ECON 208

Source credits and notes on attribution

Material contained in this textbook was taken and compiled from several sources. The full list of sources used in this textbook as well as their license and copyright information (where locatable) are outlined below.

In-text, attribution is indicated at the beginning of each section via a reference such as (Openstax, 2019).

Material that has been changed or altered from the original source has been noted where the text appears in the document.

To reproduce or reuse any materials contained in this book, consult the licensing information of the original source (listed below):

Sources

Boundless economics (version 3). (n.d.) OER2Go. Licensed Creative Commons Attribution-ShareAlike 4.0 International. © Unspecified

Boundless economics (via Lumen Learning). (n.d.). Licensed Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) © Unspecified

Curtis, D., & Irvine, I. (2016). *Principles of microeconomics* (version 2016b). © Lyryx Learning Inc, All Rights Reserved. Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Curtis, D., & Irvine, I. (2017). *Microeconomics: Markets, Methods, & Models, an Open Text.* (version 2017 - Revision A). © Lyryx Learning Inc, All Rights Reserved. Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Hutchinson, E., University of Victoria, Nicholsen, M., & Lukenchuk, B. (2017).*Principles of Microeconomics*. Licensed under Creative Commons Attribution 4.0 International License, except where otherwise noted. Note: Adaptation of *Principles of Microeconomics* which is © 2016 OpenStax and is licensed under a Creative Commons Attribution License 4.0 except for the following changes and additions, which are © 2017 by the University of Victoria, and are licensed under a Creative Commons-Attribution 4.0 International license.

Intermediate good .(n.d.).In *Wikipedia* Retrieved March 21, 2019 from https://en.wikipedia.org/wiki/Intermediate_good. Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported License © Wikipedia editors and contributors.

International economics: Theory and policy (v.1.0). (2012). Publisher: Saylor Academy Licensed Creative Commons Attribution-NonCommercial-ShareAlike 3.0 © Unspecified

Institute for Humane Studies. (2017). *Elements of economics* [video].
© Oct 17, 2017 Institute for Humane Studies. Textbook content produced by Institute for Humane Studies is licensed under a Creative Commons Attribution License 4.0 license.

Lynham, J. (2018). *Principles of microeconomics* (Hawaii Edition). Text adapted originally from OpenStax *Principles of microeconomics* (1st edition).© Jan 4, 2017 OpenStax. Textbook content produced by OpenStax is licensed under a Creative Commons Attribution License 4.0 license.

Macroeconomics principles (v.1.1).(2012). Licensed under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) © Unspecified publisher (see http://jsmith.cis.byuh.edu/attribution.html?utm_source=header)

OpenStax. (2019). *Principles of economics* (2nd ed). OpenStax. Textbook content produced by OpenStax is licensed under a Creative Commons Attribution License 4.0 license. © Mar 13, 2019

OpenStax. (2019). *Principles of microeconomics* (2nd edition). Textbook content produced by OpenStax is licensed under a Creative Commons Attribution License 4.0 license. © Feb 11, 2019 OpenStax.

Policy and theory of international trade (v.1.0). (2012). Licensed under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) © Unspecified publisher (see https://2012books.lardbucket.org/attribution.html?utm_source=header)

Theory and applications of economics (v.1.0). (2012). Licensed under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0) © Unspecified publisher (see http://jsmith.cis.byuh.edu/attribution.html?utm_source=header)

Funding acknowledgement: This open textbook compilation was sponsored by the Student Society of McGill University (SSMU) as a part of the SSMU Library Improvement Fund

CHAPTER 1: INTRODUCTION TO ECONOMICS

1.1 WHAT IS ECONOMICS, AND WHY IS IT IMPORTANT?

Source: Hutchison, 2018, Section 1.1, CC-BY 4.0 (Original source, p. 1, paragraphs 1 & 2 only)

At its core, Economics is the study of how humans make decisions in the face of scarcity. These can be individual decisions, family decisions, business decisions or societal decisions. If you look around carefully, you will see that scarcity is a fact of life. **Scarcity** means that human wants for goods, services and resources exceed what is available. Resources, such as labor, tools, land, and raw materials are necessary to produce the goods and services we want but they exist in limited supply. Of course, the ultimate scarce resource is time – everyone, rich or poor, has just 24 hours in the day to try to acquire the goods they want. At any point in time, there is only a finite amount of resources available.

Think about it this way: In 2016, the labor force in Canada contained 19.4 million workers, according to Statistics Canada. The total area of the Canada is 9.99 million square kilometres. These are large numbers for such crucial resources, however, they are limited. Because these resources are limited, so are the numbers of goods and services we produce with them. Combine this with the fact that human wants seem to be virtually infinite, and you can see why scarcity is a problem.

Economics is the study of how humans make decisions in the face of scarcity.

Figure 1. Scarcity of Resources. Homeless people are a stark reminder that scarcity of resources is real. (Credit: "daveynin"/Flickr Creative Commons)

Introduction to Choice in a World of Scarcity

Source: Lynham, 2018, Chapter 2 Introduction, CC-BY 4.0 (Original Source, paragraphs 2 & 4 removed)

Figure 1. *Choices and Tradeoffs. In general, the higher the degree, the higher the salary. So why aren't more people pursuing higher degrees? The short answer: choices and tradeoffs. (Credit: University of Hawai'i News/Flickr Creative Commons)*

Choices … To What Degree?

In 2015, the median income for workers who hold master's degrees varies from males to females. The average of the two is $2,951 (USD) weekly. Multiply this average by 52 weeks, and you get an average salary of $153,452. Compare that to the median weekly earnings for a full-time worker over 25 with no higher than a bachelor's degree: $1,224 weekly and $63,648 a year. What about those with no higher than a high school diploma in 2015? They earn just $664 weekly and $34,528 over 12 months. In other words, says the Bureau of Labor Statistics (BLS), earning a bachelor's degree boosted salaries 54% over what you would have earned if you had stopped your education after high school. A master's degree yields a salary almost double that of a high school diploma.

Given these statistics, we might expect a lot of people to choose to go to college and at least earn a bachelor's degree. Assuming that people want to improve their material well-being, it seems like they would make those choices that give them the greatest opportunity to consume goods and services. As it turns out, the analysis is not nearly as simple as this. In fact, in 2014, the BLS reported that while almost 88% of the population in the United States had a high school diploma, only 33.6% of 25–65 year olds had bachelor's degrees, and only 7.4% of 25–65 year olds in 2014 had earned a master's.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

This brings us to the subject of this chapter: why people make the choices they make and how economists go about explaining those choices.

You will learn quickly when you examine the relationship between economics and scarcity that choices involve tradeoffs. Every choice has a cost.

Because people live in a world of scarcity, they cannot have all the time, money, possessions, and experiences they wish. Neither can society.

Source: Curtis & Irvine, 2016, Section 1.3, CC-BY-NC-SA 3.0 (Original source, p. 1, pararaph 1 only)

Opportunity cost and the market Individuals face choices at every turn: In deciding to go to the hockey game tonight, you may have to forgo a concert; or you will have to forgo some leisure time this week in order to earn additional income for the hockey game ticket. Indeed, there is no such thing as a free lunch, a free hockey game or a free concert. In economics we say that these limits or constraints reflect opportunity cost. The **opportunity cost** of a choice is what must be sacrificed when a choice is made. That cost may be financial; it may be measured in time, or simply the alternative foregone.

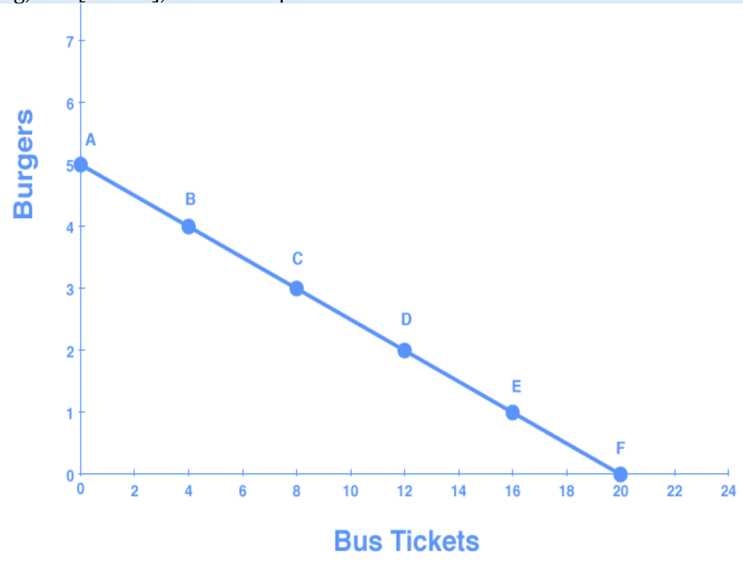
Source: Lynham, 2018, section 2.1, CC-BY 4.0 (Original Source, paragraphs 5 only)

Economists use the term **opportunity cost** to indicate what must be given up to obtain something that is desired. The idea behind opportunity cost is that the cost of one item is the lost opportunity to do or consume something else; in short, opportunity cost is the value of the next best alternative. For Alphonso, the opportunity cost of a burger is the four bus tickets he would have to give up. He would decide whether or not to choose the burger depending on whether the value of the burger exceeds the value of the forgone alternative—in this case, bus tickets. Since people must choose, they inevitably face tradeoffs in which they have to give up things they desire to get other things they desire more.

**Source (Figure only): "Boundless Economics", Lumen Learning, n.d. [section], CC-BY-SA 4.0 International

Figure 1: Choosing Between Burgers and Bus Tickets

Bus tickets are $0.50 and burgers are $2. Each point on the line represents all combinations of bus tickets and burgers that can be bought with $10. There is an opportunity cost, if the entire $10 is to be spent on burgers and bus tickets. The slope of the line represents such an opportunity cost. The opportunity cost of one extra burger is four bus tickets and the opportunity cost of one extra bus ticket is ¼ of a burger.



[2] Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Source: Lynham, 2018, section 2.1, CC-BY 4.0 (Original source, paragraphs 6, 8, & 10 only)

A fundamental principle of economics is that every choice has an opportunity cost. If you sleep through your economics class (not recommended, by the way), the opportunity cost is the learning you miss from not attending class. If you spend your income on video games, you cannot spend it on movies. If you choose to marry one person, you give up the opportunity to marry anyone else. In short, opportunity cost is all around us and part of human existence.

Opportunity cost: what must be sacrificed when a choice is made.
Every choice has an opportunity cost.

Identifying Opportunity Cost

In many cases, it is reasonable to refer to the opportunity cost as the **price**. If your cousin buys a new bicycle for $300, then $300 measures the amount of "other consumption" that he has given up. For practical purposes, there may be no special need to identify the specific alternative product or products that could have been bought with that $300, but sometimes the price as measured in dollars may not accurately capture the true opportunity cost. This problem can loom especially large when costs of time are involved.

Attending college is another case where the opportunity cost exceeds the monetary cost. The out-of-pocket costs of attending college include tuition, books, room and board, and other expenses. But in addition, during the hours that you are attending class and studying, it is impossible to work at a paying job. Thus, college imposes both an out-of-pocket cost and an opportunity cost of lost earnings.

Source: Lynham, 2018, section 2.2, CC-BY 4.0 (Original source, paragraphs 1-8 only, sections/lines on *budget constraints* removed in paragraphs 3, 4, 5)

Production Possibilities Frontier and Social Choices Just as individuals cannot have everything they want and must instead make choices, society as a whole cannot have everything it might want, either. This section of the chapter will explain the constraints faced by society, using a model called the **production possibilities frontier (PPF)**. There are more similarities than differences between individual choice and social choice. As you read this section, focus on the similarities.

Because society has limited resources (e.g., labor, land, capital, raw materials) at any point in time, there is a limit to the quantities of goods and services it can produce. Suppose a society desires two products, healthcare and education. This situation is illustrated by the production possibilities frontier in Figure 1.

In Figure 1, healthcare is shown on the vertical axis and education is shown on the horizontal axis. If the society were to allocate all of its resources to healthcare, it could produce at point A. But it would not have any resources to produce education. If it were to allocate all of its resources to education, it could produce at point F. Alternatively, the

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

society could choose to produce any combination of healthcare and education shown on the production possibilities frontier.

Most important, the production possibilities frontier clearly shows the tradeoff between healthcare and education. Suppose society has chosen to operate at point B, and it is considering producing more education. Because the PPF is downward sloping from left to right, the only way society can obtain more education is by giving up some healthcare. That is the tradeoff society faces. Suppose it considers moving from point B to point C. What would the opportunity cost be for the additional education? The opportunity cost would be the healthcare society has to give up. The opportunity cost is shown by the slope of the production possibilities frontier.

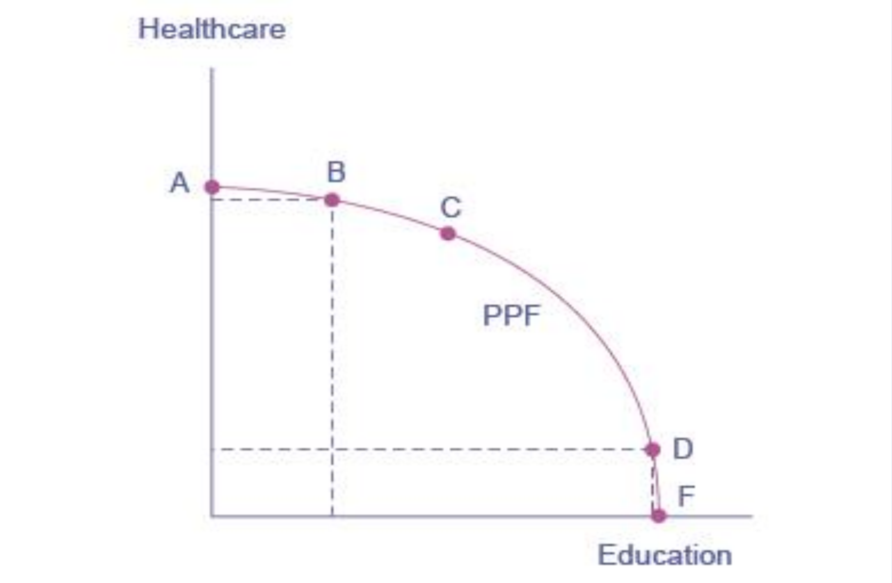


Figure 1. *A Healthcare vs. Education Production Possibilities Frontier. This production possibilities frontier shows a tradeoff between devoting social resources to healthcare and devoting them to education. At A all resources go to healthcare and at B, most go to healthcare. At D most resources go to education, and at F, all go to education.*

The Shape of the PPF (section on Law of Diminishing Returns removed)

To understand why the PPF is curved, start by considering point A at the top left-hand side of the PPF. At point A, all available resources are devoted to healthcare and none are left for education. This situation would be extreme and even ridiculous. Now imagine that some of these resources are diverted from healthcare to education, so that the economy is at point B instead of point A. Diverting some resources away from A to B causes relatively little reduction in health because the last few marginal dollars going into healthcare services are not producing much additional gain in health. However, putting those marginal dollars into education, which is completely without resources at point A, can produce relatively large gains. For this reason, the shape of the PPF from A to B is relatively flat, representing a relatively small drop-off in health and a relatively large gain in education.

Now consider the other end, at the lower right, of the production possibilities frontier. Imagine that society starts at choice D, which is devoting nearly all resources to education and very few to healthcare, and moves to point F, which is devoting all spending to education and none to healthcare. For the sake of concreteness, you can imagine that in the movement from D to F, the last few doctors must become high school science teachers, the last few nurses must become school librarians rather than dispensers of vaccinations, and the last few emergency rooms are turned into kindergartens. The gains to education from adding these last few resources to education are very small. However, the opportunity cost lost to health will be fairly large, and thus the slope of the PPF between D and F is steep, showing a large drop in health for only a small gain in education.

The lesson is not that society is likely to make an extreme choice like devoting no resources to education at point A or no resources to health at point F. Instead, the lesson is that the gains from committing additional marginal resources to education depend on how much is already being spent. If on the one hand, very few resources are currently committed to education, then an increase in resources used can bring relatively large gains. On the other hand, if a large number of resources are already committed to education, then committing additional resources will bring relatively smaller gains.

Microeconomics and Macroeconomics

Source: Lynham, 2018, Section 1.2, CC-BY 4.0 (Original source, paragraphs 2, 6, 7, 8 only)

It should be clear by now that economics covers a lot of ground. That ground can be divided into two parts: **Microeconomics** focuses on the actions of individual agents within the economy, like households, workers, and businesses; **Macroeconomics** looks at the economy as a whole. It focuses on broad issues such as growth of production, the number of unemployed people, the inflationary increase in prices, government deficits, and levels of exports and imports. Microeconomics and macroeconomics are not separate subjects, but rather complementary perspectives on the overall subject of the economy.

Microeconomics
What determines how households and individuals spend their budgets? What combination of goods and services will best fit their needs and wants, given the budget they have to spend? How do people decide whether to work, and if so, whether to work full time or part time? How do people decide how much to save for the future, or whether they should borrow to spend beyond their current means?

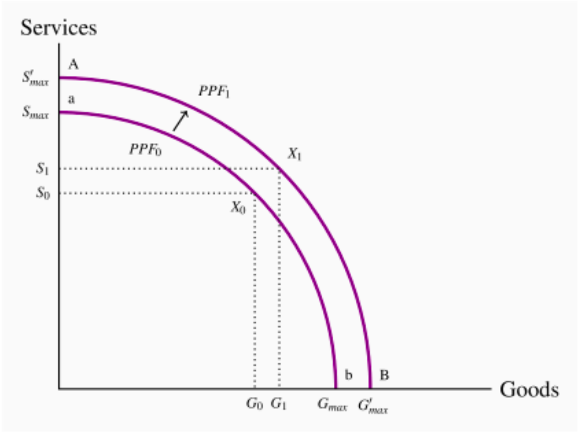
What determines the products, and how many of each, a firm will produce and sell? What determines what prices a firm will charge? What determines how a firm will produce its products? What determines how many workers it will hire? How will a firm finance its business? When will a firm decide to expand, downsize, or even close? In the

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

microeconomic part of this book, we will learn about the theory of consumer behavior and the theory of the firm.

Macroeconomics
What determines the level of economic activity in a society? In other words, what determines how many goods and services a nation actually produces? What determines how many jobs are available in an economy? What determines a nation's standard of living? What causes the economy to speed up or slow down? What causes firms to hire more workers or to lay workers off? Finally, what causes the economy to grow over the long term?

Figure 1.5 Growth and the PPF



Economic growth or an increase in the available resources can be envisioned as an outward shift in the PPF from PPF_0 to PPF_1. With PPF_1 the economy can produce more in both sectors than with PPF_0.

Source: Curtis & Irvine, 2016, Section, CC-BY-NC-SA 3.0 (Original source, p. 3 only)

1.2 MARKETS AND GOVERNMENT

Source: Curtis & Irvine, 2016, Section 1.1, CC-BY-NC-SA 3.0 (Original source, p. 6, paragraphs 3 & 7 removed)

Markets play a key role in coordinating the choices of individuals with the decisions of business. In **modern market economies** goods and services are supplied by both business and government. Hence we call them **mixed economies**. Some products or services are available through the marketplace to those who wish to buy them and have the necessary income—as in cases like coffee and wireless services. Other services are provided to all people through government programs like law enforcement and health care.

Mixed economy: goods and services are supplied both by private suppliers and government.

Markets offer the choice of a wide range of goods and services at various prices. Individuals can use their incomes to decide the pattern of expenditures and the bundle of goods and services they prefer. Businesses sell goods and services in the expectation that the market price will cover costs and yield a profit.

Decision Makers (subtitles added)

Maximizing Decisions We will represent **individuals** and **firms** by envisaging that they have explicit objectives – to maximize their happiness or profit. However, this does not imply that individuals and firms are concerned only with such objectives. On the contrary, much of microeconomics and macroeconomics focuses upon the role of **government**: How it manages the economy through fiscal and monetary policy, how it

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

redistributes through the tax-transfer system, how it supplies information to buyers and sets safety standards for products.

Since governments perform all of these society-enhancing functions, in large measure governments reflect the social ethos of voters. So, while these voters may be maximizing at the individual level in their everyday lives, and our models of human behaviour in microeconomics certainly emphasize this optimization, economics does not see individuals and corporations as being devoid of civic virtue or compassion, nor does it assume that only market-based activity is important. Governments play a central role in modern economies, to the point where they account for more than one third of all economic activity in the modern mixed economy.

Governments supply goods and services in many spheres, for example, health and education. The provision of public education is motivated both by a concern for equality and a realization that an educated labour force increases the productivity of an economy. Likewise, the provision of law and order, through our legal system broadly defined, represents more than a commitment to a just society at the individual level; without a legal system that enforces contracts and respects property rights, the private sector of the economy would diminish dramatically as a result of corruption, uncertainty and insecurity. It is the lack of such a secure environment in many of the world's economies that inhibits their growth and prosperity.

Source: OpenStax, 2019, Section 1.3, *Principles of microeconomics*, 2e, CC-BY 4.0 (Original source, paragraphs 4 & 6 only)

The Flow of Income and Expenditure A good model to start with in economics is the **circular flow diagram** (Figure 1.6). It pictures the economy as consisting of two groups—households and firms—that interact in two markets: the **goods and services market** in which firms sell and households buy and the **labor market** in which households sell labor to business firms or other employees.

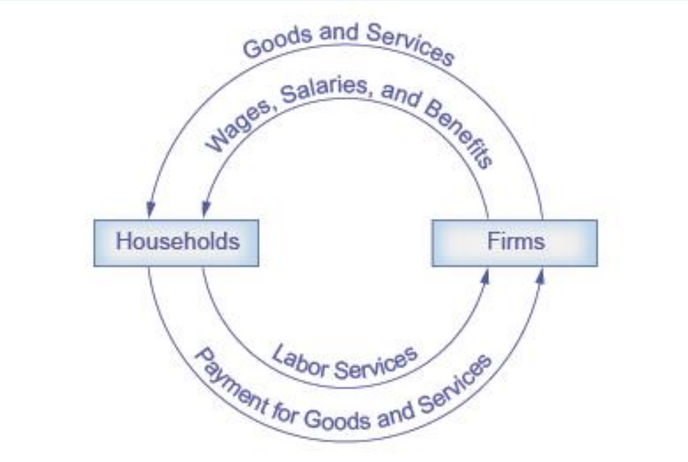


Figure 1.6. *The Circular Flow Diagram. The circular flow diagram shows how households and firms interact in the goods and services market, and in the labor market. The direction of the arrows shows that in the goods and services market, households receive goods and services and pay firms for them. In the labor market, households provide labor and receive payment from firms through wages, salaries, and benefits.*

[8] Download for free at http://cnx.org/contents/5c09762c-b540-47d3-9541-dda1f44f16e5@11.13.

Of course, in the real world, there are many different markets for goods and services and markets for many different types of labor. The circular flow diagram simplifies this to make the picture easier to grasp. In the diagram, firms produce goods and services, which they sell to households in return for revenues. This is shown in the outer circle, and represents the two sides of the product market (for example, the market for goods and services) in which households demand and firms supply. Households sell their labor as workers to firms in return for wages, salaries and benefits. This is shown in the inner circle and represents the two sides of the labor market in which households supply and firms demand.

Production & Trade

Source: Lynham, 2018, Section 1.1, CC-BY 4.0 (Original source, paragraphs 1-5, 13, 15-17 removed, subtitles modified)

The Division of Labour

The formal study of economics began when Adam Smith (1723–1790) published his famous book *The Wealth of Nations* in 1776. Many authors had written on economics in the centuries before Smith, but he was the first to address the subject in a comprehensive way. In the first chapter, Smith introduces the **division of labor**, which means that the way a good or service is produced is divided into a number of tasks that are performed by different workers, instead of all the tasks being done by the same person.

To illustrate the division of labor, Smith counted how many tasks went into making a pin: drawing out a piece of wire, cutting it to the right length, straightening it, putting a head on one end and a point on the other, and packaging pins for sale, to name just a few. Smith counted 18 distinct tasks that were often done by different people—all for a pin, believe it or not!

Modern businesses divide tasks as well. Even a relatively simple business like a restaurant divides up the task of serving meals into a range of jobs like top chef, sous chefs, less-skilled kitchen help, servers to wait on the tables, a greeter at the door, janitors to clean up, and a business manager to handle paychecks and bills—not to mention the economic connections a restaurant has with suppliers of food, furniture, kitchen equipment, and the building where it is located. A complex business like a large manufacturing factory or a hospital can have hundreds of job classifications.

Specialization

When the tasks involved with producing a good or service are divided and subdivided, workers and businesses can produce a greater quantity of output. In his observations of pin factories, Smith observed that one worker alone might make 20 pins in a day, but that a small business of 10 workers (some of whom would need to do two or three of the 18 tasks involved with pin-making), could make 48,000 pins in a day. How can a group of workers, each specializing in certain tasks, produce so much more than the same number of workers who try to produce the entire good or service by themselves? Smith offered three reasons.

First, **specialization** in a particular small job allows workers to focus on the parts of the production process where they have an advantage. (In later chapters, we will develop this idea by discussing **comparative advantage**.) People have different skills, talents, and interests, so they will be better at some jobs than at others. The particular advantages may be based on educational choices, which are in turn shaped by interests and talents. Only those with medical degrees qualify to become doctors, for instance. For some goods, specialization will be affected by geography—it is easier to be a wheat farmer in North Dakota than in Florida, but easier to run a tourist hotel in Florida than in North Dakota. If you live in or near a big city, it is easier to attract enough customers to operate a successful dry cleaning business or movie theater than if you live in a sparsely populated rural area. Whatever the reason, if people specialize in the production of what they do best, they will be more productive than if they produce a combination of things, some of which they are good at and some of which they are not.

Second, workers who specialize in certain tasks often learn to produce more quickly and with higher quality. This pattern holds true for many workers, including assembly line laborers who build cars, stylists who cut hair, and doctors who perform heart surgery. In fact, specialized workers often know their jobs well enough to suggest innovative ways to do their work faster and better.

A similar pattern often operates within businesses. In many cases, a business that focuses on one or a few products (sometimes called its “**core competency**”) is more successful than firms that try to make a wide range of products.

Trade and Markets Specialization only makes sense, though, if workers can use the pay they receive for doing their jobs to purchase the other goods and services that they need. In short, specialization requires trade.

Source: Lynham, 2018, Section 1.4, CC-BY 4.0 (Original source, paragraphs 9 & 10 only)

The Rise of Globalization Recent decades have seen a trend toward **globalization**, which is the expanding cultural, political, and economic connections between people around the world. One measure of this is the increased buying and selling of goods, services, and assets across national borders—in other words, international trade and financial capital flows.

