

existing firms may be high even though they have relatively high market shares and industry concentration measures.

LOS 9.h: Identify the type of market structure within which a firm operates.

The identification of the type of market structure within which a firm is operating is based on the characteristics we outlined earlier. Our earlier table is repeated here in Figure 9.21. Because the analyst is attempting to determine the degree of pricing power firms in the industry have, the focus is on number of firms in the industry, the importance of barriers to entry, the nature of substitute products, and the nature of industry competition. Significant interdependence among firm pricing and output decisions is always a characteristic of an oligopoly market, although some interdependence is present under monopolistic competition, even with many more firms than for an oligopoly structure.

The following table illustrates the differences in characteristics among the various market structures.

Figure 9.21: Characteristics of Market Structures

	Perfect Competition	Monopolistic Competition	Oligopoly	Monopoly
Number of sellers	Many firms	Many firms	Few firms	Single firm
Barriers to entry	Very low	Low	High	Very high
Nature of substitute products	Very good substitutes	Good substitutes but differentiated	Very good substitutes or differentiated	No good substitutes
Nature of competition	Price only	Price, marketing, features	Price, marketing, features	Advertising
Pricing power	None	Some	Some to significant	Significant



MODULE QUIZ 9.4

1. Which of the following statements *most accurately* describes a significant difference between a monopoly firm and a perfectly competitive firm? A perfectly competitive firm:
 - A. minimizes costs; a monopolistic firm maximizes profit.
 - B. maximizes profit; a monopolistic firm maximizes price.
 - C. takes price as given; a monopolistic firm must search for the best price.
2. A monopolist will expand production until $MR = MC$ and charge a price determined by:
 - A. the demand curve.
 - B. the marginal cost curve.
 - C. the average total cost curve.
3. When a regulatory agency requires a monopolist to use average cost pricing, the intent is to price the product where:
 - A. the ATC curve intersects the MR curve.
 - B. the MR curve intersects the demand curve.
 - C. the ATC curve intersects the demand curve.

4. Which of the following is *most likely* an advantage of the Herfindahl-Hirschman Index relative to the N -firm concentration ratio? The Herfindahl-Hirschman Index:
 - A. is simpler to calculate.
 - B. considers barriers to entry.
 - C. is more sensitive to mergers.
5. A market characterized by low barriers to entry, good substitutes, limited pricing power, and marketing of product features is *best* characterized as:
 - A. oligopoly.
 - B. perfect competition.
 - C. monopolistic competition.

KEY CONCEPTS

LOS 9.a

Perfect competition is characterized by:

- Many firms, each small relative to the market.
- Very low barriers to entry into or exit from the industry.
- Homogeneous products that are perfect substitutes, no advertising or branding.
- No pricing power.

Monopolistic competition is characterized by:

- Many firms.
- Low barriers to entry into or exit from the industry.
- Differentiated products, heavy advertising and marketing expenditure.
- Some pricing power.

Oligopoly markets are characterized by:

- Few sellers.
- High barriers to entry into or exit from the industry.
- Products that may be homogeneous or differentiated by branding and advertising.
- Firms that may have significant pricing power.

Monopoly is characterized by:

- A single firm that comprises the whole market.
- Very high barriers to entry into or exit from the industry.
- Advertising used to compete with substitute products.
- Significant pricing power.

LOS 9.b

Perfect competition:

- Price = marginal revenue = marginal cost (in equilibrium).
- Perfectly elastic demand, zero economic profit in equilibrium.

Monopolistic competition:

- Price > marginal revenue = marginal cost (in equilibrium).

- Zero economic profit in long-run equilibrium.

Oligopoly:

- Price > marginal revenue = marginal cost (in equilibrium).
- May have positive economic profit in long-run equilibrium, but moves toward zero economic profit over time.

Monopoly:

- Price > marginal revenue = marginal cost (in equilibrium).
- May have positive economic profit in long-run equilibrium, profits may be zero because of expenditures to preserve monopoly.

LOS 9.c

Under perfect competition, a firm's short-run supply curve is the portion of the firm's short-run marginal cost curve above average variable cost. A firm's long-run supply curve is the portion of the firm's long-run marginal cost curve above average total cost.

Firms operating under monopolistic competition, oligopoly, and monopoly do not have well-defined supply functions, so neither marginal cost curves nor average cost curves are supply curves in these cases.

LOS 9.d

All firms maximize profits by producing the quantity of output for which marginal cost equals marginal revenue. Under perfect competition (perfectly elastic demand), marginal revenue also equals price.

Firms in monopolistic competition or that operate in oligopoly or monopoly markets all face downward-sloping demand curves. Selling price is determined from the price on the demand curve for the profit maximizing quantity of output.

LOS 9.e

Whether a firm operates in perfect competition, monopolistic competition, or is a monopoly, profits are maximized by producing and selling the quantity for which marginal revenue equals marginal cost. Under perfect competition, price equals marginal revenue. Under monopolistic competition or monopoly, firms face downward-sloping demand curves so that marginal revenue is less than price, and the price charged at the profit-maximizing quantity is the price from the firm's demand curve at the optimal (profit-maximizing) level of output.

Under oligopoly, the pricing strategy is not clear. Because firm decisions are interdependent, the optimal pricing and output strategy depends on the assumptions made about other firms' cost structures and about competitors' responses to a firm's price changes.

LOS 9.f

An increase (decrease) in demand will increase (decrease) economic profits in the short run under all market structures. Positive economic profits result in entry of firms into the industry unless barriers to entry are high. Negative economic profits result in exit of firms from the industry unless barriers to exit are high. When firms enter (exit) an industry, market supply increases (decreases), resulting in a decrease (increase) in market price and an increase (decrease) in the equilibrium quantity traded in the market.

LOS 9.g

A concentration ratio for N firms is calculated as the percentage of market sales accounted for by the N largest firms in the industry and is used as a simple measure of market structure and market power.

The Herfindahl-Hirschman Index measure of concentration is calculated as the sum of the squared market shares of the largest N firms in an industry and better reflects the effect of mergers on industry concentration.

Neither measure actually measures market power directly. Both can be misleading measures of market power when potential competition restricts pricing power.

LOS 9.h

To identify the market structure in which a firm is operating, we need to examine the number of firms in its industry, whether products are differentiated or other types of non-price competition exist, and barriers to entry, and compare these to the characteristics that define each market structure.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 9.1

1. **A** When a firm operates under conditions of pure competition, MR always equals price. This is because, in pure competition, demand is perfectly elastic (a horizontal line), so MR is constant and equal to price. (LOS 9.a)
2. **B** The supply function is not well-defined in markets other than those that can be characterized as perfect competition. (LOS 9.c)
3. **A** In a purely competitive market, economic losses indicate that firms are overproducing, causing prices to fall below average total costs. This can occur in the short run. In the long run, however, market supply will decrease as firms exit the industry, and prices will rise to the point where economic profits are zero. (LOS 9.b)
4. **C** A purely competitive firm will tend to expand its output so long as the market price is greater than MC. In the short run and long run, profit is maximized when $P = MC$. (LOS 9.d)
5. **C** If price is greater than average variable cost, a firm will continue to operate in the short run because it is covering at least some of its fixed costs. (LOS 9.d)

Module Quiz 9.2

1. **B** The demand for products from firms competing in monopolistic competition is relatively elastic due to the availability of many close substitutes. If a firm increases its product price, it will lose customers to firms selling substitute products at lower prices. (LOS 9.b)
2. **B** Monopolistic competition is likely to result in a higher price and lower quantity of output compared to perfect competition. (LOS 9.d)
3. **C** The profit-maximizing output is the quantity at which marginal revenue equals marginal cost. In a price-searcher industry structure (i.e., any structure that is not perfect competition), price is greater than marginal revenue. (LOS 9.d, 9.e, 9.f)

Module Quiz 9.3

1. **C** An oligopolistic industry has a great deal of interdependence among firms. One firm's pricing decisions or advertising activities will affect the other firms. (LOS 9.a)
2. **C** The kinked demand model assumes that each firm in a market believes that at some price, demand is more elastic for a price increase than for a price decrease. (LOS 9.b)
3. **A** The Nash equilibrium results when each nation pursues the strategy that is best, given the strategy that is pursued by the other nation.
 - Given that Germany complies with the agreement: France will get €8 billion if it complies, but €10 billion if it defaults. Therefore, France should default.
 - Given that Germany defaults: France will get €2 billion if it complies, but €4 billion if it defaults. Therefore, France should default.
 - Because France is better off in either case by defaulting, France will default.
 - Germany will follow the same logic and reach the same conclusion.(LOS 9.f)

Module Quiz 9.4

1. **C** Monopolists must search for the profit maximizing price (and output) because they do not have perfect information regarding demand. Firms under perfect competition take the market price as given and only determine the profit maximizing quantity. (LOS 9.b)
2. **A** A monopolist will expand production until $MR = MC$, and the price of the product will be determined by the demand curve. (LOS 9.d)
3. **C** When a regulatory agency requires a monopolist to use average cost pricing, the intent is to price the product where the ATC curve intersects the market demand curve. A problem in using this method is actually determining exactly what the ATC is. (LOS 9.f)
4. **C** Although the N -firm concentration ratio is simple to calculate, it can be relatively insensitive to mergers between companies with large market shares. Neither the HHI nor the N -firm concentration ratio consider barriers to entry. (LOS 9.g)
5. **C** These characteristics are associated with a market structure of monopolistic competition. Firms in perfect competition do not compete on product features. Oligopolistic markets have high barriers to entry. (LOS 9.h)

READING 10

AGGREGATE OUTPUT, PRICES, AND ECONOMIC GROWTH

EXAM FOCUS

This reading introduces macroeconomics and the measurement of aggregate economic output. The crucial concepts to grasp here are aggregate demand, short-run aggregate supply, and long-run aggregate supply. Know the factors that cause the aggregate demand and supply curves to shift and the sources of long-run economic growth. Understand the various measures of aggregate income (nominal and real GDP, national income, personal income, and personal disposable income). The interaction among saving, investment, the fiscal balance, and the trade balance will be built on in later readings on international trade and foreign exchange.

MODULE 10.1: GDP, INCOME, AND EXPENDITURES



Video covering this content is available online.

LOS 10.a: Calculate and explain gross domestic product (GDP) using expenditure and income approaches.

Gross domestic product (GDP) is the total market value of the goods and services produced in a country within a certain time period. GDP is the most widely used measure of the size of a nation's economy. GDP includes only purchases of newly produced goods and services. The sale or resale of goods produced in previous periods is excluded. Transfer payments made by the government (e.g., unemployment, retirement, and welfare benefits) are not economic output and are not included in the calculation of GDP.

The values used in calculating GDP are *market values of final goods* and services—that is, goods and services that will not be resold or used in the production of other goods and services. The value of the computer chips that Intel makes is not explicitly included in GDP; their value is included in the final prices of computers that use the chips. The value of a Rembrandt painting that sells for 10 million euros is not included in the calculation of GDP, as it was not produced during the period.

Goods and services provided by government are included in GDP even though they are not explicitly priced in markets. For example, the services provided by police and the judiciary, and goods such as roads and infrastructure improvements, are included. Because these goods and services are not sold at market prices, they are valued at their cost to the government.

GDP also includes the value of owner-occupied housing, just as it includes the value of rental housing services. Because the value of owner-occupied housing is not revealed in market transactions, the value is estimated for inclusion in GDP. The value of labor not sold, such as a homeowner's repairs to his own home, is not included in GDP. By-products of production, such as environmental damage, are not included in GDP.

GDP can be calculated as the sum of all the spending on newly produced goods and services, or as the sum of the income received as a result of producing these goods and services. Under the **expenditure approach**, GDP is calculated by summing the amounts spent on goods and services produced during the period. Under the **income approach**, GDP is calculated by summing the amounts earned by households and companies during the period, including wage income, interest income, and business profits.

For the whole economy, total expenditures and total income must be equal, so the two approaches should produce the same result. In practice, measurement issues result in different values under the two methods.

LOS 10.b: Compare the sum-of-value-added and value-of-final-output methods of calculating GDP.

So far, we have described the calculation of GDP under the expenditure approach as summing the values of all final goods and services produced. This expenditure method is termed the **value-of-final-output method**.

Under the **sum-of-value-added method**, GDP is calculated by summing the additions to value created at each stage of production and distribution. An example of the calculation for a specific product is presented in Figure 10.1.

Figure 10.1: Value Added at Stages of Production

Stage of Production	Sales Value (\$)	Value Added (\$)
Raw materials/components	\$100	\$100
Manufacturing	\$350	\$250
Retail	\$400	\$50
Sum of value added		\$400

The intuition is clear. The prices of final goods and services include, and are equal to, the additions to value at each stage of production (e.g., from mining iron ore and making steel to assembling an automobile that contains machined steel parts).

LOS 10.c: Compare nominal and real GDP and calculate and interpret the GDP deflator.

Nominal GDP is simply GDP as we have described it under the expenditures approach: the total value of all goods and services produced by an economy, valued at current market prices. For an economy with N different goods and services, we can express nominal GDP as:

$$\begin{aligned} \text{nominal GDP}_t \text{ for year } t &= \sum_{i=1}^N P_{i,t} Q_{i,t} \\ &= \sum_{i=1}^N (\text{price of good } i \text{ in year } t) \\ &\quad \times (\text{quantity of good } i \text{ produced in year } t) \end{aligned}$$

Because nominal GDP is based on current prices, inflation will increase nominal GDP even if the physical output of goods and services remains constant from one year to the next. **Real GDP** measures the output of the economy using prices from a base year, removing the effect of changes in prices so that inflation is not counted as economic growth.

Real GDP is calculated relative to a *base year*. By using base-year prices and current-year output quantities, real GDP growth reflects only increases in total output, not simply increases (or decreases) in the money value of total output.

Assuming the base year prices are those for five years ago, real GDP can be calculated as:

$$\begin{aligned} \text{real GDP for year } t &= \sum_{i=1}^N P_{i,t-5} Q_{i,t} \\ &= \sum_{i=1}^N (\text{price of good } i \text{ in year } t - 5) \\ &\quad \times (\text{quantity of good } i \text{ produced in year } t) \end{aligned}$$

The **GDP deflator** is a price index that can be used to convert nominal GDP into real GDP, taking out the effects of changes in the overall price level. The GDP deflator is based on the current mix of goods and services, using prices at the beginning and end of the period. The GDP deflator is calculated as:

$$\begin{aligned} \text{GDP deflator for year } t &= \frac{\sum_{i=1}^N P_{i,t} Q_{i,t}}{\sum_{i=1}^N P_{i, \text{base year}} Q_{i,t}} \times 100 \\ &= \frac{\text{nominal GDP in year } t}{\text{value of year } t \text{ output at base year prices}} \times 100 \end{aligned}$$

Per-capita real GDP is defined as real GDP divided by population and is often used as a measure of the economic well-being of a country's residents.

EXAMPLE: Calculating and using the GDP deflator

1. GDP in 20X2 is \$1.80 billion at 20X2 prices and \$1.65 billion when calculated using 20X1 prices. Calculate the GDP deflator using 20X1 as the base period.
2. Nominal GDP was \$213 billion in 20X6 and \$150 billion in 20X1. The 20X6 GDP deflator relative to the base year 20X1 is 122.3. Calculate real GDP for 20X6 and the compound annual real growth rate of economic output from 20X1 to 20X6.

Answer:

1. GDP deflator = $1.80 / 1.65 \times 100 = 109.1$, reflecting a 9.1% increase in the price level.
2. Real GDP 20X6 = $\$213 / 1.223 = \174.16 .

Noting that real and nominal GDP are the same for the base year, the compound real annual growth rate of economic output over the 5-year period is:

$$\left(\frac{174.16}{150}\right)^{\frac{1}{5}} - 1 = 3.03\%$$

LOS 10.d: Compare GDP, national income, personal income, and personal disposable income.

Using the expenditure approach, the major components of real GDP are consumption, investment, government spending, and **net exports** (exports minus imports). These components are summarized in the equation:

$$\text{GDP} = C + I + G + (X - M)$$

where:

C = consumption spending

I = business investment (capital equipment, inventories)

G = government purchases

X = exports

M = imports

We may also express this equation as:

$$\text{GDP} = (C + G^C) + (I + G^I) + (X - M)$$

where:

G^C = government consumption

G^I = government investment (capital goods, inventories)

Under the income approach, we have the following equation for GDP, or **gross domestic income (GDI)**:

$$\text{GDP} = \text{national income} + \text{capital consumption allowance} + \text{statistical discrepancy}$$

A **capital consumption allowance (CCA)** measures the depreciation (i.e., wear) of physical capital from the production of goods and services over a period. CCA can be thought of as the amount that would have to be reinvested to maintain the productivity of physical capital from one period to the next. The *statistical discrepancy* is an adjustment for the difference between GDP measured under the income approach and the expenditure approach because they use different data.

National income is the sum of the income received by all factors of production that go into the creation of final output:

$$\begin{aligned} \text{national income} &= \text{compensation of employees (wages and benefits)} \\ &+ \text{corporate and government enterprise profits before taxes} \\ &+ \text{interest income} \\ &+ \text{unincorporated business net income (business owners' incomes)} \\ &+ \text{rent} \\ &+ \text{indirect business taxes - subsidies (taxes and subsidies that are included in final prices)} \end{aligned}$$



PROFESSOR'S NOTE

Candidates should be aware that different countries' economic reporting bureaus may use their own terminology. For example, Statistics Canada defines *gross domestic income* as "net domestic income + consumption of fixed capital + statistical discrepancy," where net domestic income is "compensation of employees + gross operating surplus + gross mixed income + taxes less subsidies on production + taxes less subsidies on products and imports."

Personal income is a measure of the pretax income received by households and is one determinant of consumer purchasing power and consumption. Personal income differs from national income in that personal income includes all income that households receive, including government transfer payments such as unemployment or disability benefits.

Household disposable income or **personal disposable income** is personal income after taxes. Disposable income measures the amount that households have available to either save or spend on goods and services and is an important economic indicator of the ability of consumers to spend and save.

LOS 10.e: Explain the fundamental relationship among saving, investment, the fiscal balance, and the trade balance.

To show how private savings are related to investment, the government sector, and foreign trade, we will combine the income and expenditure approaches to measuring GDP.

As we have seen, total expenditures can be stated as $GDP = C + I + G + (X - M)$. Total income, which must equal total expenditures, can be stated as:

$$GDP = C + S + T$$

where:

C = consumption spending

S = household and business savings

T = net taxes (taxes paid minus transfer payments received)

Because total income equals total expenditures, we have the equality:

$$C + I + G + (X - M) = C + S + T$$

Rearranging this equation and solving for S (household and business savings), we get the following fundamental relationship:

$$S = I + (G - T) + (X - M)$$

Note that $(G - T)$ is the **fiscal balance**, or the difference between government spending and tax receipts. Recall that $(X - M)$ is net exports, or the **trade balance**. This equation shows that private savings must equal private investment, plus government borrowing or minus government savings, and minus the trade deficit or plus the trade surplus.



PROFESSOR'S NOTE

In this equation and the ones we will derive from it, a positive value for $(G - T)$ is a government budget deficit and a negative value for $(G - T)$ is a budget surplus. On the other hand, a positive value for $(X - M)$ is a trade surplus and a negative value for $(X - M)$ is a trade deficit.

If we solve this equation for the fiscal balance, we get:

$$(G - T) = (S - I) - (X - M)$$

From this equation, we can see that a government deficit ($G - T > 0$) must be financed by some combination of a trade deficit ($X - M < 0$) or an excess of private saving over private investment ($S - I > 0$).



PROFESSOR'S NOTE

In the reading on International Trade and Capital Flows, we will see that a trade deficit (current account deficit) must be associated with an inflow of foreign investment (capital account surplus). So we can interpret this equation as saying a fiscal deficit must be financed by a combination of domestic and foreign capital.



MODULE QUIZ 10.1

1. The *least appropriate* approach to calculating a country's gross domestic product (GDP) is summing for a given time period:
 - A. the value of all purchases and sales that took place within the country.
 - B. the amount spent on final goods and services produced within the country.
 - C. the income generated in producing all final goods and services produced within the country.
2. Gross domestic product does not include the value of:
 - A. transfer payments.
 - B. government services.
 - C. owner-occupied housing.
3. When GDP is calculated by the sum-of-value-added method, what is the value of a manufactured product in GDP?
 - A. The sum of the product's value at each stage of production and distribution.
 - B. The sum of the increases in the product's value at each stage of production and distribution.
 - C. The product's retail price less the value added at each stage of production and distribution.
4. Real GDP is *best* described as the value of:
 - A. current output measured at current prices.
 - B. current output measured at base-year prices.
 - C. base-year output measured at current prices.
5. The GDP deflator is calculated as 100 times the ratio of:
 - A. nominal GDP to real GDP.
 - B. base year prices to current year prices.
 - C. current year nominal GDP to base year nominal GDP.
6. Which of the following measures of income is the sum of wages and benefits, pretax profits, interest income, owners' income from unincorporated businesses, rent, and taxes net of subsidies?
 - A. Personal income.
 - B. National income.
 - C. Disposable income.
7. If a government budget deficit increases, net exports must:
 - A. increase, or the excess of private saving over private investment must decrease.
 - B. decrease, or the excess of private saving over private investment must increase.

C. decrease, or the excess of private saving over private investment must decrease.

MODULE 10.2: AGGREGATE DEMAND AND SUPPLY



Video covering this content is available online.

LOS 10.f: Explain how the aggregate demand curve is generated.

The **aggregate demand curve (AD curve)** illustrates the negative relationship between the price level and the level of real output demanded by consumers, businesses, and government. Points on the AD are combinations of the price level and real output for which the following two conditions hold:

1. *The goods market is in equilibrium:* Aggregate income equals aggregate expenditure, as we saw earlier in the fundamental relationship among saving, investment, and the fiscal and trade balances.
2. *The money market is in equilibrium:* Individuals and businesses are willing to hold the real money supply (nominal money supply adjusted for the price level).

Three effects explain why the AD curve slopes downward: a wealth effect, an interest rate effect, and a real exchange rate effect.

The **wealth effect** results from how changes in the price level affect consumers' purchasing power. For any given amount of nominal wealth (wealth stated in currency units), an increase in the price level reduces the amount of goods and services that amount of nominal wealth will purchase. A decrease in the price level increases the purchasing power of existing nominal wealth so that consumers will demand more goods and services. (People feel richer with lower prices and increase their consumption of goods and services.)

The **interest rate effect** is related to the demand for money. When the price level increases, to purchase the same amount of real goods and services, consumers need to hold greater nominal money balances. However, as we are holding the nominal money supply constant, only an increase in interest rates will restore the equilibrium between the money supply and the nominal balances people desire to hold (interest is the opportunity cost of holding money rather than investing it in interest-bearing securities). Higher interest rates will tend to decrease both demand for consumption goods (especially goods that are typically purchased on credit, such as cars and appliances) and business investment demand (a higher cost of capital reduces the number of profitable investment projects, as we explain in detail in the Corporate Issuers topic area).

The **real exchange rate effect** refers to the effect of an increase in the domestic price level on the net exports ($X - M$) component of GDP. When the domestic price level increases relative to the price level in a foreign country, the real price (the amount of foreign goods they must give up) of the domestic country's goods increases for foreigners, which reduces demand for exports. At the same time, the real price of imported goods to domestic consumers and businesses falls, increasing the quantity of imports demanded by domestic consumers. Both the decrease in

exports and the increase in imports reduce net exports ($X - M$), reducing aggregate demand for real goods and services.

To summarize, the aggregate demand curve slopes downward (a lower price level is associated with higher output) because lower price levels increase real wealth, decrease equilibrium real interest rates, make domestic goods less expensive to foreign consumers, and make foreign goods more expensive to domestic consumers and businesses, all of which increase the quantity of domestic output demanded.

LOS 10.g: Explain the aggregate supply curve in the short run and long run.

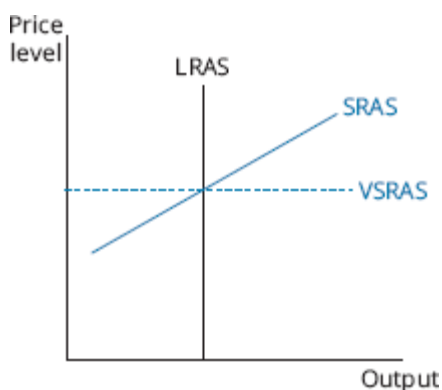
The Aggregate Supply Curve

The **aggregate supply (AS) curve** describes the relationship between the price level and the quantity of real GDP supplied, when all other factors are kept constant. That is, it represents the amount of output that firms will produce at different price levels.

We need to consider three aggregate supply curves with different time frames: the very short-run aggregate supply (VSRAS) curve, the short-run aggregate supply (SRAS) curve, and the long-run aggregate supply (LRAS) curve.

In the very short run, firms will adjust output without changing price by adjusting labor hours and intensity of use of plant and equipment in response to changes in demand. We represent this with the perfectly elastic very short run aggregate supply (VSRAS) curve in Figure 10.2.

Figure 10.2: Aggregate Supply Curves



In the short run, the SRAS curve slopes upward because some input prices will change as production is increased or decreased. We assume in the short run that *output prices* will change proportionally to the price level but that at least some *input prices* are sticky, meaning that they do not adjust to changes in the price level in the short run. When output prices increase, the price level increases, but firms see no change in input prices in the short run. Firms respond by increasing output in anticipation of greater profits from higher output prices. The result is an upward-sloping SRAS curve.

All input costs can vary in the long run, and the LRAS curve in Figure 10.2 is perfectly inelastic. In the long run, wages and other input prices change proportionally to the price level, so the

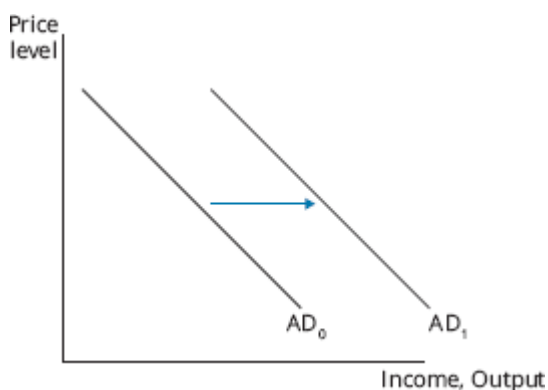
price level has no long-run effect on aggregate supply. We refer to this level of output as **potential GDP** or **full-employment GDP**.

LOS 10.h: Explain causes of movements along and shifts in aggregate demand and supply curves.

Shifts in the Aggregate Demand Curve

The aggregate demand (AD) curve reflects the total level of expenditures in an economy by consumers, businesses, governments, and foreigners. A number of factors can affect this level of expenditures and cause the AD curve to shift. Note that a *change in the price level* is represented as a *movement along the AD curve*, not a shift in the AD curve. In Figure 10.3, an increase in aggregate demand is shown by a shift to the right, indicating that the quantity of goods and services demanded is greater at any given price level.

Figure 10.3: Increase in Aggregate Demand



In trying to understand and remember the factors that affect aggregate demand, it may help to recall that, from the expenditure point of view, $GDP = C + I + G + \text{net } X$. For changes in each of the following factors that increase aggregate demand (shift AD to the right), we identify which component of expenditures is increased.

1. **Increase in consumers' wealth:** As the value of households' wealth increases (real estate, stocks, and other financial securities), the proportion of income saved decreases and spending increases, increasing aggregate demand (C increases).
2. **Business expectations:** When businesses are more optimistic about future sales, they tend to increase their investment in plant, equipment, and inventory, which increases aggregate demand (I increases).
3. **Consumer expectations of future income:** When consumers expect higher future incomes, due to a belief in greater job stability or expectations of rising wage income, they save less for the future and increase spending now, increasing aggregate demand (C increases).
4. **High capacity utilization:** When companies produce at a high percentage¹ of their capacity, they tend to invest in more plant and equipment, increasing aggregate demand (I increases).
5. **Expansionary monetary policy:** When the rate of growth of the money supply is increased, banks have more funds to lend, which puts downward pressure on interest rates. Lower

interest rates increase investment in plant and equipment because the cost of financing these investments declines. Lower interest rates and greater availability of credit will also increase consumers' spending on consumer durables (e.g., automobiles, large appliances) that are typically purchased on credit. Thus, the effect of expansionary monetary policy is to increase aggregate demand (C and I increase).

Note that if the economy is operating at potential GDP (LRAS) when the monetary expansion takes place, the increase in real output will be only for the short run. In the long run, subsequent increases in input prices decrease SRAS and return output to potential GDP.

6. **Expansionary fiscal policy:** Expansionary fiscal policy refers to a decreasing government budget surplus (or an increasing budget deficit) from decreasing taxes, increasing government expenditures, or both. A decrease in taxes increases disposable income and consumption, while an increase in government spending increases aggregate demand directly (C increases for tax cut, G increases for spending increase).



PROFESSOR'S NOTE

A complete analysis of monetary and fiscal policy as they relate to overall expenditures and GDP is presented in our reading on Monetary and Fiscal Policy.

7. **Exchange rates:** A decrease in the relative value of a country's currency will increase exports and decrease imports. Both of these effects tend to increase domestic aggregate demand (net X increases).



PROFESSOR'S NOTE

We will analyze the effect of exchange rates on exports and imports in our reading on Currency Exchange Rates.

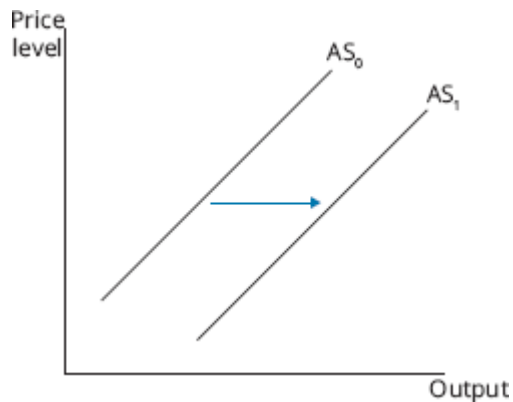
8. **Global economic growth:** GDP growth in foreign economies tends to increase the quantity of imports (domestic exports) foreigners demand. By increasing domestic export demand, this will increase aggregate demand (net X increases).

Note that for each factor, a change in the opposite direction will tend to decrease aggregate demand.

Shifts in the Short-Run Aggregate Supply Curve

The **short-run aggregate supply (SRAS) curve** reflects the relationship between output and the price level when wages and other input prices are held constant (or are slow to adjust to higher output prices). The curve shows the total level of output that businesses are willing to supply at different price levels. A number of factors can affect this level of output and cause the SRAS curve to shift. In Figure 10.4, an increase in aggregate supply is shown by a shift to the right, as the quantity supplied at each price level increases.

Figure 10.4: Increase in Aggregate Supply



In addition to changes in potential GDP (shifts in long-run aggregate supply), a number of factors can cause the SRAS curve to shift to the right:

1. **Labor productivity:** Holding the wage rate constant, an increase in labor productivity (output per hour worked) will decrease unit costs to producers. Producers will increase output as a result, increasing SRAS (shifting it to the right).
2. **Input prices:** A decrease in nominal wages or the prices of other important productive inputs will decrease production costs and cause firms to increase production, increasing SRAS. Wages are often the largest contributor to a producer's costs and have the greatest impact on SRAS.
3. **Expectations of future output prices:** When businesses expect the price of their output to increase in the future, they will expand production, increasing SRAS.
4. **Taxes and government subsidies:** Either a decrease in business taxes or an increase in government subsidies for a product will decrease the costs of production. Firms will increase output as a result, increasing SRAS.
5. **Exchange rates:** Appreciation of a country's currency in the foreign exchange market will decrease the cost of imports. To the extent that productive inputs are purchased from foreign countries, the resulting decrease in production costs will cause firms to increase output, increasing SRAS.

Again, an opposite change in any of these factors will tend to decrease SRAS.

Shifts in the Long-Run Aggregate Supply Curve

The **long-run aggregate supply (LRAS) curve** is vertical (perfectly inelastic) at the potential (full-employment) level of real GDP. Changes in factors that affect the real output that an economy can produce at full employment will shift the LRAS curve.

Factors that will shift the LRAS curve are:

1. **Increase in the supply and quality of labor:** Because LRAS reflects output at full employment, an increase in the labor force will increase full-employment output and the LRAS. An increase in the skills of the workforce, through training and education, will increase the productivity of a labor force of a given size, increasing potential real output and increasing LRAS.

2. **Increase in the supply of natural resources:** Just as with an increase in the labor force, increases in the available amounts of other important productive inputs will increase potential real GDP and LRAS.
3. **Increase in the stock of physical capital:** For a labor force of a given size, an increase in an economy's accumulated stock of capital equipment will increase potential output and LRAS.
4. **Technology:** In general, improvements in technology increase labor productivity (output per unit of labor) and thereby increase the real output that can be produced from a given amount of productive inputs, increasing LRAS.

Decreases in labor quality, labor supply, the supply of natural resources, or the stock of physical capital will all decrease LRAS (move the curve to the left). Technology does not really retreat, but a law prohibiting the use of an improved technology could decrease LRAS.

Movement Along Aggregate Demand and Supply Curves

In contrast with *shifts* in the aggregate demand and aggregate supply curves, *movements along* these curves reflect the impact of a change in the price level on the quantity demanded and the quantity supplied. Changes in the price level alone do not cause shifts in the AD and AS curves, although we have allowed that changes in expected future prices can.



MODULE QUIZ 10.2

1. The aggregate demand curve illustrates which of the following relationships?
 - A. Direct relationship between aggregate income and the price level.
 - B. Inverse relationship between aggregate income and the price level.
 - C. Direct relationship between aggregate income and the real interest rate.
2. An economy's potential output is *best* represented by:
 - A. long-run aggregate supply.
 - B. short-run aggregate supply.
 - C. long-run aggregate demand.
3. A stronger domestic currency relative to foreign currencies is *most likely* to result in:
 - A. a shift in the aggregate supply curve toward lower supply.
 - B. a shift in the aggregate demand curve toward lower demand.
 - C. a movement along the aggregate demand curve towards higher prices.
4. Which of the following factors would be *least likely* to shift the aggregate demand curve?
 - A. The price level increases.
 - B. The federal deficit expands.
 - C. Expected inflation decreases.

MODULE 10.3: MACROECONOMIC EQUILIBRIUM AND GROWTH



Video covering this content is available online.

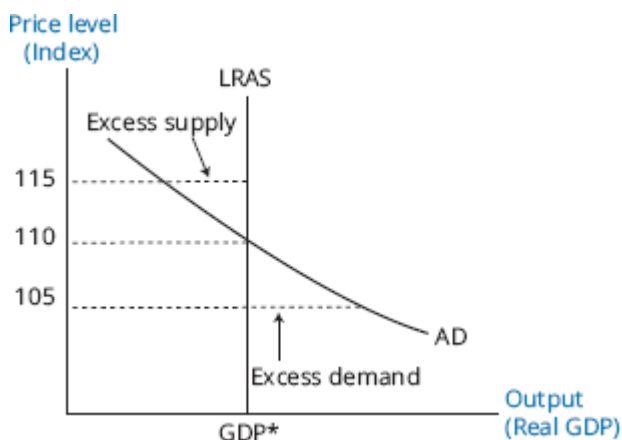
LOS 10.i: Describe how fluctuations in aggregate demand and aggregate supply cause short-run changes in the economy and the business cycle.

LOS 10.j: Distinguish among the following types of macroeconomic equilibria: long-run full employment, short-run recessionary gap, short-run inflationary gap, and short-run stagflation.

LOS 10.k: Explain how a short-run macroeconomic equilibrium may occur at a level above or below full employment.