Globalization has occurred for a number of reasons. Improvements in shipping air cargo have driven down transportation costs. Innovations in computing and telecommunications have made it easier and cheaper to manage long-distance economic connections of production and sales. Many valuable products and services in the modern economy can take the form of information—for example: computer software; financial advice; travel planning; music, books and movies; and blueprints for designing a building. These products and many others can be transported over telephones and computer networks at ever-lower costs. Finally, international agreements and treaties between countries have encouraged greater trade.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

1.3 HOW ECONOMIES CAN BE ORGANIZED: AN OVERIEW OF ECONOMIC SYSTEMS

Source: Lynham, 2018, Section 1.4, CC-BY 4.0 (Original source, paragraphs 2-8 only, subtitles added).

There are at least three ways societies have found to organize an economy.

Traditional Economies The first is the **traditional economy**, which is the oldest economic system and can be found in parts of Asia, Africa, and South America. Traditional economies organize their economic affairs the way they have always done (i.e., tradition). Occupations stay in the family. Most families are farmers who grow the crops they have always grown using traditional methods. What you produce is what you get to consume. Because things are driven by tradition, there is little economic progress or development.

Command Economies Command economies are very different. In a **command economy**, economic effort is devoted to goals passed down from a ruler or ruling class. Ancient Egypt was a good example: a large part of economic life was devoted to building pyramids. Medieval manor life is another example: the lord provided the land for growing crops and protection in the event of war. In return, vassals provided labor and soldiers to do the lord's bidding. In the last century, communism emphasized command economies.

In a command economy, the government decides what goods and services will be produced and what prices will be charged for them. The government decides what methods of production will be used and how much workers will be paid. Many necessities like healthcare and education are provided for free. Currently, Cuba and North Korea have command economies.

Free-Market Economies Although command economies have a very centralized structure for economic decisions, market economies have a very decentralized structure. A **market** is an institution that brings together buyers and sellers of goods or services, who may be either individuals or businesses. In a **market economy**, decision-making is decentralized. Market economies are based on **private enterprise**: the means of production (resources and businesses) are owned and operated by private individuals or groups of private individuals. Businesses supply goods and services based on demand. (In a command economy, by contrast, resources and businesses are owned by the government.) What goods and services are supplied depends on what is demanded. A person's income is based on his or her ability to convert resources (especially labor) into something that society values. The more society values the person's output, the higher the income (think Lady Gaga or LeBron James). In this scenario, economic decisions are determined by market forces, not governments.

Mixed Economies Most economies in the real world are mixed; they combine elements of command and market (and even traditional) systems. The U.S. economy is positioned toward the market-oriented end of the spectrum. Many countries in Europe and Latin America, while primarily market-oriented, have a greater degree of government involvement in economic decisions than does the U.S. economy. China and Russia, while

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

they are closer to having a market-oriented system now than several decades ago, remain closer to the command economy end of the spectrum.

Most economies in the real world are mixed; they combine elements of command, market, and traditional systems.

Regulations: The Rules of the Game

Markets and government regulations are always entangled. There is no such thing as an absolutely free market. Regulations always define the “rules of the game” in the economy. Economies that are primarily market-oriented have fewer regulations—ideally just enough to maintain an even playing field for participants. At a minimum, these laws govern matters like safeguarding private property against theft, protecting people from violence, enforcing legal contracts, preventing fraud, and collecting taxes. Conversely, even the most command-oriented economies operate using markets. How else would buying and selling occur? But the decisions of what will be produced and what prices will be charged are heavily regulated. Heavily regulated economies often have **underground economies**, which are markets where the buyers and sellers make transactions without the government’s approval.

The question of how to organize economic institutions is typically not a black-or-white choice between all market or all government, but instead involves a balancing act over the appropriate combination of market freedom and government rules.

CHAPTER 2: THEORIES, MODELS, AND ECONOMIC DATA

2.1 POSITIVE AND NORMATIVE STATEMENTS

Source: Lynham, 2018, Section 2.3, CC-BY 4.0 (Original source, paragraph 7 only)

First, economics is not a form of moral instruction. Rather, it seeks to describe economic behavior as it actually exists. Philosophers draw a distinction between **positive statements**, which describe the world as it is, and **normative statements**, which describe how the world should be. For example, an economist could analyze a proposed subway system in a certain city. If the expected benefits exceed the costs, he concludes that the project is worth doing—an example of positive analysis. Another economist argues for extended unemployment compensation during the Great Depression because a rich country like the United States should take care of its less fortunate citizens—an example of normative analysis.

Source: Curtis & Irvine, 2016, Section 2.3, CC-BY-NC-SA 3.0 (Original source, p. 1-2, paragraphs 1-3 only)

Positive economics studies objective or scientific explanations of how the economy functions. Its aim is to understand and generate predictions about how the economy may respond to changes and policy initiatives. In this effort economists strive to act as detached scientists, regardless of political sympathies or ethical code. Personal judgments and preferences are (ideally) kept apart. In this particular sense, economics is similar to the natural sciences such as physics or biology. To date in this chapter we have been exploring economics primarily from a positive standpoint.

In contrast, **normative economics** offers recommendations based partly on value judgments. While economists of different political persuasions can agree that raising the income tax rate would lead to some reduction in the number of hours worked, they may yet differ in their views on the advisability of such a rise. One economist may believe that the additional revenue that may come in to government coffers is not worth the disincentives to work; another may think that, if such monies can be redistributed to benefit the needy, or provide valuable infrastructure, the negative impact on the workers paying the income tax is worth it.

Positive economics studies objective or scientific explanations of how the economy functions.

Normative economics offers recommendations based partly on value judgments.

Scientific research can frequently resolve differences that arise in positive economics—not so in normative economics. For example, if we claim that "the elderly have high medical bills, and the government should cover all of the bills", we are making both a positive and a normative statement. The first part is positive, and its truth is easily established. The latter part is normative, and individuals of different beliefs may reasonably differ. Some people may believe that the money would be better spent on the environment and have the aged cover at least part of their own medical costs. Positive economics does not attempt to show that one of these views is correct and the other false. The views are based on value judgments, and are motivated by a concern for **equity**. Equity is a vital guiding principle in the formation of policy and is frequently, though not always, seen as being in competition with the drive for economic growth. Equity is driven primarily by normative considerations. Few economists would disagree with the assertion that a government should implement policies that improve the lot of the poor—but to what degree?

2.2 HOW ECONOMISTS USE THEORIES AND MODELS TO UNDERSTAND ECONOMIC ISSUES

Source: Lynham, 2018, Section 1.3, CC-BY 4.0 (Original source, paragraphs 2, 3, 8 only, italics added)

Economists see the world through a different lens than anthropologists, biologists, classicists, or practitioners of any other discipline. They analyze issues and problems with economic theories that are based on particular assumptions about human behavior, that are different than the assumptions an anthropologist or psychologist might use. A **theory** is a simplified representation of how two or more variables interact with each other. The purpose of a theory is to take a complex, real-world issue and simplify it down to its essentials. If done well, this enables the analyst to understand the issue and any problems around it. A good theory is simple enough to be understood, while complex enough to capture the key features of the object or situation being studied.

Sometimes economists use the term **model** instead of theory. *Strictly speaking, a theory is a more abstract representation, while a model is more applied or empirical representation*. Models are used to test theories, but for this course we will use the terms interchangeably.

Economists carry a set of theories in their heads like a carpenter carries around a toolkit. When they see an economic issue or problem, they go through the theories they know to see if they can find one that fits. Then they use the theory to derive insights about the issue or problem. In economics, theories are expressed as diagrams, graphs, or even as mathematical equations. (Do not worry. In this course, we will mostly use graphs.) Economists do not figure out the answer to the problem first and then draw the graph to illustrate. Rather, they use the graph of the theory to help them figure out the answer. Although at the introductory level, you can sometimes figure out the right answer without applying a model, if you keep studying economics, before too long you will run into issues and problems that you will need to graph to solve. Both micro and macroeconomics are explained in terms of theories and models. The most well-known theories are probably those of supply and demand, but you will learn a number of others.

Source: Curtis & Irvine, 2016, Section 2.1, CC-BY-NC-SA 3.0 (Original source, p. 1 only)

The analysis of behaviour necessarily involves data. Data may serve to validate or contradict a theory. Data analysis, even without being motivated by economic theory, frequently displays patterns of behaviour that merit examination. The terms *variables* and *data* are related. **Variables** are measures that can take on different magnitudes. The interest rate on a student loan, for example, is a variable with a certain value at a point in

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

time but perhaps a different value at an earlier or later date. Economic theories and models explain the causal relationships between variables. In contrast, **Data** are the recorded values of variables. Sets of data provide specific values for the variables we want to study and analyze. Knowing that gross domestic product (a variable) declined in 2009 is just a partial description of events. If the data indicate that it decreased by exactly 3%, we know a great deal more – we know that the decline was significantly large.

Variables: measures that can take on different values.

Data: recorded values of variables.

Sets of data help us to test our models or theories, but first we need to pay attention to the economic logic involved in observations and modelling. For example, if sunspots or baggy pants were found to be correlated with economic expansion, would we consider these events a coincidence or a key to understanding economic growth? The observation is based on facts or data, but it need not have any economic content. The economist's task is to distinguish between coincidence and economic causation.

While the more frequent wearing of loose clothing in the past may have been associated with economic growth because they both occurred at the same time (correlation), one could not argue on a logical basis that this behaviour causes good economic times. Therefore, the past association of these variables should be considered as no more than a coincidence. Once specified on the basis of economic logic, a model must be tested to determine its usefulness in explaining observed economic events.

Source: “Theory and applications of economics”, 2012, Section 19.2, CC-BY-NC-SA 3.0 (Original source, “Toolkit: Section 31.16” only)

An **exogenous** variable is something that comes from outside a model and is not explained in our analysis. An **endogenous** variable is one that is explained within our analysis. When using the supply-and-demand framework, price and quantity are endogenous variables; everything else is exogenous.

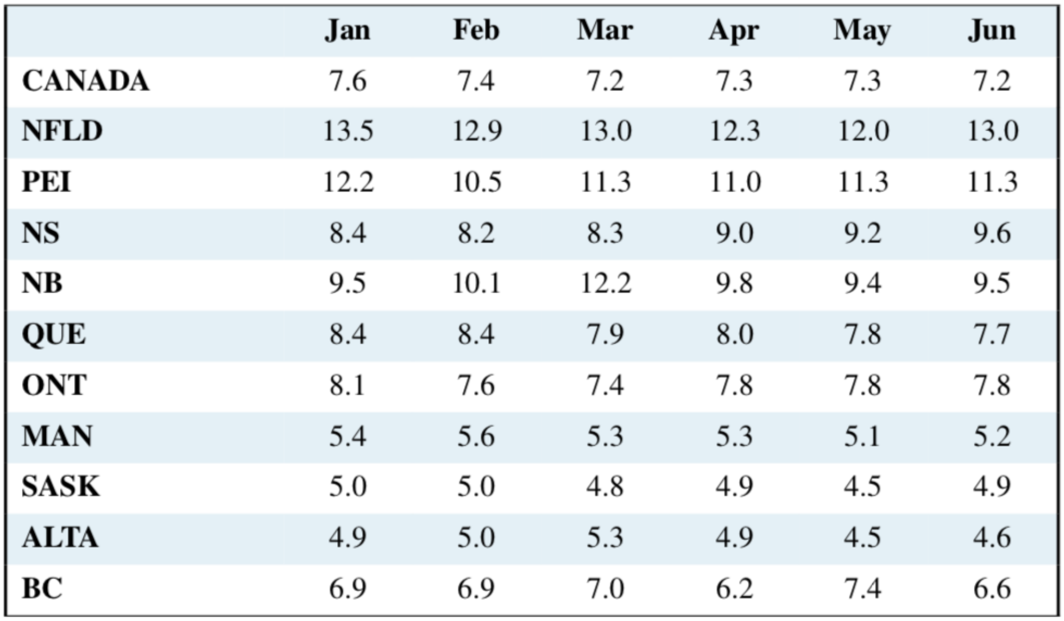
2.3 DATA, THEORY, AND ECONOMIC MODELS

Source: Curtis & Irvine, 2017, Section 2.2, CC-BY-NC-SA 3.0 (Original source, p. 34-34, only section titled *Index Numbers)*

Index Numbers

It is important in economic analysis to discuss and interpret data in a meaningful manner. **Index numbers** help us greatly in doing this. They are values of a given variable, or an average of a set of variables expressed relative to a base value. The key characteristics of indexes are that they are not dependent upon the units of measurement of the data in question, and they are interpretable easily with reference to a given base value. To illustrate, let us change the price data in column 2 of Table 2.1 into index number form.

Index number: value for a variable, or an average of a set of variables, expressed relative to a given base value.



	Jan	Feb	Mar	Apr	May	Jun
CANADA	7.6	7.4	7.2	7.3	7.3	7.2
NFLD	13.5	12.9	13.0	12.3	12.0	13.0
PEI	12.2	10.5	11.3	11.0	11.3	11.3
NS	8.4	8.2	8.3	9.0	9.2	9.6
NB	9.5	10.1	12.2	9.8	9.4	9.5
QUE	8.4	8.4	7.9	8.0	7.8	7.7
ONT	8.1	7.6	7.4	7.8	7.8	7.8
MAN	5.4	5.6	5.3	5.3	5.1	5.2
SASK	5.0	5.0	4.8	4.9	4.5	4.9
ALTA	4.9	5.0	5.3	4.9	4.5	4.6
BC	6.9	6.9	7.0	6.2	7.4	6.6

Table 2.2: Unemployment rates, Canada and Provinces, monthly 2012, seasonally adjusted

Source: Statistics Canada CANSIM Table 282-0087

The first step is to choose a base year as a reference point. This could be any one of the periods. We will simply take the first period as the year and set the price index value equal to 100 in that year. The value of 100 is usually chosen in order to make comparisons simple, but in some cases a base year value of 1.0 is used. If the base year value of 100 is used, the value of index in any year t is:

$$\text{Value of index} = \frac{\text{Absolute value in year t}}{\text{Absolute value in base year}} \times 100$$

Suppose we choose 1999 as the base year for constructing an index of the house prices given in Table 2.1 House prices in that year were $330,000. Then the index for the base year has a value:

$$\text{Index in 1999} = \frac{\$330{,}000}{\$330{,}000} \times 100 = 100$$

Applying the method to each value in column 2 yields column 3, which is now in index number form. For example, the January 2003 value is:

$$\text{Index in 2003} = \frac{\$395{,}000}{\$330{,}000} \times 100 = 119.7 \text{ 2.2.}$$

Each value in the index is interpreted relative to the value of 100, the base price in January 1999. The beauty of this column lies first in its ease of interpretation. For example, by 2003 the price increased to 119.7 points relative to a value of 100. This yields an immediate interpretation: The index has increased by 19.7 points per hundred or percent. While it is particularly easy to compute a percentage change in a data series when the base value is 100, it is not necessary that the reference point have a value of 100. By definition, a **percentage change** is given by the change in values relative to the initial value, multiplied by 100. For example, the percentage change in the price from 2006 to 2007, using the price index is: (190.91–175.76)/175.76×100 = 8.6 percent.

Percentage change = (change in values)/original value × 100.

Furthermore, index numbers enable us to make comparisons with the price patterns for other goods much more easily. If we had constructed a price index for wireless phones, which also had a base value of 100 in 1999, we could make immediate comparisons without having to compare one set of numbers defined in dollars with another defined in tens of thousands of dollars. In short, index numbers simplify the interpretation of data.

Data Analysis

Source: Curtis & Irvine, 2016, Section 2.1, CC-BY-NC-SA 3.0 (Original source, p. 3-7 only)

Data Types Data come in several forms. One form is **time-series**, which reflects a set of measurements made in sequence at different points in time. The first column in Table 2.1 reports the values for house prices in North Vancouver for the first quarter of each year, between 2001 and 2011. Evidently this is a time series. Annual data report one observation per year. We could, alternatively, have presented the data in monthly, weekly, or even daily form. The frequency we use depends on the purpose: If we are interested in the longer-term trend in house prices, then the annual form suffices. In contrast, financial economists, who study the behaviour of stock prices, might not be content with daily or even hourly prices; they may need prices minute-by-minute. Such data are called **high-frequency data**, whereas annual data are **low-frequency data**.

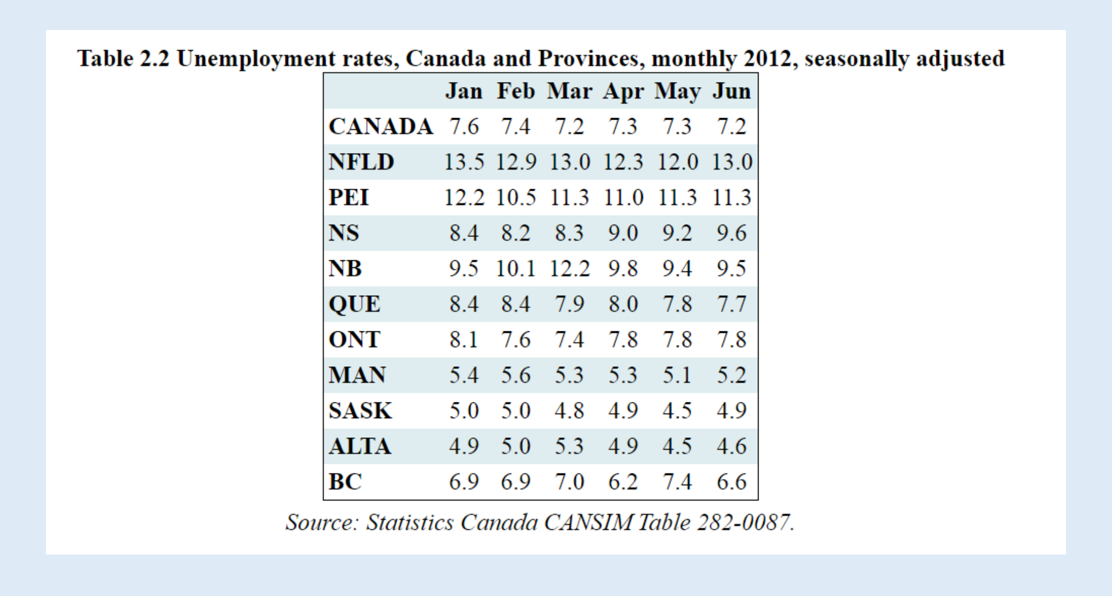


Table 2.2 Unemployment rates, Canada and Provinces, monthly 2012, seasonally adjusted

	Jan	Feb	Mar	Apr	May	Jun
CANADA	7.6	7.4	7.2	7.3	7.3	7.2
NFLD	13.5	12.9	13.0	12.3	12.0	13.0
PEI	12.2	10.5	11.3	11.0	11.3	11.3
NS	8.4	8.2	8.3	9.0	9.2	9.6
NB	9.5	10.1	12.2	9.8	9.4	9.5
QUE	8.4	8.4	7.9	8.0	7.8	7.7
ONT	8.1	7.6	7.4	7.8	7.8	7.8
MAN	5.4	5.6	5.3	5.3	5.1	5.2
SASK	5.0	5.0	4.8	4.9	4.5	4.9
ALTA	4.9	5.0	5.3	4.9	4.5	4.6
BC	6.9	6.9	7.0	6.2	7.4	6.6

Source: Statistics Canada CANSIM Table 282-0087.

Time-series: a set of measurements made sequentially at different points in time.

High (low) frequency data: series with short (long) intervals between observations.

In contrast to time-series data, cross-section data record the values of different variables at a point in time. Table 2.2 contains a cross-section of unemployment rates for Canada and Canadian provinces economies. For January 2012 we have a snapshot of the provincial economies at that point in time, likewise for the months until June. This table therefore contains **repeated cross-sections**.

When the unit of observation is the same over time such repeated cross sections are called longitudinal data. For example, a health survey that followed and interviewed the same individuals over time would yield longitudinal data. If the individuals differ each time the survey is conducted, the data are repeated cross sections**. Longitudinal data** therefore follow the same units of observation through time.

Cross-section data: values for different variables recorded at a point in time.

Repeated cross-section data: cross-section data recorded at regular or irregular intervals.

Longitudinal data: follow the same units of observation through time.

Graphing the data Data can be presented in graphical as well as tabular form. Figure 2.1 plots the house price data from the second column of Table 2.1. Each asterisk in the figure represents a price value and a corresponding time period. The horizontal axis reflects time, the vertical axis price in dollars. The graphical presentation of data simply provides a visual rather than numeric perspective. It is immediately evident that house prices increased consistently during this 11-year period, with a single downward 'correction' in 2009. We have plotted the data a second time in Figure 2.2 to illustrate the need to read graphs carefully. The greater *apparent* slope in Figure 2.1 might easily be interpreted to mean that prices increased more steeply than suggested in Figure 2.2. But a careful reading of the axes reveals that this is not so; using different scales when plotting data or constructing diagrams can mislead the unaware viewer.

Figure 2.1 House prices in dollars 1999-2012

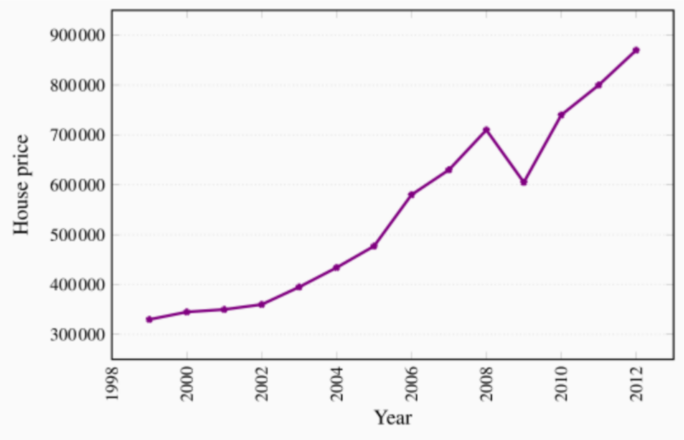
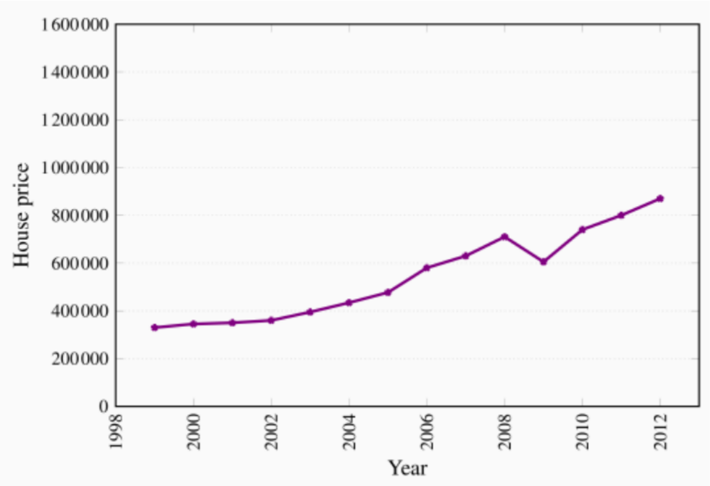


Figure 2.2 House prices in dollars 1999-2012



Example: Road Fatalities Data

Source: Curtis & Irvine, 2016, Section 2.2, CC-BY-NC-SA 3.0 (Original source, p. 4-5 only)

Road fatalities – theory, evidence and inference

Table 2.3 contains data on annual road fatalities per 100,000 drivers for various age groups. In the background, we have a *theory*, proposing that driver fatalities depend upon the age of the driver, the quality of roads and signage, speed limits, the age of the automobile stock and perhaps some other variables. Our model focuses upon a subset of these variables, and in order to present the example in graphical terms we specify fatalities as being dependent upon a single variable – age of driver.

Table 2.3 Non-linearity: Driver fatality rates Canada, 2009

Age of driver	Fatality rate per 100,000 drivers
20-24	9.8
25-34	4.4
35-44	2.7
45-54	2.4
55-64	1.9
65+	2.9

Source: Transport Canada, Canadian motor vehicle traffic collision statistics, 2009.

The scatter diagram is presented in Figure 2.4. Two aspects of this plot stand out. First, there is an exceedingly steep decline in the fatality rate when we go from the youngest age group to the next two age groups. The decline in fatalities between the youngest and second youngest groups is about 20 points, whereas the decline between the third and fourth age groups is less than 2 points. This suggests that behaviour is not the same throughout the age distribution. Second, we notice that fatalities increase for the oldest age group, perhaps indicating that the oldest drivers are not as good as middle-aged drivers.

These two features suggest that the relationship between fatalities and age differs across the age spectrum. Accordingly, a straightline would not be an accurate way of representing the behaviours in these data. A straight line through the plot implies that a given change in age should have a similar impact on fatalities, no matter the age group. Accordingly we have an example of a *non-linear relationship*. Such a non-linear relationship might be represented by the curve going through the plot. Clearly the slope of this line varies as we move from one age category to another.

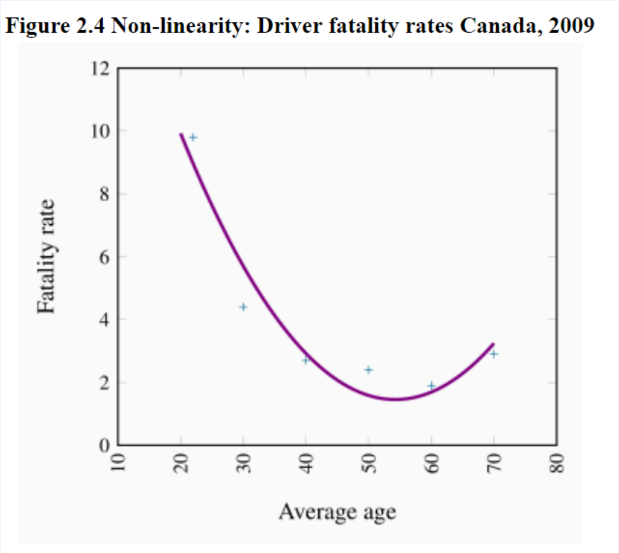


Figure 2.4 Non-linearity: Driver fatality rates Canada, 2009

Fatality rates vary non-linearly with age: At first they decline, then increase again, relative to the youngest age group.

2.4 THE USE OF MATHEMATICS IN PRINCIPLES OF ECONOMICS

Source: Lynham, 2018, Appendix A, CC-BY 4.0. (Original source, "Algebraic Models" only, removed sections *Solving Models with Algebra* and *Solving Models with Graphs)*

Functions Often economic models (or parts of models) are expressed in terms of mathematical functions. What is a function? A **function** describes a relationship. Sometimes the relationship is a definition. For example (using words), your professor is Adam Smith. This could be expressed as Professor = Adam Smith. Or Friends = Bob + Shawn + Margaret.

Often in economics, functions describe cause and effect. The variable on the left-hand side is what is being explained ("the effect"). On the right-hand side is what is doing the explaining ("the causes"). For example, suppose your GPA was determined as follows:

$$GPA = 0.25 \times combined\_SAT + 0.25 \times class\_attendance + 0.50 \times hours\_spent\_studying$$

This equation states that your GPA depends on three things: your combined SAT score, your class attendance, and the number of hours you spend studying. It also says that study

time is twice as important (0.50) as either combined_SAT score (0.25) or class_attendance (0.25). If this relationship is true, how could you raise your GPA? By not skipping class and studying more. Note that you cannot do anything about your SAT score, since if you are in college, you have (presumably) already taken the SATs.
Of course, economic models express relationships using economic variables, like Budget = money_spent_on_econ_books + money_spent_on_music, assuming that the only things you buy are economics books and music.

Most of the relationships we use in this course are expressed as linear equations of the form: $y = b + mx$

Expressing Equations Graphically Graphs are useful for two purposes. The first is to express equations visually, and the second is to display statistics or data. This section will discuss expressing equations visually.

To a mathematician or an economist, a **variable** is the name given to a quantity that may assume a range of values. In the equation of a line presented above, x and y are the variables, with x on the horizontal axis and y on the vertical axis, and b and m representing factors that determine the shape of the line. To see how this equation works, consider a numerical example:

$$y = 9 + 3x$$

In this equation for a specific line, the b term has been set equal to 9 and the m term has been set equal to 3. Table 1 shows the values of x and y for this given equation. Figure 1 shows this equation, and these values, in a graph. To construct the table, just plug in a series of different values for x, and then calculate what value of y results. In the figure, these points are plotted and a line is drawn through them.

x	y
0	9
1	12
2	15
3	18
4	21
5	24
6	27

Table 1. Values for the Slope Intercept Equation

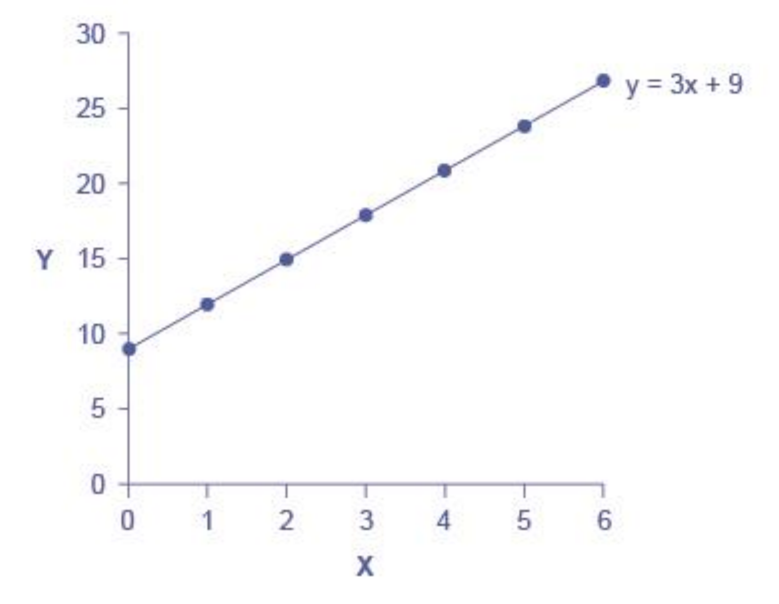


Figure 1. *Slope and the Algebra of Straight Lines. This line graph has x on the horizontal axis and y on the vertical axis. The y-intercept—that is, the point where the line intersects the y-axis—is 9. The slope of the line is 3; that is, there is a rise of 3 on the vertical axis for every increase of 1 on the horizontal axis. The slope is the same all along a straight line.*

This example illustrates how the b and m terms in an equation for a straight line determine the shape of the line. The b term is called the y-intercept. The reason for this name is that, if x = 0, then the b term will reveal where the line intercepts, or crosses, the y-axis. In this example, the line hits the vertical axis at 9. The m term in the equation for the line is the slope. Remember that **slope** is defined as rise over run; more specifically, the slope of a line from one point to another is the change in the vertical axis divided by the change in the horizontal axis. In this example, each time the x term increases by one (the run), the y term rises by three. Thus, the slope of this line is three. Specifying a y-intercept and a slope—that is, specifying b and m in the equation for a line—will identify a specific line. Although it is rare for real-world data points to arrange themselves as an exact straight line, it often turns out that a straight line can offer a reasonable approximation of actual data.

Interpreting the Slope

The concept of slope is very useful in economics, because it measures the relationship between two variables. A **positive slope** means that two variables are positively related; that is, when x increases, so does y, or when x decreases, y decreases also. Graphically, a positive slope means that as a line on the line graph moves from left to right, the line rises. The length-weight relationship, shown in Figure 3 later in this Appendix, has a positive slope. We will learn in other chapters that price and quantity supplied have a positive relationship; that is, firms will supply more when the price is higher.

A **negative slope** means that two variables are negatively related; that is, when x increases, y decreases, or when x decreases, y increases. Graphically, a negative slope means that, as the line on the line graph moves from left to right, the line falls. The altitude-air density relationship, shown in Figure 4 later in this appendix, has a negative slope. We will learn that price and quantity demanded have a negative relationship; that is, consumers will purchase less when the price is higher.

A slope of zero means that there is no relationship between x and y. Graphically, the line is flat; that is, zero rise over the run. Figure 5 of the unemployment rate, shown later in this appendix, illustrates a common pattern of many line graphs: some segments where the slope is positive, other segments where the slope is negative, and still other segments where the slope is close to zero.

The slope of a straight line between two points can be calculated in numerical terms. To calculate slope, begin by designating one point as the "starting point" and the other point as the "end point" and then calculating the rise over run between these two points. As an example, consider the slope of the air density graph between the points representing an altitude of 4,000 meters and an altitude of 6,000 meters:

Rise: Change in variable on vertical axis (end point minus original point)

$$\begin{aligned} &= 0.100 \; - \; 0.307 \\ &= -0.207 \end{aligned}$$

Run: Change in variable on horizontal axis (end point minus original point)

$$\begin{aligned} &= 6,000 \; - \; 4,000 \\ &= 2,000 \end{aligned}$$

Thus, the slope of a straight line between these two points would be that from the altitude of 4,000 meters up to 6,000 meters, the density of the air decreases by approximately 0.1 kilograms/cubic meter for each of the next 1,000 meters

Suppose the slope of a line were to increase. Graphically, that means it would get steeper. Suppose the slope of a line were to decrease. Then it would get flatter. These conditions are true whether or not the slope was positive or negative to begin with. A higher positive slope means a steeper upward tilt to the line, while a smaller positive slope means a flatter upward tilt to the line. A negative slope that is larger in absolute value (that is, more negative) means a steeper downward tilt to the line. A slope of zero is a horizontal flat line. A vertical line has an infinite slope.

Suppose a line has a larger intercept. Graphically, that means it would shift out (or up) from the old origin, parallel to the old line. If a line has a smaller intercept, it would shift in (or down), parallel to the old line.

CHAPTER 3: Market Demand, Supply, and Price

3.1 DEMAND FOR GOODS AND SERVICES

Quantity Demanded

Source: Lynham, 2018, Section 3.1, CC-BY 4.0. (Original source, paragraphs 1-5 only, subtitles added, "Is demand the same as quantity demanded" moved up)

Economists use the term **demand** to refer to the amount of some good or service consumers are willing and able to purchase at each price. Demand is based on needs and wants—a consumer may be able to differentiate between a need and a want, but from an economist's perspective they are the same thing. Demand is also based on ability to pay. If you cannot pay for it, you have no effective demand.

What a buyer pays for a unit of the specific good or service is called **price**. The total number of units purchased at that price is called the **quantity demanded**. A rise in price of a good or service almost always decreases the quantity demanded of that good or service. Conversely, a fall in price will increase the quantity demanded. When the price of a gallon of gasoline goes up, for example, people look for ways to reduce their consumption by combining several errands, commuting by carpool or mass transit, or taking weekend or vacation trips closer to home. Economists call this inverse relationship between price and quantity demanded the **law of demand**. The law of demand assumes that all other variables that affect demand (to be explained in the next module) are held constant.

Is demand the same as quantity demanded?

In economic terminology, demand is not the same as quantity demanded. When economists talk about demand, they mean the relationship between a range of prices and the quantities demanded at those prices, as illustrated by a demand curve or a demand schedule. When economists talk about quantity demanded, they mean only a certain point on the demand curve, or one quantity on the demand schedule. In short, demand refers to the curve and quantity demanded refers to the (specific) point on the curve.

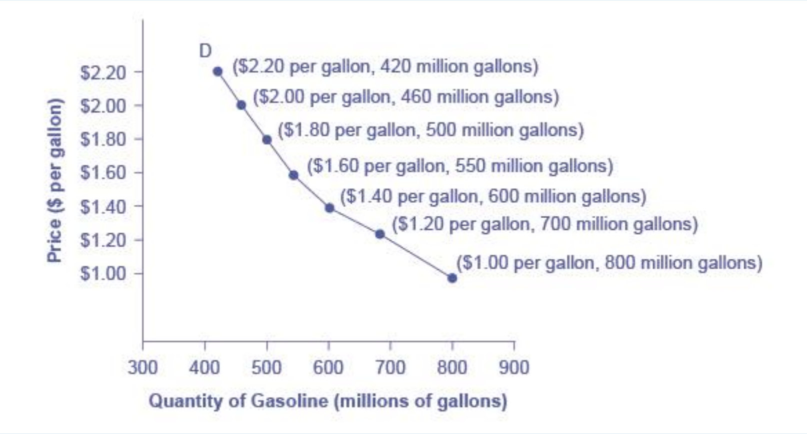
The Demand Curve

An example from the market for gasoline can be shown in the form of a table or a graph. A table that shows the quantity demanded at each price, such as Table 1, is called a **demand schedule**. Price in this case is measured in dollars per gallon of gasoline. The quantity demanded is measured in millions of gallons over some time period (for example, per day or per year) and over some geographic area (like a state or a country). A **demand curve** shows the relationship between price and quantity demanded on a graph like Figure 1, with quantity on the horizontal axis and the price per gallon on the vertical axis. (Note that this is an exception to the normal rule in mathematics that the independent variable (x) goes on the horizontal axis and the dependent variable (y) goes on the vertical. Economics is not math.)

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

The demand schedule shown by Table 1 and the demand curve shown by the graph in Figure 1 are two ways of describing the same relationship between price and quantity demanded.

Figure 1. *A Demand Curve for Gasoline. The demand schedule shows that as price rises, quantity demanded decreases, and vice versa. These points are then graphed, and the line connecting them is the demand curve (D). The downward slope of the demand curve again illustrates the law of demand—the inverse relationship between prices and quantity demanded.*



Price (per gallon)	Quantity Demanded (millions of gallons)
$1.00	800
$1.20	700
$1.40	600
$1.60	550
$1.80	500
$2.00	460
$2.20	420

Table 1. Price and Quantity Demanded of Gasoline

Demand curves will appear somewhat different for each product. They may appear relatively steep or flat, or they may be straight or curved. Nearly all demand curves share the fundamental similarity that they slope down from left to right. So demand curves embody the law of demand: As the price increases, the quantity demanded decreases, and conversely, as the price decreases, the quantity demanded increases.

Source: Curtis & Irvine, 2016, Section 2.2, CC-BY-NC-SA 3.0 (Original source, p. 3, "demand curve" definition only)

The **demand curve** is a graphical expression of the relationship between price and quantity demanded, with other influences remaining unchanged.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Shifts in Demand Any of the following variables (other than the product's price) influence demand and will shift the demand curve:

Source: Curtis & Irvine, 2016, Section 2.2, CC-BY-NC-SA 3.0 (Original source; p. 1, paragraph 1 removed; p. 5-6 removed; list numbers added; "5. Changes in the Composition of the Population" added from other source)

1. The prices of related goods – oil and gas, Kindle and paperbacks We expect that the price of other forms of energy would impact the price of natural gas. For example, if electricity, oil or coal becomes less expensive we would expect some buyers to switch to these other products. Alternatively, if gas-burning furnaces experience a technological breakthrough that makes them more efficient and cheaper we would expect some users of other fuels to move to gas. Among these examples, it is clear that oil and electricity are substitute fuels for gas; in contrast the efficient new gas furnace complements the use of gas. We use these terms, substitutes and complements, to describe products that influence the demand for the primary good.

Substitute goods: when a price reduction (rise) for a related product reduces (increases) the demand for a primary product, it is a substitute for the primary product.

Complementary goods: when a price reduction (rise) for a related product increases (reduces) the demand for a primary product, it is a complement for the primary product.

Clearly electricity is a substitute for gas in the power market, whereas a gas furnace is a complement for gas as a fuel. The words substitutes and complements immediately suggest the nature of the relationships. Every product has complements and substitutes. As another example: Electronic readers such as Kindle, Nook and Kobo are substitutes for paper-form books; a rise in the price of paper books should increase the demand for electronic readers at any given price for electronic readers. In graphical terms, the demand curve shifts in response to changes in the prices of other goods – an increase in the price of paper-form books will shift the demand for electronic readers outward, because more electronic readers will be demanded at any price.

2. Buyer incomes – which goods to buy The demand for most goods increases in response to income increases. Given this, the demand curve for gas will shift outward if household incomes in the economy increase. Household incomes may increase either because there are more households in the economy or because the incomes of the existing households grow.

Most goods are demanded in greater quantity in response to higher incomes at any given price. But there are exceptions. For example, public transit demand may decline at any price when household incomes rise, because some individuals move to cars. Or the demand for laundromats may decline in response to higher incomes, as households purchase more of their own consumer durables – washers and driers. We use the term inferior good to define these cases: An inferior good is one whose demand declines in response to increasing incomes, whereas a normal good experiences an increase in demand in response to rising incomes.

An **inferior good** is one whose demand falls in response to higher incomes.

A **normal good** is one whose demand increases in response to higher incomes.

There is a further sense in which consumer incomes influence demand, and this relates to how the incomes are distributed in the economy. In the discussion above we stated that higher total incomes shift demand curves outwards when goods are normal. But think of the difference in the demand for electronic readers between Portugal and Saudi Arabia. These economies have roughly the same average per-person income, but incomes are distributed more unequally in Saudi Arabia. It does not have a large middle class that can afford electronic readers or iPads, despite the huge wealth held by the elite. In contrast, Portugal has a relatively larger middle class that can afford such goods. Consequently, the distribution of income can be an important determinant of the demand for many commodities and services.

3. Tastes and networks – hemlines and homogeneity While demand functions are drawn on the assumption that tastes are constant, in an evolving world they are not. We are all subject to peer pressure, the fashion industry, marketing, and a desire to maintain our image. If the fashion industry dictates that lapels or long skirts are de rigueur for the coming season, some fashion-conscious individuals will discard a large segment of their wardrobe, even though the clothes may be in perfectly good condition: Their demand is influenced by the dictates of current fashion.

Correspondingly, the items that other individuals buy or use frequently determine our own purchases. Businesses frequently decide that all of their employees will have the same type of computer and software on account of network economies: It is easier to communicate if equipment is compatible, and it is less costly to maintain infrastructure where the variety is less.

4. Expectations – betting on the future In our natural gas example, if households expected that the price of natural gas was going to stay relatively low for many years – perhaps on account of the discovery of large deposits – then they would be tempted to purchase a gas burning furnace rather than an oil burning furnace. In this example, it is more than the current price that determines choices; the prices that are expected to prevail in the future also determine current demand.

Expectations are particularly important in stock markets. When investors anticipate that corporations will earn high rewards in the future they will buy a stock today. If enough people believe this, the price of the stock will be driven upward on the market, even before profitable earnings are registered.

Source: Lynham, 2018, Section 3.2, CC-BY 4.0. (Original source, paragraph 13 only)

5. Changes in the Composition of the Population The proportion of elderly citizens in the United States population is rising. It rose from 9.8% in 1970 to 12.6% in 2000, and will be a projected (by the U.S. Census Bureau) 20% of the population by 2030. A society

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

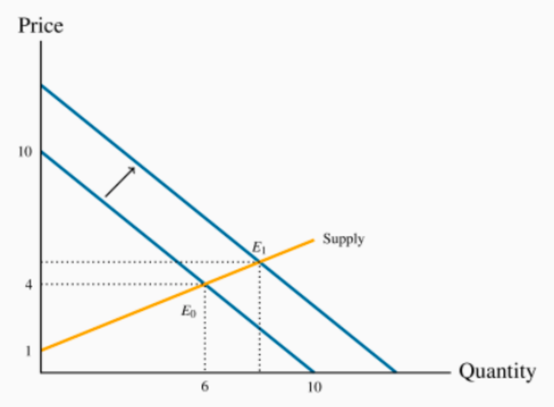
with relatively more children, like the United States in the 1960s, will have greater demand for goods and services like tricycles and day care facilities. A society with relatively more elderly persons, as the United States is projected to have by 2030, has a higher demand for nursing homes and hearing aids. Similarly, changes in the size of the population can affect the demand for housing and many other goods. Each of these changes in demand will be shown as a shift in the demand curve.

END OF THIS SOURCE: Lynham, 2018, Section 3.1, CC-BY 4.0.

Shifts in demand

The demand curve in Figure 3.2 is drawn for a given level of other prices, incomes, tastes, and expectations. Movements along the demand curve reflect solely the impact of different prices for the good in question, holding other influences constant. But changes in any of these other factors will change the position of the demand curve. Figure 3.3 illustrates a shift in the demand curve. This shift could result from a rise in household incomes that increase the quantity demanded at every price. This is illustrated by an outward shift in the demand curve. With supply conditions unchanged, there is a new equilibrium at E1, indicating a greater quantity of purchases accompanied by a higher price. The new equilibrium reflects a change in quantity supplied and a change in demand.

Figure 3.3 Demand shift and new equilibrium



The outward shift in demand leads to a new equilibrium E_1.

Source: Lynham, 2018, Section 3.2, CC-BY 4.0. (Original source, “Summing Up Factors That Change Demand” only, Figure 5 moved up)

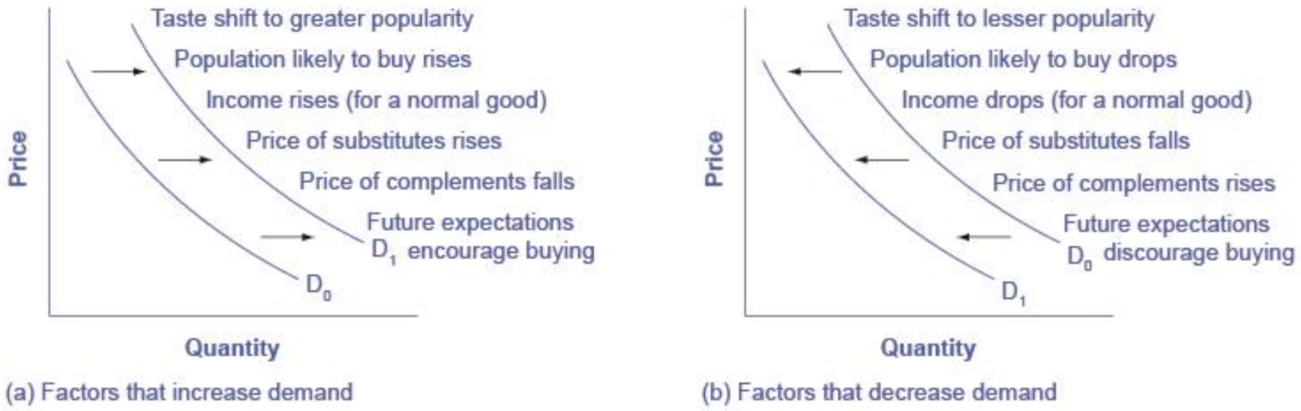

Taste shift to greater popularity
Population likely to buy rises
Income rises (for a normal good)
Price of substitutes rises
Price of complements falls
Future expectations encourage buying
D_1
D_0
Price
Quantity
(a) Factors that increase demand
Taste shift to lesser popularity
Population likely to buy drops
Income drops (for a normal good)
Price of substitutes falls
Price of complements rises
Future expectations discourage buying
D_0
D_1
Price
Quantity
(b) Factors that decrease demand

Figure 5. *Factors That Shift Demand Curves. (a) A list of factors that can cause an increase in demand from D_0 to D_1. (b) The same factors, if their direction is reversed, can cause a decrease in demand from D_0 to D_1.*

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Six factors that can shift demand curves are summarized in Figure 5. The direction of the arrows indicates whether the demand curve shifts represent an increase in demand or a decrease in demand. Notice that a change in the price of the good or service itself is not listed among the factors that can shift a demand curve. A change in the price of a good or service causes a movement along a specific demand curve, and it typically leads to some change in the quantity demanded, but it does not shift the demand curve.

When a demand curve shifts, it will then intersect with a given supply curve at a different equilibrium price and quantity. We are, however, getting ahead of our story. Before discussing how changes in demand can affect equilibrium price and quantity, we first need to discuss shifts in supply curves.

3.2 Supply of Goods and Services

Source: Lynham, 2018, Section 3.1, CC-BY 4.0. (Original source, paragraphs 6-9 only, subtitles added)

Quantity Supplied

When economists talk about **supply**, they mean the amount of some good or service a producer is willing to supply at each price. Price is what the producer receives for selling one unit of a **good** or **service**. A rise in price almost always leads to an increase in the **quantity supplied** of that good or service, while a fall in price will decrease the quantity supplied. When the price of gasoline rises, for example, it encourages profit-seeking firms to take several actions: expand exploration for oil reserves; drill for more oil; invest in more pipelines and oil tankers to bring the oil to plants where it can be refined into gasoline; build new oil refineries; purchase additional pipelines and trucks to ship the gasoline to gas stations; and open more gas stations or keep existing gas stations open longer hours. Economists call this positive relationship between price and quantity supplied—that a higher price leads to a higher quantity supplied and a lower price leads to a lower quantity supplied—the **law of supply**. The law of supply assumes that all other variables that affect supply (to be explained in the next module) are held constant.

Still unsure about the different types of supply? See the following Clear It Up feature.

Is supply the same as quantity supplied?

In economic terminology, supply is not the same as quantity supplied. When economists refer to supply, they mean the relationship between a range of prices and the quantities supplied at those prices, a relationship that can be illustrated with a supply curve or a supply schedule. When economists refer to quantity supplied, they mean only a certain point on the supply curve, or one quantity on the supply schedule. In short, supply refers to the curve and quantity supplied refers to the (specific) point on the curve.

The Supply Curve

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Figure 2 illustrates the law of supply, again using the market for gasoline as an example. Like demand, supply can be illustrated using a table or a graph. A **supply schedule** is a table, like Table 2, that shows the quantity supplied at a range of different prices. Again, price is measured in dollars per gallon of gasoline and quantity supplied is measured in millions of gallons. A **supply curve** is a graphic illustration of the relationship between price, shown on the vertical axis, and quantity, shown on the horizontal axis. The supply schedule and the supply curve are just two different ways of showing the same information. Notice that the horizontal and vertical axes on the graph for the supply curve are the same as for the demand curve.

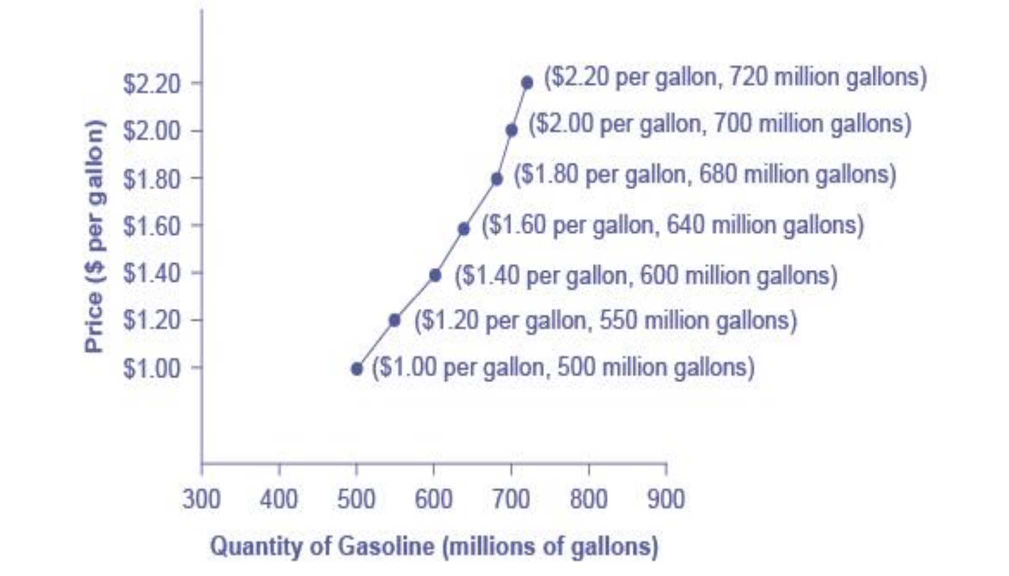


Figure 2. *A Supply Curve for Gasoline. The supply schedule is the table that shows quantity supplied of gasoline at each price. As price rises, quantity supplied also increases, and vice versa. The supply curve (S) is created by graphing the points from the supply schedule and then connecting them. The upward slope of the supply curve illustrates the law of supply—that a higher price leads to a higher quantity supplied, and vice versa.*

Price (per gallon)	Quantity Supplied (millions of gallons)
$1.00	500
$1.20	550
$1.40	600
$1.60	640
$1.80	680
$2.00	700
$2.20	720

Table 2. Price and Supply of Gasoline

The shape of supply curves will vary somewhat according to the product: steeper, flatter, straighter, or curved. Nearly all supply curves, however, share a basic similarity: they slope up from left to right and illustrate the law of supply: as the price rises, say, from $1.00 per gallon to $2.20 per gallon, the quantity supplied increases from 500 gallons to 720 gallons. Conversely, as the price falls, the quantity supplied decreases.

Source: Lynham, 2018, Section 3.1, CC-BY 4.0. (Original source, paragraphs 18-20, 25-30, numbered list added)

How Production Costs Affect Supply

Shifts in the Supply Curve

1. Input Prices In thinking about the factors that affect supply, remember what motivates firms: profits, which are the difference between revenues and costs. Goods and services are produced using combinations of labor, materials, and machinery, or what we call **inputs** or **factors of production**. If a firm faces lower costs of production, while the prices for the good or service the firm produces remain unchanged, a firm's profits go up. When a firm's profits increase, it is more motivated to produce output, since the more it produces the more profit it will earn. So, when costs of production fall, a firm will tend to supply a larger quantity at any given price for its output. This can be shown by the supply curve shifting to the right.

Take, for example, a messenger company that delivers packages around a city. The company may find that buying gasoline is one of its main costs. If the price of gasoline falls, then the company will find it can deliver messages more cheaply than before. Since lower costs correspond to higher profits, the messenger company may now supply more of its services at any given price. For example, given the lower gasoline prices, the company can now serve a greater area, and increase its supply.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Conversely, if a firm faces higher costs of production, then it will earn lower profits at any given selling price for its products. As a result, a higher cost of production typically causes a firm to supply a smaller quantity at any given price. In this case, the supply curve shifts to the left.