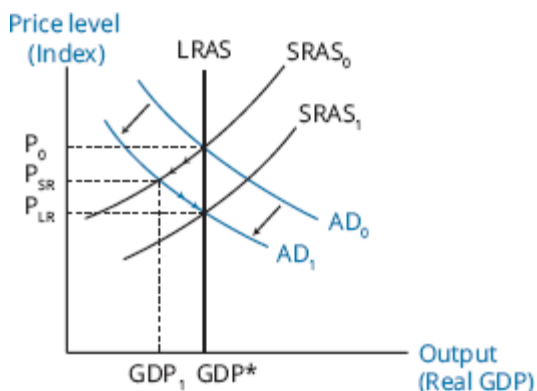
Having explained the factors that cause shifts in the aggregate demand and aggregate supply curves, we now turn our attention to the effects of fluctuations in aggregate supply and demand on real GDP and the business cycle. Our starting point is an economy that is in *long-run full-employment equilibrium*, as illustrated in Figure 10.5.

Figure 10.5: Long-Run Equilibrium Real Output



First consider a decrease in aggregate demand, which can result from a decrease in the growth rate of the money supply, an increase in taxes, a decrease in government spending, lower equity and house prices, or a decrease in the expectations of consumers and businesses for future economic growth. As illustrated in Figure 10.6, a decrease in aggregate demand will reduce both real output and the price level in the short run. The new short-run equilibrium output, GDP_1 , is less than full employment (potential) GDP. The decrease in aggregate demand has resulted in both lower real output and a lower price level.

Figure 10.6: Adjustment to a Decrease in Aggregate Demand



Because real GDP is less than full employment GDP, we say there is a **recessionary gap**. A recession is a period of declining GDP and rising unemployment. Classical economists believed that unemployment would drive down wages, as workers compete for available jobs, which in turn would increase SRAS and return the economy to its full employment level of real GDP. Keynesian economists, on the other hand, believe that this might be a slow and economically

painful process and that increasing aggregate demand through government action is the preferred alternative. Both expansionary fiscal policy (increasing government spending or decreasing taxes) and expansionary monetary policy (increasing the growth rate of the money supply to reduce interest rates) are methods to increase aggregate demand and return real GDP to its full employment (potential) level.

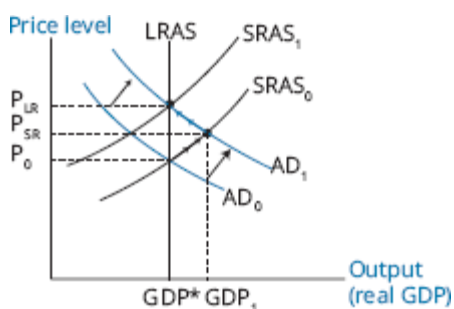


PROFESSOR'S NOTE

We will describe Classical, Keynesian, and other business cycle theories in the reading on Understanding Business Cycles.

A second case to consider is an increase in aggregate demand that results in an equilibrium at a level of GDP greater than full-employment GDP in the short run, as illustrated in Figure 10.7. Note that both GDP and the price level are increased. The economy can operate at a level of GDP greater than full-employment GDP in the short run, as workers work overtime and maintenance of productive equipment is delayed, but output greater than full-employment GDP cannot be maintained in the long run. In the long run, the economy always returns to full-employment GDP along the LRAS curve.

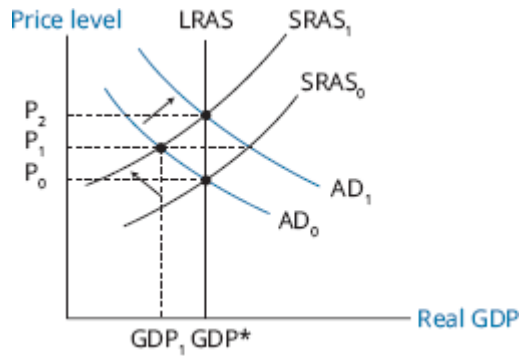
Figure 10.7: Adjustment to an Increase in Aggregate Demand



We term the difference between GDP_1 and full-employment GDP in Figure 10.7 an **inflationary gap** because the increase in aggregate demand from its previous level causes upward pressure on the price level. Competition among producers for workers, raw materials, and energy may shift the SRAS curve to the left, returning the economy to full-employment GDP but at a price level that is higher still. Alternatively, government policy makers can reduce aggregate demand by decreasing government spending, increasing taxes, or slowing the growth rate of the money supply, in order to move the economy back to the initial long run equilibrium at full-employment GDP.

Changes in wages or the prices of other important productive inputs can shift the SRAS curve, affecting real GDP and the price level in the short run. An important case to consider is a decrease in SRAS caused by an increase in the prices of raw materials or energy. As illustrated in Figure 10.8, the new short-run equilibrium is at lower GDP and a higher overall price level for goods and services compared to the initial long-run equilibrium. This combination of declining economic output and higher prices is termed **stagflation** (stagnant economy with inflation).

Figure 10.8: Stagflation

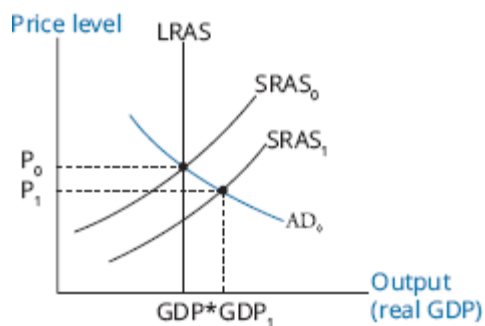


A subsequent decrease in input prices can return the economy to its long-run equilibrium output. An increase in aggregate demand from either expansionary fiscal or monetary policy can also return the economy to its full employment level, but at a price level that is higher still compared to the initial equilibrium.

Stagflation is an especially difficult situation for policy makers because actions to increase aggregate demand to restore full employment will also increase the price level even more. Conversely, a decision by policy makers to fight inflation by decreasing aggregate demand will decrease GDP even further. A decrease in wages and the prices of other productive inputs may be expected to increase SRAS and restore full-employment equilibrium. However, this process may be quite slow and doing nothing may be a very risky strategy for a government when voters expect action to restore economic growth or stem inflationary pressures.

The fourth case to consider is an increase in SRAS due to a decrease in the price of important productive inputs. As illustrated in Figure 10.9, the resulting new short-run equilibrium is at a level of GDP greater than full-employment GDP and a lower overall price level.

Figure 10.9: Decrease in Input Prices



In Figure 10.10, we present a summary of the short-run effects of shifts in aggregate demand and in aggregate supply on real GDP, unemployment, and the price level.

Figure 10.10: Short-Run Macroeconomic Effects

Type of Change	Real GDP	Unemployment	Price Level
Increase in AD	Increase	Decrease	Increase
Decrease in AD	Decrease	Increase	Decrease
Increase in AS	Increase	Decrease	Decrease
Decrease in AS	Decrease	Increase	Increase

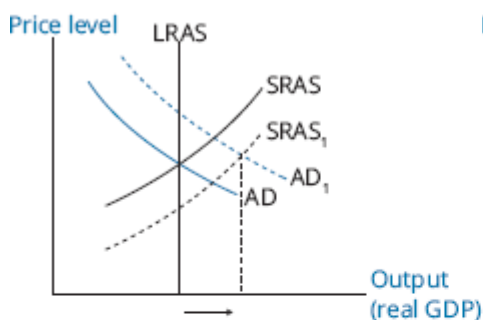
LOS 10.I: Analyze the effect of combined changes in aggregate supply and demand on the economy.

When both aggregate supply and aggregate demand change, the effects on equilibrium output and the price level may be clear when the effects on the variable are in the same direction (or ambiguous when the effects on the variable are in opposite directions). We summarize the effects of combined changes in demand and supply in Figure 10.11.

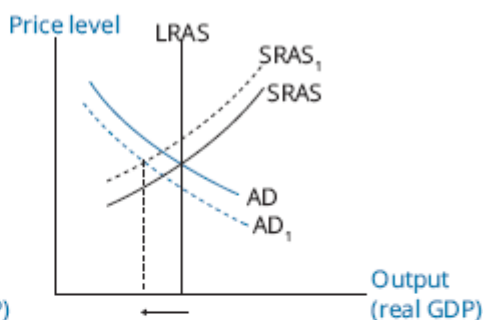
- When *aggregate demand and aggregate supply both increase*, real GDP increases but the effect on the price level depends on the relative magnitudes of the changes because their price effects are in opposite directions [Panel (a) of Figure 10.11].
- When *aggregate demand and aggregate supply both decrease*, real GDP decreases but the effect on the price level depends on the relative magnitudes of the changes because their price effects are in opposite directions [Panel (b) of Figure 10.11].
- When *aggregate demand increases and aggregate supply decreases*, the price level will increase but the effect on real GDP depends on the relative magnitudes of the changes because their effects on economic output are in opposite directions [Panel (c) of Figure 10.11].
- When *aggregate demand decreases and aggregate supply increases*, the price level will decrease but the effect on real GDP depends on the relative magnitudes of the changes because their effects on economic output are in opposite directions [Panel (d) of Figure 10.11].

Figure 10.11: Changes in Aggregate Supply and Aggregate Demand

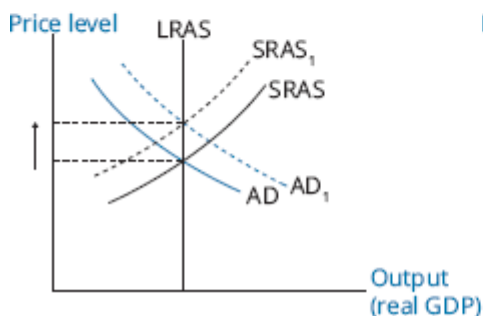
(a) AD and SRAS increase



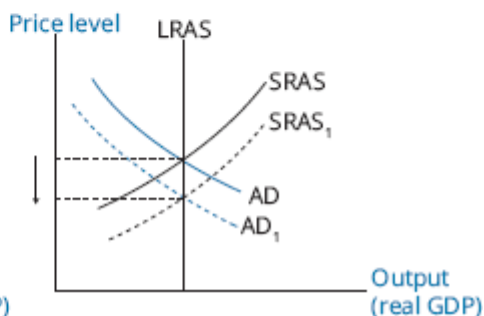
(b) AD and SRAS decrease



(c) AD increases, SRAS decreases



(d) AD decreases, SRAS increases



LOS 10.m: Describe sources, measurement, and sustainability of economic growth.

Economic growth can best be explained by examining five important **sources of economic growth**:

1. **Labor supply.** The **labor force** is the number of people over the age of 16 who are either working or available for work but currently unemployed. It is affected by population growth, net immigration, and the labor force participation rate (described in our reading on Understanding Business Cycles). Growth of the labor force is an important source of economic growth.
2. **Human capital.** The education and skill level of a country's labor force can be just as important a determinant of economic output as the size of the labor force. Because workers who are skilled and well-educated (possess more human capital) are more productive and better able to take advantage of advances in technology, investment in human capital leads to greater economic growth.
3. **Physical capital stock.** A high rate of investment increases a country's stock of physical capital. As noted earlier, a larger capital stock increases labor productivity and potential GDP. An increased rate of investment in physical capital can increase economic growth.
4. **Technology.** As noted previously, improvements in technology increase productivity and potential GDP. More rapid improvements in technology lead to greater rates of economic growth.
5. **Natural resources.** Raw material inputs, such as oil and land, are necessary to produce economic output. These resources may be *renewable* (e.g., forests) or *non-renewable* (e.g., coal). Countries with large amounts of productive natural resources can achieve greater rates of economic growth.

Sustainability of Economic Growth

One way to view potential GDP is with the following equation:

$$\text{potential GDP} = \text{aggregate hours worked} \times \text{labor productivity}$$

Or, stated in terms of economic growth:

$$\text{growth in potential GDP} = \text{growth in labor force} + \text{growth in labor productivity}$$

An economy's sustainable growth rate can be estimated by estimating the growth rate of labor productivity and the growth rate of the labor force. For example, if Japan's labor force is projected to shrink by 1%, while its labor productivity is expected to grow by 2%, then we would estimate the growth in potential GDP as: $-1\% + 2\% = 1\%$.

The **sustainable rate of economic growth** is important because long-term equity returns are highly dependent on economic growth over time. A country's sustainable rate of economic growth is the rate of increase in the economy's productive capacity (potential GDP).

LOS 10.n: Describe the production function approach to analyzing the sources of economic growth.

A **production function** describes the relationship of output to the size of the labor force, the capital stock, and productivity.

Economic output can be thought of as a function of the amounts of labor and capital that are available and their productivity, which depends on the level of technology available. That is:

$$Y = A \times f(L, K)$$

where:

Y = aggregate economic output

L = size of labor force

K = amount of capital available

A = total factor productivity

The multiplier, A, is referred to as **total factor productivity** and quantifies the amount of output growth that is not explained by increases in the size of the labor force and capital. Total factor productivity is closely related to technological advances. Generally, total factor productivity cannot be observed directly and must be inferred based on the other factors.

The production function can be stated on a per-worker basis by dividing by L :

$$Y/L = A \times f(K/L)$$

where:

Y/L = output per worker (labor productivity)

K/L = physical capital per worker

This relationship suggests that labor productivity can be increased by either improving technology or increasing physical capital per worker.

We assume that the production function exhibits **diminishing marginal productivity** for each individual input, meaning the amount of additional output produced by each additional unit of input declines (holding the quantities of other inputs constant). For this reason, sustainable long-term growth cannot necessarily be achieved simply by **capital deepening investment**—that is to say, increasing physical capital per worker over time. Productivity gains and growth of the labor force are also necessary for long-term sustainable growth.

LOS 10.o: Define and contrast input growth with growth of total factor productivity as components of economic growth.

A well-known model (the *Solow model* or *neoclassical model*) of the contributions of technology, labor, and capital to economic growth is:

growth in potential GDP = growth in technology + W_L (growth in labor) + W_C (growth in capital)

where W_L and W_C are labor's percentage share of national income and capital's percentage share of national income. Like the multiplier, A, in a production function, the additional growth in potential GDP from "growth in technology" represents the change in total factor productivity, the growth of output that is not explained by the growth of labor and capital. Growth in technology is the primary driver of the growth in total factor productivity.

Consider a developed country where $W_L = 0.7$ and $W_C = 0.3$. For that country, a 1% increase in the labor force will lead to a much greater increase in economic output than a 1% increase in the capital stock. Similarly, sustained growth of the labor force will result in greater economic growth over time than sustained growth of the capital stock of an equal magnitude.

Sometimes the relationship between potential GDP, improvements in technology, and capital growth is written on a per-capita basis² as:

$$\text{growth in per-capita potential GDP} = \text{growth in technology} + W_C (\text{growth in the capital-to-labor ratio})$$

With $W_C = 0.25$, for example, each 1% increase in capital per worker will increase GDP per worker by 0.25%. In developed economies, where capital per worker is already relatively high, growth of technology will be the primary source of growth in GDP per worker. At higher levels of capital per worker, an economy will experience diminishing marginal productivity of capital and must look to advances in technology for strong economic growth.



MODULE QUIZ 10.3

- Starting from short-run equilibrium, if aggregate demand is increasing faster than long-run aggregate supply:
 - the price level is likely to increase.
 - downward pressure on wages should ensue.
 - supply will increase to meet the additional demand.
- A short-run macroeconomic equilibrium in which output must decrease to restore long-run equilibrium is most accurately characterized as:
 - stagflation.
 - a recessionary gap.
 - an inflationary gap.
- Which of the following combinations of changes in aggregate demand and aggregate supply is *most likely* to result in decreasing prices? Aggregate demand:
 - decreases while aggregate supply increases.
 - decreases while aggregate supply decreases.
 - increases while aggregate supply decreases.
- Labor productivity is *most likely* to increase as a result of:
 - an increase in physical capital.
 - a decrease in net immigration.
 - an increase in the labor force participation rate.
- Long-term sustainable growth of an economy is *least likely* to result from growth in:
 - the supply of labor.
 - capital per unit of labor.
 - output per unit of labor.
- In a production function model of economic output, total factor productivity represents the output growth that can be accounted for by:
 - capital growth but not labor growth.
 - neither labor growth nor capital growth.
 - the combined effects of labor growth and capital growth.
- In a developed economy, the primary source of growth in potential GDP is:
 - capital investment.
 - labor supply growth.
 - technology advances.

KEY CONCEPTS

LOS 10.a

Gross domestic product (GDP) is the market value of all final goods and services produced within a country during a certain time period.

Using the expenditure approach, GDP is calculated as the total amount spent on goods and services produced in the country during a time period.

Using the income approach, GDP is calculated as the total income earned by households and businesses in the country during a time period.

LOS 10.b

The expenditure approach to measuring GDP can use the sum-of-value-added method or the value-of-final-output method.

- Sum-of-value-added: GDP is calculated by summing the additions to value created at each stage of production and distribution.
- Value-of-final-output: GDP is calculated by summing the values of all final goods and services produced during the period.

LOS 10.c

Nominal GDP values goods and services at their current prices. Real GDP measures current year output using prices from a base year.

The GDP deflator is a price index that can be used to convert nominal GDP into real GDP by removing the effects of changes in prices.

LOS 10.d

The four components of gross domestic product are consumption spending, business investment, government spending, and net exports.

$$\text{GDP} = C + I + G + (X - M).$$

National income is the income received by all factors of production used in the creation of final output.

Personal income is the pretax income received by households.

Household disposable income is personal income after taxes.

LOS 10.e

Private saving and investment are related to the fiscal balance and the trade balance. A fiscal deficit must be financed by some combination of a trade deficit or an excess of private saving over private investment.

$$(G - T) = (S - I) - (X - M).$$

LOS 10.f

The aggregate demand curve slopes downward because higher price levels reduce real wealth, increase real interest rates, and make domestically produced goods more expensive compared to goods produced abroad, all of which reduce the quantity of domestic output demanded.

LOS 10.g

The short-run aggregate supply curve shows the positive relationship between real GDP supplied and the price level, when other factors are held constant. Holding some input costs such as wages fixed in the short run, the curve slopes upward because higher output prices result in greater output (real wages fall).

Because all input prices are assumed to be flexible in the long run, the long-run aggregate supply curve is perfectly inelastic (vertical). Long-run aggregate supply represents potential GDP, the full employment level of economic output.

LOS 10.h

Changes in the price level cause movement along the aggregate demand or aggregate supply curves.

Shifts in the aggregate demand curve are caused by changes in household wealth, business and consumer expectations, capacity utilization, fiscal policy, monetary policy, currency exchange rates, and global economic growth rates.

Shifts in the short-run aggregate supply curve are caused by changes in nominal wages or other input prices, expectations of future prices, business taxes, business subsidies, and currency exchange rates, as well as by the factors that affect long-run aggregate supply.

Shifts in the long-run aggregate supply curve are caused by changes in labor supply and quality, the supply of physical capital, the availability of natural resources, and the level of technology.

LOS 10.i

The short-run effects of changes in aggregate demand and in aggregate supply are summarized in the following table:

Type of Change	Real GDP	Unemployment	Price Level
Increase in AD	Increase	Decrease	Increase
Decrease in AD	Decrease	Increase	Decrease
Increase in AS	Increase	Decrease	Decrease
Decrease in AS	Decrease	Increase	Increase

LOS 10.j

In long-run equilibrium, real GDP is equal to full-employment (potential) GDP. An increase in aggregate demand can result in a short-run equilibrium with GDP greater than full-employment GDP, termed an inflationary gap. A decrease in aggregate demand can result in a short-run equilibrium with GDP less than full-employment, termed a recessionary gap. When short-run aggregate supply decreases, the resulting short-run equilibrium is with GDP reduced to less than full-employment GDP but with an increase in the price level, termed stagflation.

LOS 10.k

From a situation of long-run equilibrium: an increase in either aggregate demand or aggregate supply can result in a short-run equilibrium with real GDP greater than full employment GDP; a decrease in either aggregate demand or aggregate supply can result in a short-run equilibrium with real GDP less than full-employment GDP.

LOS 10.l

Short-run effects of shifts in both aggregate demand and aggregate supply on the price level and real GDP:

Aggregate Demand	Aggregate Supply	Change in Real GDP	Change in Price Level
Increase	Increase	Increase	May increase or decrease
Decrease	Decrease	Decrease	May increase or decrease
Increase	Decrease	May increase or decrease	Increase
Decrease	Increase	May increase or decrease	Decrease

LOS 10.m

Sources of economic growth include increases in the supply of labor, increases in human capital, increases in the supply of physical capital, increasing availability of natural resources, and advances in technology.

The sustainable rate of economic growth is determined by the rate of increase in the labor force and the rate of increase in labor productivity.

LOS 10.n

A production function relates economic output to the supply of labor, the supply of capital, and total factor productivity. Total factor productivity is a residual factor, which represents that part of economic growth not accounted for by increases in the supply of labor and capital. Increases in total factor productivity can be attributed to advances in technology.

LOS 10.o

In developed countries, where a high level of capital per worker is available and capital inputs experience diminishing marginal productivity, technological advances that increase total factor productivity are the main source of sustainable economic growth.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 10.1

- A** Adding all purchases and sales is not appropriate because these would include goods that were produced before the time period in question. All purchases and sales could also result in double-counting intermediate goods. GDP is the market value of all final goods and services produced in a country in a certain period of time. GDP can be calculated either by totaling the amount spent on goods and services produced in the economy (the expenditure approach), or the income generated in producing these goods and services (the income approach). (LOS 10.b)
- A** Owner-occupied housing and government services are included in GDP at imputed (estimated) values. Transfer payments are excluded from the calculation of GDP. (LOS 10.a)
- B** Using the sum-of-value-added method, GDP can be calculated by summing the value added at each stage in the production and distribution process. Summing the value of the product at each stage of production would count the value added at earlier stages multiple times. The value added at earlier stages would not be included in GDP if it was deducted from the retail price. (LOS 10.b)
- B** Real GDP is the value of current period output calculated using prices from a base year. (LOS 10.c)

5. **A** The GDP deflator is the ratio of nominal GDP to real GDP, or equivalently the ratio of current year prices to base year prices. (LOS 10.c)
6. **B** National income is the income received by all factors of production used in the generation of final output. Personal income measures the pretax income that households receive. Disposable income is personal income after taxes. (LOS 10.d)
7. **B** The fundamental relationship among saving, investment, the fiscal balance, and the trade balance is described by the following equation: $(G - T) = (S - I) - (X - M)$. If the government budget deficit $(G - T)$ increases, the larger budget deficit must be financed by some combination of an increase in the excess of private saving over private investment $(S - I)$ or a decrease in net exports $(X - M)$. (LOS 10.e)

Module Quiz 10.2

1. **B** The inverse relationship between aggregate income (or output) and the price level is the aggregate demand curve. (LOS 10.f)
2. **A** The LRAS curve is vertical at the level of potential GDP. (LOS 10.g)
3. **B** Strengthening of the domestic currency should cause exports to decrease and imports to increase, causing the AD curve to shift to the left (lower demand). At the same time, the cost of raw material inputs should decrease in domestic currency terms, causing the SRAS curve to shift to the right (greater supply). Changes in the price level cause movement along the AD and AS curves; in this case, any shifts along these curves will be towards lower prices. (LOS 10.h)
4. **A** Since the y-axis of the aggregate supply/demand model is the price level, a change in the price level is a movement along the AD curve. As long as inflation expectations are unchanged, an increase in the price level will not shift the aggregate demand curve. (LOS 10.h)

Module Quiz 10.3

1. **A** If AD is increasing faster than LRAS, the economy is expanding faster than its full-employment rate of output. This will cause pressure on wages and resource prices and lead to an increase in the price level. The SRAS curve will shift to the left—a decrease in supply for any given price level—until the rate of output growth slows to its full-employment potential. (LOS 10.i)
2. **C** If output must decrease to restore long-run equilibrium, the short-run equilibrium must be at an output level greater than long-run aggregate supply. This describes an inflationary gap. (LOS 10.j, 10.k)
3. **A** Decreasing aggregate demand combined with increasing aggregate supply will result in decreasing prices. Increasing aggregate demand combined with decreasing aggregate supply will result in increasing prices. A decrease or an increase in both aggregate demand and aggregate supply may either increase or decrease prices. (LOS 10.l)
4. **A** Increased investment in physical capital can increase labor productivity. Labor force participation rates and net immigration affect the size of the labor force and the aggregate number of hours worked, but do not necessarily affect labor productivity. (LOS 10.m)
5. **B** The sustainable rate of economic growth is a measurement of the rate of increase in the economy's productive capacity. An economy's sustainable rate of growth depends on the growth rate of the labor supply and the growth rate of labor productivity. Due to diminishing marginal productivity, an economy generally cannot achieve long-term sustainable growth through continually increasing the stock of capital relative to labor (i.e., capital deepening). (LOS 10.m)
6. **B** Total factor productivity represents output growth in excess of that resulting from the growth in labor and capital. (LOS 10.n)

7. **C** For developed economies, advances in technology are likely to be the primary source of growth in potential GDP because capital per worker is already high enough to experience diminishing marginal productivity of capital. (LOS 10.o)

¹ According to the Federal Reserve, "Industrial plants usually operate at capacity utilization rates that are well below 100 percent... For total industry and total manufacturing, utilization rates have exceeded 90 percent only in wartime." (Federal Reserve Statistical Release G.17, "Industrial Production and Capacity Utilization," www.federalreserve.gov/releases/g17/current/g17.pdf)

² Paul R. Kutasovic, CFA, and Richard G. Fritz, *Aggregate Output, Prices, and Economic Growth*, CFA® Program Level I 2022 Curriculum, Volume 2.

READING 11

UNDERSTANDING BUSINESS CYCLES

EXAM FOCUS

The phase of the business cycle is the starting point for top-down financial analysis. Candidates need to know how to interpret the many economic indicators that are available and why various indicators tend to lead, coincide with, or lag behind changes in economic activity. Indicators of unemployment and inflation are crucial for understanding fiscal and monetary policy actions.

MODULE 11.1: BUSINESS CYCLE PHASES



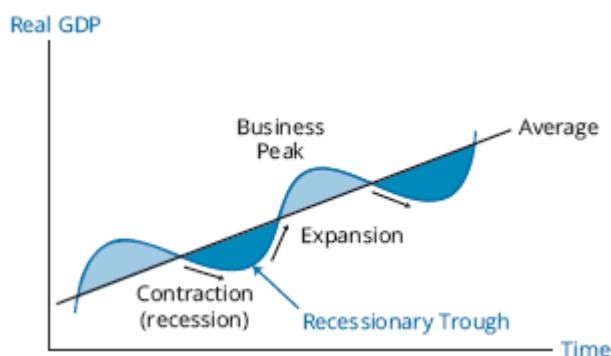
Video covering this content is available online.

LOS 11.a: Describe the business cycle and its phases.

The **business cycle** is characterized by fluctuations in economic activity. Real gross domestic product (GDP) and the rate of unemployment are the key variables used to determine the current phase of the cycle.

The business cycle has four phases: **expansion** (real GDP is increasing), **peak** (real GDP stops increasing and begins decreasing), **contraction** or **recession** (real GDP is decreasing), and **trough** (real GDP stops decreasing and begins increasing). The phases are illustrated in Figure 11.1.

Figure 11.1: Business Cycle



An expansion features growth in most sectors of the economy, with increasing employment, consumer spending, and business investment. As an expansion approaches its peak, the rates of increase in spending, investment, and employment slow but remains positive, while inflation accelerates.

A contraction or recession is associated with declines in most sectors, with inflation typically decreasing. When the contraction reaches a trough and the economy begins a new expansion or **recovery**, economic growth becomes positive again and inflation is typically moderate, but employment growth may not start to increase until the expansion has taken hold convincingly.

A common rule of thumb is to consider two consecutive quarters of growth in real GDP as the beginning of an expansion and two consecutive quarters of declining real GDP as indicating the beginning of a contraction. Statistical agencies that date expansions and recessions, such as the National Bureau of Economic Research in the United States, look at a wider variety of economic data such as employment, industrial production, and real personal income to identify turning points in the business cycle.

A key aspect of business cycles is that they recur, but not at regular intervals. Past business cycles have been as short as a year or longer than a decade.

The idea of a business cycle applies to economies that consist mainly of businesses. For economies that are mostly subsistence agriculture or dominated by state planning, fluctuations in activity are not really “business cycles” in the sense we are discussing here.

LOS 11.b: Describe credit cycles.

Credit cycles refer to cyclical fluctuations in interest rates and the availability of loans (credit). Typically, lenders are more willing to lend and tend to offer lower interest rates during economic expansions and are less willing to lend and require higher interest rates when the economy is slowing (contracting).

Credit cycles may amplify business cycles. Widely available or “loose” credit conditions during expansions can lead to “bubbles” (prices based on implausible expectations) in the markets for some assets, such as subprime mortgages in the period leading up to the financial crisis of 2007–2009. Some research suggests that expansions tend to be stronger, and contractions deeper and longer lasting, when they coincide with credit cycles. They do not always coincide, however, as historical data suggests credit cycles have been longer in duration than business cycles on average.

LOS 11.c: Describe how resource use, consumer and business activity, housing sector activity, and external trade sector activity vary as an economy moves through the business cycle.

Business Activity and Resource Use Fluctuation

Inventories are an important business cycle indicator. Firms try to keep enough inventory on hand to meet sales demand but do not want to keep too much of their capital tied up in inventory. As a result, the ratio of inventory to sales in many industries trends toward a normal level in times of steady economic growth.

When an expansion is approaching its peak, sales growth begins to slow, and unsold inventories accumulate. This can be seen in an increase in the **inventory-sales ratio** above its normal level.

Firms respond to an unplanned increase in inventory by reducing production, which is one of the causes of the subsequent contraction in the economy. An increase in inventories is counted in the GDP statistics as economic output, whether the increase is planned or unplanned. An analyst who looks only at GDP growth, rather than the inventory-sales ratio, might see economic strength rather than the beginning of weakness.

The opposite occurs when a contraction reaches its trough. Having reduced their production levels to adjust for lower sales demand, firms find their inventories becoming depleted more quickly once sales growth begins to accelerate. This causes the inventory-sales ratio to decrease below its normal level. To meet the increase in demand, firms will increase output, and the inventory-sales ratio will increase toward normal levels.

One of the ways firms react to fluctuations in business activity is by adjusting their utilization of labor and physical capital. Adding and subtracting workers in lockstep with changes in economic growth would be costly for firms, in terms of both direct expenses and the damage it would do to employee morale and loyalty. Instead, firms typically begin by changing how they utilize their current workers, producing less or more output per hour or adjusting the hours they work by adding or removing overtime. Only when an expansion or contraction appears likely to persist will they hire or lay off workers.

Similarly, because it is costly to adjust production levels by frequently buying and selling plant and equipment, firms first adjust their production levels by using their existing physical capital more or less intensively. As an expansion persists, firms will increase their production capacity by investing more in plant and equipment. During contractions, however, firms will not necessarily sell plant and equipment outright. They can reduce their physical capacity by spending less on maintenance or by delaying the replacement of equipment that is near the end of its useful life.

Consumer Sector Activity

Consumer spending, the largest component of gross domestic product, depends on the level of consumers' current incomes and their expectations about their future incomes. As a result, consumer spending increases during expansions and decreases during contractions.

Consumer spending in some sectors is more sensitive to business cycle phases than spending in other sectors. Spending on **durable goods** is highly cyclical because they are often higher-value purchases. Consumers are more willing to purchase high-value durable goods (e.g., appliances, furniture, automobiles) during expansions, when incomes are increasing and economic confidence is high. During contractions (and sometimes extending into the early stages of expansions), consumers often postpone durable goods purchases until they are more confident about their employment status and prospects for income growth.

Consumer spending on **services** is also positively correlated with business cycle phases, but not to the same extent as durable goods spending. Services include spending that is more discretionary, such as for lodging or food away from home, but also includes spending that is less discretionary, such as spending on telecommunications, health care, and insurance. The more discretionary a service is, the more cyclical consumer spending on it tends to be. Spending on **nondurable goods**, such as food at home or household products for everyday use, remains relatively stable over the business cycle.

Housing Sector Activity

Although the housing sector is a small part of the economy relative to overall consumer spending, cyclical swings in activity in the housing market can be large so that the effect on overall economic activity is greater than it otherwise would be. Important determinants of the level of economic activity in the housing sector are:

1. **Mortgage rates:** Low interest rates tend to increase home buying and construction while high interest rates tend to reduce home buying and construction.
2. **Housing costs relative to income:** When incomes are cyclically high (low) relative to home costs, including mortgage financing costs, home buying and construction tend to increase (decrease). Housing activity can decrease even when incomes are rising late in a cycle if home prices are rising faster than incomes, leading to decreases in purchase and construction activity in the housing sector.
3. **Speculative activity:** As we saw in the housing sector in 2007 and 2008 in many economies, rising home prices can lead to purchases based on expectations of further gains. Higher prices led to more construction and eventually excess building. This resulted in falling prices that decreased or eliminated speculative demand and led to dramatic decreases in housing activity overall.
4. **Demographic factors:** The proportion of the population in the 25- to 40-year-old segment is positively related to activity in the housing sector because these are the ages of greatest household formation. In China, a strong population shift from rural areas to cities as manufacturing activity has grown has required large increases in construction of new housing to accommodate those needs.

External Trade Sector Activity

The most important factors determining the level of a country's imports and exports are domestic GDP growth, GDP growth of trading partners, and currency exchange rates. Increasing growth of domestic GDP leads to increases in purchases of foreign goods (imports), while decreasing domestic GDP growth reduces imports. Exports depend on the growth rates of GDP of other economies (especially those of important trading partners). Increasing foreign incomes increase sales to foreigners (exports) and decreasing economic growth in foreign countries decreases domestic exports.

An increase in the value of a country's currency makes its goods more expensive to foreign buyers and foreign goods less expensive to domestic buyers, which tends to decrease exports and increase imports. A decrease in the value of a country's currency has the opposite effect, increasing exports and decreasing imports. Currencies affect import and export volumes over time in response to persistent trends in foreign exchange rates, rather than in response to short-term changes which can be quite volatile.

Currency effects can differ in direction from GDP growth effects and change in response to a complex set of variables. The effects of changes in GDP levels and growth rates are more direct and immediate.

Typical business cycle characteristics may be summarized as follows:

Trough:

- GDP growth rate changes from negative to positive.
- High unemployment rate, increasing use of overtime and temporary workers.
- Spending on consumer durable goods and housing may increase.
- Moderate or decreasing inflation rate.

Expansion:

- GDP growth rate increases.
- Unemployment rate decreases as hiring accelerates.
- Investment increases in producers' equipment and home construction.
- Inflation rate may increase.
- Imports increase as domestic income growth accelerates.

Peak:

- GDP growth rate decreases.
- Unemployment rate decreases but hiring slows.
- Consumer spending and business investment grow at slower rates.
- Inflation rate increases.

Contraction/recession:

- GDP growth rate is negative.
- Hours worked decrease, unemployment rate increases.
- Consumer spending, home construction, and business investment decrease.
- Inflation rate decreases with a lag.
- Imports decrease as domestic income growth slows.

LOS 11.d: Describe theories of the business cycle.

The causes of business cycles are a subject of considerable debate among economists.

Neoclassical school economists believe shifts in both aggregate demand and aggregate supply are primarily *driven by changes in technology* over time. They also believe that the economy has a strong tendency toward full-employment equilibrium, as recession puts downward pressure on the money wage rate, or as over-full employment puts upward pressure on the money wage rate. They conclude that business cycles result from *temporary deviations from long-run equilibrium*.

The Great Depression of the 1930s did not support the beliefs of the neoclassical economists. The economy in the United States operated significantly below its full-employment level for many years. Additionally, business cycles in general have been more severe and more prolonged than the neoclassical model would suggest.

British economist John Maynard Keynes attempted to explain the Depression and the nature of business cycles. He provided policy recommendations for moving the economy toward full-employment GDP and reducing the severity and duration of business cycles. Keynes believed

that *shifts in aggregate demand due to changes in expectations* were the primary cause of business cycles. **Keynesian school** economists believe these fluctuations are primarily due to swings in the level of optimism of those who run businesses. They overinvest and overproduce when they are too optimistic about future growth in potential GDP, and they underinvest and underproduce when they are too pessimistic or fearful about the future growth in potential GDP.

Keynesians argue that wages are “downward sticky,” reducing the ability of a decrease in money wages to increase short-run aggregate supply and move the economy from recession (or depression) back toward full employment. The policy prescription of Keynesian economists is to increase aggregate demand directly, through monetary policy (increasing the money supply) or through fiscal policy (increasing government spending, decreasing taxes, or both).

The **New Keynesian school** added the assertion that the prices of productive inputs other than labor are also “downward sticky,” presenting additional barriers to the restoration of full-employment equilibrium.

A third view of macroeconomic equilibrium is that held by the **Monetarist school**. Monetarists believe the variations in aggregate demand that cause business cycles are due to variations in the rate of growth of the money supply, likely from *inappropriate decisions by the monetary authorities*. Monetarists believe that recessions can be caused by external shocks or by inappropriate decreases in the money supply. They suggest that to keep aggregate demand stable and growing, the central bank should follow a policy of steady and predictable increases in the money supply.

Economists of the **Austrian school** believe business cycles are caused by *government intervention in the economy*. When policymakers force interest rates down to artificially low levels, firms invest too much capital in long-term and speculative lines of production, compared to actual consumer demand. When these investments turn out poorly, firms must decrease output in those lines, which causes a contraction.



PROFESSOR'S NOTE

Austrian school economists refer to this misdirection of capital as “malinvestment.” The theory is related closely to the credit cycles discussed earlier.

New Classical school economists introduced **real business cycle theory (RBC)**. RBC emphasizes the effect of real economic variables such as *changes in technology and external shocks*, as opposed to monetary variables, as the cause of business cycles. RBC applies utility theory, which we described in the readings on microeconomic analysis, to macroeconomics. Based on a model in which individuals and firms maximize expected utility, New Classical economists argue that policymakers should not try to counteract business cycles because expansions and contractions are efficient market responses to real external shocks.



MODULE QUIZ 11.1

1. In the early part of an economic expansion, inventory-sales ratios are *most likely* to:
 - A. increase because sales are unexpectedly low.
 - B. increase because businesses plan for expansion.
 - C. decrease because of unexpected increases in sales.
2. The contraction phase of the business cycle is *least likely* accompanied by decreasing:
 - A. unemployment.

- B. inflation pressure.
 - C. economic output.
3. According to which business cycle theory should expansionary monetary policy be used to fight a recession?
- A. Keynesian school.
 - B. Monetarist school.
 - C. New Classical school.

MODULE 11.2: INFLATION AND INDICATORS



LOS 11.e: Interpret a set of economic indicators and describe their uses and limitations.

Video covering this content is available online.

Economic indicators can be classified into three categories: **leading indicators** that have been known to change direction before peaks or troughs in the business cycle, **coincident indicators** that change direction at roughly the same time as peaks or troughs, and **lagging indicators** that don't tend to change direction until after expansions or contractions are already underway.

The Conference Board publishes indexes of leading, coincident, and lagging indicators for several countries. Their indexes for the United States include the following components:

- *Leading indicators*: Average weekly hours in manufacturing; initial claims for unemployment insurance; manufacturers' new orders for consumer goods; manufacturers' new orders for non-defense capital goods ex-aircraft; Institute for Supply Management new orders index; building permits for new houses; S&P 500 equity price index; Leading Credit Index; 10-year Treasury to Fed funds interest rate spread; and consumer expectations.
- *Coincident indicators*: Employees on nonfarm payrolls; real personal income; index of industrial production; manufacturing and trade sales.
- *Lagging indicators*: Average duration of unemployment; inventory-sales ratio; change in unit labor costs; average prime lending rate; commercial and industrial loans; ratio of consumer installment debt to income; change in consumer price index.

Other sources, such as the Organization for Economic Cooperation and Development (OECD) and the Economic Cycle Research Institute (ECRI), also publish indexes of economic indicators for the world's major economies.

Analysts should use leading, coincident, and lagging indicators together to determine the phase of the business cycle. They should also use the composite indexes to confirm what is indicated by individual indicators. If a widely followed leading indicator, such as stock prices or initial claims for unemployment insurance, changes direction, but most other leading indicators have not, an analyst should not yet conclude that a peak or trough is imminent.

EXAMPLE: Interpreting economic indicators

Karen Trumbull, CFA, gathers the following economic reports for the United States in the most recent two months:

	Latest Month	Prior Month
Building permits	+1.8%	+0.7%
Commercial and industrial loans	-0.9%	-1.6%
Consumer price index	-0.1%	-0.2%
Index of industrial production	+0.2%	0.0%
New orders for consumer goods	+2.2%	+1.6%
Real personal income	0.0%	-0.4%

Based on these indicators, what should Trumbull conclude about the phase of the business cycle?

Answer:

Commercial and industrial loans and the consumer price index are lagging indicators. Industrial production and real personal income are coincident indicators. These indicators suggest the business cycle has been in the contraction phase.

Building permits and orders for consumer goods are leading indicators. Increases in both of these in the latest two months suggest an economic expansion may be emerging.

Taken together, these data indicate that the business cycle may be at or just past its trough.

Analysts should be aware that the classifications *leading*, *coincident*, and *lagging* indicators reflect tendencies in the timing of their turning points, not exact relationships with the business cycle. Not all changes in direction of leading indicator indexes have been followed by corresponding changes in the business cycle, and even when they have, the lead time has varied. This common criticism is summed up in the often repeated comment, “Declines in stock prices have predicted nine of the last four recessions.”



PROFESSOR'S NOTE

Analysts who use economic indicators in forecasting models must guard against look-ahead bias. The data are not available immediately. For example, data for May are typically first released in mid- to late June and may be revised in July and August.

LOS 11.f: Describe types of unemployment and compare measures of unemployment.

Unemployment can be divided into three categories:

1. **Frictional unemployment** results from the time lag necessary to match employees who seek work with employers needing their skills. Frictional unemployment is always with us as employers expand or contract their businesses and workers move, are fired, or quit to seek other opportunities.
2. **Structural unemployment** is caused by long-run changes in the economy that eliminate some jobs while generating others for which unemployed workers are not qualified. Structural unemployment differs from frictional unemployment in that the unemployed workers do not currently have the skills needed to perform the jobs that are available.

3. **Cyclical unemployment** is caused by changes in the general level of economic activity. Cyclical unemployment is positive when the economy is operating at less than full capacity and can be negative when an expansion leads to employment temporarily over the full employment level.

A person who is not working is considered to be **unemployed** if he is actively searching for work.¹ One who has been seeking work unsuccessfully for several months is referred to as *long-term unemployed*.

The **unemployment rate** is the percentage of people in the labor force who are unemployed. The **labor force** includes all people who are either employed or unemployed. People who choose not to be in the labor force are said to be *voluntarily unemployed* and are not included in the calculation of the unemployment rate.

A person who is employed part time but would prefer to work full time or is employed at a low-paying job despite being qualified for a significantly higher-paying one is said to be **underemployed**. Identification of the number of underemployed is somewhat subjective and not easily discernible from employment statistics.

The **participation ratio** (also referred to as the *activity ratio* or *labor force participation rate*) is the percentage of the working-age population who are either employed or actively seeking employment.

Short-term fluctuations in the participation ratio can occur because of changes in the number of **discouraged workers**, those who are available for work but are neither employed nor actively seeking employment. The participation rate tends to increase when the economy expands and decrease during recessions. Discouraged workers who stopped seeking jobs during a recession are motivated to seek work again once the expansion takes hold and they believe their prospects of finding work are better.

This movement of discouraged workers out of and back into the labor force causes the unemployment rate to be a lagging indicator of the business cycle. Early in an expansion when hiring prospects begin to improve, the number of discouraged workers who re-enter the labor force is greater than the number who are hired immediately. This causes the unemployment rate to increase even though employment is expanding. To gauge the current state of the labor market, analysts should include other widely available indicators such as the number of employees on payrolls.

Earlier, we noted that firms tend to be slow to hire or lay off workers at business cycle turning points. This also causes the unemployment rate to lag the business cycle. The effect can also be seen in data on **productivity**, or output per hour worked. Productivity declines early in contractions as firms try to keep employees on despite producing less output. Productivity increases early in expansions as firms try to produce more output but are not yet ready to hire new workers.

When comparing unemployment rates across countries, analysts should note that different reporting agencies may use somewhat dissimilar methods for calculating the statistics. Also, all of the employment indicators mentioned here apply only to legal employment. Participants in illegal sectors of the economy are not reflected in employment data.

LOS 11.g: Explain inflation, hyperinflation, disinflation, and deflation.

Inflation is a persistent increase in the price level over time. If the price level increases in a single jump but does not continue rising, the economy is not experiencing inflation. An increase in the price of a single good, or in relative prices of some goods, is not inflation. If inflation is present, the prices of almost all goods and services are increasing.

Inflation erodes the purchasing power of a currency. Inflation favors borrowers at the expense of lenders because when the borrower returns the principal to the lender, it is worth less in terms of goods and services (in real terms) than it was worth when it was borrowed. Inflation that accelerates out of control is referred to as **hyperinflation**, which can destroy a country's monetary system and bring about social and political upheavals.

The **inflation rate** is the percentage increase in the price level, typically compared to the prior year. Analysts can use the inflation rate as a business cycle indicator and to anticipate changes in central bank monetary policy. As we will see in the reading on Monetary and Fiscal Policy, an objective of central banks is to keep inflation within some target range. **Disinflation** refers to an inflation rate that is decreasing over time but remains greater than zero.

A persistently decreasing price level (i.e., a negative inflation rate) is called **deflation**. Deflation is commonly associated with deep recessions. When most prices are decreasing, consumers delay purchases because they believe they can buy the same goods more cheaply in the future. For firms, deflation results in decreasing revenue and increasing real fixed costs.



PROFESSOR'S NOTE

Values stated as “real” are adjusted for inflation over some defined period. This makes values at different points in time comparable in terms of purchasing power.

LOS 11.h: Explain the construction of indexes used to measure inflation.

To calculate a rate of inflation, we need to use a **price index** as a proxy for the price level. A price index measures the average price for a defined basket of goods and services. The **consumer price index (CPI)** is the best-known indicator of U.S. inflation. Many countries use indexes similar to the CPI.

The CPI basket represents the purchasing patterns of a typical urban household. Weights for the major categories in the CPI are shown in Figure 11.2.

Figure 11.2: Relative Importance in the CPI as of April 2016

Category	Percent of Index
Food	13.9%
Energy	6.6%
All items less food and energy	79.5%
Commodities less food and energy commodities:	
Apparel	3.2%
New vehicles	3.8%
Used cars and trucks	2.1%
Medical care commodities	1.8%
Alcoholic beverages	1.0%
Tobacco and smoking products	0.7%
Services less energy services:	
Shelter	33.3%
Medical care services	6.6%
Transportation services	5.9%

Source: Bureau of Labor Statistics, U.S. Department of Labor (stats.bls.gov)

To calculate the CPI, the Bureau of Labor Statistics compares the cost of the CPI basket today with the cost of the basket in an earlier *base period*. The value of the index is as follows:

$$\text{CPI} = \frac{\text{cost of basket at current prices}}{\text{cost of basket at base period prices}} \times 100$$

EXAMPLE: Calculating a price index

The following table shows price information for a simplified basket of goods:

Item	Quantity	Price in Base Period	Current Price
Cheeseburgers	200	2.50	3.00
Movie tickets	50	7.00	10.00
Gasoline (in gallons)	300	1.50	3.00
Digital watches	100	12.00	9.00

Calculate the change in the price index for this basket from the base period to the current period.

Answer:

Reference base period:

Cheeseburgers	$200 \times 2.50 =$	500
Movie tickets	$50 \times 7.00 =$	350
Gasoline	$300 \times 1.50 =$	450
Watches	$100 \times 12.00 =$	<u>1,200</u>
Cost of basket		2,500

Current period:

Cheeseburgers	$200 \times 3.00 =$	600
Movie tickets	$50 \times 10.00 =$	500
Gasoline	$300 \times 3.00 =$	900
Watches	$100 \times 9.00 =$	<u>900</u>
Cost of basket		2,900

$$\text{price index} = \frac{2,900}{2,500} \times 100 = 116$$

The price index is up $\frac{116}{100} - 1 = 16\%$ over the period.



PROFESSOR'S NOTE

The LOS requires you to “explain the construction of” price indexes but does not require you to calculate them.

Analysts who compare price indexes for different countries should be aware of differences in their composition. The weights assigned to each good and service reflect the typical consumer’s purchasing patterns, which are likely to be significantly different across countries and regions. There can also be differences in how the data are collected. In the United States, for example, the most frequently cited CPI measure is based on the purchases typical of “all urban consumers.” Other countries may survey a different set of consumers and consequently use different baskets of goods.

An alternative measure of consumer price inflation is the *price index for personal consumption expenditures*. In the United States, this index is created by surveying businesses rather than consumers. The *GDP deflator*, which we described in an earlier reading, is another widely used inflation measure.

Analysts who look for emerging trends in consumer prices are often interested in the prices of goods in process. Widespread price increases for producers’ goods may be passed along to consumers. For most major economies, a **producer price index (PPI)** or **wholesale price index (WPI)** is available. Analysts can observe the PPI for different stages of processing (raw materials, intermediate goods, and finished goods) to watch for emerging price pressure. Sub-indexes of the PPI are also useful for identifying changes in relative prices of producers’ inputs, which may indicate shifts in demand among industries.

For both consumer and producer prices, analysts and policymakers often distinguish between **headline inflation** and **core inflation**. Headline inflation refers to price indexes for all goods. Core inflation refers to price indexes that exclude food and energy. Food and energy prices are typically more volatile than those of most other goods. Thus, core inflation can sometimes be a more useful measure of the underlying trend in prices.

LOS 11.i: Compare inflation measures, including their uses and limitations.

The price index we calculated in our example is a **Laspeyres index**, which uses a constant basket of goods and services. Most countries calculate consumer price inflation this way.

Three factors cause a Laspeyres index of consumer prices to be biased upward as a measure of the cost of living:

- *New goods.* Older products are often replaced by newer, but initially more expensive, products. New goods are periodically added to the market basket, and the older goods they replace are reduced in weight in the index. This biases the index upward.
- *Quality changes.* If the price of a product increases because the product has improved, the price increase is not due to inflation but still increases the price index.
- *Substitution.* Even in an inflation-free economy, prices of goods relative to each other change all the time. When two goods are substitutes for each other, consumers increase their purchases of the relatively cheaper good and buy less of the relatively more expensive good. Over time, such changes can make a Laspeyres index's fixed basket of goods a less accurate measure of typical household spending.

A technique known as **hedonic pricing** can be used to adjust a price index for product quality. To address the bias from substitution, reporting agencies can use a *chained* or *chain-weighted* price index such as a **Fisher index**. A Fisher index is the geometric mean of a Laspeyres index and a **Paasche index**. A Paasche index uses the current consumption weights, prices from the base period, and prices in the current period.

EXAMPLE: Paasche index

Continuing the example we presented earlier, assume the basket of goods has changed as follows:

Item	Quantity in base period	Price in base period	Quantity in current period	Current price
Cheeseburgers	200	2.50	205	3.00
Movie tickets	50	7.00	45	10.00
Gasoline (in gallons)	300	1.50	295	3.00
Digital watches	100	12.00	105	9.00

Calculate a Paasche index for the current period, compare it to the Laspeyres index (previously calculated as 116), and explain the difference.

Answer:

Reference base period:

Cheeseburgers	$205 \times 2.50 =$	512.50
Movie tickets	$45 \times 7.00 =$	315.00
Gasoline	$295 \times 1.50 =$	442.50
Watches	$105 \times 12.00 =$	<u>1,260.00</u>
Cost of basket		2,530.00

Current period:

Cheeseburgers	$205 \times 3.00 =$	615.00
Movie tickets	$45 \times 10.00 =$	450.00
Gasoline	$295 \times 3.00 =$	885.00
Watches	$105 \times 9.00 =$	<u>945.00</u>
Cost of basket		2,895.00

$$\text{Paasche index} = \frac{2,895}{2,530} \times 100 = 114.43$$

The Paasche index is less than 116 because, compared to the base period, consumers have substituted away from the two goods with the largest percentage price increases (gasoline and movie tickets).



PROFESSOR'S NOTE

The LOS does not require you to calculate these indexes. We show these examples to illustrate how substitution of goods by consumers can affect index values.

LOS 11.j: Contrast cost-push and demand-pull inflation.

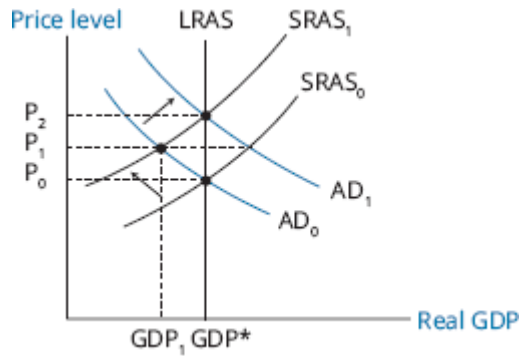
The two types of inflation are cost-push and demand-pull. **Cost-push inflation** results from a decrease in aggregate supply, while **demand-pull inflation** results from an increase in aggregate demand.

Cost-Push Inflation

Inflation can result from an initial decrease in aggregate supply caused by an increase in the real price of an important factor of production, such as wages or energy. Figure 11.3 illustrates the effect on output and the price level of a decrease in aggregate supply. The reduction from $SRAS_0$ to $SRAS_1$ increases the price level to P_1 , and with no initial change in aggregate demand, reduces output to GDP_1 .

If the decline in GDP brings a policy response that stimulates aggregate demand so output returns to its long-run potential, the result would be a further increase in the price level to P_2 .

Figure 11.3: Cost-Push Inflation



Because labor is the most important cost of production, wage pressure can be a source of cost-push inflation (sometimes called *wage-push inflation* when it occurs). Upward pressure on wages is more likely to emerge when cyclical unemployment is low, but it can occur even when cyclical unemployment is present. Because every individual provides a different type and quality of labor, some segments of the economy may have trouble finding enough qualified workers even during a contraction. As a result, the **non-accelerating inflation rate of unemployment (NAIRU)**, also called the **natural rate of unemployment (NARU)**, can be higher than the rate associated with the absence of cyclical unemployment. NARU or NAIRU can vary over time and is likely different across countries.

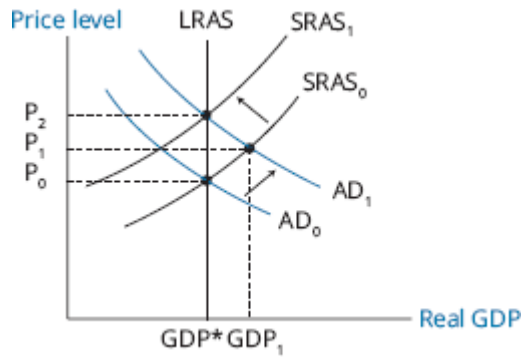
Analysts can use publicly available data on hourly and weekly earnings and labor productivity to identify signs of potential wage pressure. Wage increases are not inflationary as long as they remain in line with gains in productivity. A useful indicator of wages and benefits in terms of productivity is **unit labor costs**, the ratio of total labor compensation per hour to output units per hour.

An additional source of wage pressure is **expected inflation**. If workers expect inflation to increase, they will increase their wage demands accordingly. One indicator analysts use to gauge expected inflation is the difference in yield between inflation-indexed bonds, such as Treasury Inflation-Protected Securities, and otherwise similar non-indexed Treasury bonds.

Demand-Pull Inflation

Demand-pull inflation can result from an increase in the money supply, increased government spending, or any other change that increases aggregate demand. Figure 11.4 shows the effect on the price level when the economy is at full employment and aggregate demand increases (shifts to the right). In Figure 11.4, the economy is initially at full-employment equilibrium, with output at GDP^* and the price level at P_0 , so that the aggregate demand and short-run aggregate supply curves are AD_0 and $SRAS_0$. Real GDP is equal to potential GDP, which is represented by the long-run aggregate supply curve LRAS.

Figure 11.4: Demand-Pull Inflation



Now suppose the central bank increases the money supply, which increases aggregate demand to AD_1 . With no initial change in aggregate supply, output increases to GDP_1 , and the price level increases to P_1 . Prices rise, and real GDP is above potential (full-employment) GDP.

With real GDP above its full-employment level, the increase in GDP is not sustainable. Unemployment falls below its natural rate, which puts upward pressure on real wages. Rising real wages result in a decrease in short-run aggregate supply (the curve shifts left to $SRAS_1$) until real GDP reverts back to full-employment GDP. Output falls back to GDP^* , and the price level increases further to P_2 .

In the absence of other changes, the economy would reach a new equilibrium price level at P_2 . But what would happen if the central bank tried to keep GDP above the full-employment level with further increases in the money supply? The same results would occur repeatedly. Output cannot remain above its potential in the long run, but the induced increase in aggregate demand and the resulting pressure on wages would keep the price level rising even higher. Demand-pull inflation would persist until the central bank reduced the growth rate of the money supply and allowed the economy to return to full-employment equilibrium at a level of real GDP equal to potential GDP.

Economists often use the capacity utilization rate of industry to indicate the potential for demand-pull inflation. High rates of capacity utilization suggest the economy is producing at or above potential GDP and may experience inflationary pressure.

The impact on output is the key difference between the demand-pull and cost-push effects. The demand-pull effect increases GDP above full-employment GDP, while with cost-push inflation, a decrease in aggregate supply initially decreases GDP.



MODULE QUIZ 11.2

- An economic indicator that has turning points which tend to occur after the turning points in the business cycle is classified as:
 - a lagging indicator.
 - a leading indicator.
 - a trailing indicator.
- The unemployment rate is defined as the number of unemployed as a percentage of:
 - the labor force.
 - the number of employed.
 - the working-age population.
- A country's year-end consumer price index over a 5-year period is as follows:

Year 1 106.5

Year 2 114.2

Year 3 119.9

Year 4 124.8

Year 5 128.1

The behavior of inflation as measured by this index is *best* described as:

- A. deflation.
 - B. disinflation.
 - C. hyperinflation.
4. Core inflation is *best* described as an inflation rate:
- A. for producers' raw materials.
 - B. the central bank views as acceptable.
 - C. that excludes certain volatile goods prices.
5. Which of the following is *least likely* to reduce substitution bias in a consumer price index?
- A. Use a chained index.
 - B. Use a Paasche index.
 - C. Adjust for the bias directly using hedonic pricing.
6. In which of the following inflation scenarios does short-run aggregate supply decrease due to increasing wage demands?
- A. Cost-push inflation.
 - B. Demand-pull inflation.
 - C. Both cost-push and demand-pull inflation.

KEY CONCEPTS

LOS 11.a

The business cycle has four phases:

1. Expansion: Real GDP is increasing.
2. Peak: Real GDP stops increasing and begins decreasing.
3. Contraction: Real GDP is decreasing.
4. Trough: Real GDP stops decreasing and begins increasing.

Expansions feature increasing output, employment, consumption, investment, and inflation. Contractions are characterized by decreases in these indicators.

Business cycles are recurring but do not occur at regular intervals, can differ in strength or severity, and do not persist for specific lengths of time.

LOS 11.b

Credit cycles are cyclical fluctuations in interest rates and credit availability. Credit cycles may amplify business cycles and cause bubbles in the markets for some assets.

LOS 11.c

Inventory to sales ratios typically increase late in expansions when sales slow and decrease near the end of contractions when sales begin to accelerate. Firms decrease or increase production to restore their inventory-sales ratios to their desired levels.

Because hiring and laying off employees have high costs, firms prefer to adjust their utilization of current employees. As a result, firms are slow to lay off employees early in contractions and slow to add employees early in expansions.

Firms use their physical capital more intensively during expansions, investing in new capacity only if they believe the expansion is likely to continue. They use physical capital less intensively during contractions, but they are more likely to reduce capacity by deferring maintenance and not replacing equipment than by selling their physical capital.

Consumer spending fluctuates with the business cycle. Durable goods spending is highly sensitive to business cycles and spending on services is somewhat sensitive, but spending on nondurable goods is relatively insensitive to business cycles.

The level of activity in the housing sector is affected by mortgage rates, demographic changes, the ratio of income to housing prices, and investment or speculative demand for homes resulting from recent price trends.

Domestic imports tend to rise with increases in GDP growth and domestic currency appreciation, while increases in foreign incomes and domestic currency depreciation tend to increase domestic export volumes.

LOS 11.d

Neoclassical economists believe business cycles are temporary and driven by changes in technology, and that rapid adjustments of wages and other input prices cause the economy to move to full-employment equilibrium.

Keynesian economists believe excessive optimism or pessimism among business managers causes business cycles and that contractions can persist because wages are slow to move downward. New Keynesians believe input prices other than wages are also slow to move downward.

Monetarists believe inappropriate changes in the rate of money supply growth cause business cycles, and that money supply growth should be maintained at a moderate and predictable rate to support the growth of real GDP.

Austrian-school economists believe business cycles are initiated by government intervention that drives interest rates to artificially low levels.

Real business cycle theory holds that business cycles can be explained by utility-maximizing actors responding to real economic forces such as external shocks and changes in technology, and that policymakers should not intervene in business cycles.

LOS 11.e

Leading indicators have turning points that tend to precede those of the business cycle.

Coincident indicators have turning points that tend to coincide with those of the business cycle.

Lagging indicators have turning points that tend to occur after those of the business cycle.

A limitation of using economic indicators to predict business cycles is that their relationships with the business cycle are inexact and can vary over time.

LOS 11.f

Frictional unemployment results from the time it takes for employers looking to fill jobs and employees seeking those jobs to find each other. Structural unemployment results from long-term economic changes that require workers to learn new skills to fill available jobs. Cyclical unemployment is positive (negative) when the economy is producing less (more) than its potential real GDP.

A person is considered unemployed if he is not working, is available for work, and is actively seeking work. The labor force includes all people who are either employed or unemployed. The unemployment rate is the percentage of labor force participants who are unemployed.

LOS 11.g

Inflation is a persistent increase in the price level over time. An inflation rate is a percentage increase in the price level from one period to the next.

Disinflation is a decrease in the inflation rate over time. Deflation refers to a persistent decrease in the price level (i.e., a negative inflation rate).

LOS 11.h

A price index measures the cost of a specific basket of goods and services relative to its cost in a prior (base) period. The inflation rate is most often calculated as the annual percentage change in a price index.

The most widely followed price index is the consumer price index (CPI), which is based on the purchasing patterns of a typical household. The GDP deflator and the producer or wholesale price index are also used as measures of inflation.

Headline inflation is a percentage change in a price index for all goods. Core inflation is calculated by excluding food and energy prices from a price index because of their high short-term volatility.

LOS 11.i

A Laspeyres price index is based on the cost of a specific basket of goods and services that represents actual consumption in a base period. New goods, quality improvements, and consumers' substitution of lower-priced goods for higher-priced goods over time cause a Laspeyres index to be biased upward.

A Paasche price index uses current consumption weights for the basket of goods and services for both periods and thereby reduces substitution bias. A Fisher price index is the geometric mean of a Laspeyres and a Paasche index.

LOS 11.j

Cost-push inflation results from a decrease in aggregate supply caused by an increase in the real price of an important factor of production, such as labor or energy.

Demand-pull inflation results from persistent increases in aggregate demand that increase the price level and temporarily increase economic output above its potential or full-employment level.

The non-accelerating inflation rate of unemployment (NAIRU) represents the unemployment rate below which upward pressure on wages is likely to develop.

Wage demands reflect inflation expectations.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 11.1

1. **C** Early in an expansion, inventory-sales ratios typically decrease below their normal levels as accelerating sales draw down inventories of produced goods. (LOS 11.c)
2. **A** An economic contraction is likely to feature increasing unemployment (i.e., decreasing employment), along with declining economic output and decreasing inflation pressure. (LOS 11.a)
3. **A** Keynesian school economists recommend monetary or fiscal policy action to stimulate aggregate demand and restore full employment. Monetarists believe the rate of money supply growth should be kept stable and predictable. The New Classical school recommends against monetary or fiscal policy intervention because recessions reflect individuals' and firms' utility-maximizing response to real factors in the economy. (LOS 11.d)

Module Quiz 11.2

1. **A** Lagging indicators have turning points that occur after business cycle turning points. (LOS 11.e)
2. **A** The unemployment rate is the number of unemployed as a percentage of the labor force. (LOS 11.f)
3. **B** The yearly inflation rate is as follows:
Year 2 $(114.2 - 106.5) / 106.5 = 7.2\%$
Year 3 $(119.9 - 114.2) / 114.2 = 5.0\%$
Year 4 $(124.8 - 119.9) / 119.9 = 4.1\%$
Year 5 $(128.1 - 124.8) / 124.8 = 2.6\%$
The inflation rate is decreasing, but the price level is still increasing. This is best described as disinflation. (LOS 11.g)
4. **C** Core inflation is measured using a price index that excludes food and energy prices. (LOS 11.h)
5. **C** Adopting a chained price index method addresses substitution bias, as does using a Paasche index. Hedonic pricing adjusts for improvements in the quality of products over time, not substitution bias. (LOS 11.i)
6. **C** Both inflation scenarios can involve a decrease in short-run aggregate supply due to increasing wage demands. In a wage-push scenario, which is a form of cost-push inflation, the decrease in aggregate supply causes real GDP to fall below full employment. In a demand-pull inflation scenario, an increase in aggregate demand causes real GDP to increase beyond full employment, which creates wage pressure that results in a decrease in short-run aggregate supply. (LOS 11.j)

READING 12

MONETARY AND FISCAL POLICY

EXAM FOCUS

This reading covers the supply and demand for money, as well as fiscal and monetary policy. This is a lot of material, but you really need to get it all down to be prepared for the exam. Concentrate initially on all the definitions and the basics of expansionary and contractionary fiscal and monetary policy. When you read it the second time, try to understand every cause-and-effect relationship so you can trace the effects of a policy change through the economy. In this way, you will be able to answer questions about the effect of, for example, open market purchases of securities by the central bank on interest rates, consumption, saving, private investment, and, of course, real GDP in the short and long run. You should understand the role of the central bank in a developed economy, including its limitations in achieving its stated objectives.

MODULE 12.1: MONEY AND INFLATION



Video covering this content is available online.

LOS 12.a: Compare monetary and fiscal policy.

Fiscal policy refers to a government's use of spending and taxation to influence economic activity. The budget is said to be *balanced* when tax revenues equal government expenditures. A **budget surplus** occurs when government tax revenues exceed expenditures, and a **budget deficit** occurs when government expenditures exceed tax revenues.

Monetary policy refers to the central bank's actions that affect the quantity of money and credit in an economy in order to influence economic activity. Monetary policy is said to be **expansionary** (or *accommodative* or *easy*) when the central bank increases the quantity of money and credit in an economy. Conversely, when the central bank is reducing the quantity of money and credit in an economy, the monetary policy is said to be **contractionary** (or *restrictive* or *tight*).

Both monetary and fiscal policies are used by policymakers with the goals of maintaining stable prices and producing positive economic growth. Fiscal policy can also be used as a tool for redistribution of income and wealth.

LOS 12.b: Describe functions and definitions of money.

Money is most commonly defined as a generally accepted medium of exchange. Rather than exchanging goods and services directly (bartering), using money facilitates indirect exchange.

Money has three primary functions:

- Money serves as a **medium of exchange** or **means of payment** because it is accepted as payment for goods and services.
- Money also serves as a **unit of account** because prices of all goods and services are expressed in units of money: dollars, yen, rupees, pesos, and so forth. This allows us to determine how much of any good we are foregoing when consuming another.
- Money provides a **store of value** because money received for work or goods now can be saved to purchase goods later.

Narrow money is the amount of notes (currency) and coins in circulation in an economy plus balances in checkable bank deposits. **Broad money** includes narrow money plus any amount available in liquid assets, which can be used to make purchases.

Measures of money differ among monetary authorities, but there is consistency in that broad measures of money include money that is less liquid (immediately spendable) than that included in narrow money measures. We have included definitions of narrow and broad monetary aggregates used by the U.S. Federal Reserve and by the European Central Bank as examples.

According to the Federal Reserve Bank of New York:

The money supply measures reflect the different degrees of liquidity—or spendability—that different types of money have. The narrowest measure, M1, is restricted to the most liquid forms of money; it consists of currency in the hands of the public; travelers checks; demand deposits, and other deposits against which checks can be written. M2 includes M1, plus savings accounts, time deposits of under \$100,000, and balances in retail money market mutual funds.

The European Central Bank describes their monetary aggregates as follows:

	M1	M2	M3
Currency in circulation	X	X	X
Overnight deposits	X	X	X
Deposits with an agreed maturity of up to 2 years		X	X
Deposits redeemable at notice of up to 3 months		X	X
Repurchase agreements			X
Money market fund shares/units			X
Debt securities issued with a maturity of up to 2 years			X

LOS 12.c: Explain the money creation process.

In the early stages of money development, **promissory notes** were developed. When customers deposited gold (or other precious metal) with early bankers, they were issued a promissory note, which was a promise by the banker to return that gold on demand from the depositor. Promissory notes themselves then became a medium of exchange. Bankers, recognizing that all the deposits would never be withdrawn at the same time, started lending a portion of deposits to earn interest. This led to what is called **fractional reserve banking**.

In a fractional reserve banking system, a bank holds a proportion of deposits in reserve. In most countries, banks are required to hold a minimum percentage of deposits as reserves.

When cash is deposited in a bank, the portion that is not required to be held in reserve can be loaned out. When a bank makes a cash loan and the borrower spends the money, the sellers who receive this cash may deposit it in banks as well. These funds can now be loaned out by these banks, except for the portion that must be held as reserves by each bank. This process of lending, spending, and depositing can continue until deposits are some multiple of the original cash amount.

Consider a bank that has \$1,000 in **excess reserves** (cash not needed for reserves) that it lends. Assume the required reserve ratio is 25%. If the borrower of the \$1,000 deposits the cash in a second bank, the second bank will be able to lend its excess reserves of \$750 ($0.75 \times \$1,000$). Those funds may be deposited in a third bank, which can then lend its excess reserve of \$563 ($0.75 \times \750). If this lending and depositing continues, the money supply can expand to \$4,000 [$(1 / 0.25) \times \$1,000$]. One dollar of excess reserves can generate a \$4 increase in the money supply.

The total amount of money that can be created is calculated as:

$$\text{money created} = \frac{\text{new deposit}}{\text{reserve requirement}} = \frac{1,000}{0.25} = \$4,000$$

With 25% of deposits held as reserves, the original deposit can result in total deposits four times as large, and we say that the **money multiplier** is four.

$$\text{money multiplier} = \frac{1}{\text{reserve requirement}} = \frac{1}{0.25} = 4$$

If the required reserve percentage is decreased, the money multiplier increases, and the quantity of money that can be created increases. If the reserve requirement was reduced from 25% to 10%, the money multiplier would increase from 4 to 10.