Other Factors That Affect Supply

2. Natural Conditions The cost of production for many agricultural products will be affected by changes in natural conditions. For example, in 2014 the Manchurian Plain in Northeastern China, which produces most of the country's wheat, corn, and soybeans, experienced its most severe drought in 50 years. A drought decreases the supply of agricultural products, which means that at any given price, a lower quantity will be supplied; conversely, especially good weather would shift the supply curve to the right.

3. Technology When a **firm** discovers a new technology that allows the firm to produce at a lower cost, the supply curve will shift to the right, as well. For instance, in the 1960s a major scientific effort nicknamed the Green Revolution focused on breeding improved seeds for basic crops like wheat and rice. By the early 1990s, more than two-thirds of the wheat and rice in low-income countries around the world was grown with these Green Revolution seeds—and the harvest was twice as high per acre. A technological improvement that reduces costs of production will shift supply to the right, so that a greater quantity will be produced at any given price.

4. Government Taxes & Subsidies Government policies can affect the cost of production and the supply curve through taxes, regulations, and subsidies. For example, the U.S. government imposes a tax on alcoholic beverages that collects about $8 billion per year from producers. Taxes are treated as costs by businesses. Higher costs decrease supply for the reasons discussed above. Other examples of policy that can affect cost are the wide array of government regulations that require firms to spend money to provide a cleaner environment or a safer workplace; complying with regulations increases costs.

A government subsidy, on the other hand, is the opposite of a tax. A subsidy occurs when the government pays a firm directly or reduces the firm's taxes if the firm carries out certain actions. From the firm's perspective, taxes or regulations are an additional cost of production that shifts supply to the left, leading the firm to produce a lower quantity at every given price. Government subsidies reduce the cost of production and increase supply at every given price, shifting supply to the right. The following Work It Out feature shows how this shift happens.

Summing Up Factors That Change Supply

Changes in the cost of inputs, natural disasters, new technologies, and the impact of government decisions all affect the cost of production. In turn, these factors affect how much firms are willing to supply at any given price.

Figure 11 summarizes factors that change the supply of goods and services. Notice that a change in the price of the product itself is not among the factors that shift the supply curve. Although a change in price of a good or service typically causes a change in quantity supplied or a movement along the supply curve for that specific good or service, it does not cause the supply curve itself to shift.

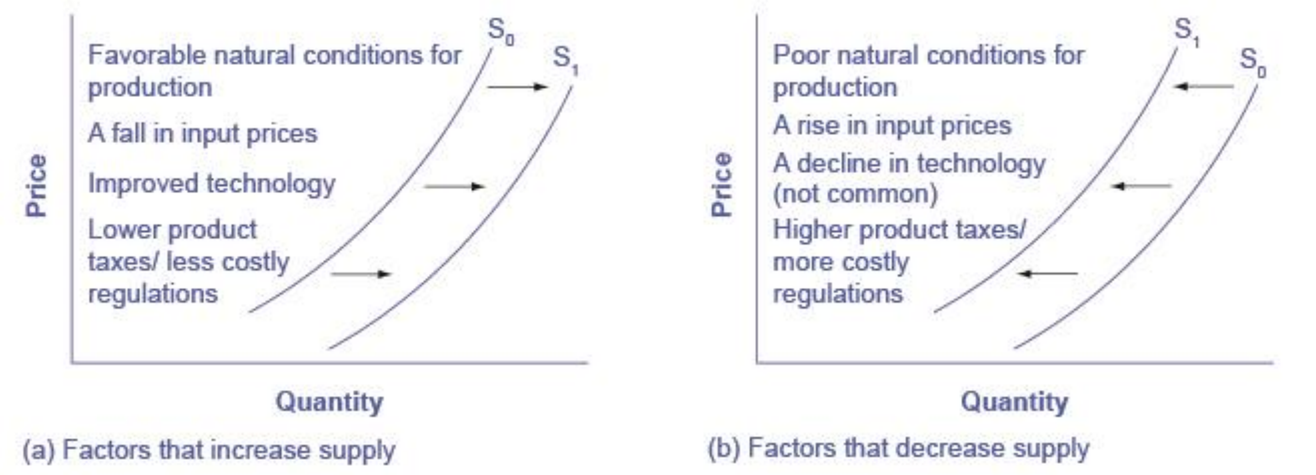


Figure 11. *Factors That Shift Supply Curves. (a) A list of factors that can cause an increase in supply from $S_0$to S_1. (b) The same factors, if their direction is reversed, can cause a decrease in supply from S_0 to S*

3.3 Price Formation & Market Equilibrium

The Marketplace

Source: Curtis & Irvine, 2016, Section 3.1, CC-BY-NC-SA 3.0 (Original source, paragraphs 3-5 only)

The importance of the marketplace springs from its role as an allocating mechanism. Elevated prices effectively send a signal to suppliers that the buyers in the market place a high value on the product being traded; conversely when prices are low. Accordingly, suppliers may decide to cease supplying markets where prices do not remunerate them sufficiently, and redirect their energies and the productive resources under their control to other markets – markets where the product being traded is more highly valued, and where the buyer is willing to pay more.

Whatever their form, the marketplace is central to the economy we live in. Not only does it facilitate trade, it also provides a means of earning a livelihood. Suppliers must hire resources – human and non-human in order to bring their supplies to market and these resources must be paid a return – income is generated.

In this chapter we will examine the process of price formation – how the prices that we observe in the marketplace come to be what they are. We will illustrate that the price for a good is inevitably linked to the quantity of a good; price and quantity are different sides of the same coin and cannot generally be analyzed separately. To understand this process more fully, we need to model a typical market. The essentials are demand and supply.

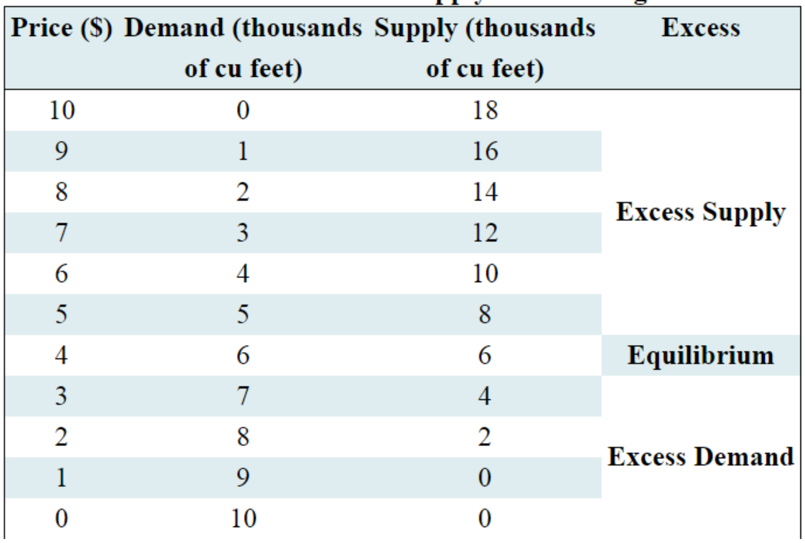
Market Equilibrium

Source: Curtis & Irvine, 2016, Section 3.2, CC-BY-NC-SA 3.0 (Original source, pages 5-6 only, Table 3.1 from p. 2 added, last 2 paragraphs removed)

Let us now bring the demand and supply schedules together in an attempt to analyze what the marketplace will produce – will a single price emerge that will equate supply and demand? We will keep other things constant for the moment, and explore what materializes at different prices. At low prices, the data in Table 3.1 indicate that the quantity demanded exceeds the quantity supplied – for example, verify what happens when the price is $3 per unit. The opposite occurs when the price is high – what would happen if the price were $8? Evidently, there exists an intermediate price, where the quantity demanded equals the quantity supplied. At this point we say that the market is in equilibrium. The **equilibrium price** equates demand and supply – it clears the market.

The **equilibrium price** equilibrates the market. It is the price at which quantity demanded equals the quantity supplied.

Table 3.1 Demand and supply for natural gas



Price ($)	Demand (thousands of cu feet)	Supply (thousands of cu feet)	Excess
10	0	18	Excess Supply
9	1	16	
8	2	14	
7	3	12	
6	4	10	
5	5	8	
4	6	6	Equilibrium
3	7	4	Excess Demand
2	8	2	
1	9	0	
0	10	0	

In Table 3.1 the equilibrium price is $4, and the equilibrium quantity is 6 thousand cubic feet of gas (we will use the notation 'k' to denote thousands). At higher prices there is an **excess supply**—suppliers wish to sell more than buyers wish to buy. Conversely, at lower prices there is an **excess demand**. Only at the equilibrium price is the quantity supplied equal to the quantity demanded.

Excess supply exists when the quantity supplied exceeds the quantity demanded at the going price.

Excess demand exists when the quantity demanded exceeds the quantity supplied at the going price.

Does the market automatically reach equilibrium? To answer this question, suppose initially that the sellers choose a price of $10. Here suppliers would like to supply 18k cubic feet, but there are no buyers—a situation of extreme excess supply. At the price of $7 the excess supply is reduced to 9k, because both the quantity demanded is now higher at 3k units, and the quantity supplied is lower at 12k. But excess supply means that there are suppliers willing to supply at a lower price, and this willingness exerts continual downward pressure on any price above the price that equates demand and supply.

At prices below the equilibrium there is, conversely, an excess demand. In this situation, suppliers could force the price upward, knowing that buyers will continue to buy at a price at which the suppliers are willing to sell. Such upward pressure would continue until the excess demand is eliminated.

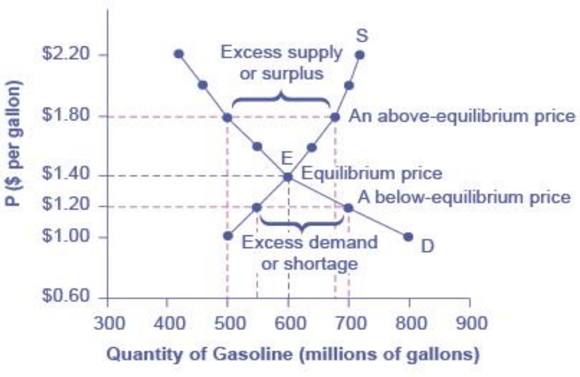
In general then, above the equilibrium price excess supply exerts downward pressure on price, and below the equilibrium excess demand exerts upward pressure on price. This process implies that the buyers and sellers have information on the various elements that make up the marketplace.

Source: Lynham, 2018, Section 3.1, CC-BY 4.0 (Original source, paragraphs 9-10 only)

Because the graphs for demand and supply curves both have price on the vertical axis and quantity on the horizontal axis, the demand curve and supply curve for a particular good or service can appear on the same graph. Together, demand and supply determine the price and the quantity that will be bought and sold in a market.

Figure 3 illustrates the interaction of demand and supply in the market for gasoline. The demand curve (D) is identical to Figure 1. The supply curve (S) is identical to Figure 2. Table 3 contains the same information in tabular form.

Figure 3. *Demand and Supply for Gasoline. The demand curve (D) and the supply curve (S) intersect at the equilibrium point E, with a price of $1.40 and a quantity of 600. The equilibrium is the only price where quantity demanded is equal to quantity supplied. At a price above equilibrium like $1.80, quantity supplied exceeds the quantity demanded, so there is excess supply. At a price below equilibrium such as $1.20, quantity demanded exceeds quantity supplied, so there is excess demand.*



Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Price (per gallon)	Quantity demanded (millions of gallons)	Quantity supplied (millions of gallons)
$1.00	800	500
$1.20	700	550
$1.40	**600**	**600**
$1.60	550	640
$1.80	500	680
$2.00	460	700
$2.20	420	720

Table 3. Price, Quantity Demanded, and Quantity Supplied

Numerical Example: We can determine the market equilibrium mathematically by using the supply and demand functions.

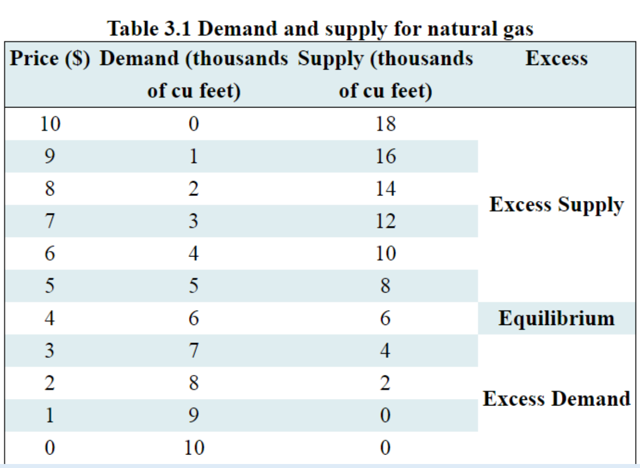
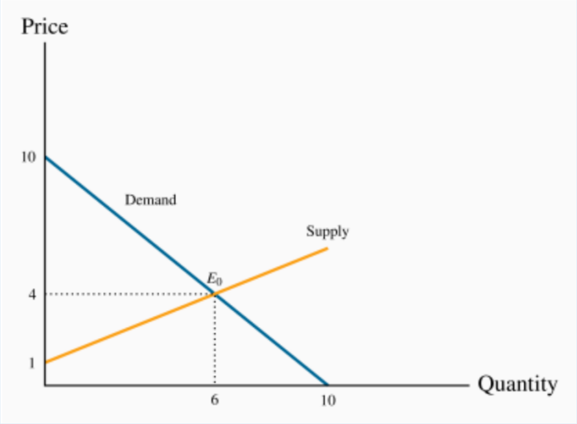


Table 3.1 Demand and supply for natural gas

Price ($)	Demand (thousands of cu feet)	Supply (thousands of cu feet)	Excess
10	0	18	Excess Supply
9	1	16	
8	2	14	
7	3	12	
6	4	10	
5	5	8	
4	6	6	Equilibrium
3	7	4	Excess Demand
2	8	2	
1	9	0	
0	10	0	

Figure 3.2 Supply, demand, equilibrium



Source: Curtis & Irvine, 2016, Section 3.2 & 3.3, CC-BY-NC-SA 3.0 (Original source, Table 3.1 from 3.2 page 2. Figure 3.2 from 3.3 page 3)

Source: Curtis & Irvine, 2016, Section 3.9, CC-BY-NC-SA 3.0 (Original source, first line and last sentences removed, subtitle added)

Demand: P=10−Q

Supply: P=1+(1/2)Q

A straight line is represented completely by the intercept and slope. In particular, if the variable P is on the vertical axis and Q on the horizontal axis, the straight-line equation relating P and Q is defined by P=a+bQ. Where the line is negatively sloped, as in the demand equation, the parameter b must take a negative value. By observing either the data in Table 3.1 or Figure 3.2 it is clear that the vertical intercept, a, takes a value of $10. The vertical intercept corresponds to a zero-value for the Q variable. Next we can see from Figure 3.2 that the slope (given by the rise over the run) is 10/10 and hence has a value of −1. Accordingly the demand equation takes the form P=10−Q.

On the supply side the price-axis intercept, from either the figure or the table, is clearly 1. The slope is one half, because a two-unit change in quantity is associated with a one-unit change in price. This is a positive relationship obviously so the supply curve can be written as P=1+(1/2)Q.

Where the supply and demand curves intersect is the market equilibrium; that is, the price-quantity combination is the same for both supply and demand where the supply curve takes on the same values as the demand curve. This unique price-quantity combination is obtained by equating the two curves: If Demand=Supply, then

$$10-Q=1+(1/2)Q.$$

Gathering the terms involving Q to one side and the numerical terms to the other side of the equation results in 9=1.5Q. This implies that the equilibrium quantity must be 6 units. And this quantity must trade at a price of $4. That is, when the price is $4 both the quantity demanded and the quantity supplied take a value of 6 units.

Changes Market Equilibrium

To illustrate the impact of market interventions examined in Section 3.7 on our numerical market model for natural gas, suppose that the government imposes a minimum price of $6 – above the equilibrium price obviously. We can easily determine the quantity supplied and demanded at such a price. Given the supply equation

$$P=1+(1/2)Q,$$

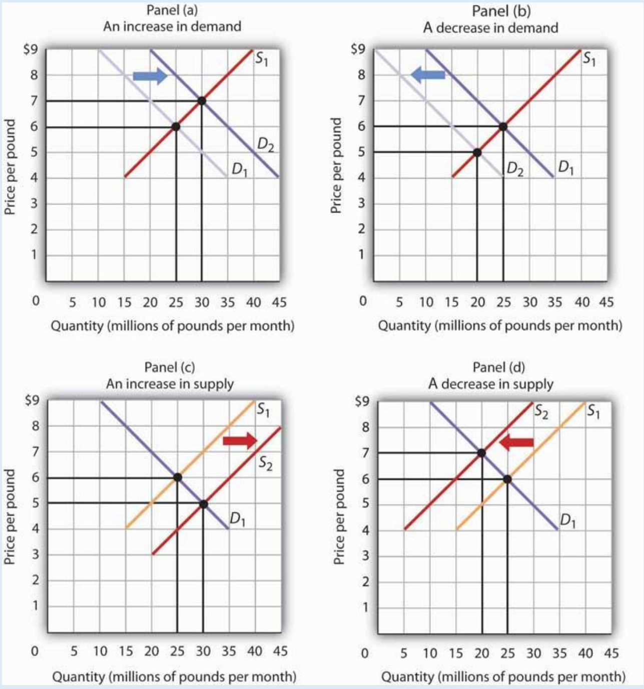
it follows that at P=6 the quantity supplied is 10. This follows by solving the relationship 6=1+(1/2)Q for the value of Q. Accordingly, suppliers would like to supply 10 units at this price.

Correspondingly on the demand side, given the demand curve

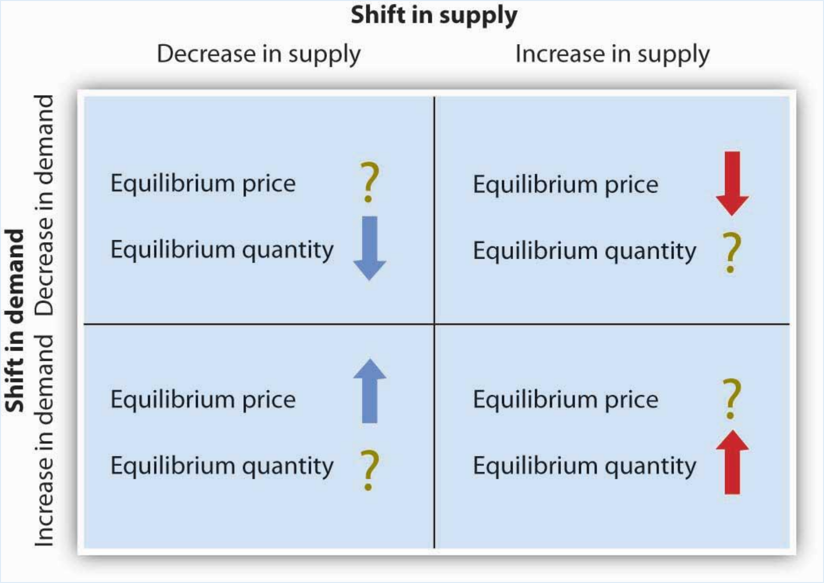
$$P=10-Q,$$

with a price given by P=$6, it must be the case that Q=4. So buyers *would like to buy* 4 units at that price: There is excess supply. But we know that the short side of the market will win out, and so the actual amount traded at this restricted price will be 4 units.

Source: “Macroeconomic principles, 2012, Section 3.3, CC-BY-NC-SA 3.0 (Original source Figure 3.17 & Figure 3.20 only)



A change in demand or in supply changes the equilibrium solution in the model. Panels (a) and (b) show an increase and a decrease in demand, respectively; Panels (c) and (d) show an increase and a decrease in supply, respectively.



		Shift in supply	
		Decrease in supply	Increase in supply
Shift in demand	Decrease in demand	Equilibrium price ? Equilibrium quantity ↓	Equilibrium price ↓ Equilibrium quantity ?
	Increase in demand	Equilibrium price ↑ Equilibrium quantity ?	Equilibrium price ? Equilibrium quantity ↑

If simultaneous shifts in demand and supply cause equilibrium price or quantity to move in the same direction, then equilibrium price or quantity clearly moves in that direction. If the shift in one of the curves causes equilibrium price or quantity to rise while the shift in the other curve causes equilibrium price or quantity to fall, then the relative amount by which each curve shifts is critical to figuring out what happens to that variable.

CHAPTER 4: Elasticity
4.1 PRICE ELASTICITY OF DEMAND

Source: Curtis & Irvine, 2016, Section 4.1, CC-BY-NC-SA 3.0 (Original source, pages 5-7 removed)

Put yourself in the position of an entrepreneur. One of your many challenges is to price your product appropriately. You may be Michael Dell choosing a price for your latest computer, or the local restaurant owner pricing your table d'ho ̂te, or you may be pricing your part-time snow-shoveling service. A key component of the pricing decision is to know how *responsive* your market is to variations in your pricing. How we measure responsiveness is the subject matter of this chapter.

We begin by analyzing the responsiveness of consumers to price changes. For example, consumers tend not to buy much more or much less food in response to changes in the general price level of food. This is because food is a pretty basic item for our existence. In contrast, if the price of textbooks becomes higher, students may decide to search for a second-hand copy, or make do with lecture notes from their friends or downloads from the course web site. In the latter case students have ready alternatives to the new text book, and so their expenditure patterns can be expected to reflect these options, whereas it is hard to find alternatives to food. In the case of food consumers are not very responsive to price changes; in the case of textbooks they are. The word 'elasticity' that appears in this chapter title is just another term for this concept of responsiveness. Elasticity has many different uses and interpretations, and indeed more than one way of being measured in any given situation. Let us start by developing a suitable numerical measure.

The slope of the demand curve suggests itself as one measure of responsiveness: If we lowered the price of a good by $1, for example, how many more units would we sell? The difficulty with this measure is that it does not serve us well when comparing different products. One dollar may be a substantial part of the price of your morning coffee and croissant, but not very important if buying a computer or tablet. Accordingly, when goods and services are measured in different units (croissants versus tablets), or when their prices are very different, it is often best to use a percentage change measure, which is *unit-free*.

The **price elasticity of demand** is measured as the percentage change in quantity demanded, divided by the percentage change in price. Although we introduce several other elasticity measures later, when economists speak of the demand elasticity they invariably mean the price elasticity of demand defined in this way.

The **price elasticity of demand** is measured as the percentage change in quantity demanded, divided by the percentage change in price.

The price elasticity of demand can be written in different forms. We will use the Greek letter epsilon, ε, as a shorthand symbol, with a subscript d to denote demand, and the capital delta, Δ, to denote a change. Therefore, we can write

$$\text{Price elasticity of demand} = \varepsilon d = \frac{\textit{Percentage change in quantity demanded}}{\textit{Percentage change in price}}$$

or, using a shortened expression,

$$\varepsilon d = \frac{\%\Delta Q}{\%\Delta P}$$

Calculating the value of the elasticity is not difficult. If we are told that a 10 percent price increase reduces the quantity demanded by 20 percent, then the elasticity value is −20%/10%=−2. The negative sign denotes that price and quantity move in opposite directions, but for brevity the negative sign is often omitted.

Consider now the data in Table 4.1 and the accompanying Figure 4.1. These data reflect the demand relation for natural gas that we introduced in Chapter 3. Note first that, when the price and quantity change, we must decide what *reference price and quantity* to use in the percentage change calculation in the definition above. We could use the initial or final price-quantity combination, or an average of the two. Each choice will yield a slightly different numerical value for the elasticity. The best convention is to *use the midpoint of the price values and the corresponding midpoint of the quantity values*. This ensures that the elasticity value is the same regardless of whether we start at the higher price or the lower price. Using the subscript 1 to denote the initial value and 2 the final value:

Average quantity Q= (Q1+Q2)/2

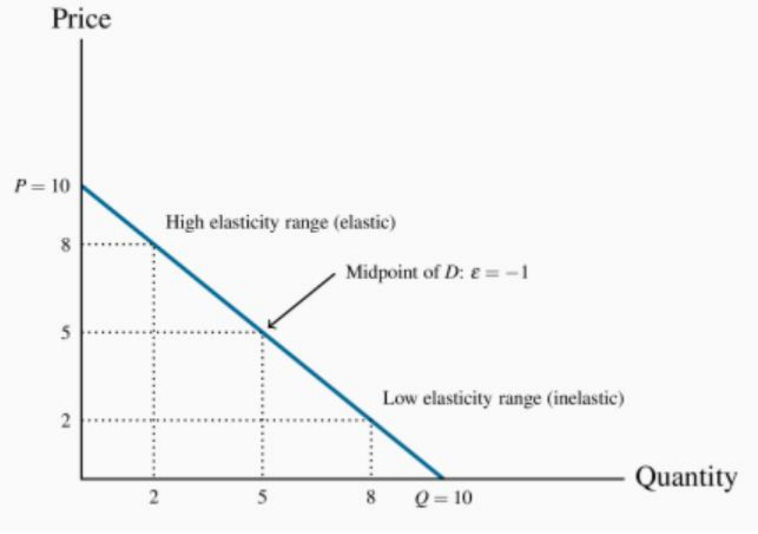
Average price P= (P1+P2)/2

Table 4.1 The demand for natural gas: Elasticities and revenue

Price ($)	Quantity demanded	Elasticity value	Total revenue ($)
10	0		0
9	1	-9.0	9
8	2		16
7	3	-2.33	21
6	4		24
5	5	-1.0	25
4	6		24
3	7	-0.43	21
2	8		16
1	9	-0.11	9
0	10		0

Elasticity calculations are based upon $2 price changes.

Figure 4.1 Elasticity variation with linear demand



In the high-price region of the demand curve the elasticity takes on a high value. At the midpoint of a linear demand curve the elasticity takes on a value of one, and at lower prices the elasticity value continues to fall.

Using this rule, consider now the value of εd when price drops from $10.00 to $8.00. The change in price is $2.00 and the average price is therefore $9.00 [=($10.00+$8.00)/2]. On the quantity side, demand goes from zero to 2 units (measured in thousands of cubic feet), and the average quantity demanded is therefore (0+2)/2=1. Putting these numbers into the formula yields:

$$\varepsilon d=\frac{(Q2-Q1)/Q}{(P2-P1)/P}=\frac{(2/1)}{-(2/9)}=-\left(\frac{2}{1}\right)\text{ x }\left(\frac{9}{2}\right)=-9.$$

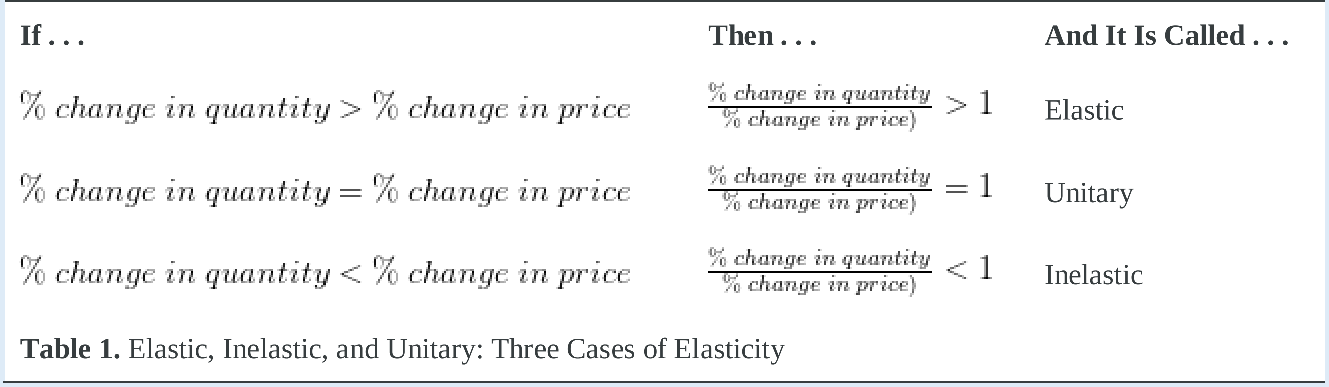
Note that the price has declined in this instance and thus the change in price is negative. Continuing down the table in this fashion yields the full set of elasticity values in the third column.

The demand elasticity is said to be *high* if it is a large negative number; the large number denotes a high degree of sensitivity. Conversely, the elasticity is *low* if it is a small negative number. High and low refer to the size of the number, ignoring the negative sign. The term is also used to define what we have just measured, indicating that it defines consumer *arc elasticity* responsiveness over a segment or arc of the demand curve.

It is helpful to analyze this numerical example by means of the corresponding demand curve that is plotted in Figure 4.1, and which we used in Chapter 3. It is a straight-line demand curve; but, despite this, the elasticity is not constant. At high prices the elasticity is high; at low prices it is low. The intuition behind this pattern is as follows. When the price is high, a given price change represents a small percentage change, because the average price in the price-term denominator is large. At high prices the quantity demanded is small and therefore the percentage quantity change tends to be large due to the small quantity value in its denominator. In sum, at high prices the elasticity is large. By the same reasoning, at low prices the elasticity is small.

We can carry this reasoning one step further to see what happens when the demand curve intersects the axes. At the horizontal axis the average price is tending towards zero. Since this extremely small value appears in the denominator of the price term it means that the price term as a whole is extremely large. Accordingly, with an extremely large value in the denominator of the elasticity expression, that whole expression is tending towards a zero value. By the same reasoning the elasticity value at the vertical intercept is tending towards an infinitely large value.

Source: Lynham, 2018, Section 5.1, CC-BY 4.0 (Original source, Table 1 only)



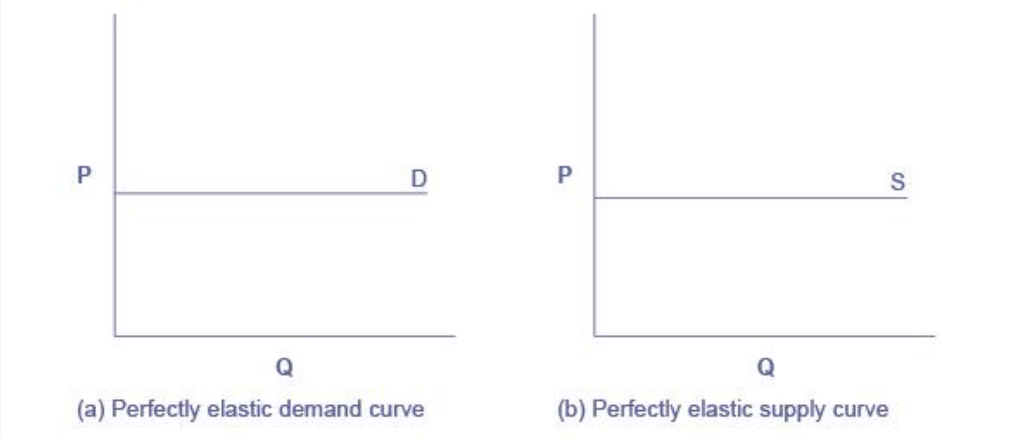
If . . .	Then . . .	And It Is Called . . .
$\%\ change\ in\ quantity > \%\ change\ in\ price$	$\frac{\%\ change\ in\ quantity}{\%\ change\ in\ price)} > 1$	Elastic
$\%\ change\ in\ quantity = \%\ change\ in\ price$	$\frac{\%\ change\ in\ quantity}{\%\ change\ in\ price)} = 1$	Unitary
$\%\ change\ in\ quantity < \%\ change\ in\ price$	$\frac{\%\ change\ in\ quantity}{\%\ change\ in\ price)} < 1$	Inelastic

Table 1. Elastic, Inelastic, and Unitary: Three Cases of Elasticity

Source: Lynham, 2018, Section 5.2, CC-BY 4.0 (Original source, paragraphs 5-6 removed, Figure 4 removed)

There are two extreme cases of elasticity: when elasticity equals zero and when it is infinite. A third case is that of constant unitary elasticity. We will describe each case.

Infinite elasticity or **perfect elasticity** refers to the extreme case where either the quantity demanded (Qd) or supplied (Qs) changes by an infinite amount in response to any change in price at all. In both cases, the supply and the **demand curve** are horizontal as shown in Figure 1. While perfectly elastic supply curves are unrealistic, goods with readily available inputs and whose production can be easily expanded will feature highly elastic supply curves. Examples include pizza, bread, books and pencils. Similarly, perfectly elastic demand is an extreme example. But luxury goods, goods that take a large share of individuals' income, and goods with many substitutes are likely to have highly elastic demand curves. Examples of such goods are Caribbean cruises and sports vehicles.

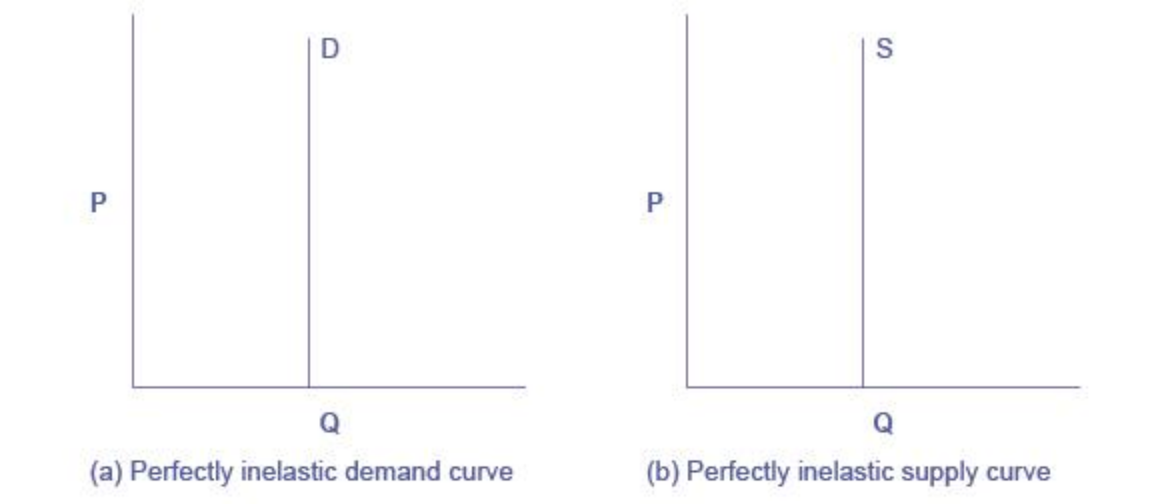


(a) Perfectly elastic demand curve

(b) Perfectly elastic supply curve

Figure 1. *Infinite Elasticity. The horizontal lines show that an infinite quantity will be demanded or supplied at a specific price. This illustrates the cases of a perfectly (or infinitely) elastic demand curve and supply curve. The quantity supplied or demanded is extremely responsive to price changes, moving from zero for prices close to P to infinite when price reach P.*

Zero elasticity or **perfect inelasticity**, as depicted in Figure 2 refers to the extreme case in which a percentage change in price, no matter how large, results in zero change in quantity. While a perfectly inelastic supply is an extreme example, goods with limited supply of inputs are likely to feature highly inelastic supply curves. Examples include diamond rings or housing in prime locations such as apartments facing Central Park in New York City. Similarly, while perfectly inelastic demand is an extreme case, necessities with no close substitutes are likely to have highly inelastic demand curves. This is the case of life-saving drugs and gasoline.



(a) Perfectly inelastic demand curve

(b) Perfectly inelastic supply curve

Figure 2. *Zero Elasticity. The vertical supply curve and vertical demand curve show that there will be zero percentage change in quantity (a) demanded or (b) supplied, regardless of the price.*

Constant unitary elasticity, in either a supply or demand curve, occurs when a price change of one percent results in a quantity change of one percent. Figure 3 shows a demand curve with constant unit elasticity. As we move down the demand curve from A to B, the price falls by 33% and quantity demanded rises by 33%; as you move from B to C, the price falls by 25% and the quantity demanded rises by 25%; as you move from C to D, the price falls by 16% and the quantity rises by 16%. Notice that in absolute value, the declines in price, as you step down the demand curve, are not identical. Instead, the price falls by $3 from A to B, by a smaller amount of $1.50 from B to C, and by a still smaller amount of $0.75 from C to D. As a result, a demand curve with constant unitary elasticity moves from a steeper slope on the left and a flatter slope on the right—and a curved shape overall.

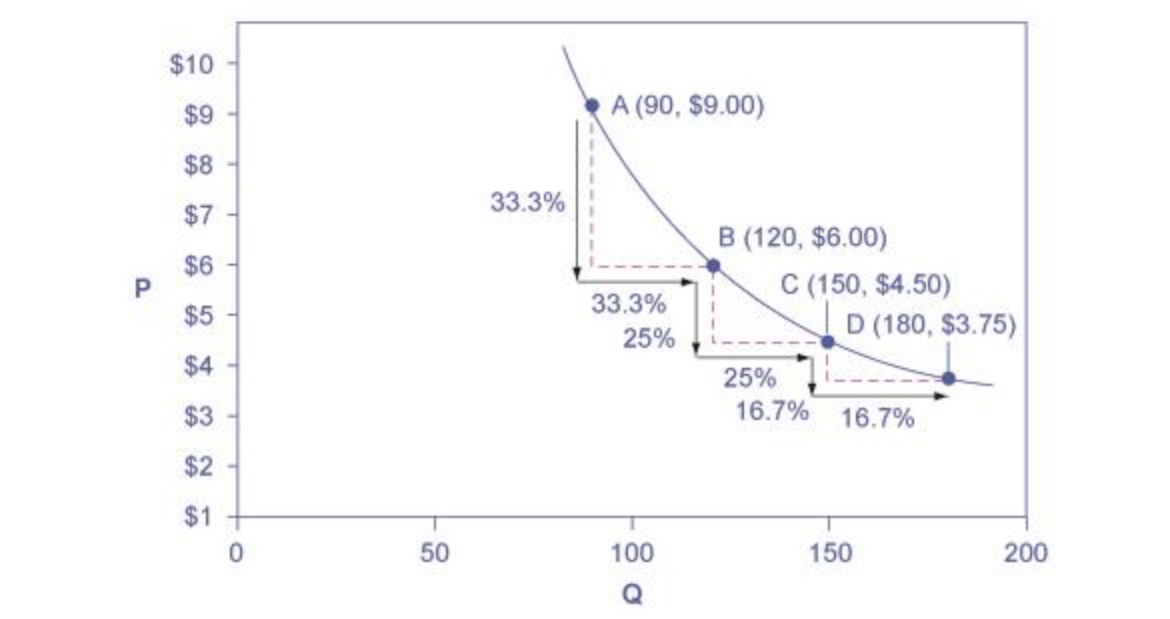


Figure 3. *A Constant Unitary Elasticity Demand Curve. A demand curve with constant unitary elasticity will be a curved line. Notice how price and quantity demanded change by an identical amount in each step down the demand curve.*

Determinants of Price Elasticity

Source: Curtis & Irvine, 2016, Section 4.1, CC-BY-NC-SA 3.0 (Original source, page 7 only, last line removed, numbered list added)

Why is it that the price elasticities for some goods and services are high and for others low?

1. Tastes One answer lies in tastes: If a good or service is a basic necessity in one's life, then price variations have a minimal effect on the quantity demanded, and these products thus have a relatively inelastic demand.

2. Availability of Substitutes A second answer lies in the ease with which we can substitute alternative goods or services for the product in question. If Apple Corporation

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

had no serious competition in the smart-phone market, it could price its products higher than in the presence of Samsung and Google, who also supply smart phones. A supplier who increases her price will lose more sales if there are ready substitutes to which buyers can switch, than if no such substitutes exist. It follows that a critical role for the marketing department in a firm is to convince buyers of the uniqueness of the firm's product.

3. Product Groups Where product groups are concerned, the price elasticity of demand for one product is necessarily higher than for the group as a whole: Suppose the price of one computer tablet brand alone falls. Buyers would be expected to substitute towards this product in large numbers – its manufacturer would find demand to be highly responsive. But if all brands are reduced in price, the increase in demand for any one will be more muted. In essence, the one tablet whose price falls has several close substitutes, but tablets in the aggregate do not.

Source: Curtis & Irvine, 2016, Section 4.3, CC-BY-NC-SA 3.0 (Original source, paragraphs 2 & 3 removed)

4. Time Dimension The price elasticity of demand is frequently lower in the short run than in the long run. For example, a rise in the price of home heating oil may ultimately induce consumers to switch to natural gas or electricity, but such a transition may require a considerable amount of time. Time is required for decision-making and investment in new heating equipment. A further example is the elasticity of demand for tobacco. Some adults who smoke may be seriously dependent and find quitting almost impossible. But if young smokers, who are not yet addicted, decide not to start on account of the higher price, then over a long period of time the percentage of the population that smokes will decline. The full impact may take decades! Accordingly when we talk of the short run and the long run, there is no simple rule for defining how long the long run actually is in terms of months or years. In some cases, adjustment may be complete in weeks, in other cases years.

Price Elasticity and Expenditure

Source: Curtis & Irvine, 2017, Section 4.2, CC-BY-NC-SA 3.0 (Original source, page 100, last 2 paragraphs removed; Table 4.1 added from page 91 added)

In Figure 4.5, we examine the expenditure or revenue impact of a price reduction in two ranges of a linear demand curve. Expenditure, or revenue, is the product of price times quantity. It is, therefore, the area of a rectangle in a price/quantity diagram. From position A, a price reduction from PA to PB has two impacts. It reduces the revenue that accrues from those QA units already being sold; the negative sign between PA and PB marks this reduction. But it increases revenue through additional sales from QA to QB. The area marked with a positive sign between QA and QB denotes this increase. Will the extra revenue caused by the quantity increase outweigh the loss in revenue associated with each unit sold before the price was reduced? It turns out that at high prices the positive impact outweighs the negative impact. The intuitive reason is that the existing sales are small and, therefore, we lose a revenue margin on a very limited quantity. The net impact on total expenditure of the price reduction is positive.

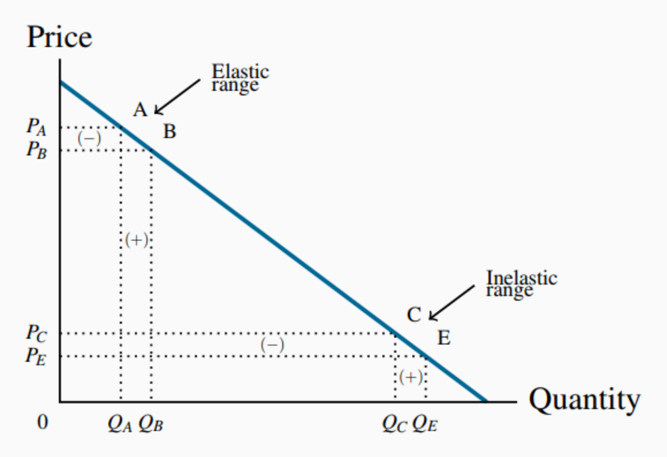


Figure 4.5: Price elasticity and revenue

When the price falls from P_A to P_B, expenditure changes from $P_A A Q_A 0$ to $P_B B Q_B 0$. In this elastic region expenditure increases, because the loss in revenue on existing units $(-)$ is less than the revenue gain $(+)$ due to the additional units sold. The opposite occurs in the inelastic region CE.

In contrast, move now to point C and consider a further price reduction from PC to PE. There is again a dual impact: a loss of revenue on existing sales, and a gain due to additional sales. But in this instance the existing sales QC are large, and therefore the loss of a price margin on these sales is more significant than the extra revenue that is generated by the additional sales. The net effect is that total expenditure falls.

So if revenue increases in response to price declines at high prices, and falls at low prices, there must be an intermediate region of the demand curve where the composite effects of the price change just offset each other, and no change in revenue results from a price change. It transpires that this occurs at the midpoint of the linear demand curve. Let us confirm this with the help of our example in Table 4.1.

Price ($)	Quantity demanded (thousands of cu ft.)	Price elasticity (arc)	Price elasticity (point)	Total revenue ($)
10.00	0	-9.0	$-\infty$	
8.00	2	-2.33	-4	16
6.00	4	-1.22	-1.5	24
5.00	5	-0.82	-1	25
4.00	6	-0.43	-0.67	24
2.00	8	-0.11	-0.25	16
0.00	10		0	0

Table 4.1: The demand for natural gas: elasticities and revenue

The fourth column of the table contains the point elasticities of demand, and the final column defines the expenditure on the good at the corresponding prices. Point elasticities are very precise; they are measured at a point rather than over a range or an arc. Note next that the point on this linear demand curve where revenue is a maximum corresponds to its midpoint—where the elasticity is unity. This is no coincidence. Price reductions increase revenue so long as demand is elastic, but as soon as demand becomes inelastic such price declines reduce revenues. When does the value become inelastic? Clearly, where the unit elasticity value is crossed. This is illustrated in Figure 4.6, which defines the relationship between total revenue (T R), or total expenditure, and quantity sold in Table 4.1. Total revenue increases initially with quantity, and this increasing quantity of sales comes about as a result of lower prices. At a quantity of 5 units the price is $5.00. This price-quantity combination corresponds to the mid-point of the demand curve.

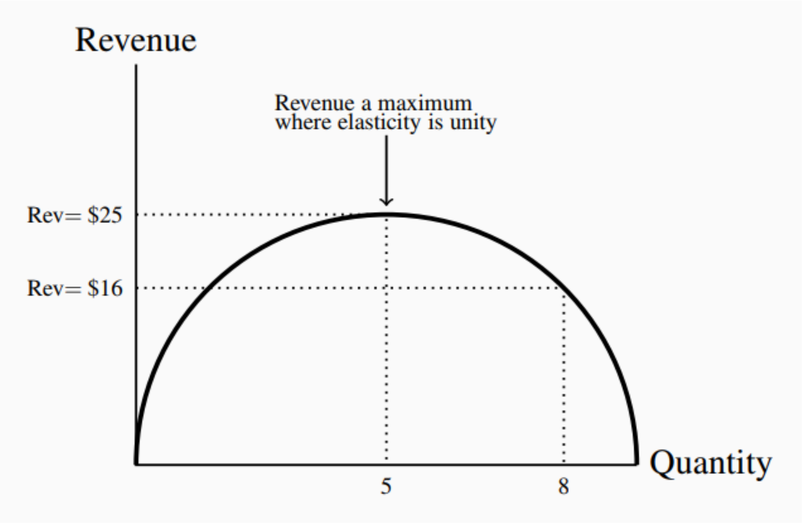


Figure 4.6: Total revenue and elasticity

Based upon the data in Table 4.1, revenue increases with quantity sold up to sales of 5 units. Beyond this output, the decline in price that must accompany additional sales causes revenue to decline.

We now have a general conclusion: In order to maximize the possible revenue from the sale of a good or service, it should be priced where the demand elasticity is unity.

Source: Lynham, 2018, Section 5.3, CC-BY 4.0 (Original source, Table 5 only)

If Demand Is . . .	**Then . . .**	**Therefore . . .**
Elastic	% change in Qd > % change in P	A given % rise in P will be more than offset by a larger % fall in Q so that total revenue (P × Q) falls.
Unitary	% change in Qd = % change in P	A given % rise in P will be exactly offset by an equal % fall in Q so that total revenue (P × Q) is unchanged.
Inelastic	% change in Qd < % change in P	A given % rise in P will cause a smaller % fall in Q so that total revenue (P × Q) rises.

Table 5. Will the Band Earn More Revenue by Changing Ticket Prices?

4.2 PRICE ELASTICITY OF SUPPLY

Source: Curtis & Irvine, 2016, Section 4.2, CC-BY-NC-SA 3.0 (Original source, figure 4.2 removed)

Now that we have developed the various dimensions of elasticity on the demand side, the analysis of elasticities on the supply side is straightforward. The **elasticity of supply** measures the responsiveness of the quantity supplied to a change in the price.

The **elasticity of supply** measures the responsiveness of quantity supplied to a change in the price.

$$\varepsilon_s = \frac{\textit{Percentage change in quantity supplied}}{\textit{Percentage change in price}} = \frac{\%\Delta Q}{\%\Delta P}$$

The subscript s denotes supply. This is exactly the same formula as for the demand curve, except that the quantities now come from a supply curve. Furthermore, and in contrast to the demand elasticity, the supply elasticity is generally a positive value because of the positive relationship between price and quantity supplied. The more elastic, or the more responsive, is supply to a given price change, the larger will be the elasticity value. In diagrammatic terms, this means that "flatter" supply curves have a greater elasticity than more "vertical" curves at a given price and quantity combination. Numerically the flatter curve has a larger value than the more vertical supply – try drawing a supply diagram similar to Figure 4.2. Technically, a completely vertical supply curve has a zero elasticity and a horizontal supply curve has an infinite elasticity – just as in the demand cases.

As always we keep in mind the danger of interpreting too much about the value of this elasticity from looking at the visual profiles of supply curves.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

4.3 ELASTICITIES AND TAX INCIDENCE

Source: Curtis & Irvine, 2016, Section 4.7, CC-BY-NC-SA 3.0 (Original source, page 5 removed, last line of page 4 removed)

Elasticity values are critical in determining the impact of a government's taxation policies. The spending and taxing activities of the government influence the use of the economy's resources. By taxing cigarettes, alcohol and fuel, the government can restrict their use; by taxing income, the government influences the amount of time people choose to work. Taxes have a major impact on almost every sector of the Canadian economy.

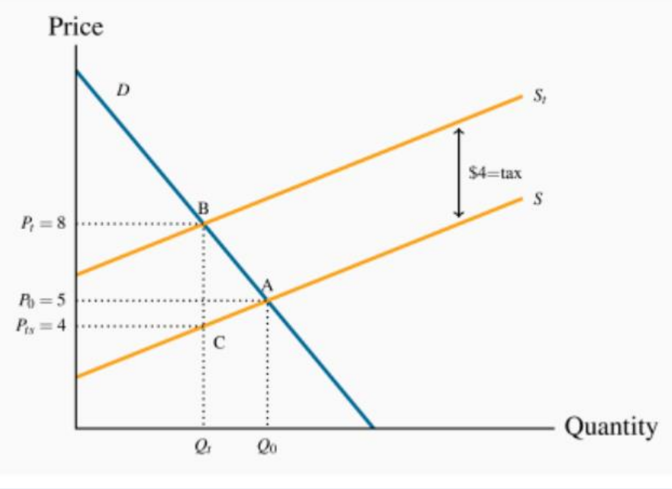
To illustrate the role played by demand and supply elasticities in tax analysis, we take the example of a sales tax. These can be of the *specific* or *ad valorem* type. A *specific* tax involves a fixed dollar levy per unit of a good sold (e.g., $10 per airport departure). An *ad valorem* tax is a percentage levy, such as Canada's Goods and Services tax (e.g., 5 percent on top of the retail price of goods and services). The impact of each type of tax is similar, and we will use the specific tax in our example below.

A layperson's view of a sales tax is that the tax is borne by the consumer. That is to say, if no sales tax were imposed on the good or service in question, the price paid by the consumer would be the same net of tax price as exists when the tax is in place. Interestingly, this is not always the case. The study of the **incidence of taxes** is the study of who really bears the tax burden, and this in turn depends upon supply and demand elasticities.

Tax Incidence describes how the burden of a tax is shared between buyer and seller.

Consider Figures 4.7 and 4.8, which define an imaginary market for inexpensive wine. Let us suppose that, without a tax, the equilibrium price of a bottle of wine is $5, and Q0 is the equilibrium quantity traded. The pre-tax equilibrium is at the point A. The government now imposes a specific tax of $4 per bottle. The impact of the tax is represented by an upward shift in supply of $4: Regardless of the price that the consumer pays, $4 of that price must be remitted to the government. As a consequence, the price paid to the supplier must be $4 less than the consumer price, and this is represented by twin supply curves: One defines the price at which the supplier is willing to supply, and the other is the tax-inclusive supply curve that the consumer faces.

Figure 4.7 Tax incidence with elastic supply



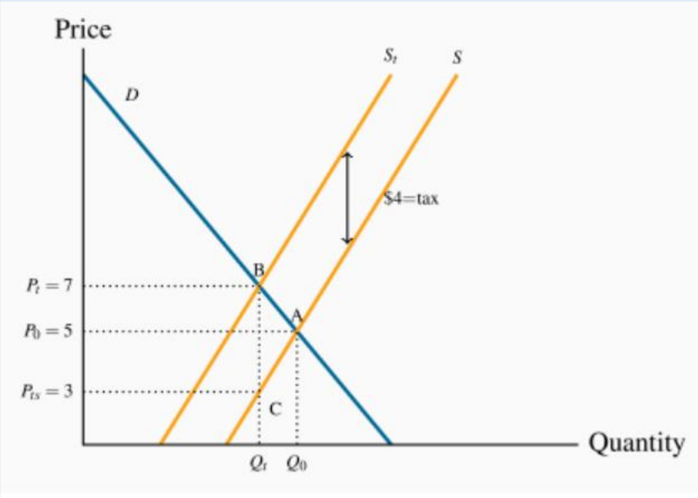
*The imposition of a specific tax of $4 shifts the supply curve vertically by $4. The final price at B (*Pt*) increases by $3 over the equilibrium price at A. At the new quantity traded,* Qt*, the supplier gets $4 per unit (*Pts*), the government gets $4 also and the consumer pays $8. The greater part of the incidence is upon the buyer, on account of the relatively elastic supply curve: His price increases by $3 of the $4 tax.*

The introduction of the tax in Figure 4.7 means that consumers now face the supply curve St. The new equilibrium is at point B. Note that the price has increased by less than the full amount of the tax—in this example it has increased by $3. This is because the reduced quantity at B is provided at a lower supply price: The supplier is willing to supply the quantity Qt at a price defined by C ($4), which is lower than A ($5).