Relationship of Money and the Price Level

The **quantity theory of money** states that quantity of money is some proportion of the total spending in an economy and implies the **quantity equation of exchange**:

$$\text{money supply} \times \text{velocity} = \text{price} \times \text{real output} \quad (MV = PY)$$

Price multiplied by real output is total spending so that **velocity** is the average number of times per year each unit of money is used to buy goods or services. The equation of exchange must hold with velocity defined in this way.

Monetarists believe that velocity and the real output of the economy change only slowly. Assuming that velocity and real output remain constant, any increase in the money supply will lead to a proportionate increase in the price level. For example, a 5% increase in the money supply will increase average prices by 5%. For this reason, monetarists argue that monetary policy can be used to control and regulate inflation. The belief that real variables (real GDP and velocity) are not affected by monetary variables (money supply and prices) is referred to as **money neutrality**.

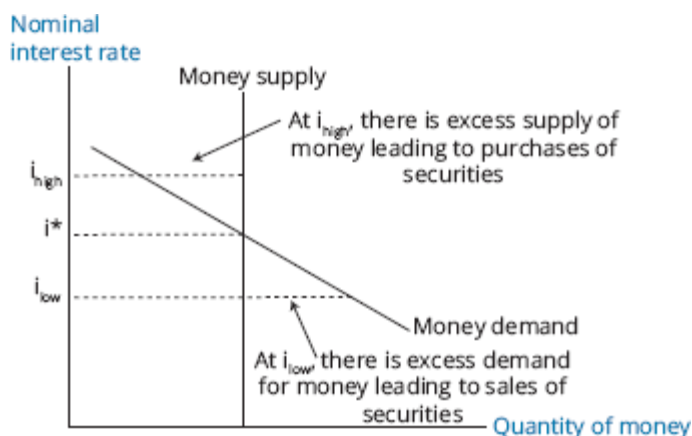
LOS 12.d: Describe theories of the demand for and supply of money.

The amount of wealth that households and firms in an economy choose to hold in the form of money is known as **demand for money**. There are three reasons for holding money:

1. *Transaction demand*: Money held to meet the need for undertaking transactions. As the level of real GDP increases, the size and number of transactions will increase, and the demand for money to carry out transactions increases.
2. *Precautionary demand*: Money held for unforeseen future needs. The demand for money for precautionary reasons is higher for large firms. In the aggregate, the total amount of precautionary demand for money increases with the size of the economy.
3. *Speculative demand*: Money that is available to take advantage of investment opportunities that arise in the future. It is inversely related to returns available in the market. As bonds and other financial instruments provide higher returns, investors would rather invest their money now than hold speculative money balances. Conversely, the demand for money for speculative reasons is positively related to perceived risk in other financial instruments. If the risk is perceived to be higher, people choose to hold money rather than invest it.

The relation between short-term interest rates and the quantity of money that firms and households demand to hold is illustrated in Figure 12.1. At lower interest rates, firms and households choose to hold more money. At higher interest rates, the opportunity cost of holding money increases, and firms and households will desire to hold less money and more interest-bearing financial assets.

Figure 12.1: The Supply and Demand for Money



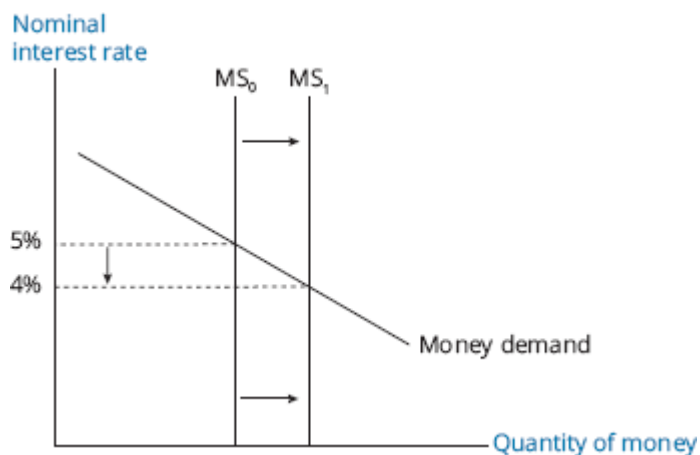
The **supply of money** is determined by the central bank (the Fed in the United States) and is independent of the interest rate. This accounts for the vertical (perfectly inelastic) supply curve in Figure 12.1.

Short-term interest rates are determined by the equilibrium between money supply and money demand. As illustrated in Figure 12.1, if the interest rate is above the equilibrium rate (i_{high}), there is excess supply of real money. Firms and households are holding more real money balances than they desire to, given the opportunity cost of holding money balances. They will

purchase securities to reduce their money balances, which will decrease the interest rate as securities prices are bid up. If interest rates are below equilibrium (i_{low}), there is excess demand for real money balances, as illustrated in Figure 12.1. Firms and households will sell securities to increase their money holdings to the desired level, decreasing securities prices and increasing the interest rate.

A central bank can affect short-term interest rates by increasing or decreasing the money supply. An increase in the money supply (shift of the money supply curve to the right) will put downward pressure on interest rates, as illustrated in Figure 12.2. With an increase in the money supply, there is excess supply of money at the previous rate of 5%. To reduce their money holdings, firms and households buy securities, increasing securities prices and decreasing the interest rate until the new equilibrium interest rate of 4% is achieved. If the central bank decreases the money supply, excess demand for money balances results in sales of securities and an increase in the interest rate.

Figure 12.2: Increase in the Money Supply



LOS 12.e: Describe the Fisher effect.

The **Fisher effect** states that the nominal interest rate is simply the sum of the real interest rate and expected inflation.

$$R_{Nom} = R_{Real} + E[I]$$

where:

R_{Nom} = nominal interest rate

R_{Real} = real interest rate

$E[I]$ = expected inflation

The idea behind the Fisher effect is that real rates are relatively stable, and changes in interest rates are driven by changes in expected inflation. This is consistent with money neutrality.

Investors are exposed to the risk that inflation and other future outcomes may be different than expected. Investors require additional return (a risk premium) for bearing this risk, which we can consider a third component of a nominal interest rate.

$$R_{\text{Nom}} = R_{\text{Real}} + E[I] + RP$$

where:

RP = risk premium for uncertainty

LOS 12.f: Describe roles and objectives of central banks.

There are several key **roles of central banks**:

1. *Sole supplier of currency*: Central banks have the sole authority to supply money. Traditionally, such money was backed by gold; the central bank stood ready to convert the money into a pre-specified quantity of gold. Later on, the gold backing was removed, and money supplied by the central bank was deemed **legal tender** by law. Money not backed by any tangible value is termed **fiat money**. As long as fiat money holds its value over time and is acceptable for transactions, it can continue to serve as a medium of exchange.
2. *Banker to the government and other banks*: Central banks provide banking services to the government and other banks in the economy.
3. *Regulator and supervisor of payments system*: In many countries, central banks may regulate the banking system by imposing standards of risk-taking allowed and reserve requirements of banks under its jurisdiction. Central banks also oversee the payments system to ensure smooth operations of the clearing system domestically and in conjunction with other central banks for international transactions.
4. *Lender of last resort*: Central banks' ability to print money allows them to supply money to banks with shortages, and this government backing tends to prevent runs on banks (i.e., large scale withdrawals) by assuring depositors their funds are secure.
5. *Holder of gold and foreign exchange reserves*: Central banks are often the repositories of the nation's gold and reserves of foreign currencies.
6. *Conductor of monetary policy*: Central banks control or influence the quantity of money supplied in an economy and growth of money supply over time.

The primary **objective of a central bank** is to *control inflation* so as to promote price stability. High inflation is not conducive to a stable economic environment. High inflation leads to **menu costs** (i.e., cost to businesses of constantly having to change their prices) and **shoe leather costs** (i.e., costs to individuals of making frequent trips to the bank so as to minimize their holdings of cash that are depreciating in value due to inflation).

In addition to price stability, some central banks have other stated goals, such as:

- Stability in exchange rates with foreign currencies.
- Full employment.
- Sustainable positive economic growth.
- Moderate long-term interest rates.

The target inflation rate in most developed countries is a range around 2% to 3%. A target of zero inflation is not used because that increases the risk of deflation, which can be very disruptive for an economy.

While most developed countries have an explicit target inflation rate, the U.S. Fed and the Bank of Japan do not. In the United States, this is because the Fed has the additional goals of maximum employment and moderate long-term interest rates. In Japan, it is because deflation, rather than inflation, has been a persistent problem in recent years.

Some developed countries, and several developing countries, choose a target level for the exchange rate of their currency with that of another country, primarily the U.S. dollar. This is referred to as **pegging** their exchange rate with the dollar. If their currency appreciates (i.e., becomes relatively more valuable), they can sell their domestic currency reserves for dollars to reduce the exchange rate. While such actions may be effective in the short run, for stability of the exchange rate over time, the monetary authorities in the pegging country must manage interest rates and economic activity to achieve their goal. This can lead to increased volatility of their money supply and interest rates. The pegging country essentially commits to a policy intended to make its inflation rate equal to the inflation rate of the country to which they peg their currency.

LOS 12.g: Contrast the costs of expected and unexpected inflation.

We turn our attention now to the costs to an economy of inflation, why central banks' target inflation rates are low, and why they care about volatility of inflation rates. At any point in time, economic agents have an expected rate of future inflation in the aggregate. The costs of inflation that is equal to the expected rate are different from the costs of inflation that differs from expectations, with the costs imposed on an economy of unanticipated inflation greater than those of perfectly anticipated inflation.

Consider an economy for which expected inflation is 6% and actual inflation will be 6% with certainty, so that inflation is *perfectly anticipated* (i.e., there is no unexpected inflation). The prices of all goods and wages could be indexed to this inflation rate so each month both wages and prices are increased approximately one-half percent. Increased demand for a product would result in monthly price increases of more than one-half percent and decreased demand would be reflected in prices that increased less than one-half percent per month.

One effect of high inflation—even when perfectly anticipated—is that the cost of holding money rather than interest-bearing securities is higher because its purchasing power decreases steadily. This will decrease the quantity of money that people willingly hold and impose some costs of more frequent movement of money from interest-bearing securities to cash or non-interest-bearing deposit accounts to facilitate transactions. To some extent, technology and the Internet have decreased these costs as movement of money between accounts has become much easier.

Much more important are the costs imposed on an economy by *unanticipated inflation*, inflation that is higher or lower than the expected rate of inflation. When inflation is higher than expected, borrowers gain at the expense of lenders as loan payments in the future are made with currency that has less value in real terms. Conversely, inflation that is less than expected will benefit lenders at the expense of borrowers. In an economy with volatile (rather than certain) inflation rates, lenders will require higher interest rates to compensate for the additional risk

they face from unexpected changes in inflation. Higher borrowing rates slow business investment and reduce the level of economic activity.

A second cost of unexpected inflation is that information about supply and demand from changes in prices becomes less reliable. Suppose that when expected inflation is 5%, a manufacturer sees that prices for his product have increased 10%. If this is interpreted as an increase in demand for the product, the manufacturer will increase capacity and production in response to the perceived increase in demand. If, in fact, general price inflation is 10% rather than the expected 5% over the recent period, the price increase in the manufacturer's product did not result from an increase in demand. The expansion of production will result in excess inventory and capacity, and the firm will decrease production, laying off workers and reducing or eliminating expenditures on increased capacity for some time. Because of these effects, unexpected inflation can increase the magnitude or frequency of business cycles. The destabilizing effects of inflation, either higher than expected or lower than expected, because of reduced information content of price changes impose real costs on an economy.



MODULE QUIZ 12.1

- Both monetary and fiscal policy are used to:
 - balance the budget.
 - achieve economic targets.
 - redistribute income and wealth.
- Which of the following statements is *least accurate*? The existence and use of money:
 - permits individuals to perform economic transactions.
 - requires the central bank to control the supply of currency.
 - increases the efficiency of transactions compared to a barter system.
- If money neutrality holds, the effect of an increase in the money supply is:
 - higher prices.
 - higher output.
 - lower unemployment.
- If the money supply is increasing and velocity is decreasing:
 - prices will decrease.
 - real GDP will increase.
 - the impact on prices and real GDP is uncertain.
- The money supply curve is perfectly inelastic because the money:
 - supply is independent of interest rates.
 - demand schedule is downward-sloping.
 - supply is dependent upon interest rates.
- The Fisher effect states that the nominal interest rate is equal to the real rate plus:
 - actual inflation.
 - average inflation.
 - expected inflation.
- A central bank's policy goals *least likely* include:
 - price stability.
 - minimizing long-term interest rates.
 - maximizing the sustainable growth rate of the economy.
- A country that targets a stable exchange rate with another country's currency *least likely*:
 - accepts the inflation rate of the other country.
 - will sell its currency if its foreign exchange value rises.
 - must also match the money supply growth rate of the other country.

MODULE 12.2: MONETARY POLICY



Video covering this content is available online.

LOS 12.h: Describe tools used to implement monetary policy.

Monetary policy is implemented using the **monetary policy tools** of the central bank. The three main policy tools of central banks are as follows:

1. *Policy rate*: In the United States, banks can borrow funds from the Fed if they have temporary shortfalls in reserves. The rate at which banks can borrow reserves from the Fed is termed the *discount rate*. For the European Central Bank (ECB), it is called the *refinancing rate*.

One way to lend money to banks is through a *repurchase agreement*. The central bank purchases securities from banks that, in turn, agree to repurchase the securities at a higher price in the future. The percentage difference between the purchase price and the repurchase price is effectively the rate at which the central bank is lending to member banks. The Bank of England uses this method, and its policy rate is called the *two-week repo (repurchase) rate*. A lower rate reduces banks' cost of funds, encourages lending, and tends to decrease interest rates overall. A higher policy rate has the opposite effect, decreasing lending and increasing interest rates.

In the United States, the *federal funds rate* is the rate that banks charge each other on overnight loans of reserves. The Fed sets a target for this market-determined rate and uses open market operations to move it to the target rate.

2. *Reserve requirements*: By increasing the reserve requirement (the percentage of deposits banks are required to retain as reserves), the central bank effectively decreases the funds that are available for lending and the money supply, which will tend to increase interest rates. A decrease in the reserve requirement will increase the funds available for lending and the money supply, which will tend to decrease interest rates. This tool only works well to increase the money supply if banks are willing to lend and customers are willing to borrow.
3. *Open market operations*: Buying and selling of securities by the central bank is referred to as open market operations. When the central bank buys securities, cash replaces securities in investor accounts, banks have excess reserves, more funds are available for lending, the money supply increases, and interest rates decrease. Sales of securities by the central bank have the opposite effect, reducing cash in investor accounts, excess reserves, funds available for lending, and the money supply, which will tend to cause interest rates to increase. In the United States, open market operations are the Fed's most commonly used tool and are important in achieving the federal funds target rate.

LOS 12.i: Describe the monetary transmission mechanism.

The **monetary transmission mechanism** refers to the ways in which a change in monetary policy, specifically the central bank's policy rate, affects the price level and inflation. There are four channels through which a change in the policy rates the monetary authorities control directly are transmitted to prices. They are transmitted through their effect on other short-term

rates, asset values, currency exchange rates, and expectations. We can examine the transmission mechanism in more detail by considering the effects of a change to a contractionary monetary policy implemented through an increase in the policy rate.

- Banks' *short-term lending rates will increase* in line with the increase in the policy rate. The higher rates will decrease aggregate demand as consumers reduce credit purchases and businesses cut back on investment in new projects.
- Bond prices, equity prices, and *asset prices in general will decrease* as the discount rates applied to future expected cash flows are increased. This may have a wealth effect because a decrease in the value of households' assets may increase the savings rate and decrease consumption.
- Both consumers and businesses may decrease their expenditures because their *expectations for future economic growth decrease*.
- The increase in interest rates may attract foreign investment in debt securities, leading to an *appreciation of the domestic currency relative to foreign currencies*. An appreciation of the domestic currency increases the foreign currency prices of exports and can reduce demand for the country's export goods.

Taken together, these effects act to decrease aggregate demand and put downward pressure on the price level. A decrease in the policy rate would affect the price level through the same channels, but in the opposite direction.

LOS 12.j: Explain the relationships between monetary policy and economic growth, inflation, interest, and exchange rates.

If money neutrality holds, changes in monetary policy and the policy rate will have no effect on real output. In the short run, however, changes in monetary policy can affect real economic growth as well as interest rates, inflation, and foreign exchange rates. The effects of a change to a more expansionary monetary policy may include any or all of the following:

- The central bank buys securities, which increases bank reserves.
- The interbank lending rate decreases as banks are more willing to lend each other reserves.
- Other short-term rates decrease as the increase in the supply of loanable funds decreases the equilibrium rate for loans.
- Longer-term interest rates also decrease.
- The decrease in real interest rates causes the currency to depreciate in the foreign exchange market.
- The decrease in long-term interest rates increases business investment in plant and equipment.
- Lower interest rates cause consumers to increase their purchases of houses, autos, and durable goods.
- Depreciation of the currency increases foreign demand for domestic goods.
- These increases in consumption, investment, and net exports all increase aggregate demand.
- The increase in aggregate demand increases inflation, employment, and real GDP.

The transmission mechanism for a decrease in interbank lending rates affects four things simultaneously:

1. Market rates decrease due to banks adjusting their lending rates for the short and long term.
2. Asset prices increase because lower discount rates are used for computing present values.
3. Firms and individuals raise their expectations for economic growth and profitability. They may also expect the central bank to follow up with further interest rate decreases.
4. The domestic currency depreciates due to an outflow of foreign money as real interest rates decline.

Together, these four factors increase domestic demand as people consume more (they have less incentive to save given lower interest rates) and increase net external demand (exports minus imports) because depreciation of the domestic currency makes exports less expensive to foreigners and imports more expensive in the domestic economy. The increase in overall demand and import prices tends to increase aggregate demand and domestic inflation.

LOS 12.k: Describe qualities of effective central banks.

For a central bank to succeed in its inflation-targeting policies, it should have **three essential qualities**:

1. *Independence*: For a central bank to be effective in achieving its goals, it should be free from political interference. Reducing the money supply to reduce inflation can also be expected to decrease economic growth and employment. The political party in power has an incentive to boost economic activity and reduce unemployment prior to elections. For this reason, politicians may interfere with the central bank's activities, compromising its ability to manage inflation. Independence should be thought of in relative terms (degrees of independence) rather than absolute terms. Even in the case of relatively independent central banks, the heads of the banks may be appointed by politicians.

Independence can be evaluated based on both **operational independence** and **target independence**. Operational independence means that the central bank is allowed to independently determine the policy rate. Target independence means the central bank also defines how inflation is computed, sets the target inflation level, and determines the horizon over which the target is to be achieved. The ECB has both target and operational independence, while most other central banks have only operational independence.

2. *Credibility*: To be effective, central banks should follow through on their stated intentions. If a government with large debts, instead of a central bank, set an inflation target, the target would not be credible because the government has an incentive to allow inflation to exceed the target level. On the other hand, a credible central bank's targets can become self-fulfilling prophecies. If the market believes that a central bank is serious about achieving a target inflation rate of 3%, wages and other nominal contracts will be based on 3% inflation, and actual inflation will then be close to that level.
3. *Transparency*: Transparency on the part of central banks aids their credibility. Transparency means central banks periodically disclose the state of the economic environment by issuing

inflation reports. Transparent central banks periodically report their views on the economic indicators and other factors they consider in their interest rate setting policy. When a central bank makes clear the economic indicators that it uses in establishing monetary policy and how they will be used, it not only gains credibility but makes policy changes easier to anticipate and implement.

LOS 12.I: Contrast the use of inflation, interest rate, and exchange rate targeting by central banks.

Central banks have used various economic variables and indicators over the years to make monetary policy decisions. In the past, some have used **interest rate targeting**, increasing the money supply when specific interest rates rose above the target band and decreasing the money supply (or the rate of money supply growth) when rates fell below the target band. Currently, **inflation targeting** is the most widely used tool for making monetary policy decisions and is, in fact, the method required by law in some countries. Central banks that currently use inflation targeting include the U.K., Brazil, Canada, Australia, Mexico, and the European Central Bank.

The most common inflation rate target is 2%, with a permitted deviation of $\pm 1\%$ so the target band is 1% to 3%. The reason the inflation target is not 0% is that variations around that rate would allow for negative inflation (i.e., deflation), which is considered disruptive to the smooth functioning of an economy. Central banks are not necessarily targeting current inflation, which is the result of prior policy and events, but inflation in the range of two years in the future.

Some countries, especially developing countries, use **exchange rate targeting**. That is, they target a foreign exchange rate between their currency and another (often the U.S. dollar), rather than targeting inflation. As an example, consider a country that has targeted an exchange rate for its currency versus the U.S. dollar. If the foreign exchange value of the domestic currency falls relative to the U.S. dollar, the monetary authority must use foreign reserves to purchase their domestic currency (which will reduce money supply growth and increase interest rates) in order to reach the target exchange rate. Conversely, an increase in the foreign exchange value of the domestic currency above the target rate will require sale of the domestic currency in currency markets to reduce its value (increasing the domestic money supply and decreasing interest rates) to move towards the target exchange rate. One result of exchange rate targeting may be greater volatility of the money supply because domestic monetary policy must adapt to the necessity of maintaining a stable foreign exchange rate.

Over the short term, the targeting country can purchase or sell its currency in the foreign exchange markets to influence the exchange rate. There are limits, however, on how much influence currency purchases or sales can have on exchange rates over time. For example, a country may run out of foreign reserves with which to purchase its currency when the exchange value of its currency is still below the target exchange rate.

The net effect of exchange rate targeting is that the targeting country will have the same inflation rate as the targeted currency and the targeting country will need to follow monetary policy and accept interest rates that are consistent with this goal, regardless of domestic economic circumstances.

LOS 12.m: Determine whether a monetary policy is expansionary or contractionary.

An economy's long-term sustainable real growth rate is called the **real trend rate** or, simply, the trend rate. The trend rate is not directly observable and must be estimated. The trend rate also changes over time as structural conditions of the economy change. For example, after a prolonged period of heavy debt use, consumers may increase saving and reduce consumption in order to reduce their levels of debt. This structural shift in the economy would reduce the trend growth rate.

The **neutral interest rate** of an economy is the growth rate of the money supply that neither increases nor decreases the economic growth rate:

$$\text{neutral interest rate} = \text{real trend rate of economic growth} + \text{inflation target}$$

When the policy rate is above (below) the neutral rate, the monetary policy is said to be **contractionary (expansionary)**. In general, contractionary policy is associated with a decrease in the *growth rate* of money supply, while expansionary policy increases its growth rate.

Monetary policy is often adjusted to reflect the source of inflation. For example, if inflation is above target due to higher aggregate demand (consumer and business spending), then contractionary monetary policy may be an appropriate response to reduce inflation. Suppose, however, that inflation is higher due to supply shocks, such as higher food or energy prices, and the economy is already operating below full employment. In such a situation, a contractionary monetary policy may make a bad situation worse.



PROFESSOR'S NOTE

In the United States, the Federal Reserve focuses on core inflation (i.e., excluding volatile food and energy prices) for this reason.

LOS 12.n: Describe limitations of monetary policy.

This transmission mechanism for monetary policy previously described does not always produce the intended results. In particular, long-term rates may not rise and fall with short-term rates because of the effect of monetary policy changes on expected inflation.

If individuals and businesses believe that a decrease in the money supply intended to reduce inflation will be successful, they will expect lower future inflation rates. Because long-term bond yields include a premium for expected inflation, long-term rates could fall (tending to increase economic growth), even while the central bank has increased short-term rates in order to slow economic activity. Conversely, increasing the money supply to stimulate economic activity could lead to an increase in expected inflation rates and long-term bond yields, even as short-term rates fall.

From a different perspective, monetary tightening may be viewed as too extreme, increasing the probability of a recession, making long-term bonds more attractive and reducing long-term interest rates. If money supply growth is seen as inflationary, higher expected future asset prices will make long-term bonds relatively less attractive and will increase long-term interest rates.

Bond market participants that act in this way have been called **bond market vigilantes**. When the central bank's policy is credible and investors believe that the inflation target rate will be maintained over time, this effect on long-term rates will be small.

Another situation in which the transmission mechanism may not perform as expected is if demand for money becomes very elastic and individuals willingly hold more money even without a decrease in short-term rates. Such a situation is called a **liquidity trap**. Increasing growth of the money supply will not decrease short-term rates under these conditions because individuals hold the money in cash balances instead of investing in interest-bearing securities. If an economy is experiencing deflation even though money supply policy has been expansionary, liquidity trap conditions may be present.

Compared to inflation, deflation is more difficult for central banks to reverse. In a deflationary environment, monetary policy needs to be expansionary. However, the central bank is limited to reducing the nominal policy rate to zero. Once it reaches zero, the central bank has limited ability to further stimulate the economy.

Another reason standard tools for increasing the money supply might not increase economic activity is that even with increasing excess reserves, banks may not be willing to lend. When what has become known as the *credit bubble* collapsed in 2008, banks around the world lost equity capital and desired to rebuild it. For this reason, they decreased their lending, even as money supplies were increased and short-term rates fell. With short-term rates near zero, economic growth still poor, and a real threat of deflation, central banks began a policy termed **quantitative easing**.

In the United Kingdom, quantitative easing entailed large purchases of British government bonds in the maturity range of three to five years. The intent was to reduce interest rates to encourage borrowing and to generate excess reserves in the banking system to encourage lending. Uncertainty about the economy's future caused banks to behave quite conservatively and willingly hold more excess reserves, rather than make loans.

In the United States, billions of dollars were made available for the Fed to buy assets other than short-term Treasury securities. Large amounts of mortgage securities were purchased from banks to encourage bank lending and to reduce mortgage rates in an attempt to revive the housing market, which had collapsed. When this program did not have the desired effect, a second round of quantitative easing (QE2) was initiated. The Fed purchased long-term Treasury bonds in large quantities (hundreds of billions of dollars) with the goal of bringing down longer-term interest rates and generating excess reserves to increase lending and economic growth. The Fed has also purchased securities with credit risk as part of its quantitative easing, improving banks' balance sheets but perhaps just shifting risk from the private sector to the public sector.

Monetary Policy in Developing Economies

Developing countries face problems in successfully implementing monetary policy. Without a liquid market in their government debt interest rate, information may be distorted and open market operations difficult to implement. In a very rapidly developing economy it may be quite difficult to determine the neutral rate of interest for policy purposes. Rapid financial innovation may change the demand to hold monetary aggregates. Central banks may lack credibility

because of past failure to maintain inflation rates in a target band and may not be given independence by the political authority.



MODULE QUIZ 12.2

1. A central bank conducts monetary policy primarily by altering:
 - A. the policy rate.
 - B. the inflation rate.
 - C. the long-term interest rate.
2. Purchases of securities in the open market by the monetary authorities are *least likely* to increase:
 - A. excess reserves.
 - B. cash in investor accounts.
 - C. the interbank lending rate.
3. An increase in the policy rate will *most likely* lead to an increase in:
 - A. business investment in fixed assets.
 - B. consumer spending on durable goods.
 - C. the foreign exchange value of the domestic currency.
4. Qualities of effective central banks include:
 - A. credibility and verifiability.
 - B. comparability and relevance.
 - C. independence and transparency.
5. If a country's inflation rate is below the central bank's target rate, the central bank is *most likely* to:
 - A. sell government securities.
 - B. increase the reserve requirement.
 - C. decrease the overnight lending rate.
6. Monetary policy is likely to be *least* responsive to domestic economic conditions if policymakers employ:
 - A. inflation targeting.
 - B. interest rate targeting.
 - C. exchange rate targeting.
7. Suppose an economy has a real trend rate of 2%. The central bank has set an inflation target of 4.5%. To achieve the target, the central bank has set the policy rate at 6%. Monetary policy is *most likely*:
 - A. balanced.
 - B. expansionary.
 - C. contractionary.
8. Monetary policy is *most likely* to fail to achieve its objectives when the economy is:
 - A. growing rapidly.
 - B. experiencing deflation.
 - C. experiencing disinflation.

MODULE 12.3: FISCAL POLICY



LOS 12.o: Describe roles and objectives of fiscal policy.

Video covering this content is available online.

Fiscal policy refers to a government's use of spending and taxation to meet macroeconomic goals. A government budget is said to be *balanced* when tax revenues equal

government expenditures. A *budget surplus* occurs when government tax revenues exceed expenditures, and a *budget deficit* occurs when government expenditures exceed tax revenues.

In general, decreased taxes and increased government spending both *increase* a budget deficit, overall demand, economic growth, and employment. Increased taxes and decreased government spending *decrease* a budget deficit, overall demand, economic growth, and employment. Budget deficits are increased in response to recessions, and budget deficits are decreased to slow growth when inflation is too high.

Keynesian economists believe that fiscal policy, through its effect on aggregate demand, can have a strong effect on economic growth when the economy is operating at less than full employment. Monetarists believe that the effect of fiscal stimulus is only temporary and that monetary policy should be used to increase or decrease inflationary pressures over time. Monetarists do not believe that monetary policy should be used in an attempt to influence aggregate demand to counter cyclical movements in the economy.

Discretionary fiscal policy refers to the spending and taxing decisions of a national government that are intended to stabilize the economy. In contrast, **automatic stabilizers** are built-in fiscal devices triggered by the state of the economy. For example, during a recession, tax receipts will fall, and government expenditures on unemployment insurance payments will increase. Both of these tend to increase budget deficits and are expansionary. Similarly, during boom times, higher tax revenues coupled with lower outflows for social programs tend to decrease budget deficits and are contractionary.

Objectives of fiscal policy may include:

- Influencing the level of economic activity and aggregate demand.
- Redistributing wealth and income among segments of the population.
- Allocating resources among economic agents and sectors in the economy.

LOS 12.p: Describe the arguments about whether the size of a national debt relative to GDP matters.

When a government runs fiscal deficits, it incurs debt that needs to be repaid as well as ongoing interest expense. Total deficits, annual deficits, and interest expense can all be evaluated relative to annual GDP. When these ratios increase beyond certain levels, it may be a cause for concern, and the solvency of the country may be questioned.

A country's **debt ratio** is the ratio of aggregate debt to GDP. Because taxes are linked to GDP, when an economy grows in real terms, tax revenues will also grow in real terms. If the real interest rate on the government's debt is higher than the real growth rate of the economy, then the debt ratio will increase over time (keeping tax rates constant). Similarly, if the real interest rate on government's debt is lower than real growth in GDP, the debt ratio will decrease (i.e., improve) over time.

Arguments *for* being concerned with the size of fiscal deficit:

- Higher deficits lead to higher future taxes. Higher future taxes will lead to disincentives to work and entrepreneurship. This leads to lower long-term economic growth.

- If markets lose confidence in the government, investors may not be willing to refinance the debt. This can lead to the government defaulting (if debt is in a foreign currency) or having to simply print money (if the debt is in local currency). Printing money would ultimately lead to higher inflation.
- Increased government borrowing will tend to increase interest rates, and firms may reduce their borrowing and investment spending as a result, decreasing the impact on aggregate demand of deficit spending. This is referred to as the **crowding-out effect** because government borrowing is taking the place of private sector borrowing.

Arguments *against* being concerned with the size of fiscal deficit:

- If the debt is primarily being held by domestic citizens, the scale of the problem is overstated.
- If the debt is used to finance productive capital investment, future economic gains will be sufficient to repay the debt.
- Fiscal deficits may prompt needed tax reform.
- Deficits would not matter if private sector savings in anticipation of future tax liabilities just offsets the government deficit (Ricardian equivalence holds).
- If the economy is operating at less than full capacity, deficits do not divert capital away from productive uses. On the contrary, deficits can aid in increasing GDP and employment.

LOS 12.q: Describe tools of fiscal policy, including their advantages and disadvantages.

Fiscal policy tools include spending tools and revenue tools.

Spending Tools

Transfer payments, also known as entitlement programs, redistribute wealth, taxing some and making payments to others. Examples include Social Security and unemployment insurance benefits. Transfer payments are not included in GDP computations.

Current spending refers to government purchases of goods and services on an ongoing and routine basis.

Capital spending refers to government spending on infrastructure, such as roads, schools, bridges, and hospitals. Capital spending is expected to boost future productivity of the economy.

Justification for spending tools:

- Provide services such as national defense that benefit all the residents in a country.
- Invest in infrastructure to enhance economic growth.
- Support the country's growth and unemployment targets by directly affecting aggregate demand.
- Provide a minimum standard of living.
- Subsidize investment in research and development for certain high-risk ventures consistent with future economic growth or other goals (e.g., green technology).

Revenue Tools

Direct taxes are levied on income or wealth. These include income taxes, taxes on income for national insurance, wealth taxes, estate taxes, corporate taxes, capital gains taxes, and Social Security taxes. Some progressive taxes (such as income and wealth taxes) generate revenue for wealth and income redistributing.

Indirect taxes are levied on goods and services. These include sales taxes, value-added taxes (VATs), and excise taxes. Indirect taxes can be used to reduce consumption of some goods and services (e.g., alcohol, tobacco, gambling).

Desirable attributes of tax policy:

- Simplicity to use and enforce.
- Efficiency; having the least interference with market forces and not acting as a deterrent to working.
- Fairness is quite subjective, but two commonly held beliefs are:
 - Horizontal equality: people in similar situations should pay similar taxes.
 - Vertical equality: richer people should pay more in taxes.
- Sufficiency, in that taxes should generate sufficient revenues to meet the spending needs of the government.

Advantages of fiscal policy tools:

- Social policies, such as discouraging tobacco use, can be implemented very quickly via indirect taxes.
- Quick implementation of indirect taxes also means that government revenues can be increased without significant additional costs.

Disadvantages of fiscal policy tools:

- Direct taxes and transfer payments take time to implement, delaying the impact of fiscal policy.
- Capital spending also takes a long time to implement. The economy may have recovered by the time its impact is felt.

Announcing a change in fiscal policy may have significant effects on expectations. For example, an announcement of future increase in taxes may immediately reduce current consumption, rapidly producing the desired goal of reducing aggregate demand. Note that not all fiscal policy tools affect economic activity equally. Spending tools are most effective in increasing aggregate demand. Tax reductions are somewhat less effective, as people may not spend the entire amount of the tax savings. Tax reductions for those with low incomes will be more effective in increasing aggregate demand, as those with lower incomes tend to spend a larger proportion of income on consumption; that is, they save a smaller proportion of income and have a higher marginal propensity to consume.

Fiscal Multiplier

Changes in government spending have magnified effects on aggregate demand because those whose incomes increase from increased government spending will in turn increase their spending, which increases the incomes and spending of others. The magnitude of the *multiplier effect* depends on the tax rate and on the marginal propensity to consume.

To understand the calculation of the multiplier effect, consider an increase in government spending of \$100 when the MPC is 80%, and the tax rate is 25%. The increase in spending increases incomes by \$100, but \$25 (100×0.25) of that will be paid in taxes. **Disposable income** is equal to income after taxes, so disposable income increases by $\$100 \times (1 - 0.25) = \75 . With an MPC of 80%, additional spending by those who receive the original \$100 increase is $\$75 \times 0.8 = \60 .

This additional spending will increase others' incomes by \$60 and disposable incomes by $\$60 \times 0.75 = \45 , from which they will spend $\$45 \times 0.8 = \36 .

Because each iteration of this process reduces the amount of additional spending, the effect reaches a limit. The **fiscal multiplier** determines the potential increase in aggregate demand resulting from an increase in government spending:

$$\text{fiscal multiplier} = \frac{1}{1 - \text{MPC}(1 - t)}$$

Here, with a tax rate of 25% and an MPC of 80%, the fiscal multiplier is $1 / [1 - 0.8(1 - 0.25)] = 2.5$, and the increase of \$100 in government spending has the potential to increase aggregate demand by \$250.