So what is the incidence of the $4 tax? Since the market price has increased from $5 to $8, and the price obtained by the supplier has fallen by $1, we say that the incidence of the tax falls mainly on the consumer: The price to the consumer has risen by three dollars and the price received by the supplier has fallen by just one dollar.

Consider now Figure 4.8, where the supply curve is less elastic, and the demand curve is unchanged. Again the supply curve must shift upward with the imposition of the $4 specific tax. But here the price received by the supplier is lower than in Figure 4.7, and the price paid by the consumer does not rise as much – the incidence is different. The consumer faces a price increase that is one-quarter, rather than three-quarters, of the tax value. The supplier faces a lower supply price, and bears a higher share of the tax.

Figure 4.8 Tax incidence with inelastic supply



*The imposition of a specific tax of $4 shifts the supply curve vertically by $4. The final price at B (*Pt*) increases by $2 over the no-tax price at A. At the new quantity traded,* Qt*, the supplier gets $3 per unit (*Pts*), the government gets $4 also and the consumer pays $7. The incidence is shared equally by suppliers and demanders.*

We can conclude from this example that, for any given demand, *the more elastic is supply, the greater is the price increase* in response to a given tax. Furthermore, a *more elastic supply curve* means that the *incidence falls more on the consumer*; while a *less elastic supply* curve means the *incidence falls more on the supplier*. This conclusion can be verified by drawing a third version of Figure 4.7 and 4.8, in which the supply curve is horizontal – perfectly elastic. When the tax is imposed the price to the consumer increases by the full value of the tax, and the full incidence falls on the buyer. While this case corresponds to the layperson's intuition of the incidence of a tax, economists recognize it as a special case of the more general outcome, where the incidence falls on both the supply side and the demand side.

These are key results in the theory of taxation. It is equally the case that *the incidence of the tax depends upon the demand elasticity*. In Figure 4.7 and 4.8 we used the same demand curve. However, it is not difficult to see that, if we were to redo the exercise with a demand curve of a different elasticity, the incidence would not be identical. At the same time, the general result on supply elasticities still holds.

4.4 INCOME & CROSS ELASTICITY OF DEMAND

Income Elasticity of Demand

Source: Curtis & Irvine, 2016, Section 4.5, CC-BY-NC-SA 3.0

In Chapter 3 we stated that higher incomes tend to increase the quantity demanded at any price. To measure the responsiveness of demand to income changes, a unit-free measure exists: The income elasticity of demand. The **income elasticity of demand** is the percentage change in quantity demanded divided by a percentage change in income.

The **income elasticity of demand** is the percentage change in quantity demanded divided by a percentage change in income.

Let us use the Greek letter eta, η, to define the income elasticity of demand and I to denote income. Then,

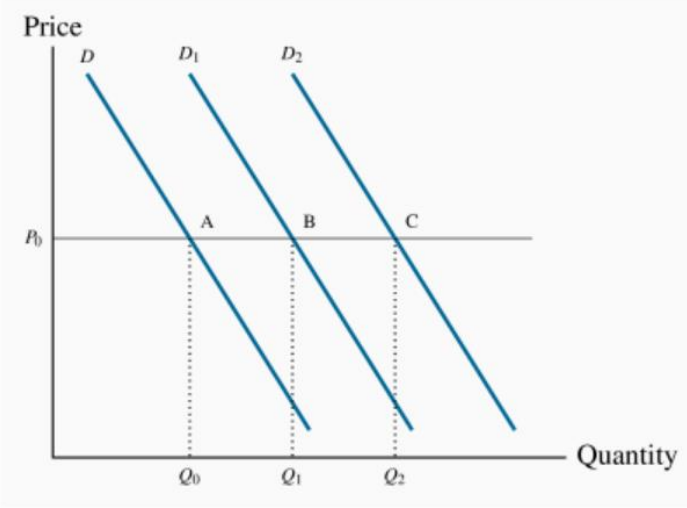
$$\eta d= \frac{Percentage\ change\ in\ quantity\ demanded}{Percentage\ change\ in\ income} = \frac{\%\Delta Q}{\%\Delta I}$$

As an example, if monthly income increases by 10 percent, and the quantity of magazines purchased increases by 15 percent, then the income elasticity of demand for magazines is 1.5 in value (=15%/10%). The income elasticity is generally positive, but not always – let us see why.

Normal, inferior, necessary, and luxury goods

The income elasticity of demand, in diagrammatic terms, is a percentage measure of how far the demand curve shifts in response to a change in income. Figure 4.6 shows two possible shifts. Suppose the demand curve is initially the one defined by D, and then income increases. In this example the supply curve is horizontal at the price P0. If the demand curve shifts to D1 as a result, the change in quantity demanded at the existing price is (Q1–Q0). However, if instead the demand curve shifts to D2, that shift denotes a larger change in quantity (Q2–Q0). Since the shift in demand denoted by D2 exceeds the shift to D1, the D2 shift is more responsive to income, and therefore implies a higher income elasticity.

Figure 4.6 Income elasticity and shifts in demand



At the price P0*, the income elasticity measures the percentage horizontal shift in demand caused by some percentage income increase. A shift from A to B reflects a lower income elasticity than a shift to C. A leftward shift in the demand curve in response to an income increase would denote a negative income elasticity – an inferior good.*

In this example, the good is a *normal good*, as defined in Chapter 3, because the demand for it increases in response to income increases. If the demand curve were to shift back to the left in response to an increase in income, then the income elasticity would be negative. In such cases the goods or services are *inferior*, as defined in Chapter 3.

Finally, we distinguish between luxuries and necessities. A **luxury** good or service is one whose income elasticity equals or exceeds unity. A **necessity** is one whose income elasticity is greater than zero but less than unity. If quantity demanded is so responsive to an income increase that the percentage increase in quantity demanded exceeds the percentage increase in income, then the elasticity value is in excess of 1, and the good or service is called a luxury. In contrast, if the percentage change in quantity demanded is less than the percentage increase in income, the value is less than unity, and we call the good or service a necessity.

A **luxury** good or service is one whose income elasticity equals or exceeds unity.

A **necessity** is one whose income elasticity is greater than zero and less than unity.

Luxuries and necessities can also be defined in terms of their share of a typical budget. An income elasticity greater than unity means that the share of an individual's budget being allocated to the product is increasing. In contrast, if the elasticity is less than unity, the budget share is falling. This makes intuitive sense—luxury cars are luxury goods by this definition because they take up a larger share of the incomes of the rich than the non-rich.

Inferior goods are those for which there exist higher-quality, more expensive, substitutes. For example, lower-income households tend to satisfy their travel needs by using public transit. As income rises, households normally reduce their reliance on public transit in favour of automobile use (despite the congestion and environmental impacts). Inferior goods, therefore, have a negative income elasticity: In the income elasticity equation definition, the numerator has a sign opposite to that of the denominator.

Inferior goods have negative income elasticity.

Empirical research indicates that goods like food and fuel have income elasticities less than 1; durable goods and services have elasticities slightly greater than 1; leisure goods and foreign holidays have elasticities very much greater than 1.

Income elasticities are useful in forecasting the demand for particular services and goods in a growing economy. Suppose real income is forecast to grow by 15% over the next five years. If we know that the income elasticity of demand for smart phones is 2.0, we could estimate the anticipated growth in demand by using the income elasticity formula: Since in this case η=2.0 and %ΔI=15 it follows that 2.0=%ΔQ/15%. Therefore the predicted demand change must be 30%.

Cross Elasticity of Demand

Source: Curtis & Irvine, 2016, Section 4.4, CC-BY-NC-SA 3.0

The price elasticity of demand tells us about consumer responses to price changes in different regions of the demand curve, holding constant all other influences. One of those influences is the price of other goods and services. A **cross-price elasticity** indicates how demand is influenced by changes in the prices of other products.

The **cross-price elasticity of demand** is the percentage change in the quantity demanded of a product divided by the percentage change in the price of another.

We write the cross price elasticity of the demand for x due to a change in the price of y as

$$\varepsilon d(x,y)=\frac{Percentage\ change\ in\ quantity\ demanded\ of\ x}{Percentage\ change\ in\ price\ of\ good\ y}=\frac{\%\Delta Qx}{\%\Delta Py}$$

For example, if the price of cable-supply internet services declines, by how much will the demand for satellite-supply services change? The cross-price elasticity may be positive or negative. These particular goods are clearly *substitutable*, and this is reflected in a *positive* value of this cross-price elasticity: The percentage change in satellite subscribers will be negative in response to a decline in the price of cable; a negative divided by a negative is positive. In contrast, a change in the price of tablets or electronic readers should induce an opposing change in the quantity of e-books purchased: Lower tablet prices will induce greater e-book purchases. In this case the price and quantity movements are in opposite directions and the elasticity is therefore negative – the goods are complements.

Application Box 4.1 Cross-price elasticity of demand between legal and illegal marijuana

In November 2016 Canada's Parliamentary Budget Office produced a research paper on the challenges associated with pricing legalized marijuana. They proposed that taxes should be low rather than high on this product, surprising many health advocates. Specifically they argued that the legal price of marijuana should be just fractionally higher than the price in the illegal market. Otherwise marijuana users would avail of the illegal market supply, which is widely available and of high quality. Effectively their research pointed to a very high cross-price elasticity of demand. This recommendation, if put into practice, means that tax revenue from marijuana sales will be small, but the size of the illegal market will decline substantially, thereby attaining a prime objective of legalization.

CHAPTER 5: Market Efficiency

5.1 Market Interventions

Source: Curtis & Irvine, 2016, Section 3.7, CC-BY-NC-SA 3.0 (Original source, 'Price ceilings' and 'Price floors' sections switched, pages 6 & 7 removed, paragraph of page 5 removed)

The freely functioning markets that we have developed certainly do not describe all markets. For example, minimum wages characterize the labour market, most agricultural markets have supply restrictions, apartments are subject to rent controls, and blood is not a freely traded market commodity in Canada. In short, price controls and quotas characterize many markets. **Price controls** are government rules or laws that inhibit the formation of market-determined prices. **Quotas** are physical restrictions on how much output can be brought to the market.

Price controls are government rules or laws that inhibit the formation of market-determined prices.

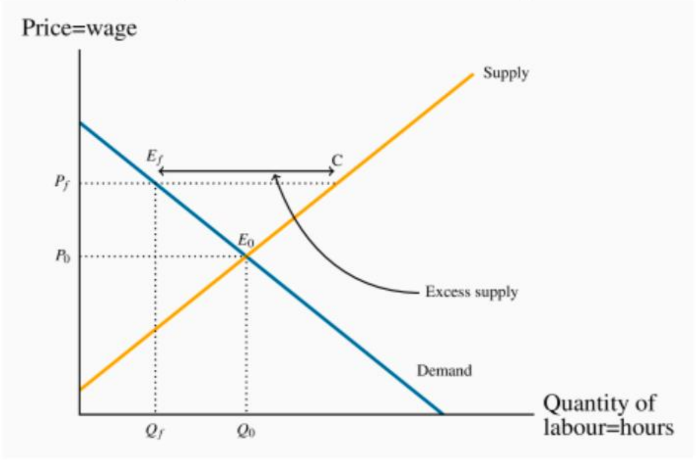
Quotas are physical restrictions on output.

Price controls come in the form of either *floors* or *ceilings*. Price floors are frequently accompanied by *marketing boards*.

Price Floors – minimum wages

An effective price floor sets the price above the market-clearing price. A minimum wage is the most widespread example in the Canadian economy. Provinces each set their own minimum, and it is seen as a way of protecting the well-being of low-skill workers. Such a floor is illustrated in Figure 3.7. The free-market equilibrium is again E0, but the effective market outcome is the combination of price and quantity corresponding to the point Ef at the price floor, Pf. In this instance, there is excess supply equal to the amount EfC.

Figure 3.7 Price floor – minimum wage



In a free market the equilibrium is E_0. A minimum wage of P_f raises the hourly wage, but reduces the hours demanded to Q_f. Thus E_fC is the excess supply.

If price floors, in the form of minimum wages, result in some workers going unemployed, why do governments choose to put them in place? The excess supply in this case corresponds to unemployment – more individuals are willing to work for the going wage than buyers (employers) wish to employ. The answer really depends upon the magnitude of the excess supply. In particular, suppose, in Figure 3.7 that the supply and demand curves going through the equilibrium E0 were more 'vertical'. This would result in a smaller excess supply than is represented with the existing supply and demand curves. This would mean in practice that a higher wage could go to workers, making them better off, without causing substantial unemployment. This is the tradeoff that governments face: With a view to increasing the purchasing power of generally lower-skill individuals, a minimum wage is set, hoping that the negative impact on employment will be small. We will return to this in the next chapter, where we examine the responsiveness of supply and demand curves to different prices.

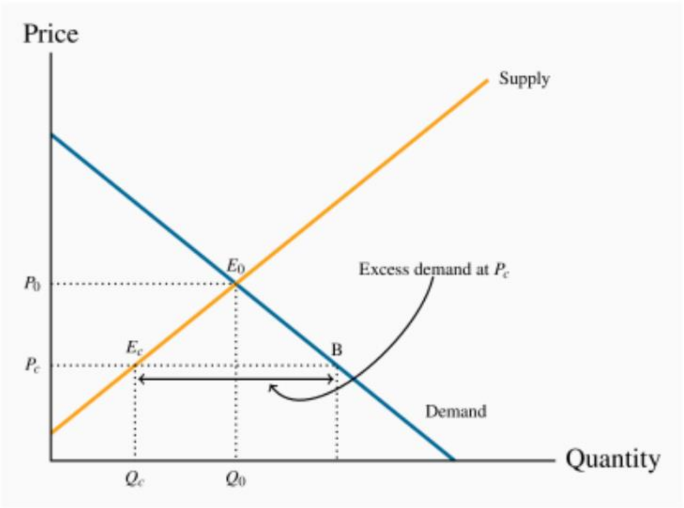
Price ceilings – rental boards

Ceilings mean that suppliers cannot legally charge more than a specific price. Limits on apartment rents are one form of ceiling. In times of emergency – such as flooding or famine, price controls are frequently imposed on foodstuffs, in conjunction with rationing, to ensure that access is not determined by who has the most income. The problem with price ceilings, however, is that they leave demand unsatisfied, and therefore they must be accompanied by some other allocation mechanism.

Consider an environment where, for some reason – perhaps a sudden and unanticipated growth in population – rents increase. Let the resulting equilibrium be defined by the point E0 in Figure 3.6. If the government were to decide that this is an unfair price because it places hardships on low- and middle-income households, it might impose a price limit, or ceiling, of Pc. The problem with such a limit is that excess demand results: Individuals want to rent more apartments than are available in the city. In a free market the price would adjust upward to eliminate the excess demand, but in this controlled environment it cannot. So some other way of allocating the available supply between demanders must evolve.

In reality, most apartments are allocated to those households already occupying them. But what happens when such a resident household decides to purchase a home or move to another city? In a free market, the landlord could increase the rent in accordance with market pressures. But in a controlled market a city's rental tribunal may restrict the annual rent increase to just a couple of percent and the demand may continue to outstrip supply. So how does the stock of apartments get allocated between the potential renters? One allocation method is well known: The existing tenant informs her friends of her plan to move, and the friends are the first to apply to the landlord to occupy the apartment. But that still leaves much unmet demand. If this is a student rental market, students whose parents live nearby may simply return 'home'. Others may choose to move to a part of the city where rents are more affordable.

Figure 3.6 The effect of a price ceiling



The free market equilibrium occurs at E_0. A price ceiling at P_c holds down the price but leads to excess demand E_cB, because Q_c is the quantity traded. A price ceiling above P_0 is irrelevant since the free market equilibrium E_0 can still be attained.

However, rent controls sometimes yield undesirable outcomes. Rent controls are widely studied in economics, and the consequences are well understood: Landlords tend not to repair or maintain their rental units in good condition if they cannot obtain the rent they believe they are entitled to. Accordingly, the residential rental stock deteriorates. In addition, builders realize that more money is to be made in building condominium units than rental units, or in converting rental units to condominiums. The frequent consequence is thus a reduction in supply and a reduced quality. Market forces are hard to circumvent because, as we emphasized in Chapter 1, economic players react to the incentives they face. These outcomes are examples of what we call the law of unintended consequences.

Source: Lynham, 2018, Section 3.4, CC-BY 4.0 (Original source, paragraph 8 only)

Price ceilings have been proposed for other products. For example, price ceilings to limit what producers can charge have been proposed in recent years for prescription drugs, doctor and hospital fees, the charges made by some automatic teller bank machines, and auto insurance rates. Price ceilings are enacted in an attempt to keep prices low for those who demand the product. But when the market price is not allowed to rise to the equilibrium level, quantity demanded exceeds quantity supplied, and thus a shortage occurs. Those who manage to purchase the product at the lower price given by the price ceiling will benefit, but sellers of the product will suffer, along with those who are not able to purchase the product at all. Quality is also likely to deteriorate.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Source: “Boundless economics”, Lumen Learning, n.d. ‘Government Intervention and Disequilibrium’, CC-BY-SA 4.0 International (Original source, paragraph 14 ‘Black Market’ only)

Black Market

Prolonged shortages caused by price ceilings can create black markets for that good. A black market is an underground network of producers that will sell consumers as much of a controlled good as they want, but at a price higher than the price ceiling. Black markets are generally illegal. However these markets provide higher profits for producers and more of a good for a consumers, so many are willing to take the risk of fines or imprisonment.

5.2 Consumer and Producer Surplus

Source: Curtis & Irvine, 2016, Section 5.2, CC-BY-NC-SA 3.0 (Original source, page 3 removed)

An understanding of economic efficiency is greatly facilitated as a result of understanding two related measures: Consumer surplus and producer surplus. Consumer surplus relates to the demand side of the market, producer surplus to the supply side. Producer surplus is also termed supplier surplus. These measures can be understood with the help of a standard example, the market for city apartments.

The market for apartments

Table 5.1 and Figure 5.1 describe the hypothetical data. We imagine first a series of city-based students who are in the market for a standardized downtown apartment. These individuals are not identical; they value the apartment differently. For example, Alex enjoys comfort and therefore places a higher value on a unit than Brian. Brian, in turn, values it more highly than Cathy or Don. Evan and Frank would prefer to spend their money on entertainment, and so on. These valuations are represented in the middle column of the demand panel in Table 5.1, and also in Figure 5.1 with the highest valuations closest to the origin. The valuations reflect the willingness to pay of each consumer.

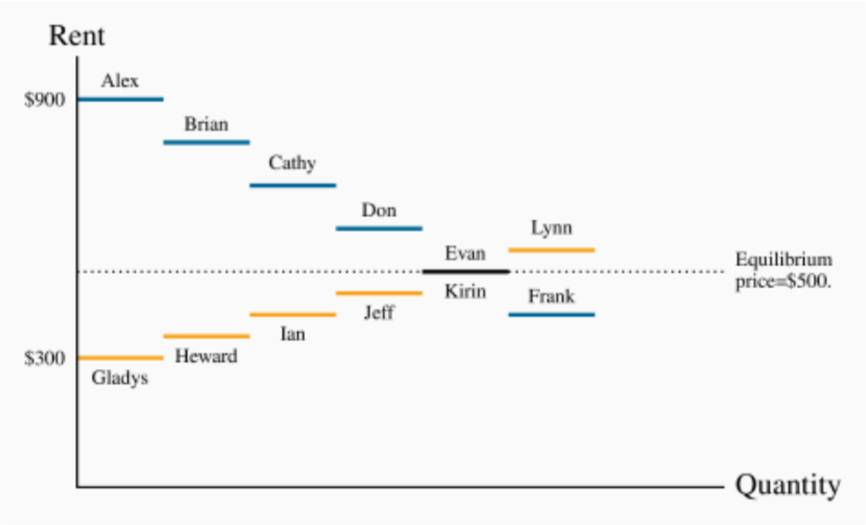
Table 5.1 Consumer and supplier surpluses

Demand			Supply		
Individual	**Demand valuation**	**Surplus**	**Individual**	**Reservation value**	**Surplus**
Alex	900	400	Gladys	300	200
Brian	800	300	Heward	350	150
Cathy	700	200	Ian	400	100
Don	600	100	Jeff	450	50
Evan	500	0	Kirin	500	0
Frank	400	0	Lynn	550	0

On the supply side we imagine the market as being made up of different individuals or owners, who are willing to put their apartments on the market for different prices. Gladys will accept less rent than Heward, who in turn will accept less than Ian. The minimum prices that the suppliers are willing to accept are called reservation prices or values, and these are given in the lower part of Table 5.1. Unless the market price is greater than their reservation price, suppliers will hold back.

By definition, as stated in Chapter 3, the demand curve is made up of the valuations placed on the good by the various demanders. Likewise, the reservation values of the suppliers form the supply curve. If Alex is willing to pay $900, then that is his demand price; if Heward is willing to put his apartment on the market for $350, he is by definition willing to supply it for that price. Figure 5.1 therefore describes the demand and supply curves in this market. The steps reflect the willingness to pay of the buyers and the reservation valuations or prices of the suppliers.

Figure 5.1 The apartment market



Demanders and suppliers are ranked in order of the value they place on an apartment. The market equilibrium is where the marginal demand value of Evan equals the marginal supply value of Kirin at $500. Five apartments are rented in equilibrium.

Demanders and suppliers are ranked in order of the value they place on an apartment. The market equilibrium is where the marginal demand value of Evan equals the marginal supply value of Kirin at $500. Five apartments are rented in equilibrium.

In this example, the equilibrium price for apartments will be $500. Let us see why. At that price the value placed on the marginal unit supplied by Kirin equals Evan's willingness to pay. Five apartments will be rented. A sixth apartment will not be rented because Lynne will let her apartment only if the price reaches $550. But the sixth potential demander is willing to pay only $400. Note that, as usual, there is just a single price in the market. Each renter pays $500, and therefore each supplier also receives $500.

The consumer and supplier surpluses can now be computed. Note that, while Don is willing to pay $600, he actually pays $500. His consumer surplus is therefore $100. In Figure 5.1, we can see that each **consumer's surplus** is the distance between the market price and the individual's valuation. These values are given in the final column of the top half of Table 5.1.

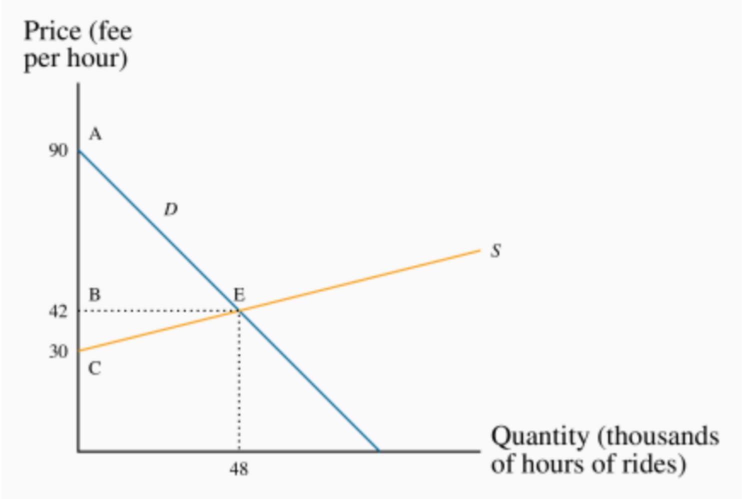
Consumer surplus is the excess of consumer willingness to pay over the market price.

Using the same reasoning, we can compute each **supplier's surplus**, which is the excess of the amount obtained for the rented apartment over the reservation price. For example, Heward obtains a surplus on the supply side of $150, while Jeff gets $50. Heward is willing to put his apartment on the market for $350, but gets the equilibrium price/rent of $500 for it. Hence his surplus is $150.

Supplier or producer surplus is the excess of market price over the reservation price of the supplier.

It should now be clear why these measures are called surpluses. *The suppliers and demanders are all willing to participate in this market because they earn this surplus.* It is a measure of their gain from being involved in the trading. The sum of each participant's surplus in the final column of Table 5.1 defines the total surplus in the market. Hence, on the demand side a total surplus arises of $1,000 and on the supply side a value of $500.

Figure 5.2 The taxi market



Consumer surplus is the area ABE, supplier surplus is the area BCE.

The demand curve represents the willingness to pay on the part of riders. The supply curve represents the willingness to supply on the part of drivers. The price per hour of rides defines the vertical axis; hours of rides (in thousands) are measured on the horizontal axis. The demand intercept of $90 says that the person who values the ride most highly is willing to pay \$90 per hour. The downward slope of the demand curve states that other buyers are willing to pay less. On the supply side no driver is willing to supply his time and vehicle unless he obtains at least \$30 per hour. To induce additional suppliers a higher price must be paid, and this is represented by the upward sloping supply curve.

The intersection occurs at a price of \$42 per hour and the equilibrium number of ride-hours supplied is 48 thousand1. Computing the surpluses is very straightforward. By definition the consumer surplus is the excess of the willingness to pay by each buyer above the uniform price. Buyers who value the ride most highly obtain the biggest surplus – the highest valuation rider gets a surplus of \$48 per hour, the difference between his willingness to pay of \$90 and the actual price of \$42. Each successive rider gets a slightly lower surplus until the final rider, who obtains zero. She pays \$42 and values the ride hours at \$42 also. On the supply side, the drivers who are willing to supply rides at the lowest reservation price (\$30 and above) obtain the biggest surplus. The 'marginal' supplier gets no surplus, because the price equals her reservation price.

From this discussion it follows that the consumer surplus is given by the area ABE and the supplier surplus by the area CBE. These are two triangular areas, and measured as half of the base by the perpendicular height. Therefore, in thousands of units:

Consumer Surplus = (demand value - price) = area ABE
=(1/2) × 48 × \$48 = \$1,152

Producer Surplus = (price - reservation supply value) = area BEC
=(1/2) × 48 × \$12 = \$288

The total surplus that arises in the market is the sum of producer and consumer surpluses, and since the units are in thousands of hours the total surplus here is
(\$1,152 + \$288) × 1,000=\$1,440,000.

1. The demand and supply functions behind these curves are P=90−1Q and P=30+(1/4)Q. Equating supply and demand yields the solutions in the text.

Efficiency

Source: Lynham, 2018, Section 3.5, CC-BY 4.0 (Original source, Figure 1, paragraphs 1, 2, 5 only)

The familiar **demand and supply diagram** holds within it the concept of economic efficiency. One typical way that economists define **efficiency** is when it is impossible to improve the situation of one party without imposing a cost on another. Conversely, if a situation is inefficient, it becomes possible to benefit at least one party without imposing costs on others.

Efficiency in the demand and supply model has the same basic meaning: The economy is getting as much benefit as possible from its scarce resources and all the possible gains from trade have been achieved. In other words, the optimal amount of each good and service is being produced and consumed.

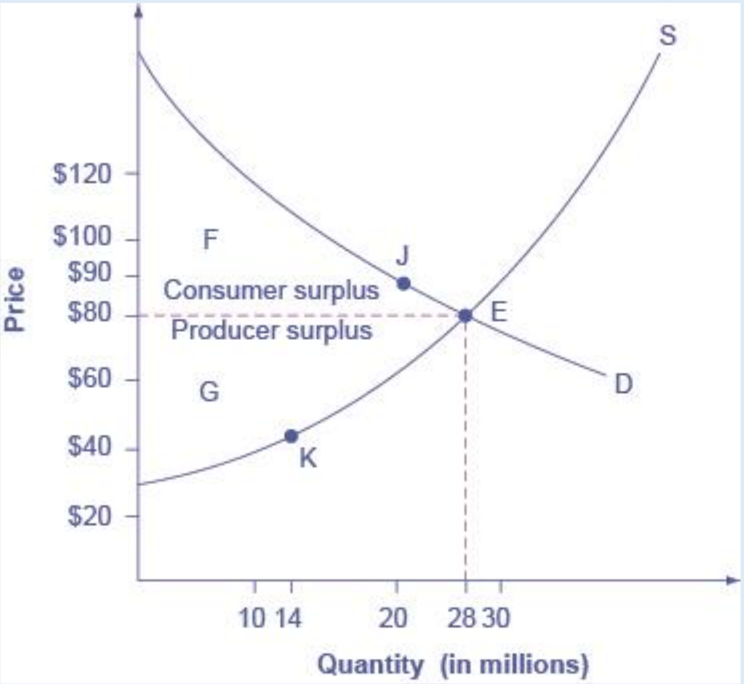


Figure 1. *Consumer and Producer Surplus. The somewhat triangular area labeled by F shows the area of consumer surplus, which shows that the equilibrium price in the market was less than what many of the consumers were willing to pay. Point J on the demand curve shows that, even at the price of $90, consumers would have been willing to purchase a quantity of 20 million. The somewhat triangular area labeled by G shows the area of producer surplus, which shows that the equilibrium price received in the market was more than what many of the producers were willing to accept for their products. For example, point K on the supply curve shows that at a price of $45, firms would have been willing to supply a quantity of 14 million.*

The sum of consumer surplus and producer surplus is **social surplus**, also referred to as **economic surplus** or **total surplus**. In Figure 1, social surplus would be shown as the area F + G. Social surplus is larger at equilibrium quantity and price than it would be at any other quantity. This demonstrates the economic efficiency of the market equilibrium. In addition, at the efficient level of output, it is impossible to produce greater consumer surplus without reducing producer surplus, and it is impossible to produce greater producer surplus without reducing consumer surplus.

Source: Curtis & Irvine, 2016, Section 5.2, CC-BY-NC-SA 3.0 (Original source, page 1 only)

The definition and measurement of the surplus is straightforward provided the supply and demand functions are known. An important characteristic of the marketplace is that

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

in certain circumstances it produces what we call an efficient outcome, or an **efficient market**. Such an outcome yields the highest possible sum of surpluses.

An **efficient market** maximizes the sum of producer and consumer surpluses.

To see that this outcome achieves the goal of maximizing the total surplus, consider what would happen if the quantity Q=48 in the taxi example were not supplied. Suppose that the city's taxi czar decreed that 50 units should be supplied, and the czar forced additional drivers on the road. If 2 additional units are to be traded in the market, consider the value of this at the margin. Suppliers value the supply more highly than the buyers are willing to pay. So on these additional 2 units negative surplus would accrue, thus reducing the total.

Second, potential buyers who would like a cheaper ride and drivers who would like a higher hourly payment do not get to participate in the market. On the demand side those individuals can take public transit, and on the supply side the those drivers can allocate their time to alternative activities. Obviously, only those who participate in the market benefit from a surplus.

Inefficiency of Price Floors and Price Ceilings

Source: Lynham, 2018, Section 3.5, CC-BY 4.0 (Original source, paragraphs 6-11 only)

The imposition of a price floor or a price ceiling will prevent a market from adjusting to its equilibrium price and quantity, and thus will create an inefficient outcome. But there is an additional twist here. Along with creating inefficiency, price floors and ceilings will also transfer some consumer surplus to producers, or some producer surplus to consumers.

Imagine that several firms develop a promising but expensive new drug for treating back pain. If this therapy is left to the market, the equilibrium price will be $600 per month and 20,000 people will use the drug, as shown in Figure 2 (a). The original level of consumer surplus is T + U and producer surplus is V + W + X. However, the government decides to impose a price ceiling of $400 to make the drug more affordable. At this price ceiling, firms in the market now produce only 15,000.

As a result, two changes occur. First, an inefficient outcome occurs and the total surplus of society is reduced. The loss in social surplus that occurs when the economy produces at an inefficient quantity is called **deadweight loss**. In a very real sense, it is like money thrown away that benefits no one. In Figure 2(a), the deadweight loss is the area U + W. When deadweight loss exists, it is possible for both consumer and producer surplus to be higher, in this case because the **price control** is blocking some suppliers and demanders from transactions they would both be willing to make.
A second change from the **price ceiling** is that some of the producer surplus is transferred to consumers. After the price ceiling is imposed, the new consumer surplus is T + V, while the new producer surplus is X. In other words, the price ceiling transfers the

area of surplus (V) from producers to consumers. Note that the gain to consumers is less than the loss to producers, which is just another way of seeing the deadweight loss.

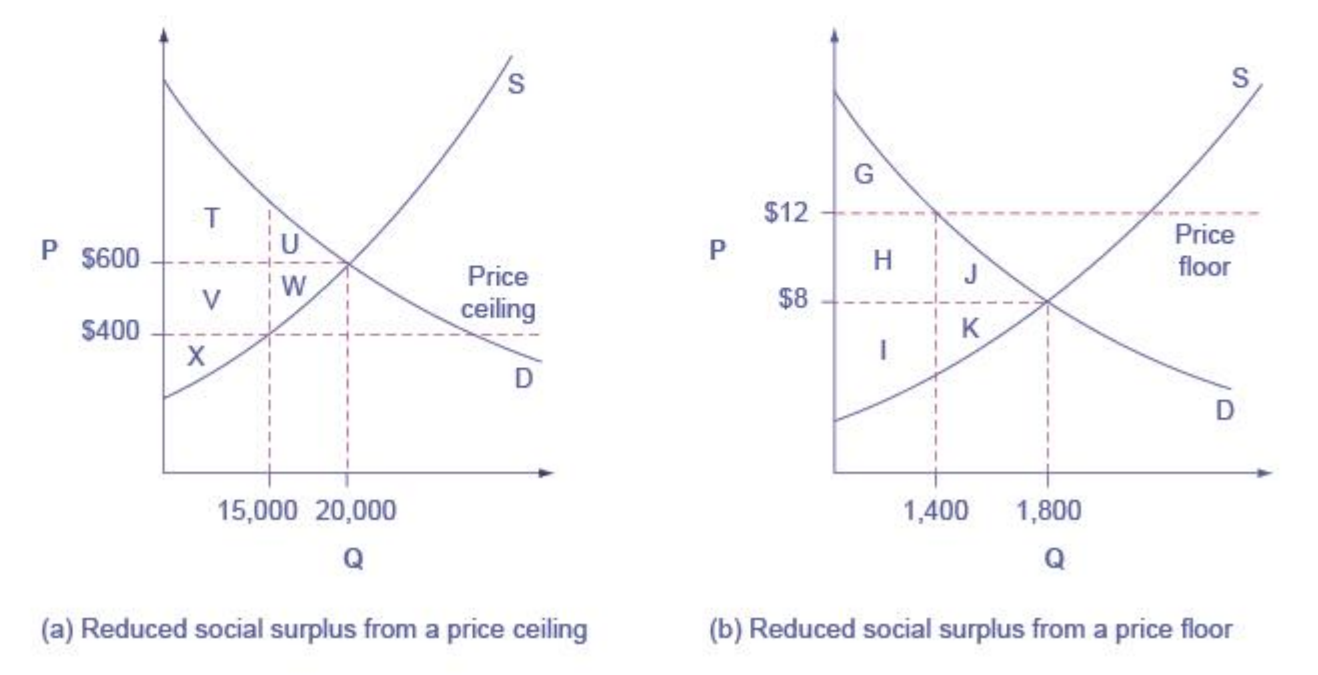


Figure 2. *(a) Efficiency and Price Floors and Ceilings. The original equilibrium price is $600 with a quantity of 20,000. Consumer surplus is T + U, and producer surplus is V + W + X. A price ceiling is imposed at $400, so firms in the market now produce only a quantity of 15,000. As a result, the new consumer surplus is T + V, while the new producer surplus is X. (b) The original equilibrium is $8 at a quantity of 1,800. Consumer surplus is G + H + J, and producer surplus is I + K. A price floor is imposed at $12, which means that quantity demanded falls to 1,400. As a result, the new consumer surplus is G, and the new producer surplus is H + I.*

Figure 2 (b) shows a price floor example using a string of struggling movie theaters, all in the same city. The current equilibrium is $8 per movie ticket, with 1,800 people attending movies. The original consumer surplus is G + H + J, and producer surplus is I + K. The city government is worried that movie theaters will go out of business, reducing the entertainment options available to citizens, so it decides to impose a price floor of $12 per ticket. As a result, the quantity demanded of movie tickets falls to 1,400. The new consumer surplus is G, and the new producer surplus is H + I. In effect, the **price floor** causes the area H to be transferred from consumer to producer surplus, but also causes a deadweight loss of J + K.

This analysis shows that a price ceiling, like a law establishing rent controls, will transfer some producer surplus to consumers—which helps to explain why consumers often favor them. Conversely, a price floor like a guarantee that farmers will receive a certain price for their crops will transfer some consumer surplus to producers, which explains why producers often favor them. However, both price floors and price ceilings block some transactions that buyers and sellers would have been willing to make, and creates deadweight loss. Removing such barriers, so that prices and quantities can adjust to their equilibrium level, will increase the economy's social surplus.

Quotas – agricultural supply

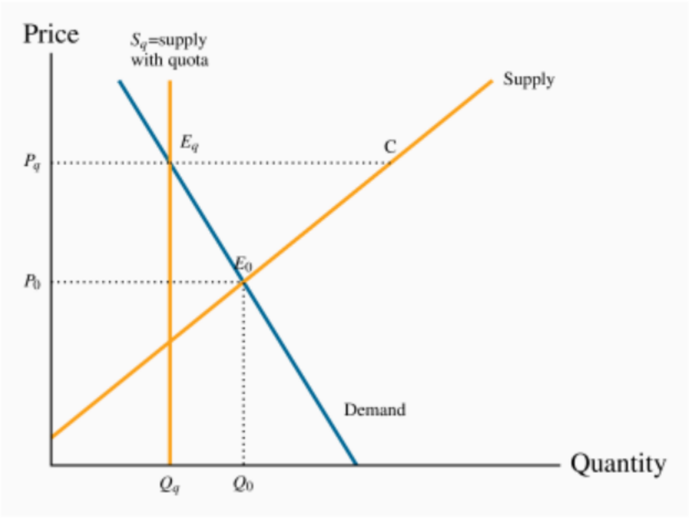
Source: Curtis & Irvine, 2016, Section 3.7, CC-BY-NC-SA 3.0 (Original source, page 6 only)

A quota represents the right to supply a specified quantity of a good to the market. It is a means of keeping prices higher than the free-market equilibrium price. As an alternative to imposing a price floor, the government can generate a high price by restricting supply.

Agricultural markets abound with examples. In these markets, farmers can supply only what they are permitted by the quota they hold, and there is usually a market for these quotas. For example, in several Canadian provinces it currently costs in the region of $30,000 to purchase a quota granting the right to sell the milk of one cow. The cost of purchasing quotas can thus easily outstrip the cost of a farm and herd. Canadian cheese importers must pay for the right to import cheese from abroad. Restrictions also apply to poultry. The impact of all of these restrictions is to raise the domestic price above the free market price.

In Figure 3.8, the free-market equilibrium is at E0. In order to raise the price above P0, the government restricts supply to Qq by granting quotas, which permit producers to supply a limited amount of the good in question. This supply is purchased at the price equal to Pq. From the standpoint of farmers, a higher price might be beneficial, even if they get to supply a smaller quantity, provided the amount of revenue they get as a result is as great as the revenue in the free market.

Figure 3.8 The effect of a quota



The government decides that the equilibrium price P_0 is too low. It decides to boost price by reducing supply from Q_0 to Q_q. It achieves this by requiring producers to have a production quota. This is equivalent to fixing supply at S_q .

CHAPTER 6: The Consumer Side of Markets

6.1 Consumer Choice

Source: Curtis & Irvine, 2016, Section 6.2, CC-BY-NC-SA 3.0 (Original source, pages 1-4 only, paragraphs 1, 4, 5 on page 3 removed, paragraph 1 of page 4 removed)

Neal loves to pump his way through the high-altitude powder at the Whistler ski and snowboard resort. His student-rate lift-ticket cost is $30 per visit. He also loves to frequent the jazz bars in downtown Vancouver, and each such visit costs him $20. With expensive passions, Neal must allocate his monthly entertainment budget carefully. He has evaluated how much satisfaction, measured in utils, he obtains from each snowboard outing and each jazz club visit. We assume that these utils are measurable, and use the term **cardinal utility** to denote this. These measurable utility values are listed in columns 2 and 3 of Table 6.1. They define the **total utility** he gets from various amounts of the two activities.

Table 6.1 Utils from snowboarding and jazz

1	2	3	4	5	6	7
Visit #	**Total snowboard utils**	**Total jazz utils**	**Marginal snowboard utils**	**Marginal jazz utils**	**Marginal snowboard utils per $**	**Marginal jazz utils per $**
1	72	52	72	52	2.4	2.6
2	132	94	60	42	2.0	2.1
3	182	128	50	34	1.67	1.7
4	224	156	42	28	1.4	1.4
5	260	180	36	24	1.2	1.2
6	292	201	32	21	1.07	1.05
7	321	220	29	19	0.97	0.95

Price of snowboard visit=$30. Price of jazz club visit=$20.

Cardinal utility is a measurable concept of satisfaction.

Total utility is a measure of the total satisfaction derived from consuming a given amount of goods and services.

Neal's total utility from each activity in this example is independent of the amount of the other activity he engages in. These total utilities are plotted in Figures 6.1 and 6.2. Clearly, more of each activity yields more utility, so the additional or **marginal utility** (MU) of each activity is positive. This positive marginal utility for any amount of the good consumed, no matter how much, reflects the assumption of non-satiation—more is always better. Note, however, that the decreasing slopes of the total utility curves show that total utility is increasing at a diminishing rate. While more is certainly better, each additional visit to Whistler or a jazz club augments Neal's utility by a smaller amount. At the margin, his additional utility declines: He has **diminishing marginal utility**. The marginal utilities associated with snowboarding and jazz are entered in columns 4 and 5 of Table 6.1. They are the differences in total utility values when consumption increases by one unit. For example, when Neal makes a sixth visit to Whistler his total utility increases from 260 utils to 292 utils. His marginal utility for the sixth unit is therefore 32 utils, as defined in column 4. In light of this example, it should be clear that we can define marginal utility as:

$$\text{Marginal Utility} = \frac{\text{additional utility}}{\text{additional consumption}} \quad \text{or,} \quad \text{MU}= \frac{\Delta \text{U}}{\Delta \text{C}}$$

where Δx denotes the change in the quantity consumed of the good or service in question.

Marginal utility is the addition to total utility created when one more unit of a good or service is consumed.

Diminishing marginal utility implies that the addition to total utility from each extra unit of a good or service consumed is declining.

Figure 6.1 TU from snowboarding

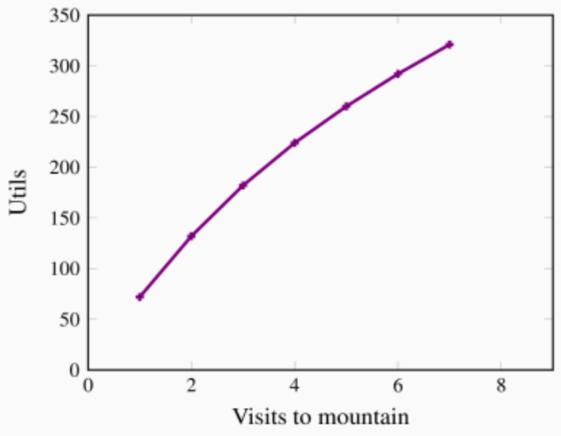


Figure 6.2 TU from jazz

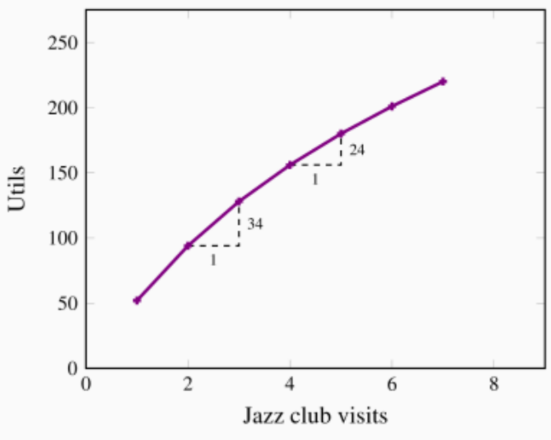


Figure 6.3 MU from snowboarding

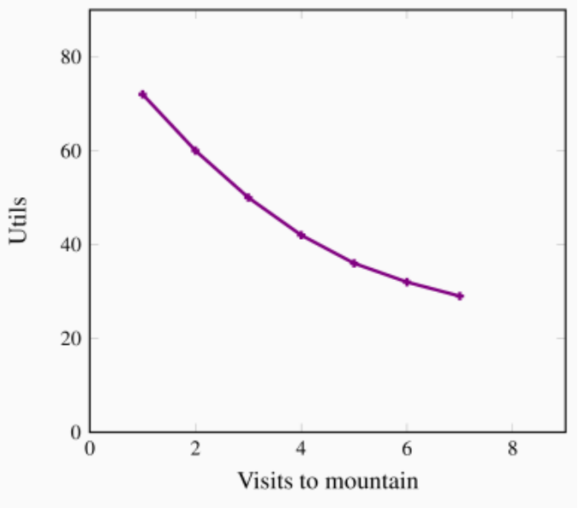
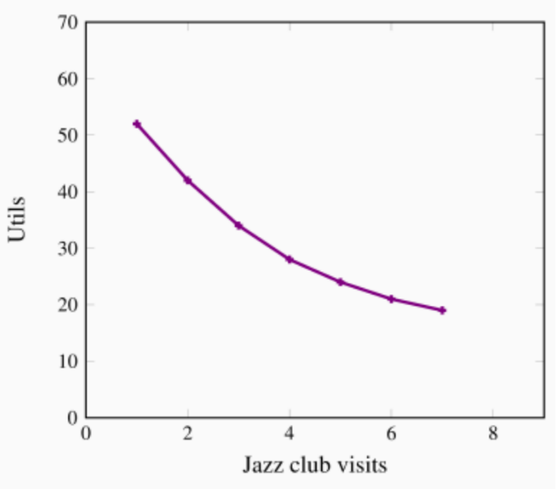


Figure 6.4 MU from jazz



Now that Neal has defined his utility schedules, he must consider the price of each activity. Ultimately, when deciding how to allocate his monthly entertainment budget, he must evaluate how much utility he gets from each dollar spent on snowboarding and jazz: What "bang for his buck" does he get? Let us see how he might go about allocating his budget. When he has fully spent his budget in the manner that will yield him greatest utility, we say that he has attained equilibrium, because he will have no incentive to change his expenditure patterns.

If he boards once, at a cost of $30, he gets 72 utils of satisfaction, which is 2.4 utils per dollar spent (=72/30). One visit to a jazz club would yield him 2.6 utils per dollar (=52/20). Initially, therefore, his dollars give him more utility per dollar when spent on jazz. His MU per dollar spent on each activity is given in the final two columns of the table. These values are obtained by dividing the MU associated with each additional unit by the good's price.

Only when the utility per dollar expended on each activity is equal at the margin will Neal be optimizing. When that condition holds, a reallocation would be of no benefit to him, because the gains from one more dollar on boarding would be exactly offset by the loss from one dollar less spent on jazz. Therefore, we can write the equilibrium condition as

$$\text{Equilibrium requires:}\sim\sim \frac{MU_s}{P_s} = \frac{MU_J}{P_J} \quad \text{or} \quad \frac{MU_s}{MU_J} = \frac{P_S}{P_J}$$

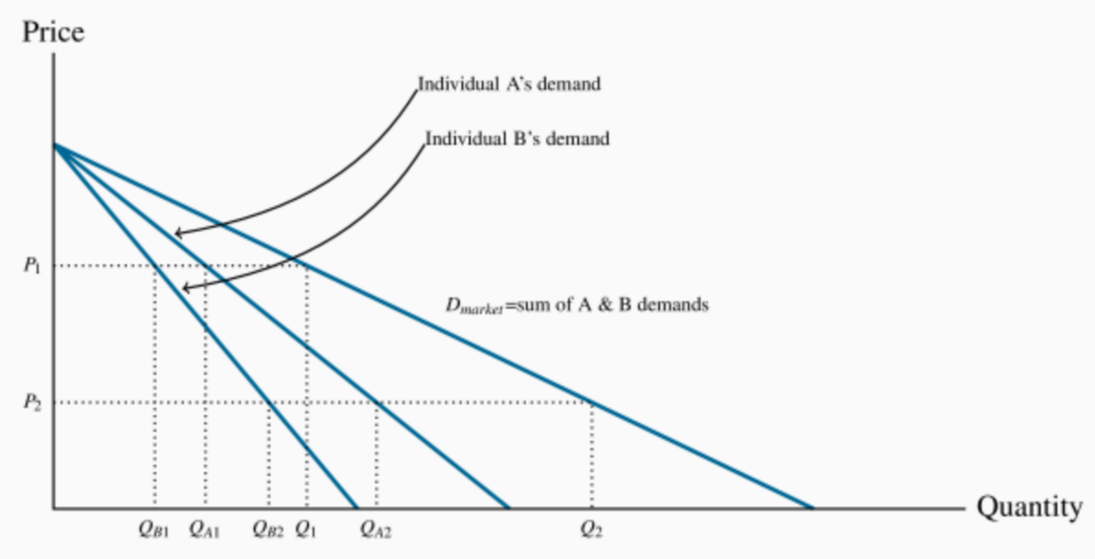
While this example has just two goods, in the more general case of many goods, this same condition must hold for all pairs of goods on which the consumer allocates his or her budget.

Individual and Market Demand Curves

Source: Curtis & Irvine, 2016, Section 3.8, CC-BY-NC-SA 3.0

Markets are made up of many individual participants on the demand and supply side. The supply and demand functions that we have worked with in this chapter are those for the total of all participants on each side of the market. But how do we arrive at such market functions when the economy is composed of individuals? We can illustrate how, with the help of Figure 3.9.

Figure 3.9 Summing individual demands



At P_1 *individual A purchases* Q_{A1} *and B purchases* Q_{B1}. *The total demand is the sum of these individual demands at this price (*Q_1*). At* P_2 *individual demands are summed to* Q_2. *Since the points* Q_1 *and* Q_2 *define the demands of the market participants it follows that market demand is the horizontal sum of these curves.*

To concentrate on the essentials, imagine that there are just two buyers of chocolate cookies in the economy. A has a stronger preference for cookies than B, so his demand is greater. To simplify, let the two demands have the same intercept on the vertical axis. The curves DA and DB indicate how many cookies A and B, respectively, will buy at each price. The market demand indicates how much they buy together at any price. Accordingly, at P1, A and B purchase the quantities QA1 and QB1 respectively. Thus Q1 = QA1 + QB1. At a price P2, they purchase QA2 and QB2. Thus Q2 = QA2 + QB2. The **market demand** is therefore the horizontal sum of the individual demands at these prices. In the figure this is defined by D_{market}.

Market demand: the horizontal sum of individual demands.

6.2 How Price Changes Affect Consumer Choices

Source: Lynham, 2018, Section 6.2, CC-BY 4.0 (Original source, paragraphs 7, 9-11 only, Figure 2 description removed)

For analyzing the possible effect of a change in price on consumption, let's again use a concrete example. Figure 2 represents the consumer choice of Sergei, who chooses between purchasing baseball bats and cameras. A price increase for baseball bats would have no effect on the ability to purchase cameras, but it would reduce the number of bats Sergei could afford to buy. Thus a price increase for baseball bats, the good on the horizontal axis, causes the budget constraint to rotate inward, as if on a hinge, from the vertical axis. As in the previous section, the point labeled M represents the originally preferred point on the original budget constraint, which Sergei has chosen after contemplating his total utility and marginal utility and the tradeoffs involved along the budget constraint. In this example, the units along the horizontal and vertical axes are not numbered, so the discussion must focus on whether more or less of certain goods will be consumed, not on numerical amounts.

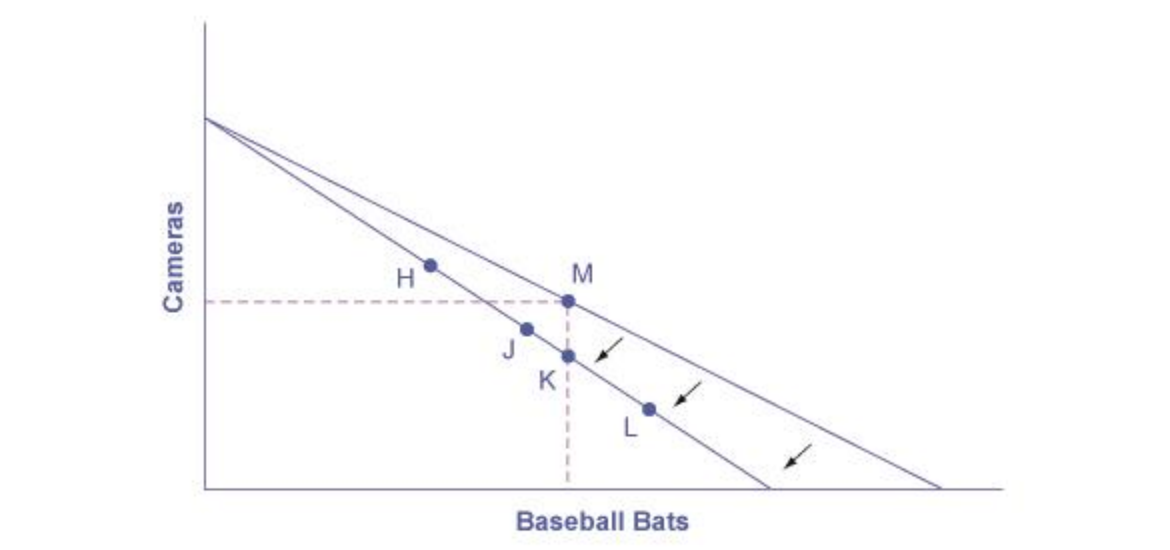


Figure 2

The typical response to higher prices is that a person chooses to consume less of the product with the higher price. This occurs for two reasons, and both effects can occur simultaneously. The **substitution effect** occurs when a price changes and consumers have an incentive to consume less of the good with a relatively higher price and more of the good with a relatively lower price. The **income effect** is that a higher price means, in effect, the buying power of income has been reduced (even though actual income has not changed), which leads to buying less of the good (when the good is normal). In this example, the higher price for baseball bats would cause Sergei to buy a fewer bats for both reasons. Exactly how much will a higher price for bats cause Sergei consumption of bats to fall? Figure 2 suggests a range of possibilities. Sergei might react to a higher price for baseball bats by purchasing the same quantity of bats, but cutting his consumption of cameras. This choice is the point K on the new budget constraint, straight below the original choice M. Alternatively, Sergei might react by dramatically reducing his purchases of bats and instead buy more cameras.