The fiscal multiplier is inversely related to the tax rate (higher tax rate decreases the multiplier) and directly related to the marginal propensity to consume (higher MPC increases the multiplier).

Balanced Budget Multiplier

In order to balance the budget, the government could increase taxes by \$100 to just offset a \$100 increase in spending. Changes in taxes also have a magnified effect on aggregate demand. An increase in taxes will decrease disposable income and consumption expenditures, thereby decreasing aggregate demand. The initial decrease in spending from a tax increase of \$100 is $100 \times \text{MPC} = 100 \times 0.8 = \80 ; beyond that, the multiplier effect is the same as we described for a direct increase in government spending, and the overall decrease in aggregate demand for a \$100 tax increase is $100(\text{MPC}) \times \text{fiscal multiplier}$, or, for our example, $100(0.8)(2.5) = \$200$.

Combining the total increase in aggregate demand from a \$100 increase in government spending with the total decrease in aggregate demand from a \$100 tax increase shows that the net effect on aggregate demand of both is an increase of $\$250 - \$200 = \$50$, so we can say that the balanced budget multiplier is positive.

If instead of a \$100 increase in taxes, we increased taxes by $100 / \text{MPC} = 100 / 0.8 = \125 and increased government spending by \$100, the net effect on aggregate demand would be zero.

Ricardian Equivalence

Increases in the current deficit mean greater taxes in the future. To maintain their preferred pattern of consumption over time, taxpayers may increase current savings (reduce current consumption) in order to offset the expected cost of higher future taxes. If taxpayers reduce current consumption and increase current saving by just enough to repay the principal and interest on the debt the government issued to fund the increased deficit, there is no effect on aggregate demand. This is known as **Ricardian equivalence** after economist David Ricardo. If taxpayers underestimate their future liability for servicing and repaying the debt, so that

aggregate demand is increased by equal spending and tax increases, Ricardian equivalence does not hold. Whether it does is an open question.

LOS 12.r: Explain the implementation of fiscal policy and difficulties of implementation.

Fiscal policy is implemented through changes in taxes and spending. This is called **discretionary fiscal policy** (as opposed to automatic stabilizers discussed previously). Discretionary fiscal policy would be designed to be expansionary when the economy is operating below full employment. Fiscal policy aims to stabilize aggregate demand. During recessions, actions can be taken to increase government spending or decrease taxes. Either change tends to strengthen the economy by increasing aggregate demand, putting more money in the hands of corporations and consumers to invest and spend. During inflationary economic booms, actions can be taken to decrease government spending or increase taxes. Either change tends to slow the economy by decreasing aggregate demand, taking money out of the hands of corporations and consumers, causing both investment and consumption spending to fall.

Discretionary fiscal policy is not an exact science. First, economic forecasts might be wrong, leading to incorrect policy decisions. Second, complications arise in practice that delay both the implementation of discretionary fiscal policy and the impact of policy changes on the economy. The lag between recessionary or inflationary conditions in the economy and the impact on the economy of fiscal policy changes can be divided into three types:

- **Recognition lag:** Discretionary fiscal policy decisions are made by a political process. The state of the economy is complex, and it may take policymakers time to recognize the nature and extent of the economic problems.
- **Action lag:** The time governments take to discuss, vote on, and enact fiscal policy changes.
- **Impact lag:** The time between the enactment of fiscal policy changes and when the impact of the changes on the economy actually takes place. It takes time for corporations and individuals to act on the fiscal policy changes, and fiscal multiplier effects occur only over time as well.

These lags can actually make fiscal policy counterproductive. For example, if the economy is in a recession phase, fiscal stimulus may be deemed appropriate. However, by the time fiscal stimulus is implemented and has its full impact, the economy may already be on a path to a recovery driven by the private sector.

Additional macroeconomic issues may hinder usefulness of fiscal policy:

- *Misreading economic statistics:* The full employment level for an economy is not precisely measurable. If the government relies on expansionary fiscal policy mistakenly at a time when the economy is already at full capacity, it will simply drive inflation higher.
- *Crowding-out effect:* Expansionary fiscal policy may crowd out private investment, reducing the impact on aggregate demand.
- *Supply shortages:* If economic activity is slow due to resource constraints (low availability of labor or other resources) and not due to low demand, expansionary fiscal policy will fail to achieve its objective and will probably lead to higher inflation.

- *Limits to deficits:* There is a limit to expansionary fiscal policy. If the markets perceive that the deficit is already too high as a proportion of GDP, funding the deficit will be problematic. This could lead to higher interest rates and actually make the situation worse.
 - *Multiple targets:* If the economy has high unemployment coupled with high inflation, fiscal policy cannot address both problems simultaneously.
-

LOS 12.s: Determine whether a fiscal policy is expansionary or contractionary.

Fiscal policy entails setting taxes and spending. A budget surplus (deficit) occurs when tax revenues exceed (fall short of) spending. Economists often focus on *changes* in the surplus or deficit to determine if the fiscal policy is expansionary or contractionary. An increase (decrease) in surplus is indicative of a contractionary (expansionary) fiscal policy. Similarly, an increase (decrease) in deficit is indicative of an expansionary (contractionary) fiscal policy.



PROFESSOR'S NOTE

For the exam, an increase (decrease) in a revenue item (e.g., sales tax) should be considered contractionary (expansionary), and an increase (decrease) in a spending item (e.g., construction of highways) should be considered expansionary (contractionary).

A government's intended fiscal policy is not necessarily obvious from just examining the current deficit. Consider an economy that is in recession so that transfer payments are increased and tax revenue is decreased, leading to a deficit. This does not necessarily indicate that fiscal policy is expansionary as, at least to some extent, the deficit is a natural outcome of the recession without any explicit action of the government. Economists often use a measure called the **structural** (or **cyclically adjusted**) **budget deficit** to gauge fiscal policy. This is the deficit that would occur based on current policies if the economy were at full employment.

LOS 12.t: Explain the interaction of monetary and fiscal policy.

Monetary policy and fiscal policy may each be either expansionary or contractionary, so there are four possible scenarios:

1. **Expansionary fiscal and monetary policy:** In this case, the impact will be highly expansionary taken together. Interest rates will usually be lower (due to monetary policy), and the private and public sectors will both expand.
2. **Contractionary fiscal and monetary policy:** In this case, aggregate demand and GDP would be lower, and interest rates would be higher due to tight monetary policy. Both the private and public sectors would contract.
3. **Expansionary fiscal policy + contractionary monetary policy:** In this case, aggregate demand will likely be higher (due to fiscal policy), while interest rates will be higher (due to increased government borrowing and tight monetary policy). Government spending as a proportion of GDP will increase.

4. **Contractionary fiscal policy + expansionary monetary policy:** In this case, interest rates will fall from decreased government borrowing and from the expansion of the money supply, increasing both private consumption and output. Government spending as a proportion of GDP will decrease due to contractionary fiscal policy. The private sector would grow as a result of lower interest rates.

Not surprisingly, the fiscal multipliers for different types of fiscal stimulus differ, and the effects of expansionary fiscal policy are greater when it is combined with expansionary monetary policy. The fiscal multiplier for direct government spending increases has been much higher than the fiscal multiplier for increases in transfers to individuals or tax reductions for workers. Within this latter category, government transfer payments to the poor have the greatest relative impact, followed by tax cuts for workers, and broader-based transfers to individuals (not targeted). For all types of fiscal stimulus, the impact is greater when the fiscal actions are combined with expansionary monetary policy. This may reflect the impact of greater inflation, falling real interest rates, and the resulting increase in business investment.



MODULE QUIZ 12.3

- Roles and objectives of fiscal policy *most likely* include:
 - controlling the money supply to limit inflation.
 - adjusting tax rates to influence aggregate demand.
 - using government spending to control interest rates.
- A government enacts a program to subsidize farmers with an expansive spending program of \$10 billion. At the same time, the government enacts a \$10 billion tax increase over the same period. Which of the following statements *best* describes the impact on aggregate demand?
 - Lower growth because the tax increase will have a greater effect.
 - No effect because the tax and spending effects just offset each other.
 - Higher growth because the spending increase will have a greater effect.
- A government reduces spending by \$50 million. The tax rate is 30%, and consumers exhibit a marginal propensity to consume of 80%. The change in aggregate demand caused by the change in government spending is *closest* to:
 - \$66 million.
 - \$114 million.
 - \$250 million.
- The size of a national debt is *most likely* to be a concern for policymakers if:
 - Ricardian equivalence holds.
 - a crowding-out effect occurs.
 - debt is used to finance capital growth.
- Sales in the retail sector have been sluggish, and consumer confidence has recently declined, indicating fewer planned purchases. In response, the president sends an expansionary government spending plan to the legislature. The plan is submitted on March 30, and the legislature refines and approves the terms of the spending plan on June 30. What type of fiscal plan is being considered, and what type of delay did the plan experience between March 30 and June 30?

<u>Fiscal plan</u>	<u>Type of lag</u>
A. Discretionary	Recognition
B. Automatic	Action
C. Discretionary	Action

- A government is concerned about the timing of the impact of fiscal policy changes and is considering requiring the compilation and reporting of economic statistics weekly, rather than

quarterly. The new reporting frequency is intended to decrease:

- A. the action lag.
- B. the impact lag.
- C. the recognition lag.

7. Fiscal policy is *most likely* to be expansionary if tax rates:

- A. and government spending both decrease.
- B. decrease and government spending increases.
- C. increase and government spending decreases.

8. In the presence of tight monetary policy and loose fiscal policy, the *most likely* effect on interest rates and the private sector share in GDP are:

<u>Interest rate</u>	<u>Share of private sector</u>
A. lower	lower
B. higher	higher
C. higher	lower

KEY CONCEPTS

LOS 12.a

Fiscal policy is a government's use of taxation and spending to influence the economy. Monetary policy deals with determining the quantity of money supplied by the central bank. Both policies aim to achieve economic growth with price level stability, although governments use fiscal policy for social and political reasons as well.

LOS 12.b

Money is defined as a widely accepted medium of exchange. Functions of money include a medium of exchange, a store of value, and a unit of account.

LOS 12.c

In a fractional reserve system, new money created is a multiple of new excess reserves available for lending by banks. The potential multiplier is equal to the reciprocal of the reserve requirement and, therefore, is inversely related to the reserve requirement.

LOS 12.d

Three factors influence money demand:

- Transaction demand, for buying goods and services.
- Precautionary demand, to meet unforeseen future needs.
- Speculative demand, to take advantage of investment opportunities.

Money supply is determined by central banks with the goal of managing inflation and other economic objectives.

LOS 12.e

The Fisher effect states that a nominal interest rate is equal to the real interest rate plus the expected inflation rate.

LOS 12.f

Central bank roles include supplying currency, acting as banker to the government and to other banks, regulating and supervising the payments system, acting as a lender of last resort, holding the nation's gold and foreign currency reserves, and conducting monetary policy.

Central banks have the objective of controlling inflation, and some have additional goals of maintaining currency stability, full employment, positive sustainable economic growth, or moderate interest rates.

LOS 12.g

High inflation, even when it is perfectly anticipated, imposes costs on the economy as people reduce cash balances because of the higher opportunity cost of holding cash. More significant costs are imposed by unexpected inflation, which reduces the information value of price changes, can make economic cycles worse, and shifts wealth from lenders to borrowers. Uncertainty about the future rate of inflation increases risk, resulting in decreased business investment.

LOS 12.h

Policy tools available to central banks include the policy rate, reserve requirements, and open market operations. The policy rate is called the discount rate in the United States, the refinancing rate by the ECB, and the 2-week repo rate in the United Kingdom.

Decreasing the policy rate, decreasing reserve requirements, and making open market purchases of securities are all expansionary. Increasing the policy rate, increasing reserve requirements, and making open market sales of securities are all contractionary.

LOS 12.i

The transmission mechanism for changes in the central bank's policy rate through to prices and inflation includes one or more of the following:

- Short-term bank lending rates.
- Asset prices.
- Expectations for economic activity and future policy rate changes.
- Exchange rates with foreign currencies.

LOS 12.j

A contractionary monetary policy (increase in policy rate) will tend to decrease economic growth, increase market interest rates, decrease inflation, and lead to appreciation of the domestic currency in foreign exchange markets. An expansionary monetary policy (decrease in policy rate) will have opposite effects, tending to increase economic growth, decrease market interest rates, increase inflation, and reduce the value of the currency in foreign exchange markets.

LOS 12.k

Effective central banks exhibit independence, credibility, and transparency.

- Independence: The central bank is free from political interference.
- Credibility: The central bank follows through on its stated policy intentions.
- Transparency: The central bank makes it clear what economic indicators it uses and reports on the state of those indicators.

LOS 12.l

Most central banks set target inflation rates, typically 2% to 3%, rather than targeting interest rates as was once common. When inflation is expected to rise above (fall below) the target band,

the money supply is decreased (increased) to reduce (increase) economic activity.

Developing economies sometimes target a stable exchange rate for their currency relative to that of a developed economy, selling their currency when its value rises above the target rate and buying their currency with foreign reserves when the rate falls below the target. The developing country must follow a monetary policy that supports the target exchange rate and essentially commits to having the same inflation rate as the developed country.

LOS 12.m

The real trend rate is the long-term sustainable real growth rate of an economy. The neutral interest rate is the sum of the real trend rate and the target inflation rate. Monetary policy is said to be contractionary when the policy rate is above the neutral rate and expansionary when the policy rate is below the neutral rate.

LOS 12.n

Reasons that monetary policy may not work as intended:

- Monetary policy changes may affect inflation expectations to such an extent that long-term interest rates move opposite to short-term interest rates.
- Individuals may be willing to hold greater cash balances without a change in short-term rates (liquidity trap).
- Banks may be unwilling to lend greater amounts, even when they have increased excess reserves.
- Short-term rates cannot be reduced below zero.
- Developing economies face unique challenges in utilizing monetary policy due to undeveloped financial markets, rapid financial innovation, and lack of credibility of the monetary authority.

LOS 12.o

Fiscal policy refers to the taxing and spending policies of the government. Objectives of fiscal policy can include (1) influencing the level of economic activity, (2) redistributing wealth or income, and (3) allocating resources among industries.

LOS 12.p

Arguments for being concerned with the size of fiscal deficit:

- Higher future taxes lead to disincentives to work, negatively affecting long-term economic growth.
- Fiscal deficits may not be financed by the market when debt levels are high.
- Crowding-out effect as government borrowing increases interest rates and decreases private sector investment.

Arguments against being concerned with the size of fiscal deficit:

- Debt may be financed by domestic citizens.
- Deficits for capital spending can boost the productive capacity of the economy.
- Fiscal deficits may prompt needed tax reform.
- Ricardian equivalence may prevail: private savings rise in anticipation of the need to repay principal on government debt.

- When the economy is operating below full employment, deficits do not crowd out private investment.

LOS 12.q

Fiscal policy tools include spending tools and revenue tools. Spending tools include transfer payments, current spending (goods and services used by government), and capital spending (investment projects funded by government). Revenue tools include direct and indirect taxation.

An advantage of fiscal policy is that indirect taxes can be used to quickly implement social policies and can also be used to quickly raise revenues at a low cost.

Disadvantages of fiscal policy include time lags for implementing changes in direct taxes and time lags for capital spending changes to have an impact.

LOS 12.r

Fiscal policy is implemented by governmental changes in taxing and spending policies. Delays in realizing the effects of fiscal policy changes limit their usefulness. Delays can be caused by:

- Recognition lag: Policymakers may not immediately recognize when fiscal policy changes are needed.
- Action lag: Governments take time to enact needed fiscal policy changes.
- Impact lag: Fiscal policy changes take time to affect economic activity.

LOS 12.s

A government has a budget surplus when tax revenues exceed government spending and a deficit when spending exceeds tax revenue.

An increase (decrease) in a government budget surplus is indicative of a contractionary (expansionary) fiscal policy. Similarly, an increase (decrease) in a government budget deficit is indicative of an expansionary (contractionary) fiscal policy.

LOS 12.t

Interaction of monetary and fiscal policies:

Monetary Policy	Fiscal Policy	Interest Rates	Output	Private Sector Spending	Public Sector Spending
Tight	Tight	higher	lower	lower	lower
Easy	Easy	lower	higher	higher	higher
Tight	Easy	higher	higher	lower	higher
Easy	Tight	lower	varies	higher	lower

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 12.1

1. **B** Both monetary and fiscal policies primarily strive to achieve economic targets such as inflation and GDP growth. Balancing the budget is not a goal for monetary policy and is a potential outcome of fiscal policy. Fiscal policy (but not monetary policy) may secondarily be used as a tool to redistribute income and wealth. (LOS 12.a)

2. **B** Money functions as a unit of account, a medium of exchange, and a store of value. Money existed long before the idea of central banking was conceived. (LOS 12.b)
3. **A** Money neutrality is the theory that changes in the money supply do not affect real output or the velocity of money. Therefore, an increase in the money supply can only increase the price level. (LOS 12.c)
4. **C** Given the equation of exchange, $MV = PY$, an increase in the money supply is consistent with an increase in nominal GDP (PY). However, a decrease in velocity is consistent with a decrease in nominal GDP. Unless we know the size of the changes in the two variables, there is no way to tell what the net impact is on real GDP (Y) and prices (P). (LOS 12.c)
5. **A** The money supply schedule is vertical because the money supply is independent of interest rates. Central banks control the money supply. (LOS 12.d)
6. **C** The Fisher effect states that nominal interest rates are equal to the real interest rate plus the expected inflation rate. (LOS 12.e)
7. **B** Central bank goals often include maximum employment, which is interpreted as the maximum sustainable growth rate of the economy; stable prices; and *moderate* (not minimum) long-term interest rates. (LOS 12.f)
8. **C** The money supply growth rate may need to be adjusted to keep the exchange rate within acceptable bounds, but is not necessarily the same as that of the other country. The other two statements are true. (LOS 12.f)

Module Quiz 12.2

1. **A** The primary method by which a central bank conducts monetary policy is through changes in the target short-term rate or policy rate. (LOS 12.h)
2. **C** Open market purchases by monetary authorities *decrease* the interbank lending rate by increasing excess reserves that banks can lend to one another and therefore increasing their willingness to lend. (LOS 12.i)
3. **C** An increase in the policy rate is likely to increase longer-term interest rates, causing decreases in consumption spending on durable goods and business investment in plant and equipment. The increase in rates, however, makes investment in the domestic economy more attractive to foreign investors, increasing demand for the domestic currency and causing the currency to appreciate. (LOS 12.i)
4. **C** The three qualities of effective central banks are independence, credibility, and transparency. (LOS 12.k)
5. **C** Decreasing the overnight lending rate would add reserves to the banking system, which would encourage bank lending, expand the money supply, reduce interest rates, and allow GDP growth and the rate of inflation to increase. Selling government securities or increasing the reserve requirement would have the opposite effect, reducing the money supply and decreasing the inflation rate. (LOS 12.j)
6. **C** Exchange rate targeting requires monetary policy to be consistent with the goal of a stable exchange rate with the targeted currency, regardless of domestic economic conditions. (LOS 12.l)
7. **B** neutral rate = trend rate + inflation target = 2% + 4.5% = 6.5%
Because the policy rate is less than the neutral rate, monetary policy is expansionary. (LOS 12.m)
8. **B** Monetary policy has limited ability to act effectively against deflation because the policy rate cannot be reduced below zero and demand for money may be highly elastic (liquidity trap). (LOS 12.n)

Module Quiz 12.3

1. **B** Influencing the level of aggregate demand through taxation and government spending is an objective of fiscal policy. Controlling inflation and interest rates are typical objectives of monetary policy. (LOS 12.o)
2. **C** The amount of the spending program exactly offsets the amount of the tax increase, leaving the budget unaffected. The multiplier for government spending is greater than the multiplier for a tax increase. Therefore,

the balanced budget multiplier is positive. All of the government spending enters the economy as increased expenditure, whereas spending is reduced by only a portion of the tax increase. (LOS 12.q)

3. **B** fiscal multiplier = $1 / [1 - MPC(1 - T)] = 1 / [1 - 0.80(1 - 0.3)] = 2.27$
change in government spending = -\$50 million
change in aggregate demand = $-(50 \times 2.27) = -\$113.64$ million (LOS 12.q)
4. **B** Crowding out refers to the possibility that government borrowing causes interest rates to increase and private investment to decrease. If government debt is financing the growth of productive capital, this should increase future economic growth and tax receipts to repay the debt. Ricardian equivalence is the theory that if government debt increases, private citizens will increase savings in anticipation of higher future taxes, and it is an argument against being concerned about the size of government debt and budget deficits. (LOS 12.p)
5. **C** The expansionary plan initiated by the president and approved by the legislature is an example of discretionary fiscal policy. The lag from the time of the submission (March 30) through time of the vote (June 30) is known as action lag. It took the legislature three months to write and pass the necessary laws. (LOS 12.r)
6. **C** More frequent and current economic data would make it easier for authorities to monitor the economy and to recognize problems. The reduction in the time between economic reports should reduce the recognition lag. (LOS 12.r)
7. **B** Increases in government spending and decreases in taxes are expansionary fiscal policy. Decreases in spending and increases in taxes are contractionary fiscal policy. (LOS 12.s)
8. **C** Tight monetary policy and loose fiscal policy both lead to higher interest rates. Tight monetary policy decreases private sector growth, while loose fiscal policy expands the public sector, reducing the overall share of private sector in the GDP. (LOS 12.t)

READING 13

INTRODUCTION TO GEOPOLITICS

EXAM FOCUS

Candidates should be familiar with the terminology from this reading as well as the framework for analysis of cooperation versus competition and the difference between globalization and nationalism. Finally, be able to demonstrate the relation between geopolitical risks and investment risks.

MODULE 13.1: GEOPOLITICS AND GEOPOLITICAL RISK



Video covering this content is available online.

LOS 13.a: Describe geopolitics from a cooperation versus competition perspective.

Geopolitics refers to interactions among nations, including the actions of **state actors** (national governments) and **non-state actors** (corporations, non-government organizations, and individuals).

Geopolitics also refers to the study of how geography affects interactions among nations and their citizens. For example, firms located in coastal countries naturally tend to be the dominant participants in international shipping.

One way to examine geopolitics is through analysis of the extent to which individual countries cooperate with one another. Potential areas for cooperation include diplomatic and military matters and economic and cultural interactions. In terms of economics, areas of cooperation include freedom of movement across borders for goods, services, and capital; agreements to harmonize tariffs; international standardization of rules; and transfers of information and technology.

While a country that engages with other countries on these matters may be considered **cooperative** and one that does not may be considered **non-cooperative**, the extent of cooperation actually varies along a spectrum. A country might be more cooperative on some issues and less cooperative on others, and its degree of cooperation can change over time or with the outcomes of the country's domestic politics. A country's current decision makers and the length of its political cycle are factors to consider when analyzing geopolitics.

A country will typically cooperate with other countries when doing so advances its national interests. For example, a country may cooperate with its neighbors in a military alliance if doing so will further its interests in protecting its citizens from foreign invaders.

We can analyze a country's national interests as a hierarchy, with its top priorities being those that ensure its survival. A country's **geophysical resource endowment** may influence its priorities. For example, a country that has mineral resources but lacks arable land needs to trade minerals for food, and therefore has an interest in cooperating with other countries to keep international trade lanes open.

Non-state actors often have interests in cooperating across borders. Individuals and firms seek to direct their resources to their highest-valued uses, and some of those uses may be in other countries. To facilitate the flow of resources, state and non-state actors may cooperate on **standardization** of regulations and processes. One key example of standardization among countries is International Financial Reporting Standards for firms presenting their accounting data to the public, which we will examine in the Financial Statement Analysis topic area.

Cultural factors, such as historical emigration patterns or a shared language, can be another influence on a country's level of cooperation. Among these cultural factors are a country's formal and informal **institutions**, such as laws, public and private organizations, or distinct customs and habits. Strong and stable institutions can make cooperation easier for state and non-state actors. For example, countries that produce and export large amounts of cultural content tend to be those with legal and ethical institutions that protect intellectual property. Cultural exchange is one means through which a country may exercise **soft power**, the ability to influence other countries without using or threatening force.

LOS 13.b: Describe geopolitics and its relationship with globalization.

Globalization refers to the long-term trend toward worldwide integration of economic activity and cultures. Data from the World Bank suggest economic openness, as measured by international trade as a percentage of total output, increased steadily from about 25% in the early 1970s to about 60% before the 2008 financial crisis, and has remained near that level since then. We may contrast globalization with **nationalism**, which in this context refers to a nation pursuing its own economic interests independently of, or in competition with, the economic interests of other countries.



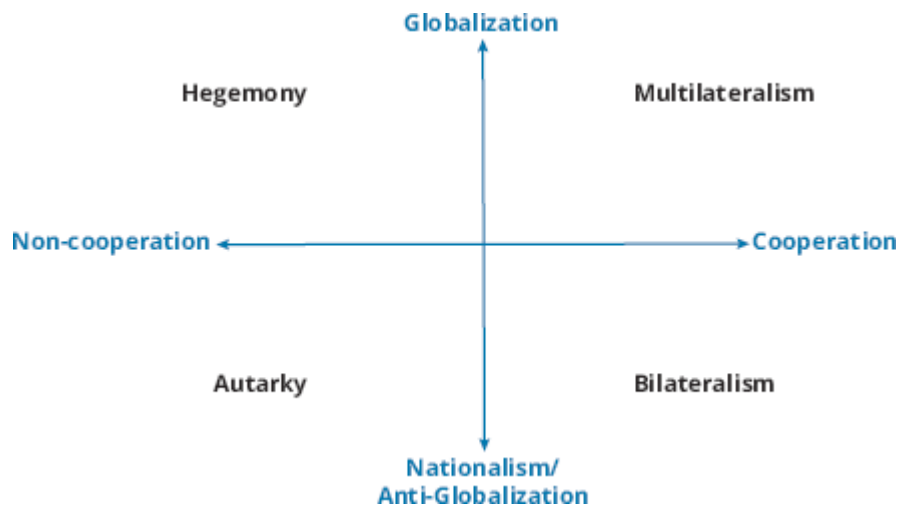
PROFESSOR'S NOTE

Debate about what the word “nationalism” means is beyond the scope of the CFA curriculum. We use it here only in the sense of opposition or resistance to globalization.

As we did with cooperation versus non-cooperation, we can think of countries' actions along a spectrum from globalization to nationalism. In general, countries that are closer to the globalization end of the spectrum are those that more actively import and export goods and services, permit freer movement of capital across borders and exchange of currencies, and are more open to cultural interaction.

In Figure 13.1 we draw each spectrum as an axis. This creates four quadrants, each of which we can associate with a type of behavior by countries. While individual countries rarely fit neatly into one of these categories, this gives us a general framework within which we can describe geopolitical actions.

Figure 13.1: Archetypes of Globalization and Cooperation



Reproduced from Level I CFA Curriculum learning module “Introduction to Geopolitics,” Exhibit 6, with permission from CFA Institute.

Characteristics we may associate with each of these categories are as follows.

- **Autarky** (non-cooperation and nationalism) refers to a goal of national self-reliance, including producing most or all necessary goods and services domestically. Autarky is often associated with a state-dominated society in general, with attributes such as government control of industry and media.
- **Hegemony** (non-cooperation and globalization) refers to countries that are open to globalization but have the size and scale to influence other countries without necessarily cooperating.
- **Bilateralism** (cooperation and nationalism) refers to cooperation between two countries. A country that engages in bilateralism may have many such relationships with other countries while tending not to involve itself in multi-country arrangements.
- **Multilateralism** (cooperation and globalization) refers to countries that engage extensively in international trade and other forms of cooperation with many other countries. Some countries may exhibit **regionalism**, cooperating multilaterally with nearby countries but less so with the world at large.

Some of the non-state actors within a country may be more oriented toward globalization than their governments. Businesses may look outside their home country for opportunities to increase profits, reduce costs, and sell to new markets. Investors may seek higher returns or diversification by investing outside their home country. Non-state actors might buy and sell foreign securities (**portfolio investment flows**) or own physical production capacity in other countries (**foreign direct investment**).

LOS 13.c: Describe tools of geopolitics and their impact on regions and economies.

We can consider **tools of geopolitics**, the means by which (primarily) state actors advance their interests in the world, as falling into three broad categories of national security, economic, and financial.

National security tools may include armed conflict, espionage, or bilateral or multilateral agreements designed to reinforce or prevent armed conflict. We can say a national security tool is *active* if a country is currently using it or *threatened* if a country is not currently using it but appears likely to do so. Armed conflict affects regions and economies by destroying productive capital and causing migration away from areas of conflict.

Economic tools can be cooperative or non-cooperative. Examples of cooperative economic tools include free trade areas, common markets, and economic and monetary unions (each of which we describe in our reading on International Trade and Capital Flows). Examples of non-cooperative economic tools include domestic content requirements, voluntary export restraints, and nationalization (i.e., the state taking control) of companies or industries.

Financial tools include foreign investment and the exchange of currencies. We can view countries as using these tools cooperatively if they allow foreign investment and the free exchange of currencies, or non-cooperatively when they restrict these activities. **Sanctions**, or restrictions on a specific geopolitical actor's financial interests, are a financial tool that state actors may use alongside national security tools.

LOS 13.d: Describe geopolitical risk and its impact on investments.

Geopolitical risk is the possibility of events that interrupt peaceful international relations. We can classify geopolitical risk into three types:

- **Event risk** refers to events about which we know the timing but not the outcome, such as national elections.
- **Exogenous risk** refers to unanticipated events, such as outbreaks of war or rebellion.
- **Thematic risk** refers to known factors that have effects over long periods, such as human migration patterns or cyber risks.

Geopolitical risk affects investment values by increasing or decreasing the risk premium investors require to hold assets in a country or region. To forecast the effect on investments of a geopolitical risk, we need to consider its probability (*likelihood*), the magnitude of its effects on investment outcomes (*impact*), and how quickly investment values would reflect these effects (*velocity*).

We can use our framework of cooperation and globalization to help estimate the **likelihood of geopolitical risk**. Countries that are more cooperative and globalized tend to have less likelihood of some geopolitical risks, such as armed conflict, but may have greater likelihood of other risks, such as the supply chain disruptions that followed the COVID-19 pandemic in 2020–2021.

To analyze the **velocity of geopolitical risk** we can classify risks as high velocity (short term), medium velocity, or low velocity (long term). Exogenous risks often have high-velocity effects on financial markets and investment values. **Black swan risk** is a term for the risk of low-likelihood exogenous events that have substantial short-term effects. Investors with longer time horizons typically do not need to react to these kinds of events, but investors with shorter horizons might find it necessary to react.

Medium-velocity risks can potentially damage specific companies or industries by increasing their costs or disrupting their production processes, while low-velocity risks tend to affect them in the “environmental, social, and governance” realm. Analyzing these kinds of risk is important for investors with long time horizons.

Because analyzing geopolitical risks requires effort, time, and resources, investors should consider whether the **impact of geopolitical risk** is likely to be high or low, and focus their analysis on risks that could have a high impact. With regard to those risks, investors should determine whether they are likely to have *discrete impacts* on a company or industry, or *broad impacts* on a country, a region, or the world. Business cycles can affect the impact of geopolitical risk, in that these risks may have greater impacts on investment values when an economy is in recession than they would have during an expansion.

Investors can use qualitative or quantitative **scenario analysis** to gauge the potential effects of geopolitical risks on their portfolios. To help identify geopolitical risks over time, investors may identify **signposts**, or data that can signal when the likelihood of an event is increasing or decreasing, such as volatility indicators in financial markets.



MODULE QUIZ 13.1

1. A state actor that is generally cooperative with other countries and primarily nationalist in pursuing its objectives is *most* accurately said to exhibit:
 - A. autarky.
 - B. hegemony.
 - C. bilateralism.
2. Which of the following tools of geopolitics is *best* described as a non-cooperative economic tool?
 - A. Voluntary export restraints.
 - B. Regional free trade agreements.
 - C. Restrictions on conversion of currencies.
3. When investing for a long time horizon, a portfolio manager should *most likely* devote resources to analyzing:
 - A. event risks.
 - B. thematic risks.
 - C. exogenous risks.

KEY CONCEPTS

LOS 13.a

Geopolitics refers to interactions among nations. On various issues ranging from diplomacy and military force to economic or cultural openness, countries lie along a spectrum from cooperative to non-cooperative.

LOS 13.b

Globalization refers to integration of economic activity and cultures among countries, and can be contrasted with nationalism, which refers to a country pursuing its own interests independently of other countries. Analysts should view geopolitical actions as being on a spectrum from nationalism to globalization.

We may describe geopolitics and its relationship with globalization using the following four broad categories: autarky (non-cooperation and nationalism); hegemony (non-cooperation and

globalization); bilateralism (cooperation and nationalism); and multilateralism (cooperation and globalization).

LOS 13.c

Tools of geopolitics include national security tools, economic tools, and financial tools.

National security tools may include armed conflict, espionage, or bilateral or multilateral national security agreements.

Cooperative economic tools include free trade areas, common markets, and economic and monetary unions. Non-cooperative economic tools include domestic content requirements, voluntary export restraints, and nationalization.

Financial tools include foreign investment, exchange of currencies, and sanctions.

LOS 13.d

Categories of geopolitical risk are event risk (when the timing is known), exogenous risk (unanticipated events), and thematic risk (known factors that have long-term effects).

Investors should analyze the likelihood of a geopolitical risk, the impact on investment values of an event if it occurs, and the velocity with which it would affect investment values.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 13.1

1. **C** Bilateralism is characterized by nationalism (as contrasted with globalization) and cooperation. Both autarky and hegemony are characterized by non-cooperation. (LOS 13.a, 13.b)
2. **A** Voluntary export restraints (exporting less of a good than the global market demands) are an example of a non-cooperative economic tool. Restrictions on the exchange of currencies are a financial tool. Free trade agreements are a cooperative economic tool. (LOS 13.c)
3. **B** Thematic risks are those that have effects over the long term. Event risks and exogenous risks are more likely to have high-velocity impacts on investment values but are less of a focus for investors with longer time horizons. (LOS 13.d)

READING 14

INTERNATIONAL TRADE AND CAPITAL FLOWS

EXAM FOCUS

International trade and currency exchange rates are key topics for both Level I and Level II. First, learn how comparative advantage results in a welfare gain from international trade and the two models of the sources of comparative advantage. Learn the types of trade restrictions and their effects on domestic price and quantity. For the balance of payments, focus on how a surplus or deficit in the broadly defined capital account must offset a deficit or surplus in the merchandise trade account. Finally, focus on how the difference between domestic income and expenditures and the difference between domestic savings and investment are related to a country's balance of trade.

MODULE 14.1: INTERNATIONAL TRADE BENEFITS



Video covering this content is available online.

Before we address specific topics and learning outcomes, it will help to define some terms as follows.

Imports: Goods and services that firms, individuals, and governments purchase from producers in other countries.

Exports: Goods and services that firms, individuals, and governments purchase from other countries from domestic producers.

Autarky or closed economy: A country that does not trade with other countries.

Free trade: A government places no restrictions or charges on import and export activity.

Trade protection: A government places restrictions, limits, or charges on exports or imports.

World price: The price of a good or service in world markets for those to whom trade is not restricted.

Domestic price: The price of a good or service in the domestic country, which may be equal to the world price if free trade is permitted or different from the world price when the domestic country restricts trade.

Net exports: The value of a country's exports minus the value of its imports over some period.

Trade surplus: Net exports are positive; the value of the goods and services a country exports are greater than the value of the goods and services it imports.

Trade deficit: Net exports are negative; the value of the goods and services a country exports is less than the value of the goods and services it imports.

Terms of trade: The ratio of an index of the prices of a country's exports to an index of the prices of its imports expressed relative to a base value of 100. If a country's terms of trade are currently 102, the prices of the goods it exports have risen relative to the prices of the goods it imports since the base period.

Foreign direct investment: Ownership of productive resources (land, factories, natural resources) in a foreign country.

Multinational corporation: A firm that has made foreign direct investment in one or more foreign countries, operating production facilities and subsidiary companies in foreign countries.

LOS 14.a: Compare gross domestic product and gross national product.

Gross domestic product over a period, typically a year, is the total value of goods and services produced within a country's borders. **Gross national product** is similar but measures the total value of goods and services produced by the labor and capital of a country's citizens. The difference is due to non-citizen incomes of foreigners working within a country, the income of citizens who work in other countries, the income of foreign capital invested within a country, and the income of capital supplied by its citizens to foreign countries. The income to capital owned by foreigners invested within a country is included in the domestic country's GDP but not in its GNP. The income of a country's citizens working abroad is included in its GNP but not in its GDP.

GDP is more closely related to economic activity within a country and so to its employment and growth.

LOS 14.b: Describe benefits and costs of international trade.

The benefits of trade are not hard to understand. As an example, consider China, and really Asia as a whole, which has had rapidly growing exports to the United States and other countries. The benefit to the importing countries has been lower-cost goods, from textiles to electronics. The benefits to the Chinese economy have been in increasing employment, increasing wages for workers, and the profits from its export products.

The costs of trade are primarily borne by those in domestic industries that compete with imported goods. Textile workers who have lost their jobs in the United States, as more and more textiles are imported, are certainly worse off in the short run. As other industries, such as health care, have grown, these workers have had to retrain to qualify for the new jobs in those fields. At the same time, U.S. firms that produce textile products using capital and technology intensive production methods have expanded. We address the reasons for this and the underlying economic theory in this reading.

Overall, economics tells us that the benefits of trade are greater than the costs for economies as a whole, so that the winners could conceivably compensate the losers and still be better off. We

now turn to the economic theory that supports this view.

LOS 14.c: Contrast comparative advantage and absolute advantage.

A country is said to have an **absolute advantage** in the production of a good if it can produce the good at a lower resource cost than another country. A country is said to have a **comparative advantage** in the production of a good if it has a lower **opportunity cost** in the production of that good, expressed as the amount of another good that could have been produced instead. Economic analysis tells us that, regardless of which country has an absolute advantage, there are potential gains from trade as long as the countries' opportunity costs of one good in terms of another are different.

This analysis is credited to David Ricardo who presented it in 1817. He used the example of the production of cloth and wine in England and Portugal. A hypothetical example of the amounts of cloth and wine these countries can produce per day of labor is presented in Figure 14.1.

Figure 14.1: Output per Unit of Labor

	Yards of Cloth	Bottles of Wine
Portugal	100	110
England	90	80

Ricardo argued that, in the absence of trading costs, England could trade cloth for wine, and Portugal could trade wine for cloth, and both countries could have more of both wine and cloth as a result. Because in Portugal a worker-day can be used to produce either 100 yards of cloth or 110 bottles of wine, its opportunity cost of a yard of cloth is $110 / 100 = 1.1$ bottles of wine and its opportunity cost of a bottle of wine is $100 / 110 = 0.91$ yards of cloth. England's opportunity cost of a yard of cloth is $80 / 90 = 0.89$ bottles of wine and its opportunity cost of a bottle of wine is $90 / 80 = 1.125$ yards of cloth.

Portugal has a comparative advantage in the production of wine as its opportunity cost is 0.91 yards of cloth compared to England's opportunity cost of 1.125 yards of cloth. As must be the case, if Portugal has a comparative advantage in wine production, England has a comparative advantage in cloth production.

To illustrate the benefits of trade, consider the output change if Portugal shifts 8 worker-days from cloth production to wine production and England shifts 10 worker-days from wine production to cloth production.

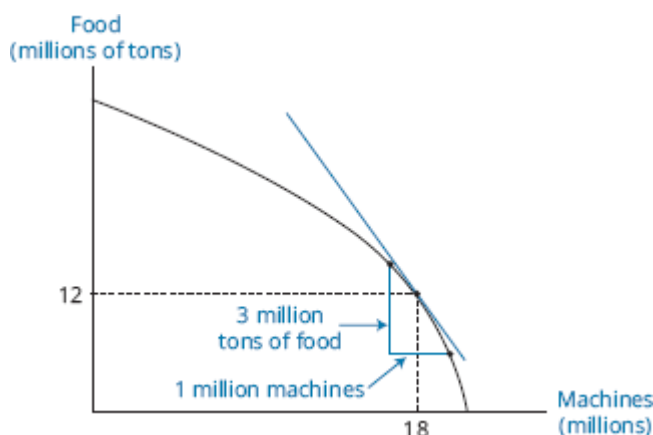
- The change in Portugal's production is $8 \times 110 = +880$ wine and $-8 \times 100 = -800$ cloth.
- The change in England's production is $-10 \times 80 = -800$ wine and $10 \times 90 = +900$ cloth.

Total production by the two countries will have increased by 80 bottles of wine and 100 yards of cloth; these are the gains from trade. The negotiated terms of trade will determine how the two countries share these gains. The important result is that *total output has increased* through trade, there's greater specialization by Portugal in wine production, and there's greater specialization by England in cloth production.

Note that Portugal has an absolute advantage in the production of both goods. However, because the countries' opportunity costs of production differ, each has a comparative advantage in one of the goods, and trade can make both countries better off.

In our simple example, we assume constant opportunity costs. As a country specializes and increases the production of an export good, increasing costs (e.g., using more marginal land for grape growing) will increase the opportunity cost of the export good. The **production possibility frontier** shown in Figure 14.2 illustrates such a situation and shows all combinations of food and machinery that an economy can produce. The slope of the frontier measures the opportunity cost of machinery in terms of food at each possible combination of food and machinery. Over a range of possible output choices around 12 million tons of food and 18 million machines, we show the slope is -3 and the opportunity cost of each million machines is 3 million tons of food. If the country were to increase the production of machinery, the amount of food production foregone would increase, as shown by the increasingly negative slope of the frontier.

Figure 14.2: A Production Possibility Frontier



LOS 14.d: Compare the Ricardian and Heckscher–Ohlin models of trade and the source(s) of comparative advantage in each model.

The **Ricardian model of trade** has only one factor of production—labor. The source of differences in production costs in Ricardo's model is *differences in labor productivity* due to differences in technology.

Heckscher and Ohlin presented a model in which there are two factors of production—capital and labor. The source of comparative advantage (differences in opportunity costs) in this model is *differences in the relative amounts of each factor* the countries possess. We can view the England and Portugal example in these terms by assuming that England has more capital (machinery) compared to labor than Portugal. Additionally, we need to assume that cloth production is more capital intensive than wine production. The result of their analysis is that the country that has more capital will specialize in the capital intensive good and trade for the less capital intensive good with the country that has relatively more labor and less capital.

In the **Heckscher-Ohlin model**, there is a redistribution of wealth within each country between labor and the owners of capital. The price of the relatively less scarce (more available)

factor of production in each country will increase so that owners of capital will earn more in England, and workers will earn more in Portugal compared to what they were without trade. This is easy to understand in the context of prices of the two goods. The good that a country imports will fall in price (that is why they import it), and the good that a country exports will rise in price. In our example, this means that the price of wine falls, and the price of cloth rises in England. Because with trade, more of the capital-intensive good, cloth, is produced in England, demand for capital and the price of capital will increase in England. As a result, capital receives more income at the expense of labor in England. In Portugal, increasing the production of wine (which is labor intensive) increases the demand for and price of labor, and workers gain at the expense of the owners of capital.



PROFESSOR'S NOTE

Remember that the model named after one economist has one factor of production, and the model named after two economists has two factors of production.



MODULE QUIZ 14.1

1. The income from a financial investment in Country P by a citizen of Country Q is *most likely* included in:
 - A. Country P's GDP but not its GNP.
 - B. Country Q's GNP and GDP.
 - C. Country P's GDP and GNP.
2. Which of the following effects is *most likely* to occur in a country that increases its openness to international trade?
 - A. Increased prices of consumer goods.
 - B. Greater specialization in domestic output.
 - C. Decreased employment in exporting industries.
3. Which of the following statements about international trade is *least accurate*? If two countries have different opportunity costs of production for two goods, by engaging in trade:
 - A. each country gains by importing the good for which it has a comparative advantage.
 - B. each country can achieve a level of consumption outside its domestic production possibility frontier.
 - C. the low opportunity cost producer of each good will export to the high opportunity cost producer of that good.
4. With regard to the Ricardian and Heckscher-Ohlin models of international trade, the amount of capital relative to labor within a country is a factor in:
 - A. both of these models.
 - B. neither of these models.
 - C. only one of these models.

MODULE 14.2: TRADE RESTRICTIONS



LOS 14.e: Compare types of trade and capital restrictions and their economic implications.

Video covering this content is available online.

There are many reasons (at least stated reasons) why governments impose trade restrictions. Some have support among economists as conceivably valid in terms of increasing a country's

welfare, while others have little or no support from economic theory. Some of the reasons for trade restrictions that have support from economists are:

- *Infant industry.* Protection from foreign competition is given to new industries to give them an opportunity to grow to an internationally competitive scale and get up the learning curve in terms of efficient production methods.
- *National security.* Even if imports are cheaper, it may be in the country's best interest to protect producers of goods crucial to the country's national defense so that those goods are available domestically in the event of conflict.

Other arguments for trade restrictions that have little support in theory are:

- *Protecting domestic jobs.* While some jobs are certainly lost, and some groups and regions are negatively affected by free trade, other jobs (in export industries or growing domestic goods and services industries) will be created, and prices for domestic consumers will be less without import restrictions.
- *Protecting domestic industries.* Industry firms often use political influence to get protection from foreign competition, usually to the detriment of consumers, who pay higher prices.

Other arguments include retaliation for foreign trade restrictions; government collection of tariffs (like taxes on imported goods); countering the effects of government subsidies paid to foreign producers; and preventing foreign exports at less than their cost of production (*dumping*).

Types of trade restrictions include:

- **Tariffs:** Taxes on imported good collected by the government.
- **Quotas:** Limits on the amount of imports allowed over some period.
- **Export subsidies:** Government payments to firms that export goods.
- **Minimum domestic content:** Requirement that some percentage of product content must be from the domestic country.
- **Voluntary export restraint:** A country voluntarily restricts the amount of a good that can be exported, often in the hope of avoiding tariffs or quotas imposed by their trading partners.

Economic Implications of Trade Restrictions

We will now examine the effects of the primary types of trade restrictions, tariffs, and subsidies.

A **tariff** placed on an imported good increases the domestic price, decreases the quantity imported, and increases the quantity supplied domestically. Domestic producers gain, foreign exporters lose, and the domestic government gains by the amount of the tariff revenues.

A **quota** restricts the quantity of a good imported to the quota amount. Domestic producers gain, and domestic consumers lose from an increase in the domestic price. The right to export a specific quantity to the domestic country is granted by the domestic government, which may or may not charge for the import licenses to foreign countries. If the import licenses are sold, the domestic government gains the revenue.

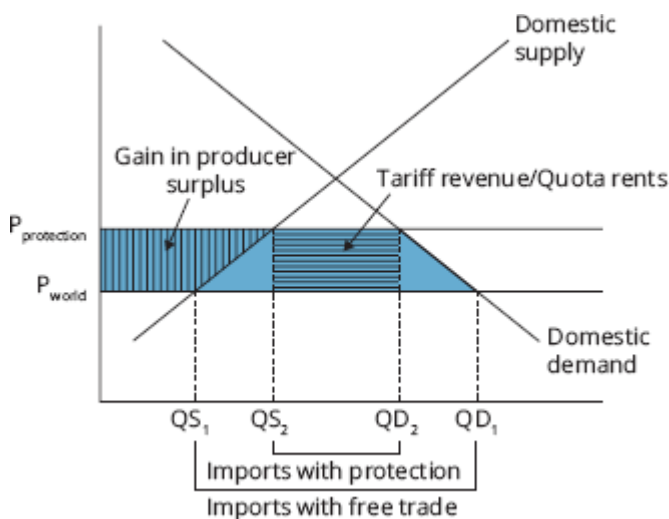
We illustrate the overall welfare effects of quotas and tariffs for a small country in Figure 14.3. We define a quota that is equivalent to a given tariff as a quota that will result in the same

decrease in the quantity of a good imported as the tariff. Defined this way, a tariff and an equivalent quota both increase the domestic price from P_{world} , the price that prevails with no trade restriction, to $P_{\text{protection}}$.

At P_{world} , prior to any restriction, the domestic quantity supplied is QS_1 , and the domestic quantity demanded is QD_1 , with the difference equal to the quantity imported, $QD_1 - QS_1$. Placing a tariff on imports increases the domestic price to $P_{\text{protection}}$, increases the domestic quantity supplied to QS_2 , and decreases the domestic quantity demanded to QD_2 . The difference is the new quantity imported. An equivalent quota will have the same effect, decreasing the quantity imported to $QD_2 - QS_2$.

The entire shaded area in Figure 14.3 represents the loss of consumer surplus in the domestic economy. The portion with vertical lines, the area to the left of the domestic supply curve between $P_{\text{protection}}$ and P_{world} , represents the gain in the producer surplus of domestic producers. The portion with horizontal lines, the area bounded by $QD_2 - QS_2$ and $P_{\text{protection}} - P_{\text{world}}$, represents the gain to the domestic government from tariff revenue. The two remaining triangular areas are the deadweight loss from the restriction on free trade.

Figure 14.3: Effects of Tariffs and Quotas



In the case of a quota, if the domestic government collects the full value of the import licenses, the result is the same as for a tariff. If the domestic government does not charge for the import licenses, this amount is a gain to those foreign exporters who receive the import licenses under the quota and are termed **quota rents**.

In terms of overall economic gains from trade, the deadweight loss is the amount of lost welfare from the imposition of the quota or tariff. From the viewpoint of the domestic country, the loss in consumer surplus is only partially offset by the gains in domestic producer surplus and the collection of tariff revenue.

If none of the quota rents are captured by the domestic government, the overall welfare loss to the domestic economy is greater by the amount of the quota rents. It is the entire difference between the gain in producer surplus and the loss of consumer surplus.

A **voluntary export restraint (VER)** is just as it sounds. It refers to a voluntary agreement by a government to limit the quantity of a good that can be exported. VERs are another way of protecting the domestic producers in the importing country. They result in a welfare loss to the importing country equal to that of an equivalent quota with no government charge for the import licenses; that is, no capture of the quota rents.

Export subsidies are payments by a government to its country's exporters. Export subsidies benefit producers (exporters) of the good but increase prices and reduce consumer surplus in the exporting country. In a small country, the price will increase by the amount of the subsidy to equal the world price plus the subsidy. In the case of a large exporter of the good, the world price decreases and some benefits from the subsidy accrue to foreign consumers, while foreign producers are negatively affected.

Most of the effects of all four of these protectionist policies are the same. With respect to the domestic (importing) country, import quotas, tariffs, and VERs all:

- Reduce imports.
- Increase price.
- Decrease consumer surplus.
- Increase domestic quantity supplied.
- Increase producer surplus.

With one exception, all will decrease national welfare. Quotas and tariffs in a large country could increase national welfare under a specific set of assumptions, primarily because for a country that imports a large amount of the good, setting a quota or tariff could reduce the world price for the good.

Capital Restrictions

Some countries impose **capital restrictions** on the flow of financial capital across borders. Restrictions include outright prohibition of investment in the domestic country by foreigners, prohibition of or taxes on the income earned on foreign investments by domestic citizens, prohibition of foreign investment in certain domestic industries, and restrictions on repatriation of earnings of foreign entities operating in a country.

Overall, capital restrictions are thought to decrease economic welfare. However, over the short term, they have helped developing countries avoid the impact of great inflows of foreign capital during periods of optimistic expansion and the impact of large outflows of foreign capital during periods of correction and market unease or outright panic. Even these short-term benefits may not offset longer-term costs if the country is excluded from international markets for financial capital flows.

LOS 14.f: Explain motivations for and advantages of trading blocs, common markets, and economic unions.

There are various types of agreements among countries with respect to trade policy. The essence of all of them is to reduce trade barriers among the countries. Reductions in trade restrictions among countries have some, by now familiar, positive and negative effects on

economic welfare. The positive effects result from increased trade according to comparative advantage, as well as increased competition among firms in member countries. The negative effects result because some firms, some industries, and some groups of workers will see their wealth and incomes decrease. Workers in affected industries may need to learn new skills to get new jobs.

On balance, economic welfare is improved by reducing or eliminating trade restrictions. Note, however, that to the extent that a trade agreement increases trade restrictions on imports from non-member countries, economic welfare gains are reduced and, in an extreme case, could be outweighed by the costs such restrictions impose. This could result if restrictions on trade with non-member countries increases a country's (unrestricted) imports from a member that has higher prices than the country's previous imports from a non-member.

We list these types of agreements, generally referred to as **trading blocs** or **regional trading agreements (RTA)**, in order of their degrees of integration.

Free Trade Areas

1. All barriers to import and export of goods and services among member countries are removed.

Customs Union

1. All barriers to import and export of goods and services among member countries are removed.
2. All countries adopt a common set of trade restrictions with non-members.

Common Market

1. All barriers to import and export of goods and services among the countries are removed.
2. All countries adopt a common set of trade restrictions with non-members.
3. All barriers to the movement of labor and capital goods among member countries are removed.

Economic Union

1. All barriers to import and export of goods and services among the countries are removed.
2. All countries adopt a common set of trade restrictions with non-members.
3. All barriers to the movement of labor and capital goods among member countries are removed.
4. Member countries establish common institutions and economic policy for the union.

Monetary Union

1. All barriers to import and export of goods and services among the countries are removed.
2. All countries adopt a common set of trade restrictions with non-members.

3. All barriers to the movement of labor and capital goods among member countries are removed.
4. Member countries establish common institutions and economic policy for the union.
5. Member countries adopt a single currency.

The North American Free Trade Agreement (NAFTA) is an example of a free trade area, the European Union (EU) is an example of an economic union, and the euro zone is an example of a monetary union.

LOS 14.g: Describe common objectives of capital restrictions imposed by governments.

Governments sometimes place restrictions on the flow of investment capital into their country, out of their country, or both. Commonly cited objectives of capital flow restrictions include the following:

- *Reduce the volatility of domestic asset prices.* In times of macroeconomic crisis, capital flows out of the country can drive down asset prices drastically, especially prices of liquid assets such as stocks and bonds. With no restrictions on inflows or outflows of foreign investment capital, the asset markets of countries with economies that are small relative to the amount of foreign investment can be quite volatile over a country's economic cycle.
 - *Maintain fixed exchange rates.* For countries with fixed exchange rate targets, limiting flows of foreign investment capital makes it easier to meet the exchange rate target and, therefore, to be able to use monetary and fiscal policy to pursue only the economic goals for the domestic economy.
 - *Keep domestic interest rates low.* By restricting the outflow of investment capital, countries can keep their domestic interest rates low and manage the domestic economy with monetary policy, as investors cannot pursue higher rates in foreign countries. China is an example of a country with a fixed exchange rate regime where restrictions on capital flows allow policymakers to maintain the target exchange rate as well as to pursue a monetary policy independent of concerns about its effect on currency exchange rates.
 - *Protect strategic industries.* Governments sometimes prohibit investment by foreign entities in industries considered to be important for national security, such as the telecommunications and defense industries.
-

LOS 14.h: Describe the balance of payments accounts including their components.

When a country's firms and individuals pay for their purchases of foreign goods, services, and financial assets, they must buy the currencies of the foreign countries in order to accomplish those transactions. Similarly, payment for sales of goods, services, and financial assets to foreigners requires them to purchase the currency of the domestic country. With adjustment for changes in foreign debt to the domestic country and domestic debt to foreign countries, these amounts must balance each other.

According to the U.S. Federal Reserve, “The BOP [**balance of payments**] includes the **current account**, which mainly measures the flows of goods and services; the **capital account**, which consists of capital transfers and the acquisition and disposal of non-produced, non-financial assets; and the **financial account**, which records investment flows.”¹

Drawing on the N.Y. Fed’s explanation, the items recorded in each account are as follows.

Current Account

The current account comprises three sub-accounts:

- **Merchandise and services.** Merchandise consists of all raw materials and manufactured goods bought, sold, or given away. Services include tourism, transportation, and business and engineering services, as well as fees from patents and copyrights on new technology, software, books, and movies.
- **Income receipts** include foreign income from dividends on stock holdings and interest on debt securities.
- **Unilateral transfers** are one-way transfers of assets, such as money received from those working abroad and direct foreign aid. In the case of foreign aid and gifts, the capital account of the donor nation is debited.

Capital Account

The capital account comprises two sub-accounts:

- **Capital transfers** include debt forgiveness and goods and financial assets that migrants bring when they come to a country or take with them when they leave. Capital transfers also include the transfer of title to fixed assets and of funds linked to the purchase or sale of fixed assets, gift and inheritance taxes, death duties, and uninsured damage to fixed assets.
- **Sales and purchases of non-financial assets** that are not produced assets include rights to natural resources and intangible assets, such as patents, copyrights, trademarks, franchises, and leases.

Financial Account

The financial account comprises two sub-accounts:

- **Government-owned assets abroad** include gold, foreign currencies, foreign securities, reserve position in the International Monetary Fund, credits and other long-term assets, direct foreign investment, and claims against foreign banks.
- **Foreign-owned assets in the country** are divided into foreign official assets and other foreign assets in the domestic country. These assets include domestic government and corporate securities, direct investment in the domestic country, domestic country currency, and domestic liabilities to foreigners reported by domestic banks.

A country that has imports valued more than its exports is said to have a *current account (trade) deficit*, while countries with more exports than imports are said to have a *current account surplus*. For a country with a trade deficit, it must be balanced by a net surplus in the capital and financial accounts. As a result, investment analysts often think of all financing flows as a single capital account that combines items in the capital and financial accounts. Thinking in this way, any deficit in the current account must be made up by a surplus in the combined

capital account. That is, the excess of imports over exports must be offset by sales of assets and debt incurred to foreign entities. A current account surplus is similarly offset by purchases of foreign physical or financial assets.

LOS 14.i: Explain how decisions by consumers, firms, and governments affect the balance of payments.

The primary influences referred to here are on the current account deficit or surplus. If a country's net savings (both government savings and private savings) are less than the amount of investment in domestic capital, this investment must be financed by foreign borrowing. Foreign borrowing results in a capital account surplus, which means there is a trade deficit.

We can write the relation between the trade deficit, saving, and domestic investment as:

$$X - M = \text{private savings} + \text{government savings} - \text{investment}$$

Lower levels of private saving, larger government deficits, and high rates of domestic investment all tend to result in or increase a current account deficit. The intuition here is that low private or government savings in relation to private investment in domestic capital requires foreign investment in domestic capital.

We can make a distinction, however, between a trade deficit resulting from high government or private consumption and one resulting from high private investment in capital. In the first case, borrowing from foreign countries to finance high consumption (low savings) increases the domestic country's liabilities without any increase to its future productive power. In the second case, borrowing from foreign countries to finance a high level of private investment in domestic capital, the added liability is accompanied by an increase in future productive power because of the investment in capital.

LOS 14.j: Describe functions and objectives of the international organizations that facilitate trade, including the World Bank, the International Monetary Fund, and the World Trade Organization.

Perhaps the best way to understand the roles of the organizations designed to facilitate trade is to examine their own statements.

According to the **International Monetary Fund** (IMF; more available at www.IMF.org):

Article I of the Articles of Agreement sets out the IMF's main goals:

- promoting international monetary cooperation;
- facilitating the expansion and balanced growth of international trade;
- promoting exchange stability;
- assisting in the establishment of a multilateral system of payments; and
- making resources available (with adequate safeguards) to members experiencing balance of payments difficulties.

According to the **World Bank** (more available at www.WorldBank.org):

The World Bank is a vital source of financial and technical assistance to developing countries around the world. Our mission is to fight poverty with passion and professionalism for lasting results and to help people help themselves and their environment by providing resources, sharing knowledge, building capacity and forging partnerships in the public and private sectors.

We are not a bank in the common sense; we are made up of two unique development institutions owned by 187 member countries: the International Bank for Reconstruction and Development (IBRD) and the International Development Association (IDA).

Each institution plays a different but collaborative role in advancing the vision of inclusive and sustainable globalization. The IBRD aims to reduce poverty in middle-income and creditworthy poorer countries, while IDA focuses on the world's poorest countries.

...Together, we provide low-interest loans, interest-free credits and grants to developing countries for a wide array of purposes that include investments in education, health, public administration, infrastructure, financial and private sector development, agriculture and environmental and natural resource management.

According to the **World Trade Organization (WTO)**; more available at www.WTO.org):

The World Trade Organization (WTO) is the only international organization dealing with the global rules of trade between nations. Its main function is to ensure that trade flows as smoothly, predictably and freely as possible.

...Trade friction is channeled into the WTO's dispute settlement process where the focus is on interpreting agreements and commitments, and how to ensure that countries' trade policies conform with them. That way, the risk of disputes spilling over into political or military conflict is reduced.

...At the heart of the system—known as the multilateral trading system—are the WTO's agreements, negotiated and signed by a large majority of the world's trading nations, and ratified in their parliaments. These agreements are the legal ground-rules for international commerce. Essentially, they are contracts, guaranteeing member countries important trade rights. They also bind governments to keep their trade policies within agreed limits to everybody's benefit.



MODULE QUIZ 14.2

1. An agreement with another country to limit the volume of goods and services sold to them is *best* described as:
 - A. a quota.
 - B. a voluntary export restraint.
 - C. a minimum domestic content rule.
2. Which of the following groups would be *most likely* to suffer losses from the imposition of a tariff on steel imports?
 - A. Domestic steel producers.
 - B. Workers in the domestic auto industry.
 - C. Workers in the domestic steel industry.
3. The *most likely* motivation for establishing a trading bloc is to:
 - A. increase economic welfare in the member countries.
 - B. increase tariff revenue for the member governments.
 - C. protect domestic industries in the member economies.
4. In which type of regional trade agreement are economic policies conducted independently by the member countries, while labor and capital are free to move among member countries?
 - A. Free trade area.
 - B. Common market.
 - C. Economic union.
5. The goal of a government that imposes restrictions on foreign capital flows is *most likely* to:
 - A. stimulate domestic interest rates.

- B. decrease domestic asset price volatility.
 - C. encourage competition with domestic industries.
6. Which of the following is *least likely* a component of the current account?
- A. Unilateral transfers.
 - B. Payments for fixed assets.
 - C. Payments for goods and services.
7. A current account deficit is *most likely* to decrease as a result of an increase in:
- A. domestic savings.
 - B. private investment.
 - C. the fiscal budget deficit.
8. Which international organization is primarily concerned with providing economic assistance to developing countries?
- A. World Bank.
 - B. World Trade Organization.
 - C. International Monetary Fund.

KEY CONCEPTS

LOS 14.a

Gross domestic product is the total value of goods and services produced within a country's borders. Gross national product measures the total value of goods and services produced by the labor and capital supplied by a country's citizens, regardless of where the production takes place.

LOS 14.b

Free trade among countries increases overall economic welfare. Countries can benefit from trade because one country can specialize in the production of an export good and benefit from economies of scale. Economic welfare can also be increased by greater product variety, more competition, and a more efficient allocation of resources.

Costs of free trade are primarily losses to those in domestic industries that lose business to foreign competition, especially less efficient producers who leave an industry. While other domestic industries will benefit from freer trade policies, unemployment may increase over the period in which workers are retrained for jobs in the expanding industries. Some argue that greater income inequality may result, but overall the gains from liberalization of trade policies are thought to exceed the costs, so that the winners could conceivably compensate the losers and still be better off.

LOS 14.c

A country is said to have an absolute advantage in the production of a good if it can produce the good at lower cost in terms of resources relative to another country.

A country is said to have a comparative advantage in the production of a good if its opportunity cost in terms of other goods that could be produced instead is lower than that of another country.

LOS 14.d

The Ricardian model of trade has only one factor of production—labor. The source of differences in production costs and comparative advantage in Ricardo's model is differences in labor productivity due to differences in technology.

Heckscher and Ohlin presented a model in which there are two factors of production—capital and labor. The source of comparative advantage (differences in opportunity costs) in this model is differences in the relative amounts of each factor that countries possess.

LOS 14.e

Types of trade restrictions include:

- Tariffs: Taxes on imported good collected by the government.
- Quotas: Limits on the amount of imports allowed over some period.
- Minimum domestic content: Requirement that some percentage of product content must be from the domestic country.
- Voluntary export restraints: A country voluntarily restricts the amount of a good that can be exported, often in the hope of avoiding tariffs or quotas imposed by their trading partners.

Within each importing country, all of these restrictions will tend to:

- Increase prices of imports and decrease quantities of imports.
- Increase demand for and quantity supplied of domestically produced goods.
- Increase producer's surplus and decrease consumer surplus.

Export subsidies decrease export prices and benefit importing countries at the expense of the government of the exporting country.

Restrictions on the flow of financial capital across borders include outright prohibition of investment in the domestic country by foreigners, prohibition of or taxes on the income earned on foreign investments by domestic citizens, prohibition of foreign investment in certain domestic industries, and restrictions on repatriation of earnings of foreign entities operating in a country.

LOS 14.f

Trade agreements, which increase economic welfare by facilitating trade among member countries, take the following forms:

- Free trade area: All barriers to the import and export of goods and services among member countries are removed.
- Customs union: Member countries *also* adopt a common set of trade restrictions with non-members.
- Common market: Member countries *also* remove all barriers to the movement of labor and capital goods among members.
- Economic union: Member countries *also* establish common institutions and economic policy for the union.
- Monetary union: Member countries *also* adopt a single currency.

LOS 14.g

Commonly cited objectives of capital flow restrictions include:

- Reducing the volatility of domestic asset prices.
- Maintaining fixed exchange rates.

- Keeping domestic interest rates low and enabling greater independence regarding monetary policy.
- Protecting strategic industries from foreign ownership.

LOS 14.h

The balance of payments refers to the fact that increases in a country's assets and decreases in its liabilities must equal (balance with) decreases in its assets and increases in its liabilities.

These financial flows are classified into three types:

- The current account includes imports and exports of merchandise and services, foreign income from dividends on stock holdings and interest on debt securities, and unilateral transfers such as money received from those working abroad and direct foreign aid.
- The capital account includes debt forgiveness, assets that migrants bring to or take away from a country, transfer of funds for the purchase or sale of fixed assets, and purchases of non-financial assets, including rights to natural resources, patents, copyrights, trademarks, franchises, and leases.
- The financial account includes government-owned assets abroad such as gold, foreign currencies and securities, and direct foreign investment and claims against foreign banks. The financial account also includes foreign-owned assets in the country, domestic government and corporate securities, direct investment in the domestic country, and domestic country currency.

Overall, any surplus (deficit) in the current account must be offset by a deficit (surplus) in the capital and financial accounts.

LOS 14.i

In equilibrium, we have the relationship:

$$\text{exports} - \text{imports} = \text{private savings} + \text{government savings} - \text{domestic investment}$$

When total savings is less than domestic investment, exports must be less than imports so that there is a deficit in the current account. Lower levels of private saving, larger government deficits, and high rates of domestic investment all tend to result in or increase a current account deficit. The intuition here is that low private or government savings in relation to private investment in domestic capital requires foreign investment in domestic capital.

LOS 14.j

The International Monetary Fund facilitates trade by promoting international monetary cooperation and exchange rate stability, assists in setting up international payments systems, and makes resources available to member countries with balance of payments problems.

The World Bank provides low-interest loans, interest-free credits, and grants to developing countries for many specific purposes. It also provides resources and knowledge and helps form private/public partnerships with the overall goal of fighting poverty.

The World Trade Organization has the goal of ensuring that trade flows freely and works smoothly. Its main focus is on instituting, interpreting, and enforcing a number of multilateral trade agreements that detail global trade policies for a large majority of the world's trading nations.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 14.1

1. **A** The income from a financial investment in Country P of a citizen of Country Q is included in Country P's GDP but not its GNP. It is included in Country Q's GNP but not its GDP. (LOS 14.a)
2. **B** Openness to international trade increases specialization as production shifts to those products in which domestic producers have a comparative advantage. Greater competition from imports will tend to decrease prices for consumer goods. Increasing international trade is likely to increase profitability and employment in exporting industries but may decrease profitability and employment in industries that compete with imported goods. (LOS 14.b)
3. **A** Each country gains by *exporting* the good for which it has a comparative advantage. (LOS 14.c)
4. **C** In the Ricardian model, labor is the only factor of production considered. In the Heckscher-Ohlin model, comparative advantage results from the relative amounts of labor and capital available in different countries. (LOS 14.d)

Module Quiz 14.2

1. **B** Voluntary export restraints are agreements to limit the volume of goods and services exported to another country. Minimum domestic content rules are limitations imposed by a government on its domestic firms. Import quotas are limitations on imports, not on exports. (LOS 14.e)
2. **B** Imposing a tariff on steel imports benefits domestic steel producers and workers by increasing the domestic price of steel and benefits the national government by increasing tax (tariff) revenue. However, the increase in the domestic price of steel would increase costs in industries that use significant amounts of steel, such as the automobile industry. The resulting increase in the price of automobiles reduces the quantity of automobiles demanded and ultimately reduces employment in that industry. (LOS 14.e)
3. **A** The motivation for trading blocs is to increase economic welfare in the member countries by eliminating barriers to trade. Joining a trading bloc may have negative consequences for some domestic industries and may decrease tariff revenue for the government. (LOS 14.f)
4. **B** These characteristics describe a common market. In a free trade area, member countries remove restrictions on goods and services trade with one another but may still restrict movement of labor and capital among member countries. In an economic union, member countries also coordinate their economic policies and institutions. (LOS 14.f)
5. **B** Decreasing the volatility of domestic asset prices may be a goal of a government that imposes capital restrictions. Other typical goals include keeping domestic interest rates low and protecting certain domestic industries, such as the defense industry. (LOS 14.g)
6. **B** Purchases and sales of fixed assets are recorded in the capital account. Goods and services trade and unilateral transfers are components of the current account. (LOS 14.h)
7. **A** Other things equal, an increase in domestic savings would tend to decrease the current account deficit, while an increase in private investment or an increase in the fiscal budget deficit would tend to increase the current account deficit. (LOS 14.i)
8. **A** The World Bank provides technical and financial assistance to economically developing countries. The World Trade Organization is primarily concerned with settling disputes among countries concerning international trade. The International Monetary Fund promotes international trade and exchange rate stability and assists member countries that experience balance of payments trouble. (LOS 14.j)

READING 15

CURRENCY EXCHANGE RATES

EXAM FOCUS

Candidates must understand spot exchange rates, forward exchange rates, and all the calculations having to do with currency appreciation and depreciation. Additionally, candidates should understand the steps a country can take to decrease a trade deficit and the requirements for these to be effective under both the elasticities and absorption approaches. Finally, candidates should make sure to know the terms for and definitions of the various exchange rate regimes countries may adopt.

MODULE 15.1: FOREIGN EXCHANGE RATES



LOS 15.a: Define an exchange rate and distinguish between nominal and real exchange rates and spot and forward exchange rates.

Video covering this content is available online.

An **exchange rate** is simply the price or cost of units of one currency in terms of another. For the purposes of this book we will write 1.416 USD/EUR to mean that each euro costs \$1.416. If you read the “/” as *per*, you will have no trouble with the notation. We say the exchange rate is \$1.416 per euro.



PROFESSOR'S NOTE

There are alternative notations for foreign exchange quotes, but expressing them as the price of the denominator currency in terms of the numerator currency is what we will use and what you can expect on the Level I exam.

In a foreign currency quotation we have the price of one currency in units of another currency. These are often referred to as the **base currency** and the **price currency**. In the quotation 1.25 USD/EUR, the USD is the price currency and the EUR is the base currency. The price of one euro (base currency) is 1.25 USD (the price currency) so 1.25 is the price of one unit of the base currency in terms of the other. It may help to remember that the euro in this example is in the bottom or “base” of the exchange rate given in terms of USD/EUR.

Sometimes an exchange rate expressed as price currency/base currency is referred to as a **direct quote** from the point of view of an investor in the price currency country and an **indirect quote** from the point of view of an investor in the base currency country. For example, a quote of 1.17 USD/EUR would be a direct exchange rate quote for a USD-based investor and an indirect quote for a EUR-based investor. Conversely, a quote of $1 / 1.17 = 0.845$ EUR/USD would

be a direct exchange rate quote for a EUR-based investor and an indirect quote for a USD-based investor.

The exchange rate at a point in time is referred to as a **nominal exchange rate**. If the nominal exchange rate (price/base) increases, the cost of a unit of the base currency in terms of the price currency has increased, so that the purchasing power of the price currency has decreased. If the USD/EUR exchange rate increases from 1.10 to 1.15, the cost of 100 euros increases from \$110 to \$115. The purchasing power of the dollar has decreased relative to the euro because the cost of 100 euros worth of goods to a consumer in the U.S. has increased over the period.

The purchasing power of one currency relative to another is also affected by changes in the price levels of the two countries. The **real exchange rate** between two currencies refers to the purchasing power of one currency in terms of the amount of goods priced in another currency, relative to an earlier (base) period.

Consider a situation in which the nominal USD/EUR exchange rate is unchanged at 1:1 over a period and the price level in the U.S. is unchanged, while prices in the Eurozone have increased by 5%. Eurozone goods that cost 100 euros at the beginning of the period cost 105 euros at the end of the period. With the nominal exchange rate unchanged, the purchasing power of the USD in the Eurozone has decreased, because exchanging 100 USD for 100 EUR will now buy only $100/105 = 95.2\%$ of the goods 100 EUR could buy at the beginning of the period.

To summarize:

- An increase in the *nominal* USD/EUR rate decreases the purchasing power of the USD in the Eurozone (and increases the purchasing power of the EUR in the U.S.); the *real* USD/EUR exchange rate has increased.
- A decrease in the *nominal* USD/EUR rate increases the purchasing power of the USD in the Eurozone (and decreases the purchasing power of the EUR in the U.S.); the *real* USD/EUR exchange has rate decreased.
- An increase in the Eurozone price level, relative to the price level in the U.S., will increase the *real* USD/EUR exchange rate, decreasing the purchasing power of the USD in the Eurozone (and increasing the purchasing power of the EUR in the U.S.).
- A decrease in the Eurozone price level, relative to the price level in the U.S., will decrease the *real* USD/EUR exchange rate, increasing the purchasing power of the USD in the Eurozone (and decreasing the purchasing power of the EUR in the U.S.).

The end-of-period real P/B exchange rate can be calculated as:

$$\text{real P/B exchange rate} = \text{nominal P/B exchange rate} \times \frac{\text{CPI}_{\text{base currency}}}{\text{CPI}_{\text{price currency}}}$$

where the CPI values are relative to base period values of 100.

We can see from the formula that:

- An increase (decrease) in the nominal exchange rate over the period increases (decreases) the end-of period real exchange rate and the purchasing power of the price currency decreases (increases).
- An increase in the price level in the price currency country relative to the price level in the base currency country will decrease the real exchange rate, increasing the purchasing power

of the price currency in terms of base country goods.

- Conversely, a decrease in the price level in the price currency country relative to the price level in the base currency country will increase the real exchange rate, decreasing the purchasing power of the price currency in terms of base country goods.

In the following example we calculate the end-of-period real \$/£ exchange rate when the nominal \$/£ exchange rate has decreased over the period, (which tends to decrease the real exchange rate and increase the purchasing power of the price currency), and when the price level in the U.K. has increased by more than the price level in the U.S. over the period (which tends to increase the real exchange rate and decrease the purchasing power of the price currency). The relative increase in U.K. prices has reduced the effects of the decrease in the nominal exchange rate on the increase in the purchasing power of the USD.

EXAMPLE: Real exchange rate

At a base period, the CPIs of the U.S. and U.K. are both 100, and the exchange rate is \$1.70/£. Three years later, the exchange rate is \$1.60/£, and the CPI has risen to 110 in the United States and 112 in the U.K. What is the real exchange rate at the end of the three-year period?

Answer:

The real exchange rate is $\$1.60/\text{£} \times 112 / 110 = \$1.629/\text{£}$ which means that U.S. goods and services that cost \$1.70 at the base period now cost only \$1.629 (in real terms) if purchased in the U.K. and the real exchange rate, \$/£, has fallen. The decrease in the real exchange rate (and the increase in the purchasing power of the USD in terms of U.K. goods) over the period is less than it would have been if the relative prices between the two countries had not changed.

A **spot exchange rate** is the currency exchange rate for immediate delivery, which for most currencies means the exchange of currencies takes place two days after the trade.

A **forward exchange rate** is a currency exchange rate for an exchange to be done in the future. Forward rates are quoted for various future dates (e.g., 30 days, 60 days, 90 days, or one year). A forward is actually an agreement to exchange a specific amount of one currency for a specific amount of another on a future date specified in the forward agreement.

A French firm that will receive 10 million GBP from a British firm six months from now has uncertainty about the amount of euros that payment will be equivalent to six months from now. By entering into a forward agreement covering 10 million GBP at the 6-month forward rate of 1.192 EUR/GBP, the French firm has agreed to exchange 10 million GBP for 11.92 million euros in six months.

LOS 15.b: Calculate and interpret the percentage change in a currency relative to another currency.

Consider a USD/EUR exchange rate that has changed from 1.42 to 1.39 USD/EUR. The percentage change in the dollar price of a euro is simply $1.39 / 1.42 - 1 = -0.0211 = -2.11\%$. Because the dollar price of a euro has fallen, the euro has *depreciated* relative to the dollar, and a

euro now buys 2.11% fewer U.S. dollars. It is correct to say that the euro has depreciated by 2.11% relative to the dollar.

On the other hand, it is *not* correct to say that the dollar has appreciated by 2.11%. To calculate the percentage appreciation of the dollar, we need to convert the quotes to EUR/USD. So our beginning quote of 1.42 USD/EUR becomes $1 / 1.42 = 0.7042$ EUR/USD, and our ending quote of 1.39 USD/EUR becomes $1 / 1.39 = 0.7194$ EUR/USD. Using these exchange rates, we can calculate the change in the euro price of a dollar as $0.7194 / 0.7042 - 1 = 0.0216 = 2.16\%$. In this case, it is correct to say that the dollar has appreciated 2.16% with respect to the euro. For the same quotes, the percentage appreciation of the dollar is not the same as the percentage depreciation in the euro.

The key point to remember is that we can correctly calculate the percentage change of the *base currency* in a foreign exchange quotation.

LOS 15.c: Describe functions of and participants in the foreign exchange market.

Foreign currency markets serve companies and individuals that purchase or sell foreign goods and services denominated in foreign currencies. An even larger market, however, exists for capital flows. Foreign currencies are needed to purchase foreign physical assets as well as foreign financial securities.

Many companies have foreign exchange risk arising from their cross-border transactions. A Japanese company that expects to receive 10 million euros when a transaction is completed in 90 days has yen/euro exchange rate risk as a result. By entering into a **forward currency contract** to sell 10 million euros in 90 days for a specific quantity of yen, the firm can reduce or eliminate its foreign exchange risk associated with the transaction. When a firm takes a position in the foreign exchange market to reduce an existing risk, we say the firm is **hedging** its risk.

Alternatively, when a transaction in the foreign exchange markets increases currency risk, we term it a **speculative** transaction or position. Investors, companies, and financial institutions, such as banks and investment funds, all regularly enter into speculative foreign currency transactions.

The primary dealers in currencies and originators of forward foreign exchange (FX) contracts are large multinational banks. This part of the FX market is often called the **sell side**. On the other hand, the **buy side** consists of the many buyers of foreign currencies and forward FX contracts. These buyers include the following:

- **Corporations** regularly engage in cross-border transactions, purchase and sell foreign currencies as a result, and enter into FX forward contracts to hedge the risk of expected future receipts and payments denominated in foreign currencies.
- **Investment accounts** of many types transact in foreign currencies, hold foreign securities, and may both speculate and hedge with currency derivatives. **Real money accounts** refer to mutual funds, pension funds, insurance companies, and other institutional accounts that do not use derivatives. **Leveraged accounts** refer to the various types of investment firms that do use derivatives, including hedge funds, firms that trade for their own accounts, and other trading firms of various types.

- **Governments** and **government entities**, including **sovereign wealth funds** and pension funds, acquire foreign exchange for transactional needs, investment, or speculation. Central banks sometimes engage in FX transactions to affect exchange rates in the short term in accordance with government policy.
- The **retail market** refers to FX transactions by households and relatively small institutions and may be for tourism, cross-border investment, or speculative trading.

LOS 15.d: Calculate and interpret currency cross-rates.

The **cross rate** is the exchange rate between two currencies implied by their exchange rates with a common third currency. Cross rates are necessary when there is no active FX market in the currency pair. The rate must be computed from the exchange rates between each of these two currencies and a third currency, usually the USD or EUR.

Let's assume that we have the following quotations for Mexican pesos and Australian dollars: MXN/USD = 10.70 and USD/AUD = 0.60. The cross rate between Australian dollars and pesos (MXN/AUD) is:

$$\text{MXN/AUD} = \text{USD/AUD} \times \text{MXN/USD} = 0.60 \times 10.70 = 6.42$$

So our MXN/AUD cross rate is 6.42 pesos per Australian dollar. The key to calculating cross rates is to note that the basis of the quotations must be such that we get the desired result algebraically. If we had started with an AUD/USD quotation of 1.67, we would have taken the inverse to get the quotation into USD/AUD terms. Another approach is to divide through, as is illustrated in the following example.

EXAMPLE: Cross rate calculation

The spot exchange rate between the Swiss franc (CHF) and the USD is CHF/USD = 1.7799, and the spot exchange rate between the New Zealand dollar (NZD) and the U.S. dollar is NZD/USD = 2.2529. Calculate the CHF/NZD spot rate.

Answer:

The CHF/NZD cross rate is:

$$(\text{CHF/USD}) / (\text{NZD/USD}) = 1.7799 / 2.2529 = 0.7900$$



MODULE QUIZ 15.1

1. One year ago, the nominal exchange rate for USD/EUR was 1.300. Since then, the real exchange rate has increased by 3%. This *most likely* implies that:
 - A. the nominal exchange rate is less than USD/EUR 1.235.
 - B. the purchasing power of the euro has increased approximately 3% in terms of U.S. goods.
 - C. inflation in the euro zone was approximately 3% higher than inflation in the United States.
2. Sell-side participants in the foreign exchange market are *most likely* to include:
 - A. banks.
 - B. hedge funds.
 - C. insurance companies.

3. Suppose that the quote for British pounds (GBP) in New York is USD/GBP 1.3110. What is the quote for U.S. dollars (USD) in London (GBP/USD)?
 - A. 0.3110.
 - B. 0.7628.
 - C. 1.3110.
4. The Canadian dollar (CAD) exchange rate with the Japanese yen (JPY) changes from JPY/CAD 75 to JPY/CAD 78. The CAD has:
 - A. depreciated by 3.8%, and the JPY has appreciated by 4.0%.
 - B. appreciated by 3.8%, and the JPY has depreciated by 4.0%.
 - C. appreciated by 4.0%, and the JPY has depreciated by 3.8%.
5. Today's spot rate for the Indonesian rupiah (IDR) is IDR/USD 2,400.00, and the New Zealand dollar trades at NZD/USD 1.6000. The NZD/IDR cross rate is:
 - A. 0.00067.
 - B. 1,492.53.
 - C. 3,840.00.
6. The NZD is trading at USD/NZD 0.3500, and the SEK is trading at NZD/SEK 0.3100. The USD/SEK cross rate is:
 - A. 0.1085.
 - B. 8.8573.
 - C. 9.2166.

MODULE 15.2: FORWARD EXCHANGE RATES



LOS 15.e: Calculate an outright forward quotation from forward quotations expressed on a points basis or in percentage terms.

Video covering this content is available online.

A forward exchange rate quote typically differs from the spot quotation and is expressed in terms of the difference between the spot exchange rate and the forward exchange rate. One way to indicate this is with points. The unit of points is the last decimal place in the spot rate quote. For a spot currency quote to four decimal places, such as 2.3481, each point is 0.0001 or 1/10,000th. A quote of +18.3 points for a 90-day forward exchange rate means that the forward rate is 0.00183 more than the spot exchange rate.

EXAMPLE: Forward exchange rates in points

The AUD/EUR spot exchange rate is 0.7313 with the 1-year forward rate quoted at +3.5 points. What is the 1-year forward AUD/EUR exchange rate?

Answer:

The forward exchange rate is $0.7313 + 0.00035 = 0.73165$.

EXAMPLE: Forward exchange rates in percent

The AUD/EUR spot rate is quoted at 0.7313, and the 120-day forward exchange rate is given as -0.062%. What is the 120-day forward AUD/EUR exchange rate?

Answer:

The forward exchange rate is $0.7313 (1 - 0.00062) = 0.7308$.

LOS 15.f: Explain the arbitrage relationship between spot rates, forward rates, and interest rates.

When currencies are freely traded and forward currency contracts exist, the percentage difference between forward and spot exchange rates is approximately equal to the difference between the two countries' interest rates. This is because there is an arbitrage trade with a riskless profit to be made when this relation does not hold.

We call this a no-arbitrage condition because if it doesn't hold there is an opportunity to make a profit without risk. The possible arbitrage is as follows: borrow Currency A at interest rate A, convert it to Currency B at the spot rate and invest it to earn interest rate B, and sell the proceeds from this investment forward at the forward rate to turn it back into Currency A. If the forward rate does not correctly reflect the difference between interest rates, such an arbitrage could generate a profit to the extent that the return from investing Currency B and converting it back to Currency A with a forward contract is greater than the cost of borrowing Currency A for the period. We consider a numerical analysis of such an arbitrage later in this reading.

For spot and forward rates expressed as price currency/base currency, the no-arbitrage relation (commonly referred to as *interest rate parity*) is:

$$\frac{\text{forward}}{\text{spot}} = \frac{(1 + \text{interest rate}_{\text{price currency}})}{(1 + \text{interest rate}_{\text{base currency}})}$$

This formula can be rearranged as necessary in order to solve for specific values of the relevant terms.

LOS 15.g: Calculate and interpret a forward discount or premium.

The **forward discount** or **forward premium** for a currency is calculated relative to the spot exchange rate. The forward discount or premium *for the base currency* is the percentage difference between the forward price and the spot price.

Consider the following spot and forward exchange rates as the price in U.S. dollars of one euro.

$$\text{USD/EUR spot} = \$1.312 \quad \text{USD/EUR 90-day forward} = \$1.320$$

The (90-day) forward premium or discount on the euro = $\text{forward/spot} - 1 = 1.320 / 1.312 - 1 = 0.609\%$. Because this is positive, it is interpreted as a forward premium on the euro of 0.609%. Since we have the forward rate for 3 months, we could annualize the discount simply by multiplying by 4 (= 12 / 3).

Because the forward quote is greater than the spot quote, it will take more dollars to buy one euro 90 days from now, so the euro is expected to appreciate versus the dollar, and the dollar is expected to depreciate relative to the euro.

If the forward quote were less than the spot quote, the calculated amount would be negative and we would interpret that as a forward discount for the euro relative to the U.S. dollar.

LOS 15.h: Calculate and interpret the forward rate consistent with the spot rate and the interest rate in each currency.

EXAMPLE: Calculating the arbitrage-free forward exchange rate

Consider two currencies, the ABE and the DUB. The spot ABE/DUB exchange rate is 4.5671, the 1-year riskless ABE rate is 5%, and the 1-year riskless DUB rate is 3%. What is the 1-year forward exchange rate that will prevent arbitrage profits?

Answer:

Rearranging our formula, we have:

$$\text{forward} = \text{spot} \left(\frac{1 + I_{\text{ABE}}}{1 + I_{\text{DUB}}} \right) \text{ and we can calculate the forward rate as}$$

$$\text{forward} = 4.5671 \left(\frac{1.05}{1.03} \right) = 4.6558 \text{ ABE/DUB}$$

Note that the forward rate is greater than the spot rate by $4.6558 / 4.5671 - 1 = 1.94\%$. This is approximately equal to the interest rate differential of $5\% - 3\% = 2\%$. The currency with the higher interest rate should depreciate over time by approximately the amount of the interest rate differential.

If we are calculating a 90-day or 180-day forward exchange rate, we need to use interest rates for 90-day or 180-day periods rather than annual rates. Note that these shorter-term rates are quoted as annualized money market yields.

EXAMPLE: Calculating the arbitrage-free forward exchange rate with 90-day interest rates

The spot ABE/DUB exchange rate is 4.5671, the 90-day riskless ABE rate is 5%, and the 90-day riskless DUB rate is 3%. What is the 90-day forward exchange rate that will prevent arbitrage profits?

Answer:

$$\text{forward} = 4.5671 \left[\frac{1 + \frac{0.05}{4}}{1 + \frac{0.03}{4}} \right] = 4.5671 \left(\frac{1.0125}{1.0075} \right) = 4.5898 \text{ ABE/DUB}$$

In our previous example, we calculated the no-arbitrage one-year forward ABE/DUB exchange rate as 4.6558. Here, we illustrate the arbitrage profit that could be gained if the forward exchange rate differs from this no-arbitrage rate. Consider a forward rate of 4.6000 so that the depreciation in the ABE is less than that implied by interest rate parity. This makes the ABE attractive to a DUB investor who can earn a riskless profit as follows:

- Borrow 1,000 DUB for one year at 3% to purchase ABE and get 4,567.1 ABE.
- Invest the 4,567.1 ABE at the ABE rate of 5% to have $1.05(4,567.1) = 4,795.45$ ABE at the end of one year.
- Enter into a currency forward contract to exchange 4,795.45 ABE in one year at the forward rate of 4.6000 ABE/DUB in order to receive $4,795.45 / 4.6000 = 1,042.49$ DUB.

The investor has ended the year with a 4.249% return on his 1,000 DUB investment, which is higher than the 3% 1-year DUB interest rate. After repaying the 1,000 DUB loan plus interest (1,030 DUB), the investor has a profit of $1,042.49 - 1,030 = 12.49$ DUB with no risk and no initial out-of-pocket investment (i.e., a pure arbitrage profit).

Arbitrageurs will pursue this opportunity, buying ABE (driving down the spot ABE/DUB exchange rate) and selling ABE forward (driving up the forward ABE/DUB exchange rate), until the interest rate parity relation is restored and arbitrage profits are no longer available.



MODULE QUIZ 15.2

1. The spot CHF/GBP exchange rate is 1.3050. In the 180-day forward market, the CHF/GBP exchange rate is -42.5 points. The 180-day forward CHF/GBP exchange rate is *closest* to:
 - A. 1.2625.
 - B. 1.3008.
 - C. 1.3093.
2. The spot rate on the New Zealand dollar (NZD) is NZD/USD 1.4286, and the 180-day forward rate is NZD/USD 1.3889. This difference means:
 - A. interest rates are lower in the United States than in New Zealand.
 - B. interest rates are higher in the United States than in New Zealand.
 - C. it takes more NZD to buy one USD in the forward market than in the spot market.
3. The current spot rate for the British pound in terms of U.S. dollars is \$1.533 and the 180-day forward rate is \$1.508. Relative to the pound, the dollar is trading *closest* to a 180-day forward:
 - A. discount of 1.63%.
 - B. premium of 1.66%.
 - C. discount of 1.66%.
4. The annual interest rates in the United States (USD) and Sweden (SEK) are 4% and 7% per year, respectively. If the current spot rate is SEK/USD 9.5238, then the 1-year forward rate in SEK/USD is:
 - A. 9.2568.
 - B. 9.7985.
 - C. 10.2884.
5. The annual risk-free interest rate is 10% in the United States (USD) and 4% in Switzerland (CHF), and the 1-year forward rate is USD/CHF 0.80. Today's USD/CHF spot rate is *closest* to:
 - A. 0.7564.
 - B. 0.8462.
 - C. 0.8888.

MODULE 15.3: MANAGING EXCHANGE RATES



LOS 15.i: Describe exchange rate regimes.

Video covering this content is available online.

The IMF categorizes **exchange rate regimes** into the following types, two for countries that do not issue their own currencies and seven for countries that issue their own currencies.

Countries That Do Not Have Their Own Currency

- A country can use the currency of another country (**formal dollarization**). The country cannot have its own monetary policy, as it does not create money/currency.
- A country can be a member of a **monetary union** in which several countries use a common currency. Within the European Union, for example, most countries use the euro. While individual countries give up the ability to set domestic monetary policy, they all participate in determining the monetary policy of the European Central Bank.

Countries That Have Their Own Currency

- A **currency board arrangement** is an explicit commitment to exchange domestic currency for a specified foreign currency at a fixed exchange rate. A notable example of such an arrangement is Hong Kong. In Hong Kong, currency is (and may be) only issued when fully backed by holdings of an equivalent amount of U.S. dollars. The Hong Kong Monetary Authority can earn interest on its U.S. dollar balances. With dollarization, there is no such income, as the income is earned by the U.S. Federal Reserve when it buys interest-bearing assets with the U.S. currency it issues. While the monetary authority gives up the ability to conduct independent monetary policy and essentially imports the inflation rate of the outside currency, there may be some latitude to affect interest rates over the short term.
- In a **conventional fixed peg arrangement**, a country pegs its currency within margins of $\pm 1\%$ versus another currency or a basket that includes the currencies of its major trading or financial partners. The monetary authority can maintain exchange rates within the band by purchasing or selling foreign currencies in the foreign exchange markets (*direct intervention*). In addition, the country can use *indirect intervention*, including changes in interest rate policy, regulation of foreign exchange transactions, and convincing people to constrain foreign exchange activity. The monetary authority retains more flexibility to conduct monetary policy than with dollarization, a monetary union, or a currency board. However, changes in policy are constrained by the requirements of the peg.
- In a system of pegged exchange rates within horizontal bands or a **target zone**, the permitted fluctuations in currency value relative to another currency or basket of currencies are wider (e.g., $\pm 2\%$). Compared to a conventional peg, the monetary authority has more policy discretion because the bands are wider.
- With a **crawling peg**, the exchange rate is adjusted periodically, typically to adjust for higher inflation versus the currency used in the peg. This is termed a *passive crawling peg*, as opposed to an *active crawling peg* in which a series of exchange rate adjustments over time is announced and implemented. An active crawling peg can influence inflation expectations, adding some predictability to domestic inflation. Monetary policy is restricted in much the same way it is with a fixed peg arrangement.
- With **management of exchange rates within crawling bands**, the width of the bands that identify permissible exchange rates is increased over time. This method can be used to transition from a fixed peg to a floating rate when the monetary authority's lack of credibility

makes an immediate change to floating rates impractical. Again, the degree of monetary policy flexibility increases with the width of the bands.

- With a system of **managed floating exchange rates**, the monetary authority attempts to influence the exchange rate in response to specific indicators such as the balance of payments, inflation rates, or employment without any specific target exchange rate or predetermined exchange rate path. Intervention may be direct or indirect. Such management of exchange rates may induce trading partners to respond in ways that reduce stability.
- When a currency is **independently floating**, the exchange rate is market-determined, and foreign exchange market intervention is used only to slow the rate of change and reduce short-term fluctuations, not to keep exchange rates at a target level.

LOS 15.j: Explain the effects of exchange rates on countries' international trade and capital flows.

We address the question of how a change in exchange rates affects a country's balance of trade using two approaches. The **elasticities approach** focuses on the impact of exchange rate changes on the total value of imports and on the total value of exports. Because a trade deficit (surplus) must be offset by a surplus (deficit) in the capital account, we can also view the effects of a change in exchange rates on capital flows rather than on goods flows. The **absorption approach** to analyzing the effect of a change in exchange rates focuses on capital flows.

The relation between the balance of trade and capital flows is expressed by the identity we presented in the reading on Aggregate Output, Prices, and Economic Growth. This identity is:

$$\text{exports} - \text{imports} \equiv (\text{private savings} - \text{investment in physical capital}) + (\text{tax revenue} - \text{government spending})$$

or

$$X - M \equiv (S - I) + (T - G)$$

The intuition is that a trade deficit ($X - M < 0$) means that the right-hand side must also be negative so that the total savings (private savings + government savings) is less than domestic investment in physical capital. The additional amount to fund domestic investment must come from foreigners, so there is a surplus in the capital account to offset the deficit in the trade account. Another thing we can see from this identity is that any government deficit not funded by an excess of domestic saving over domestic investment is consistent with a trade deficit (imports > exports) which is offset by an inflow of foreign capital (a surplus in the capital account).

Elasticities Approach

This approach to understanding the impact of exchange rate changes on the balance of trade focuses on how exchange rate changes affect total expenditures on imports and exports. Consider an initial situation in which a country has a merchandise trade deficit (i.e., its imports exceed its exports). Depreciation of the domestic currency will make imports more expensive in domestic currency terms and exports less expensive in foreign currency terms. Thus, depreciation of the domestic currency will increase exports and decrease imports and would seem to unambiguously reduce the trade deficit. However, it is not the *quantity* of imports and

exports, but the total *expenditures* on imports and exports that must change in order to affect the trade deficit. Thus, the elasticity of demand for export goods and import goods is a crucial part of the analysis.

The condition under which a depreciation of the domestic currency will decrease a trade deficit are given in what is called the generalized **Marshall-Lerner condition**:

$$W_X \epsilon_X + W_M (\epsilon_M - 1) > 0$$

where:

W_X = proportion of total trade that is exports

W_M = proportion of total trade that is imports

ϵ_X = absolute value of price elasticity of demand for exports

ϵ_M = absolute value of price elasticity of demand for imports

In the case where import expenditures and export revenues are equal, $W_X = W_M$, this condition reduces to $\epsilon_X + \epsilon_M > 1$, which is most often cited as the classic Marshall-Lerner condition.

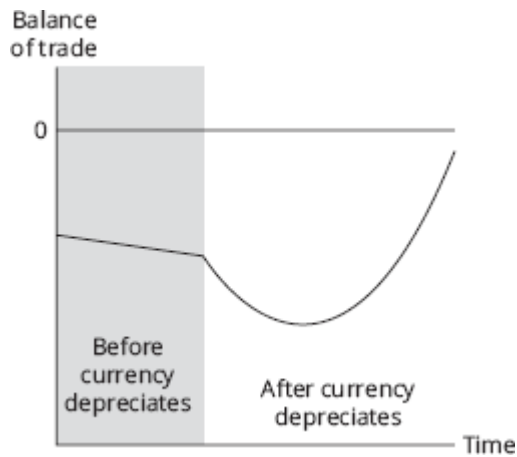
The elasticities approach tells us that currency depreciation will result in a greater improvement in the trade deficit when either import or export demand is elastic. For this reason, the compositions of export goods and import goods are an important determinant of the success of currency depreciation in reducing a trade deficit. In general, elasticity of demand is greater for goods with close substitutes, goods that represent a high proportion of consumer spending, and luxury goods in general. Goods that are necessities, have few or no good substitutes, or represent a small proportion of overall expenditures tend to have less elastic demand. Thus, currency depreciation will have a greater effect on the balance of trade when import or export goods are primarily luxury goods, goods with close substitutes, and goods that represent a large proportion of overall spending.

The J-Curve

Because import and export contracts for the delivery of goods most often require delivery and payment in the future, import and export quantities may be relatively insensitive to currency depreciation in the short run. This means that a currency depreciation may worsen a trade deficit initially. Importers adjust over time by reducing quantities. The Marshall-Lerner conditions take effect and the currency depreciation begins to improve the trade balance.

This short-term increase in the deficit followed by a decrease when the Marshall-Lerner condition is met is referred to as the **J-curve** and is illustrated in Figure 15.1.

Figure 15.1: J-Curve Effect



The Absorption Approach

One shortcoming of the elasticities approach is that it only considers the microeconomic relationship between exchange rates and trade balances. It ignores capital flows, which must also change as a result of a currency depreciation that improves the balance of trade. The absorption approach is a macroeconomic technique that focuses on the capital account and can be represented as:

$$BT = Y - E$$

where:

Y = domestic production of goods and services or national income

E = domestic absorption of goods and services, which is total expenditure

BT = balance of trade

Viewed in this way, we can see that income relative to expenditure must increase (domestic absorption must fall) for the balance of trade to improve in response to a currency depreciation. For the balance of trade to improve, domestic saving must increase relative to domestic investment in physical capital (which is a component of E). Thus, for a depreciation of the domestic currency to improve the balance of trade towards surplus, it must increase national income relative to expenditure. We can also view this as a requirement that national saving increase relative to domestic investment in physical capital.

Whether a currency depreciation has these effects depends on the current level of capacity utilization in the economy. When an economy is operating at less than full employment, the currency depreciation makes domestic goods and assets relatively more attractive than foreign goods and assets. The resulting shift in demand away from foreign goods and assets and towards domestic goods and assets will increase both expenditures and income. Because part of the income increase will be saved, national income will increase more than total expenditure, improving the balance of trade.

In a situation where the economy is operating at full employment (capacity), an increase in domestic spending will translate to higher domestic prices, which can reverse the relative price changes of the currency depreciation, resulting in a return to the previous deficit in the balance of trade. A currency depreciation at full capacity does result in a decline in the value of domestic assets. This decline in savers' real wealth will induce an increase in saving to rebuild wealth, initially improving the balance of trade from the currency depreciation. As the real

wealth of savers increases, however, the positive impact on saving will decrease, eventually returning the economy to its previous state and balance of trade.



MODULE QUIZ 15.3

1. The monetary authority of The Stoddard Islands will exchange its currency for U.S. dollars at a one-for-one ratio. As a result, the exchange rate of the Stoddard Islands currency with the U.S. dollar is 1.00, and many businesses in the Islands will accept U.S. dollars in transactions. This exchange rate regime is *best* described as:
 - A. a fixed peg.
 - B. dollarization.
 - C. a currency board.
2. A country that wishes to narrow its trade deficit devalues its currency. If domestic demand for imports is perfectly price-inelastic, whether devaluing the currency will result in a narrower trade deficit is *least likely* to depend on:
 - A. the size of the currency devaluation.
 - B. the country's ratio of imports to exports.
 - C. price elasticity of demand for the country's exports.
3. A devaluation of a country's currency to improve its trade deficit would *most likely* benefit a producer of:
 - A. luxury goods for export.
 - B. export goods that have no close substitutes.
 - C. an export good that represents a relatively small proportion of consumer expenditures.
4. Other things equal, which of the following is *most likely* to decrease a country's trade deficit?
 - A. Increase its capital account surplus.
 - B. Decrease expenditures relative to income.
 - C. Decrease domestic saving relative to domestic investment.

KEY CONCEPTS

LOS 15.a

Currency exchange rates are given as the price of one unit of currency in terms of another. A nominal exchange rate of 1.44 USD/EUR is interpreted as \$1.44 per euro. We refer to the USD as the price currency and the EUR as the base currency.

An increase (decrease) in an exchange rate represents an appreciation (depreciation) of the base currency relative to the price currency.

A spot exchange rate is the rate for immediate delivery. A forward exchange rate is a rate for exchange of currencies at some future date.

A real exchange rate measures changes in relative purchasing power over time.

$$\text{real exchange rate} = \text{nominal exchange rate} \times \left(\frac{\text{CPI}_{\text{base currency}}}{\text{CPI}_{\text{price currency}}} \right)$$

LOS 15.b

For a change in an exchange rate, we can calculate the percentage appreciation (price goes up) or depreciation (price goes down) of the base currency. For example, a decrease in the USD/EUR exchange rate from 1.44 to 1.42 represents a depreciation of the EUR relative to the USD of 1.39% ($1.42 / 1.44 - 1 = -0.0139$) because the price of a euro has fallen 1.39%.

To calculate the appreciation or depreciation of the price currency, we first invert the quote so it is now the base currency and then proceed as above. For example, a decrease in the USD/EUR exchange rate from 1.44 to 1.42 represents an appreciation of the USD relative to the EUR of 1.41%: $(1 / 1.42) / (1 / 1.44) - 1 = \frac{1.44}{1.42} - 1 = 0.0141$

The appreciation is the inverse of the depreciation, $\frac{1}{(1 - 0.0139)} - 1 = 0.0141$.

LOS 15.c

The market for foreign exchange is the largest financial market in terms of the value of daily transactions and has a variety of participants, including large multinational banks (the sell side) and corporations, investment fund managers, hedge fund managers, investors, governments, and central banks (the buy side).

Participants in the foreign exchange markets are referred to as hedgers if they enter into transactions that decrease an existing foreign exchange risk and as speculators if they enter into transactions that increase their foreign exchange risk.

LOS 15.d

Given two exchange rate quotes for three different currencies, we can calculate a currency cross rate. If the MXN/USD quote is 12.1 and the USD/EUR quote is 1.42, we can calculate the cross rate of MXN/EUR as $12.1 \times 1.42 = 17.18$.

LOS 15.e

Points in a foreign currency quotation are in units of the last digit of the quotation. For example, a forward quote of +25.3 when the USD/EUR spot exchange rate is 1.4158 means that the forward exchange rate is $1.4158 + 0.00253 = 1.41833$ USD/EUR.

For a forward exchange rate quote given as a percentage, the percentage (change in the spot rate) is calculated as forward / spot - 1. A forward exchange rate quote of +1.787%, when the spot USD/EUR exchange rate is 1.4158, means that the forward exchange rate is $1.4158 (1 + 0.01787) = 1.4411$ USD/EUR.

LOS 15.f

If a forward exchange rate does not correctly reflect the difference between the interest rates for two currencies, an arbitrage opportunity for a riskless profit exists. In this case, borrowing one currency, converting it to the other currency at the spot rate, investing the proceeds for the period, and converting the end-of-period amount back to the borrowed currency at the forward rate will produce more than enough to pay off the initial loan, with the remainder being a riskless profit on the arbitrage transaction.

LOS 15.g

To calculate a forward premium or forward discount for Currency B using exchange rates quoted as units of Currency A per unit of Currency B, use the following formula:

$$(\text{forward} / \text{spot}) - 1$$

LOS 15.h

The condition that must be met so that there is no arbitrage opportunity available is:

$$\frac{\text{forward}}{\text{spot}} = \frac{(1 + i_{\text{price currency}})}{(1 + i_{\text{base currency}})} \text{ so that } \text{forward} = \text{spot} \times \frac{(1 + i_{\text{price currency}})}{(1 + i_{\text{base currency}})}$$

If the spot exchange rate for the euro is 1.25 USD/EUR, the euro interest rate is 4% per year, and the dollar interest rate is 3% per year, the no-arbitrage one-year forward rate can be calculated as:

$$1.25 \times (1.03 / 1.04) = 1.238 \text{ USD/EUR.}$$

LOS 15.i

Exchange rate regimes for countries that do not have their own currency:

- With *formal dollarization*, a country uses the currency of another country.
- In a *monetary union*, several countries use a common currency.

Exchange rate regimes for countries that have their own currency:

- A *currency board arrangement* is an explicit commitment to exchange domestic currency for a specified foreign currency at a fixed exchange rate.
- In a *conventional fixed peg arrangement*, a country pegs its currency within margins of $\pm 1\%$ versus another currency.
- In a system of *pegged exchange rates within horizontal bands* or a *target zone*, the permitted fluctuations in currency value relative to another currency or basket of currencies are wider (e.g., $\pm 2\%$).
- With a *crawling peg*, the exchange rate is adjusted periodically, typically to adjust for higher inflation versus the currency used in the peg.
- With *management of exchange rates within crawling bands*, the width of the bands that identify permissible exchange rates is increased over time.
- With a system of *managed floating exchange rates*, the monetary authority attempts to influence the exchange rate in response to specific indicators, such as the balance of payments, inflation rates, or employment without any specific target exchange rate.
- When a currency is *independently floating*, the exchange rate is market-determined.

LOS 15.j

Elasticities (ϵ) of export and import demand must meet the Marshall-Lerner condition for a depreciation of the domestic currency to reduce an existing trade deficit:

$$W_X \epsilon_X + W_M (\epsilon_M - 1) > 0$$

Under the absorption approach, national income must increase relative to national expenditure in order to decrease a trade deficit. This can also be viewed as a requirement that national saving must increase relative to domestic investment in order to decrease a trade deficit.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 15.1

1. **B** An increase in the real exchange rate USD/EUR (the number of USD per one EUR) means a euro is worth more in purchasing power (real) terms in the United States. Changes in a real exchange rate depend on the change in the nominal exchange rate relative to the difference in

inflation. By itself, a real exchange rate does not indicate the directions or degrees of change in either the nominal exchange rate or the inflation difference. (LOS 15.a)

2. **A** Large multinational banks make up the sell side of the foreign exchange market. The buy side includes corporations, real money and leveraged investment accounts, governments and government entities, and retail purchasers of foreign currencies. (LOS 15.c)
3. **B** $1 / 1.311 = 0.7628$ GBP/USD. (LOS 15.a)
4. **C** The CAD has appreciated because it is worth a larger number of JPY. The percent appreciation is $(78 - 75) / 75 = 4.0\%$. To calculate the percentage depreciation of the JPY against the CAD, convert the exchange rates to make JPY the base currency: $1 / 75 = 0.0133$ CAD/JPY and $1 / 78 = 0.0128$ CAD/JPY. Percentage depreciation = $(0.0128 - 0.0133) / 0.0133 = -3.8\%$. (LOS 15.b)
5. **A** Start with one NZD and exchange for $1 / 1.6 = 0.625$ USD. Exchange the USD for $0.625 \times 2,400 = 1,500$ IDR. We get a cross rate of 1,500 IDR/NZD or $1 / 1,500 = 0.00067$ NZD/IDR. (LOS 15.d)
6. **A** USD/NZD $0.3500 \times$ NZD/SEK $0.3100 =$ USD/SEK 0.1085 .

Notice that the NZD term cancels in the multiplication. (LOS 15.d)

Module Quiz 15.2

1. **B** The 180-day forward exchange rate is $1.3050 - 0.00425 =$ CHF/GBP 1.30075 . (LOS 15.e)
2. **B** Interest rates are higher in the United States than in New Zealand. It takes fewer NZD to buy one USD in the forward market than in the spot market. (LOS 15.f)
3. **B** To calculate a percentage forward premium or discount for the U.S. dollar, we need the dollar to be the base currency. The spot and forward quotes given are U.S. dollars per British pound (USD/GBP), so we must invert them to GBP/USD. The spot GBP/USD price is $1 / 1.533 = 0.6523$ and the forward GBP/USD price is $1 / 1.508 = 0.6631$. Because the forward price is greater than the spot price, we say the dollar is at a forward premium of $0.6631 / 0.6523 - 1 = 1.66\%$. Alternatively, we can calculate this premium with the given quotes as spot/forward - 1 to get $1.533 / 1.508 - 1 = 1.66\%$. (LOS 15.g)
4. **B** The forward rate in SEK/USD is $9.5238 \left(\frac{1.07}{1.04} \right) = 9.7985$. Since the SEK interest rate is the higher of the two, the SEK must depreciate approximately 3%. (LOS 15.h)
5. **A** We can solve interest rate parity for the spot rate as follows: With the exchange rates quoted as USD/CHF, the spot is $0.80 \left(\frac{1.04}{1.10} \right) = 0.7564$. Since the interest rate is higher in the United States, it should take fewer USD to buy CHF in the spot market. In other words, the forward USD must be depreciating relative to the spot. (LOS 15.h)

Module Quiz 15.3

1. **C** This exchange rate regime is a currency board arrangement. The country has not formally dollarized because it continues to issue a domestic currency. A conventional fixed peg allows for a small degree of fluctuation around the target exchange rate. (LOS 15.i)
2. **A** With perfectly inelastic demand for imports, currency devaluation of any size will increase total expenditures on imports (same quantity at higher prices in the home currency). The trade deficit will narrow only if the increase in export revenues is larger than the increase in import spending. To satisfy the Marshall-Lerner condition when import demand elasticity is zero, export demand elasticity must be larger than the ratio of imports to exports in the country's international trade. (LOS 15.j)
3. **A** A devaluation of the currency will reduce the price of export goods in foreign currency terms. The greatest benefit would be to producers of goods with more elastic demand. Luxury goods tend to have higher elasticity of demand, while goods that have no close substitutes or represent a small proportion of consumer expenditures tend to have low elasticities of demand. (LOS 15.j)

4. **B** An improvement in a trade deficit requires that domestic savings increase relative to domestic investment, which would decrease a capital account surplus. Decreasing expenditures relative to income means domestic savings increase. Decreasing domestic saving relative to domestic investment is consistent with a larger capital account surplus (an increase in net foreign borrowing) and a greater trade deficit. (LOS 15.j)

TOPIC QUIZ: ECONOMICS

You have now finished the Economics topic section. Please log into your Schweser online dashboard and take the Topic Quiz on Economics. The Topic Quiz provides immediate feedback on how effective your study has been for this material. The number of questions on this quiz is approximately the number of questions for the topic on one-half of the actual Level I CFA exam. Questions are more exam-like than typical Module Quiz or QBank questions; a score of less than 70% indicates that your study likely needs improvement. These tests are best taken timed; allow 1.5 minutes per question.

After you've completed this Topic Quiz, select "Performance Tracker" to view a breakdown of your score. Select "Compare with Others" to display how your score on the Topic Quiz compares to the scores of others who entered their answers.

FORMULAS

nominal risk-free rate = real risk-free rate + expected inflation rate

required interest rate on a security = nominal risk-free rate
+ default risk premium
+ liquidity premium
+ maturity risk premium

effective annual rate = $(1 + \text{periodic rate})^m - 1$

continuous compounding: $e^r - 1 = \text{EAR}$

$$\text{PV}_{\text{perpetuity}} = \frac{\text{PMT}}{I/Y}$$

$$\text{FV} = \text{PV}(1 + I/Y)^N$$

$$\text{population mean: } \mu = \frac{\sum_{i=1}^N X_i}{N}$$

$$\text{sample mean: } \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

geometric mean return (R_G): $1 + R_G = \sqrt[n]{(1 + R_1) \times (1 + R_2) \times \dots \times (1 + R_n)}$

$$\text{harmonic mean: } X_H = \frac{N}{\sum_{i=1}^N \frac{1}{X_i}}$$

$$\text{weighted mean: } \bar{X}_w = \sum_{i=1}^n w_i X_i$$

position of the observation at a given percentile, y : $L_y = (n + 1) \frac{y}{100}$

range = maximum value – minimum value

excess kurtosis = sample kurtosis – 3

$$\text{MAD} = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}$$

$$\text{population variance} = \sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N},$$

where μ = population mean and N = number of possible outcomes

$$\text{sample variance} = s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1},$$

where \bar{X} = sample mean and n = sample size

$$\text{coefficient of variation: } CV = \frac{s_x}{\bar{X}} = \frac{\text{standard deviation of } x}{\text{average value of } x}$$

$$\text{target downside deviation: } s_{\text{target}} = \sqrt{\frac{\sum_{\text{all } X_i < B} (X_i - B)^2}{n - 1}},$$

$$\text{joint probability: } P(AB) = P(A | B) \times P(B)$$

$$\text{addition rule: } P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

$$\text{multiplication rule: } P(A \text{ and } B) = P(A) \times P(B)$$

$$\text{total probability rule: } P(R) = P(R | S_1) \times P(S_1) + P(R | S_2) \times P(S_2) + \dots + P(R | S_N) \times P(S_N)$$

$$\text{expected value: } E(X) = \sum P(x_i)x_i = P(x_1)x_1 + P(x_2)x_2 + \dots + P(x_n)x_n$$

$$\text{variance from a probability model: } \text{Var}(X) = E\{[X - E(X)]^2\}$$

$$\text{Cov}(R_i, R_j) = E\{[R_i - E(R_i)][R_j - E(R_j)]\}$$

$$\text{Corr}(R_i, R_j) = \frac{\text{Cov}(R_i, R_j)}{\sigma(R_i)\sigma(R_j)}$$

$$\text{portfolio expected return: } E(R_p) = \sum_{i=1}^N w_i E(R_i) = w_1 E(R_1) + w_2 E(R_2) + \dots + w_n E(R_n)$$

$$\text{portfolio variance: } \text{Var}(R_p) = \sum_{i=1}^N \sum_{j=1}^N w_i w_j \text{Cov}(R_i, R_j)$$

$$\text{where } w_i = \frac{\text{market value of investment in asset } i}{\text{market value of the portfolio}}$$

Bayes' formula:

updated probability = $\frac{\text{probability of new information for a given event}}{\text{unconditional probability of new information}} \times \text{prior probability of event}$

combination (binomial) formula: ${}_n C_r = \frac{n!}{(n-r)!r!}$

permutation formula: ${}_n P_r = \frac{n!}{(n-r)!}$

binomial probability: $p(x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$

for a binomial random variable: $E(X) = np$; variance = $np(1-p)$

for a normal variable:

90% confidence interval for \bar{X} is $\bar{X} - 1.65s$ to $\bar{X} + 1.65s$

95% confidence interval for \bar{X} is $\bar{X} - 1.96s$ to $\bar{X} + 1.96s$

99% confidence interval for \bar{X} is $\bar{X} - 2.58s$ to $\bar{X} + 2.58s$

$$z = \frac{\text{observation} - \text{population mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

$$\text{SFRatio} = \frac{[E(R_p) - R_L]}{\sigma_p}$$

continuously compounded rate of return: $r_{cc} = \ln\left(\frac{S_1}{S_0}\right) = \ln(1 + \text{HPR})$

for a uniform distribution: $P(x_1 \leq X \leq x_2) = \frac{(x_2 - x_1)}{(b - a)}$

sampling error of the mean = sample mean - population mean = $\bar{x} - \mu$

standard error of the sample mean, known population variance: $\sigma_x = \frac{\sigma}{\sqrt{n}}$

standard error of the sample mean, unknown population variance: $s_x = \frac{s}{\sqrt{n}}$

confidence interval: point estimate \pm (reliability factor \times standard error)

confidence interval for the population mean: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

tests for population mean = μ_0 : z-statistic = $\frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$, t-statistic = $\frac{\bar{x} - \mu_0}{s / \sqrt{n}}$

test for equality of variances: $F = \frac{s_1^2}{s_2^2}$, where $s_1^2 > s_2^2$

paired comparisons test: t -statistic = $\frac{\bar{d} - \mu_{dz}}{s_d}$

test for differences in means:

t -statistic = $\frac{(\bar{x}_1 - \bar{x}_2)}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{1/2}}$ (sample variances assumed unequal)

t -statistic = $\frac{(\bar{x}_1 - \bar{x}_2)}{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)^{1/2}}$ (sample variances assumed equal)

test for correlation: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

regression slope: $\hat{b}_1 = \frac{\text{Cov}_{XY}}{\sigma_X^2}$

coefficient of determination: $R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$

standard error of estimate = $\sqrt{\text{mean squared error}}$

own-price elasticity = $\frac{\% \text{ change in quantity demanded}}{\% \text{ change in own price}}$

income elasticity = $\frac{\% \text{ change in quantity demanded}}{\% \text{ change in income}}$

cross-price elasticity = $\frac{\% \text{ change in quantity demanded}}{\% \text{ change in price of related good}}$

breakeven points:

perfect competition: $AR = ATC$

imperfect competition: $TR = TC$

short-run shutdown points:

perfect competition: $AR < AVC$

imperfect competition: $TR < TVC$

$$\begin{aligned} \text{nominal GDP}_t \text{ for year } t &= \sum_{i=1}^N P_{i,t} Q_{i,t} \\ &= \sum_{i=1}^N \left(\text{price of good } i \text{ in year } t \right) \\ &\quad \times \left(\text{quantity of good } i \text{ produced in year } t \right) \end{aligned}$$

$$\begin{aligned} \text{real GDP for year } t &= \sum_{i=1}^N P_{i, \text{base year}} Q_{i,t} \\ &= \sum_{i=1}^N \left(\text{price of good } i \text{ in base year} \right) \\ &\quad \times \left(\text{quantity of good } i \text{ produced in year } t \right) \end{aligned}$$

GDP deflator for year t

$$= \frac{\sum_{i=1}^N P_{i,t} Q_{i,t}}{\sum_{i=1}^N P_{i, \text{base year}} Q_{i,t}} \times 100 = \frac{\text{nominal GDP in year } t}{\text{value of year } t \text{ output at base year prices}} \times 100$$

GDP, expenditure approach:

$$\text{GDP} = C + I + G + (X - M)$$

where:

C = consumption spending

I = business investment (capital equipment, inventories)

G = government purchases

X = exports

M = imports

GDP, income approach:

$$\text{GDP} = \text{national income} + \text{capital consumption allowance} + \text{statistical discrepancy}$$

national income = compensation of employees (wages and benefits)
 + corporate and government enterprise profits before taxes
 + interest income
 + unincorporated business net income (business owners' incomes)
 + rent
 + indirect business taxes – subsidies (taxes and subsidies that are included in final prices)

$$\text{growth in potential GDP} = \text{growth in technology} + W_L(\text{growth in labor}) + W_C(\text{growth in capital})$$

where:

W_L = labor's percentage share of national income

W_C = capital's percentage share of national income

growth in per-capita potential GDP = growth in technology + W_C (growth in the capital-to-labor ratio)

where:

W_C = capital's percentage share of national income

consumer price index = $\frac{\text{cost of basket at current prices}}{\text{cost of basket at base period prices}} \times 100$

money multiplier = $\frac{1}{\text{reserve requirement}}$

equation of exchange: money supply \times velocity = price \times real output (MV = PY)

Fisher effect: nominal interest rate = real interest rate + expected inflation rate

neutral interest rate = real trend rate of economic growth + inflation target

fiscal multiplier:

$$\frac{1}{1 - \text{MPC}(1 - t)}$$

where:

t = tax rate

MPC = marginal propensity to consume

real exchange rate = nominal exchange rate \times $\left(\frac{\text{CPI}_{\text{base currency}}}{\text{CPI}_{\text{price currency}}} \right)$

real exchange rate = $\frac{\text{nominal exchange rate}}{\left(\frac{\text{CPI}_{\text{price currency}}}{\text{CPI}_{\text{base currency}}} \right)}$

forward premium (+) or discount (-) for the base currency:

$$\frac{\text{forward}}{\text{spot}} - 1$$

interest rate parity:

$$\frac{\text{forward}}{\text{spot}} = \frac{(1 + \text{interest rate}_{\text{price currency}})}{(1 + \text{interest rate}_{\text{base currency}})}$$

Marshall-Lerner condition:

$$W_X \epsilon_X + W_M (\epsilon_M - 1) > 0$$

where:

W_M = proportion of trade that is imports

W_X = proportion of trade that is exports

ϵ_M = elasticity of demand for imports

ϵ_X = elasticity of demand for exports

APPENDICES

APPENDIX A: AREAS UNDER THE NORMAL CURVE

Most of the examples in this book have used one version of the z-table to find the area under the normal curve. This table provides the cumulative probabilities (or the area under the entire curve to left of the z-value).

Probability Example

Assume that the annual earnings per share (EPS) for a large sample of firms is normally distributed with a mean of \$5.00 and a standard deviation of \$1.50. What is the approximate probability of an observed EPS value falling between \$3.00 and \$7.25?

If $EPS = x = \$7.25$, then $z = (x - \mu)/\sigma = (\$7.25 - \$5.00)/\$1.50 = +1.50$

If $EPS = x = \$3.00$, then $z = (x - \mu)/\sigma = (\$3.00 - \$5.00)/\$1.50 = -1.33$

Solving Using The Cumulative Z-Table

For z-value of 1.50: Use the row headed 1.5 and the column headed 0 to find the value 0.9332. This represents the area under the curve to the left of the critical value 1.50.

For z-value of -1.33: Use the row headed 1.3 and the column headed 3 to find the value 0.9082. This represents the area under the curve to the left of the critical value +1.33. The area to the left of -1.33 is $1 - 0.9082 = 0.0918$.

The area between these critical values is $0.9332 - 0.0918 = 0.8414$, or 84.14%.

Hypothesis Testing—One-Tailed Test Example

A sample of a stock's returns on 36 nonconsecutive days results in a mean return of 2.0%. Assume the population standard deviation is 20.0%. Can we say with 95% confidence that the mean return is greater than 0%?

$H_0: \mu \leq 0.0\%$, $H_a: \mu > 0.0\%$. The test statistic = z-statistic

$$= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = (2.0 - 0.0) / (20.0 / 6) = 0.60$$

The significance level = $1.0 - 0.95 = 0.05$, or 5%. Because we are interested in a return greater than 0.0%, this is a one-tailed test.

Using the Cumulative Z-Table

Because this is a one-tailed test with an alpha of 0.05, we need to find the value 0.95 in the cumulative z-table. The closest value is 0.9505, with a corresponding critical z-value of 1.65. Because the test statistic is less than the critical value, we fail to reject H_0 .

Hypothesis Testing—Two-Tailed Test Example

Using the same assumptions as before, suppose that the analyst now wants to determine if he can say with 99% confidence that the stock's return is not equal to 0.0%.

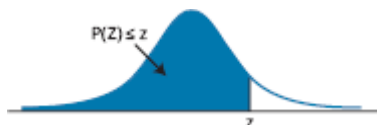
$H_0: \mu = 0.0\%$, $H_a: \mu \neq 0.0\%$. The test statistic (z-value) = $(2.0 - 0.0) / (20.0 / 6) = 0.60$. The significance level = $1.0 - 0.99 = 0.01$, or 1%. Because we are interested in whether or not the stock return is nonzero, this is a two-tailed test.

Using the Cumulative Z-Table

Because this is a two-tailed test with an alpha of 0.01, there is a 0.005 rejection region in both tails. Thus, we need to find the value 0.995 ($1.0 - 0.005$) in the table. The closest value is 0.9951, which corresponds to a critical z-value of 2.58. Because the test statistic is less than the critical value, we fail to reject H_0 and conclude that the stock's return equals 0.0%.

CUMULATIVE Z-TABLE

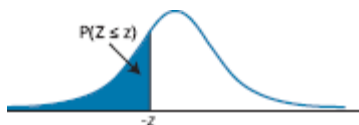
Standard Normal Distribution



$$P(Z \leq z) = N(z) \text{ for } z \geq 0$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Standard Normal Distribution



$P(Z \leq z) = N(z)$ for $z \geq 0$

z	0.00	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
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1057	0.1038	0.1020	0.1003	0.0985
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.4	0.0082	0.0080	0.0078	0.0076	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

APPENDIX B: STUDENT'S *T*-DISTRIBUTION

Level of Significance for One-Tailed Test						
df	0.100	0.050	0.025	0.01	0.005	0.0005
Level of Significance for Two-Tailed Test						
df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.294
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

APPENDIX C: *F*-TABLE AT 5% (UPPER TAIL)

F-Table, Critical Values, 5% in Upper Tail

Degrees of freedom for the numerator along top row

Degrees of freedom for the denominator along side row

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	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	8.74	8.70	8.66	8.64	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39

APPENDIX D: F^2 TABLE AT 2.5% (UPPER TAIL)

F -Table, Critical Values, 2.5% in Upper Tails

Degrees of freedom for the numerator along top row

Degrees of freedom for the denominator along side row

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	648	799	864	900	922	937	948	957	963	969	977	985	993	997	1001	1006
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48

APPENDIX E: CHI-SQUARED TABLE

Values of χ^2 (Degrees of Freedom, Level of Significance)

Probability in Right Tail

Degrees of Freedom	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

INDEX

A

absolute advantage, 361
absolute frequency, 32
absorption approach, 390
action lag, 340
addition rule of probability, 73, 74, 75
additivity principle, 19
advertising expenses, 248
aggregate demand curve (AD curve), 276
aggregate supply curve (AS curve), 277
alternative hypothesis, 156
analysis of variance (ANOVA), 202
annuities, 7
annuities due, 7
a priori probability, 72
arithmetic mean, 45, 49
Austrian school, 303
autarky, 353, 359
automatic stabilizers, 335
average cost pricing, 258

B

balance of payments, 369
bar chart, 38
Bayes' formula, 86
Bernoulli random variable, 105
biased estimator, 53
bilateralism, 353
bimodal, 47
bimodal data set, 47
binomial distribution, 105
binomial formula, 88

binomial random variable, 105
black swan risk, 355
bond market vigilantes, 333
bootstrap, 146
box and whisker plot, 51, 52
brand names, 248
breakeven point, 229
broad money, 320
bubble line chart, 42
budget deficit, 319
budget surplus, 319
business cycle, 297
business expectations, 278
buy side, 383

C

capital account, 369
capital consumption allowance (CCA), 273
capital deepening investment, 288
capital restrictions, 367
capital spending, 337
capital transfers, 370
cartel, 251
cash flow additivity principle, 18
categorical data, 30
central limit theorem, 136
chi-square distribution, 178
chi-square distribution (c^2), 124
closed economy, 359
clustered bar chart, 39
cluster sampling, 135
coefficient of determination, 204
coefficient of variation (CV), 54
combination formula, 88
comparative advantage, 361
comparisons, 44, 65
complements, 222
compounding, 3
compounding frequency, 20

compound interest, 1, 3
concentration measures, 260
conditional expected values, 80
conditional probability, 73
confidence interval, 111, 141
confidence interval for the population mean, 142
confusion matrix, 35
consistent estimator, 139
consumer price index (CPI), 308
contingency table, 35
continuous compounding, 120
continuous data, 29
continuous distribution, 101
continuous random variable, 100
continuous uniform distribution, 104
contraction, 297
contractionary monetary policy, 320, 332
convenience sampling, 135
conventional fixed peg arrangement, 389
cooperative, 351
core inflation, 310
corporations, 383
correlation, 110
correlation coefficient, 61
cost of capital, 6
cost-push inflation, 312
Cournot model, 250
covariance, 60, 82
covariance matrix, 83
crawling peg, 389
credit cycles, 298
critical values, 157
cross-price elasticity of demand, 222
cross rate, 383
cross-sectional data, 30
crowding-out effect, 336
cumulative absolute frequency, 34
cumulative distribution function (cdf), 101
cumulative frequency distribution chart, 38
cumulative relative frequency, 34
currency board arrangement, 389

current account, 369
current spending, 337
cyclically adjusted budget deficit, 341
cyclical unemployment, 306

D

data snooping, 147
data-snooping bias, 147
data table, 32
data types, 29
debt ratio, 336
decile, 51, 66
decision rule, 157, 161
default risk, 4
default risk premium, 4
deflation, 307
degrees of freedom (df), 122, 203, 214, 215
demand for money, 322
demand-pull inflation, 312
dependent variable, 193
desirable properties of an estimator, 138
diminishing marginal productivity, 227, 288
diminishing marginal returns, 227
direct quote, 380
direct taxes, 337
discount factor, 6
discounting, 3, 6
discount rate, 3, 6, 16
discouraged workers, 306
discrete compounding, 120
discrete data, 29
discrete distribution, 100
discretely compounded returns, 120
discrete random variable, 100
discrete uniform random variable, 103
discretionary fiscal policy, 335
diseconomies of scale, 232
disinflation, 307
dispersion, 52

disposable income, 338
distribution function, 101
distributions, 44, 65
dollarization, 389
domestic price, 360
dominant firm model, 252
downside risk, 55
durable goods, 300

E

economic significance, 163
economic tools, 354
effective annual rate (EAR), 23
effective annual yield (EAY), 23
efficient estimator, 139
elasticities approach, 390
elasticity along a linear demand curve, 221
empirical probability, 72
equality of variances, 182
equation of exchange, 322
event, 71
event risk, 354
excess capacity, 247
excess kurtosis, 59
excess reserves, 321
exchange rate regimes, 388
exchange rates, 279, 280, 379
exchange rate targeting, 331
exhaustive events, 71
exogenous risk, 355
expansion, 297
expansionary fiscal policy, 279
expansionary monetary policy, 279, 319, 332
expected inflation, 313
expected value, 77, 78, 106
expenditure approach, 270
exports, 359
export subsidies, 365, 366

F

factorial, 87
factors of production, 226
false positives, 165
F-distribution, 124, 181
financial account, 369
financial tools, 354
fiscal balance, 274
fiscal multiplier, 338
fiscal policy, 319, 335
fiscal policy tools, 337
Fisher effect, 324
Fisher index, 310
foreign direct investment, 354, 360
foreign-owned assets, 370
formal dollarization, 389
forward currency contract, 382
forward discount, 386
forward exchange rate, 381
forward premium, 386
fractional reserve banking, 321
free trade, 359
frequency distribution, 32
frequency polygon, 38
frictional unemployment, 306
full-employment GDP, 278
future value, 1
future value factor, 5
future value (FV), 1, 5, 7, 10, 14
 of an annuity due, 10
 of an ordinary annuity, 7
 of an uneven cash flow series, 14
 of a single sum, 5
future value interest factor, 5

G

GDP deflator, 271
geometric mean, 48, 49

geophysical resource endowment, 352
geopolitical risk, 354
geopolitics, 351
Giffen good, 226
global economic growth, 279
globalization, 352
government entities, 383
government-owned assets abroad, 370
gross domestic income (GDI), 273
gross domestic product (GDP), 269
gross national product, 360
grouped bar chart, 39

H

harmonic mean, 49
headline inflation, 310
heat map, 43
Heckscher-Ohlin model, 363
hedging, 382
hedonic pricing, 310
hegemony, 353
Herfindahl-Hirschman Index (HHI), 260
heteroskedasticity, 199
histogram, 37
homoskedasticity, 199
human capital, 287
hyperinflation, 307
hypothesis, 155
Hypothesis testing, steps, 156

I

impact lag, 340
impact of geopolitical risk, 355
imports, 359
income approach, 270
income effect, 225
income elasticity, 222

income receipts, 370
independent events, 75, 76
independently floating exchange rate, 390
independent variable, 194
indirect quote, 380
indirect taxes, 337
inferior good, 222
inflationary gap, 284
inflation premium, 4
inflation rate, 307
inflation reports, 331
inflation targeting, 331
input prices, 280
institutions, 352
intercept, 196
interest on interest, 1
interest rate effect, 276
interest rate targeting, 331
International Monetary Fund, 371
interquartile range, 51
interval, 32
inventory-sales ratio, 299
investment accounts, 383

J

jackknife, 146
J-curve, 391
joint frequencies, 35
joint probability, 73
judgmental sampling, 136

K

Keynesian school, 302
kinked demand curve model, 249
kurtosis, 59, 60
 interpretation, 60

L

labeling, 87
labor force, 287, 306
labor productivity, 280
labor supply, 287
Laspeyres index, 310
leptokurtic distributions, 59
leveraged accounts, 383
likelihood, 73
likelihood of geopolitical risk, 355
line charts, 41
lin-log model, 210
liquidity risk, 4
liquidity risk premium, 4
liquidity trap, 333
location of the mean, median, and mode, 57
log-lin model, 210
log-log model, 211
lognormal distribution, 119
long run, 228
long-run aggregate supply (LRAS) curve, 281
long-run shutdown point, 229
look-ahead bias, 147

M

managed floating exchange rates, 390
management of exchange rates within crawling bands, 389
marginal cost pricing, 258
marginal frequency, 35
marginal probability, 73
markup, 247
Marshall-Lerner condition, 391
maturity risk, 4
maturity risk premium, 4
mean absolute deviation (MAD), 52
mean differences, 172
means of payment, 320
measures of central tendency, 45

measures of location, 51
median, 47
medium of exchange, 320
menu costs, 325
merchandise and services trade, 370
mesokurtic distributions, 59
minimum domestic content requirement, 365
minimum efficient scale, 231
modal interval, 33
mode, 47
Monetarist school, 303
monetary policy, 319
monetary policy tools, 328
monetary transmission mechanism, 328
monetary union, 389
money, 320
money multiplier, 322
money neutrality, 322
monopolistic competition, 238, 246
monopoly, 238
Monte Carlo simulation, 124
multilateralism, 354
multinational corporation, 360
multiplication rule of counting, 89
multiplication rule of probability, 73, 74
multivariate distribution, 110
multivariate normal distribution, 110
mutually exclusive events, 71, 75

N

narrow money, 320
Nash equilibrium, 251
national income, 273
nationalism, 352
national security tools, 354
natural monopoly, 238, 257
natural rate of unemployment (NARU), 312
natural resources, 287
negative skew, 57

Neoclassical school, 302
net exports, 272, 360
New Classical school, 303
New Keynesian school, 303
N-firm concentration ratio, 260
nominal data, 30
nominal exchange rate, 380
nominal GDP, 271
nominal risk-free rate, 4
non-accelerating inflation rate of unemployment (NAIRU), 312
non-cooperative, 351
nondurable goods, 300
nonparametric tests, 183
non-probability sampling, 133
non-state actors, 351
normal distribution, 109
normal good, 222
null hypothesis, 156
numerical data, 29

O

objective of a central bank, 325
objective probabilities, 72
odds, 72
oligopoly, 238
one-dimensional array, 31
one-stage cluster sampling, 135
one-tailed hypothesis test, 158
one-tailed test, 157, 205
opportunity cost, 3, 6
ordinal data, 30
ordinary annuities, 7
ordinary least squares, 196
outcome, 71
outliers, 46, 57, 201

P

Paasche index, 310
paired comparisons test, 174
panel data, 30
parametric tests, 183
participation ratio, 306
peak, 297
pegging, 325
per-capita real GDP, 272
percentile, 51, 66
perfect competition, 238
periodic rates, 23
permutation formula, 88
perpetuity, 12
personal disposable income, 273
personal income, 273
physical capital stock, 287
platykurtic distributions, 59
point estimates, 141
population mean, 45
portfolio expected return, 82
portfolio investment flows, 354
portfolio variance, 83
positive skew, 57
potential GDP, 278
power of a test, 161
predicted value, 195, 208
present value, 1
present value factor, 6
present value interest factor, 6
present value (PV), 1, 6, 8, 11, 12, 14
 of an annuity due, 11
 of an ordinary annuity, 8
 of an uneven cash flow series, 14
 of a perpetuity, 12
 of a single sum, 6
price discrimination, 255
price index, 308
price relative, 120
probability distribution, 99
probability function, 100
probability sampling, 133

probability tree, 80
producer price index (PPI), 310
product innovation, 247
production function, 226, 288
productivity, 307
promissory notes, 321
properties of an estimator, 138
properties of covariance, 82
properties of probability, 71, 72
p-value, 165

Q

qualitative data, 30
quantile, 50
quantitative data, 29
quantitative easing, 333
quantity equation of exchange, 322
quantity theory of money, 322
quartile, 66
quartiles, 50, 51
quintile, 50, 66
quota rents, 366
quotas, 365

R

random sampling, 133
random variable, 71
range, 52
real business cycle theory (RBC), 303
real exchange rate, 380
real exchange rate effect, 276
real GDP, 271
real money accounts, 383
real risk-free rate, 4
recession, 297
recessionary gap, 283
recognition lag, 340

recovery, 298
regionalism, 354
regression line, 195
rejection points, 157
relationships, 44, 65
relative dispersion, 54
relative frequency, 34
rent seeking, 257
required rate of return, 3, 6
residual, 195
retail market, 383
revenue tools, 337
Ricardian equivalence, 339
Ricardian model of trade, 363
risk, types of, 4
roles of central banks, 324
Roy's safety-first criterion, 116

S

sample covariance, 60, 83
sample kurtosis, 60
sample mean, 45
sample selection bias, 147
sample skewness, 58
sample standard deviation, 54
sample variance, 53
sampling and estimation, 133
sampling distribution, 134
sampling error, 134
sanctions, 354
scatter plot, 42
scatter plot matrix, 43
scatter plots, 61
scenario analysis, 355
sell side, 383
services, 300
shoe leather costs, 325
Shortfall risk, 116
short run, 228

short-run aggregate supply (SRAS) curve, 280
short-run shutdown point, 229
shutdown point, 242
significance level, 157
signposts, 355
simple linear regression, 193
simple random sampling, 133, 134
skew, skewness, 57, 59
 interpretation of the skewness measure, 59
slope coefficient, 196
soft power, 352
sources of economic growth, 287
sovereign wealth funds, 383
Spearman rank correlation test, 184
speculative foreign exchange transactions, 382
spending tools, 337
spot exchange rate, 381
spurious correlation, 63
stacked bar chart, 39
Stackelberg model, 250
stagflation, 284
standard deviation, 54
standard error of the sample mean, 137
standardization, 352
standardizing, 113
standard normal distribution, 113
state actors, 351
stated annual interest rates, 23
statistical significance, 163
stratified random sampling, 134
structural budget deficit, 341
structural unemployment, 306
structured data, 31
Student's t-distribution, 121
subjective probability, 72
subsidies, 280
substitutes, 222
substitution effect, 225
sum of squared errors (SSE), 195, 202
sum of squares regression (SSR), 202
sum-of-value-added method, 270

supply of money, 323
survivorship bias, 147
sustainable rate of economic growth, 287
symmetrical distributions, 57
systematic sampling, 134

T

target downside deviation, 55
target semideviation, 55
target zone, 389
tariffs, 365
taxes, 280
t-distribution, 121
technology, 281, 287
terms of trade, 360
test statistic, 160
thematic risk, 355
threshold level, 116
time line, 3
time-period bias, 148
time series, 30
time value of money, 1
tools of geopolitics, 354
total factor productivity, 288
total probability rule, 77
total sum of squares (SST), 202
total variation, 202
trade balance, 274
trade deficit, 360
trade protection, 359
trade surplus, 360
trading blocs, 368
transfer payments, 337
tree map, 40
trimmed mean, 46, 49
trimodal, 47
trimodal data set, 47
trough, 297
t-test, 167, 206

two-dimensional array, 32
two-stage cluster sampling, 135
two-tailed test, 157
Type I error, 160
Type II error, 160

U

unbiased estimator, 139
unconditional probability, 73
underemployed, 306
unemployed, 306
unemployment rate, 306
unilateral transfers, 370
unimodal, 47
unimodal data set, 47
unit labor costs, 312
unit of account, 320
univariate distributions, 110
unstructured data, 31

V

value-of-final-output method, 270
Veblen good, 226
velocity of geopolitical risk, 355
volatility, 79
voluntary export restraint (VER), 365, 366

W

wealth effect, 276
weighted average, 46
weighted mean, 46
wholesale price index (WPI), 310
winsorized mean, 46, 49
word cloud, 40
World Bank, 371

world price, 360

World Trade Organization, 372

Z

z-test, 168