The key is that it would be imprudent to assume that a change in the price of baseball bats will only or primarily affect the good whose price is changed, while the quantity consumed of other goods remains the same. Since Sergei purchases all his products out of the same budget, a change in the price of one good can also have a range of effects, either positive or negative, on the quantity consumed of other goods.

In short, a higher price typically causes reduced consumption of the good in question, but it can affect the consumption of other goods as well.

Link to Video: Paradox of Value

Institute for Humane Studies. (2017). *Elements of economics* [video].

CHAPTER 7: The Producer Side of Markets- Short Run Production & Costs
7.1 Firms

Business Organization

Source: Curtis & Irvine, 2016, Section 7.1, CC-BY-NC-SA 3.0 (Original source, pages 1-2 only)

Many different types of suppliers provide goods and services to the marketplace. Some are small, some are large. But, whatever their size, suppliers choose an organizational form that is appropriate for their business: Aircraft and oil rigs are produced by large corporations; dental services are provided by individual professionals or private partnerships.

The initial material of this chapter addresses organizational forms, their goals and their operation. We then examine why individuals choose to invest in firms, and propose that such investment provides individual investors with a means both to earning a return on their savings and to managing the risk associated with investing.

Understanding the way firms and capital markets function is crucial to understanding our economic history and how different forms of social and economic institutions interact. For example, seventeenth-century Amsterdam had a thriving bourgeoisie, well-developed financial markets, and investors with savings. This environment facilitated the channeling of investors' funds to firms specializing in trade and nautical conquest. This tiny state was then the source of some of the world's leading explorers and traders, and it had colonies stretching to Indonesia. The result was economic growth and prosperity.

In contrast, for much of the twentieth century, the Soviet Union dominated a huge territory covering much of Asia and Europe. But capital markets were non-existent, independent firms were stifled, and economic decline ultimately ensued. Much of the enormous difference in the respective patterns of economic development can be explained by the fact that one state fostered firms, capital markets, and legal institutions, while the other did not. In terms of our production possibility frontier: One set of institutional arrangements was conducive to expanding the possibilities; the other was not. Sustainable new businesses invariably require investors at an early point in the lifecycle of the business. Accordingly, financial institutions that facilitate the flow of savings and financial investment into new enterprises perform a vital function in the economy.

Businesses, or firms, have several different forms. At the smallest scale, a business takes the form of a **sole proprietor** or sole trader who is the exclusive owner. A sole trader gets all of the revenues from the firm and incurs all of the costs. Hence he may make profits or be personally liable for the losses. In the latter case his business or even personal assets may be confiscated to cover debts. Personal bankruptcy may result.

Sole proprietor is the single owner of a business.

If a business is to grow, **partners** may be required. Such partners can inject money in exchange for a share of future profits. Firms where trust is involved, such as legal or accounting firms, typically adopt this structure. A firm is given credibility when customers see that partners invest their own wealth in it.

Partnership: a business owned jointly by two or more individuals, who share in the profits and are jointly responsible for losses.

If a business is to grow to a significant size it will generally need cash and thus partners. Firms that provide legal services or dental services rely primarily on human expertise, and therefore they need relatively little physical capital. Hence their cash start-up needs tend to be modest. But firms that produce aircraft need vast amounts of money to construct assembly facilities; pharmaceuticals may need a billion dollars worth of research and development to bring a new drug to the marketplace. Such businesses form corporations – also known as companies.

Large organizations have several inherent advantages over small organizations when a high output level is required. Specialization in particular tasks leads to increased efficiency for production workers. At the same time, non-production workers can perform a multitude of different tasks. If a large corporation decided to contract out every task involved in bringing its product to market, the costs of such agreements could be prohibitively high. In addition, synergies can arise from teamwork. New ideas and better work flow are more likely to materialize when individuals work in close proximity than when working as isolated units, no matter how efficient they may be individually. A key aspect of such large organizations is that *they have a legal identity separate from the managers and owners.*

Corporation or company is an organization with a legal identity separate from its owners that produces and trades.

Ownership and Corporate Goals

Source: Curtis & Irvine, 2016, Section 7.2, CC-BY-NC-SA 3.0 (Original source, page 1, paragraphs 1 & 2 only)

As economists, we believe that profit maximization accurately describes a typical firm's objective. However, since large firms are not run by their owners but by their executives or agents, it is frequently hard for the shareholders to know exactly what happens within a company. Even the board of directors—the guiding managerial group—may not be fully aware of the decisions, strategies, and practices of their executives and managers. Occasionally things go wrong, and managers decide to follow their own interests rather than the interests of the company. In technical terms, the interests of the corporation and its shareholders might not be aligned with the interests of its managers. For example, managers might have a short horizon and take steps to increase their own income in the short term, knowing that they will move to another job before the long-term effects of their decisions impact the firm.

At the same time, the marketplace for the ownership of corporations exerts a certain discipline: If firms are not as productive or profitable as possible, they may become subject to takeover by other firms. Fear of such takeover can induce executives and boards to maximize profits.

7.2 Costs & Profits

Source: Lynham, 2018, Section 7.1, CC-BY 4.0 (Original source, paragraphs 1 & 7 removed)

Each of these businesses, regardless of size or complexity, tries to earn a profit:

$$Profit = Total\ Revenue\ -\ Total\ Cost$$

Total **revenue** is the income brought into the firm from selling its products. It is calculated by multiplying the price of the product times the quantity of output sold:

$$Total\ Revenue = Price\ \times\ Quantity$$

We will see in the following chapters that revenue is a function of the demand for the firm's products.

We can distinguish between two types of cost: explicit and implicit. **Explicit costs** are out-of-pocket costs, that is, payments that are actually made. Wages that a firm pays its employees or rent that a firm pays for its office are explicit costs. **Implicit costs** are more subtle, but just as important. They represent the opportunity cost of using resources already owned by the firm. Often for small businesses, they are resources contributed by the owners; for example, working in the business while not getting a formal salary, or using the ground floor of a home as a retail store. Implicit costs also allow for depreciation of goods, materials, and equipment that are necessary for a company to operate. (See the Work it Out feature for an extended example.)

These two definitions of cost are important for distinguishing between two conceptions of profit, accounting profit and economic profit. **Accounting profit** is a cash concept. It means total revenue minus explicit costs—the difference between dollars brought in and dollars paid out. **Economic profit** is total revenue minus total cost, including both explicit and implicit costs. The difference is important because even though a business pays income taxes based on its accounting profit, whether or not it is economically successful depends on its economic profit.

Calculating Implicit Costs

Consider the following example. Fred currently works for a corporate law firm. He is considering opening his own legal practice, where he expects to earn $200,000 per year once he gets established. To run his own firm, he would need an office and a law clerk. He has found the perfect office, which rents for $50,000 per year. A law clerk could be hired for $35,000 per year. If these figures are accurate, would Fred's legal practice be profitable?

Step 1. First you have to calculate the costs. You can take what you know about explicit costs and total them:

Office rental:	$50, 000
Law clerk's salary:	+$35, 000
Total explicit costs	$85, 000

Step 2. Subtracting the explicit costs from the revenue gives you the accounting profit.

Revenues:	$200, 000
Explicit Costs:	-$85, 000
Accounting profit:	$115, 000

But these calculations consider only the explicit costs. To open his own practice, Fred would have to quit his current job, where he is earning an annual salary of $125,000. This would be an implicit cost of opening his own firm.

Step 3. You need to subtract both the explicit and implicit costs to determine the true economic profit:

$$\textit{Economic profit} = \textit{total revenues} - \textit{explicit costs} - \textit{implicit costs}$$

$$= \$200,000 - \$85,000 - \$125,000$$

$$= -\$10,000 \textit{ per year}$$

Fred would be losing $10,000 per year. That does not mean he would not want to open his own business, but it does mean he would be earning $10,000 less than if he worked for the corporate firm.

Implicit costs can include other things as well. Maybe Fred values his leisure time, and starting his own firm would require him to put in more hours than at the corporate firm. In this case, the lost leisure would also be an implicit cost that would subtract from economic profits.

Source: Curtis & Irvine, 2016, Section 7.2, CC-BY-NC-SA 3.0 (Original source, page 5 only)

We can summarize this: **Accounting profit** is the difference between revenues and explicit costs. **Economic profit** is the difference between revenue and the sum of explicit and implicit costs. **Explicit costs** are the measured financial costs; **Implicit costs** represent the opportunity cost of the resources used in production.

Accounting profit: is the difference between revenues and explicit costs.

Economic profit: is the difference between revenue and the sum of explicit and implicit costs.

Explicit costs: are the measured financial costs.

Implicit costs: represent the opportunity cost of the resources used in production.

Production

Source: Curtis & Irvine, 2016, Section 8.1, CC-BY-NC-SA 3.0 (Original source, page 1 only, paragraph 1 removed)

The remuneration of managers in virtually all corporations is linked to profitability. Efficient production, a.k.a. cost reduction, is critical to achieving this goal. In this chapter we will examine cost management and efficient production from the ground up – by exploring how a small entrepreneur brings his or her product to market in the most efficient way possible. As we shall see, efficient production and cost minimization amount to the same thing: Cost minimization is the financial reflection of efficient production.

Efficient production is critical in any budget-driven organization, not just in the private sector. Public institutions equally are, and should be, concerned with costs and efficiency.

Entrepreneurs employ factors of production (capital and labour) in order to transform raw materials and other inputs into goods or services. The relationship between output and the inputs used in the production process is called a production function. It specifies how much output can be produced with given combinations of inputs. A **production function** is not restricted to profit-driven organizations. Municipal road repairs are carried out with labour and capital. Students are educated with teachers, classrooms, computers, and books. Each of these is a production process.

Production function: a technological relationship that specifies how much output can be produced with specific amounts of inputs.

“Intermediate good”, n.d., CC-BY-SA 3.0 Unported

Intermediate goods or producer goods or semi-finished products are goods, such as partly finished goods, used as inputs in the production of other goods including final goods. A firm may make and then use intermediate goods, or make and then sell, or buy then use them. In the production process, intermediate goods either become part of the final product, or are changed beyond recognition in the process. This means intermediate goods are resold among industries.

The Time Frame

Source: Curtis & Irvine, 2016, Section 8.2, CC-BY-NC-SA 3.0

We distinguish initially between the **short run** and the **long run**. When discussing technological change, we use the term very long run. These concepts have little to do with clocks or calendars; rather, they are defined by the degree of flexibility an entrepreneur or manager has in her production process. A key decision variable is capital.

A customary assumption is that a producer can hire more labour immediately, if necessary, either by taking on new workers (since there are usually some who are unemployed and looking for work), or by getting the existing workers to work longer hours. In contrast, getting new capital in place is generally more time consuming: The entrepreneur may have to place an order for new machinery, which will involve a production and delivery time lag. Or she may have to move to a more spacious location in order to accommodate the added capital. Whether this calendar time is one week, one month, or one year is of no concern to us. We define the long run as a period of sufficient length to enable the entrepreneur to adjust her capital stock, whereas in the short run at least one factor of production is fixed. Note that it matters little whether it is labour or capital that is fixed in the short run. A software development company may be able to install new capital (computing power) instantaneously but have to train new developers. In such a case capital is variable and labour is fixed in the short run. The definition of the short run is that one of the factors is fixed, and in our examples we will assume that it is capital.

Short run: a period during which at least one factor of production is fixed. If capital is fixed, then more output is produced by using additional labour.

Long run: a period of time that is sufficient to enable all factors of production to be adjusted.

Very long run: a period sufficiently long for new technology to develop.

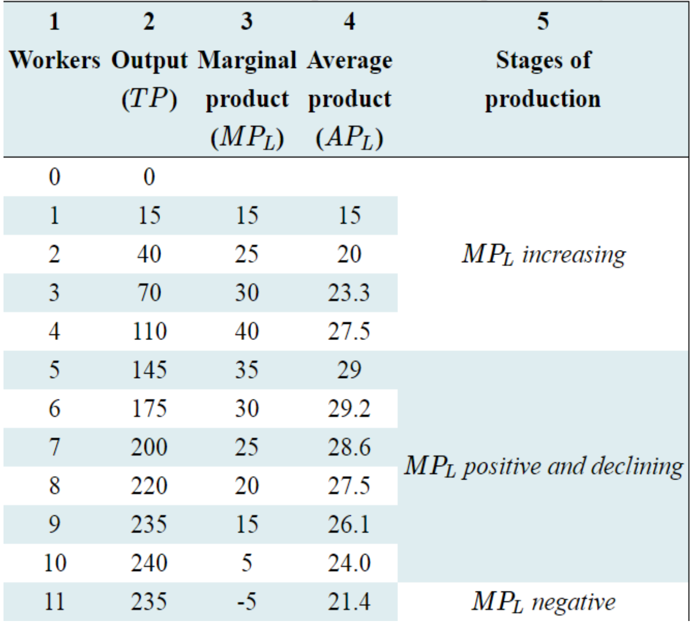
7.3 PRODUCTION IN THE SHORT RUN

Source: Curtis & Irvine, 2016, Section 8.2, CC-BY-NC-SA 3.0

Black Diamond Snowboards (BDS) is a start-up snowboard producing enterprise. Its founder has invented a new lamination process that gives extra strength to his boards. He has set up a production line in his garage that has four workstations: Laminating, attaching the steel edge, waxing, and packing.

With this process in place, he must examine how productive his firm can be. After extensive testing, he has determined exactly how his productivity depends upon the number of workers. If he employs only one worker, then that worker must perform several tasks, and will encounter 'down time' between workstations. Extra workers would therefore not only increase the total output; they could, in addition, increase output per worker. He also realizes that once he has employed a critical number of workers, additional workers may not be so productive: Because they will have to share the fixed amount of machinery in his garage, they may have to wait for another worker to finish using a machine. At such a point, the productivity of his plant will begin to fall off, and he may want to consider capital expansion. But for the moment he is constrained to using this particular assembly plant. Testing leads him to formulate the relationship between workers and output that is described in Table 8.1.

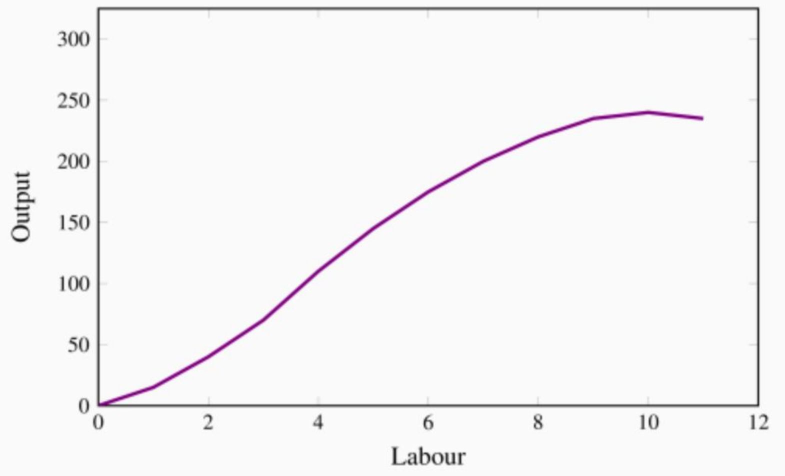
Table 8.1 Snowboard production and productivity



1	2	3	4	5
Workers	Output (TP)	Marginal product (MP_L)	Average product (AP_L)	Stages of production
0	0			MP_L increasing
1	15	15	15	
2	40	25	20	
3	70	30	23.3	
4	110	40	27.5	
5	145	35	29	MP_L positive and declining
6	175	30	29.2	
7	200	25	28.6	
8	220	20	27.5	
9	235	15	26.1	
10	240	5	24.0	
11	235	-5	21.4	MP_L negative

By increasing the number of workers in the plant, BDS produces more boards. The relationship between these two variables in columns 1 and 2 in the table is plotted in Figure 8.1. This is called the **total product function** (TP), and it defines the output produced with different amounts of labour in a plant of fixed size.

Figure 8.1 Total product curve



Output increases with the amount of labour used. Initially the increase in output due to using more labour is high, subsequently it is lower. The initial phase characterizes increasing productivity, the later phase defines declining productivity.

Total product is the relationship between total output produced and the number of workers employed, for a given amount of capital.

This relationship is positive, indicating that more workers produce more boards. But the curve has an interesting pattern. In the initial expansion of employment it becomes progressively steeper – its curvature is slightly convex; following this phase the function's increase becomes progressively less steep – its curvature is concave. These different stages in the TP curve tell us a great deal about productivity in BDS. To see this, consider the additional number of boards produced by each worker. The first worker produces 15. When a second worker is hired, the total product rises to 40, so the additional product attributable to the second worker is 25. A third worker increases output by 30 units, and so on. We refer to this additional output as the marginal product (MP) of an additional worker, because it defines the incremental, or marginal, contribution of the worker. These values are entered in column 3.

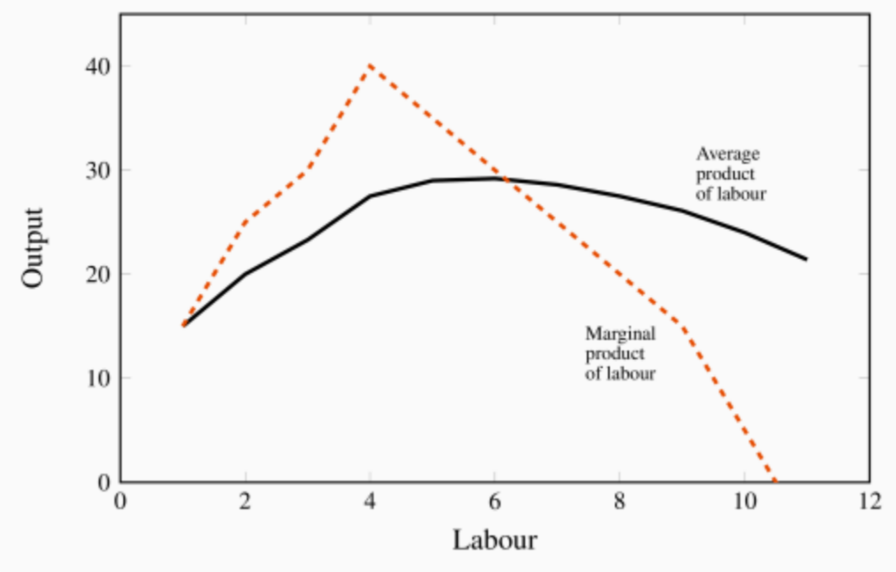
More generally the MP of labour is defined as the change in output divided by the change in the number of units of labour employed. Using, as before, the Greek capital delta (Δ) to denote a change, we can define

$$\text{MPL}=\frac{\text{Change in output produced}}{\text{Change in labour employed}}=\frac{\Delta \text{Q}}{\Delta \text{L}}$$

In this example the change in labour is one unit at each stage and hence the **marginal product of labour** is simply the corresponding change in output. It is also the case that the MPL is the slope of the TP curve – the change in the value on the vertical axis due to a change in the value of the variable on the horizontal axis.

Marginal product of labour is the addition to output produced by each additional worker. It is also the slope of the total product curve.

Figure 8.2 Average and marginal product curves



The productivity curves initially rise and then decline, reflecting increasing and decreasing productivity. The MP_L *curves must intersect the* AP_L *curve at the maximum of the* AP_L*: The average must increase if the marginal exceeds the average and must decline if the marginal is less than the average.*

During the initial stage of production expansion, the marginal product of each worker is increasing. It increases from 15 to 40 as BDS moves from having one employee to four employees. This increasing MP is made possible by the fact that each worker is able to spend more time at his workstation, and less time moving between tasks. But, at a certain point in the employment expansion, the MP reaches a maximum and then begins to tail off. At this stage – in the concave region of the TP curve – additional workers continue to produce additional output, but at a diminishing rate. For example, while the fourth worker adds 40 units to output, the fifth worker adds 35, the sixth worker 30, and so on. This declining MP is due to the constraint of a fixed number of machines: All workers must share the same capital. The MP function is plotted in Figure 8.2.

The phenomenon we have just described has the status of a law in economics: The **law of diminishing returns** states that, in the face of a fixed amount of capital, the contribution of additional units of a variable factor must eventually decline.

Law of diminishing returns: when increments of a variable factor (labour) are added to a fixed amount of another factor (capital), the marginal product of the variable factor must eventually decline.

The relationship between Figures 8.1 and 8.2 should be noted. First, the MPL reaches a maximum at an output of 4 units – where the slope of the TP curve is greatest. The MPL curve remains positive beyond this output, but declines: The TP curve reaches a maximum when the tenth unit of labour is employed. An eleventh unit actually reduces total output; therefore, the MP of this eleventh worker is negative! In Figure 8.2, the MP curve becomes negative at this point. The garage is now so crowded with workers that they are beginning to obstruct the operation of the production process. Thus the producer would never employ an eleventh unit of labour.

Next, consider the information in the fourth column of the table. It defines the average product of labour (APL)—the amount of output produced, on average, by workers at different employment levels:

$$\text{APL} = \frac{\text{Total output produced}}{\text{Total amount of labour employed}} = \frac{Q}{L}$$

This function is also plotted in Figure 8.2. Referring to the table: The AP column indicates, for example, that when two units of labour are employed and forty units of output are produced, the average production level of each worker is 20 units (=40/2). When three workers produce 70 units, their average production is 23.3 (=70/3), and so forth. Like the MP function, this one also increases and subsequently decreases, reflecting exactly the same productivity forces that are at work on the MP curve.

Average product of labour is the number of units of output produced per unit of labour at different levels of employment.

The AP and MP functions intersect at the point where the AP is at its peak. This is no accident, and has a simple explanation. Imagine a baseball player who is batting .280 coming into today's game—he has been hitting his way onto base 28 percent of the time when he goes up to bat, so far this season. This is his average product, AP.

In today's game, if he bats .500 (he hits his way to base on half of his at-bats), then he will improve his average. Today's batting (his MP) at .500 therefore pulls up his season's AP. Accordingly, whenever the MP exceeds the AP, the AP is pulled up. By the same reasoning, if his MP is less than his season average, his average will be pulled down. It follows that the two functions must intersect at the peak of the AP curve. To summarize:

> If the MP exceeds the AP, then the AP increases;
> If the MP is less than the AP, then the AP declines.

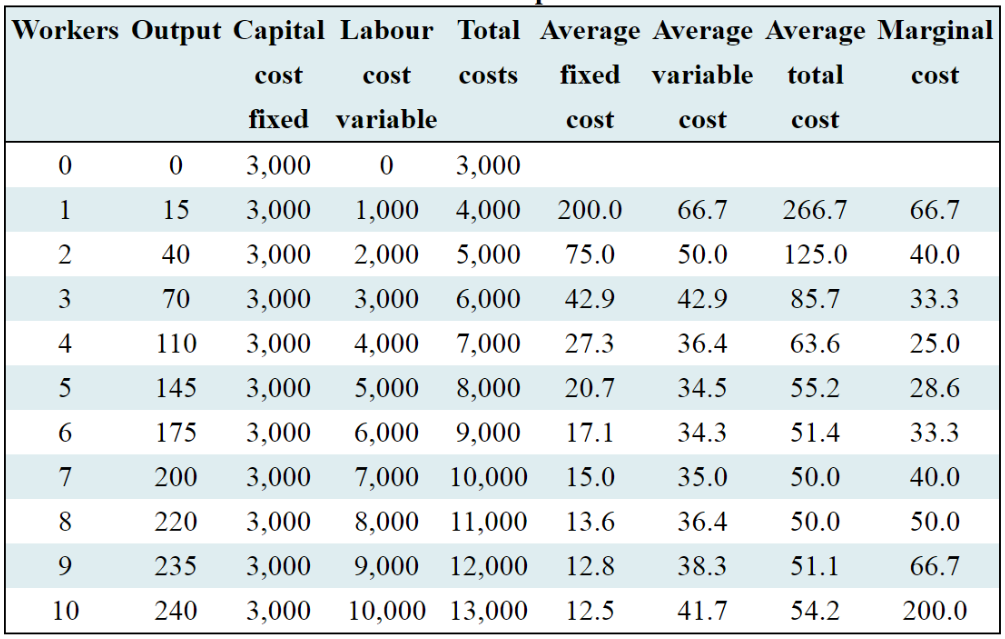
While the owner of BDS may understand his productivity relations, his ultimate goal is to make profit, and for this he must figure out how productivity translates into cost.

7.3 COSTS IN THE SHORT RUN

Source: Curtis & Irvine, 2016, Section 8.2, CC-BY-NC-SA 3.0 (Original source, Figure 8.4 replaced, 'Teams and services' section removed)

The cost structure for the production of snowboards at Black Diamond is illustrated in Table 8.2. Employees are skilled and are paid a weekly wage of $1,000. The cost of capital is $3,000 and it is fixed, which means that it does not vary with output. As in Table 8.1, the number of employees and the output are given in the first two columns. The following three columns define the capital costs, the labour costs, and the sum of these in producing different levels of output. We use the terms **fixed**, **variable**, and **total costs** to define the cost structure of a firm. Fixed costs do not vary with output, whereas variable costs do, and total costs are the sum of fixed and variable costs. To keep this example as simple as possible, we will ignore the cost of raw materials. We could add an additional column of costs, but doing so will not change the conclusions.

Table 8.2 Snowboard production costs



Workers	Output	Capital cost fixed	Labour cost variable	Total costs	Average fixed cost	Average variable cost	Average total cost	Marginal cost
0	0	3,000	0	3,000				
1	15	3,000	1,000	4,000	200.0	66.7	266.7	66.7
2	40	3,000	2,000	5,000	75.0	50.0	125.0	40.0
3	70	3,000	3,000	6,000	42.9	42.9	85.7	33.3
4	110	3,000	4,000	7,000	27.3	36.4	63.6	25.0
5	145	3,000	5,000	8,000	20.7	34.5	55.2	28.6
6	175	3,000	6,000	9,000	17.1	34.3	51.4	33.3
7	200	3,000	7,000	10,000	15.0	35.0	50.0	40.0
8	220	3,000	8,000	11,000	13.6	36.4	50.0	50.0
9	235	3,000	9,000	12,000	12.8	38.3	51.1	66.7
10	240	3,000	10,000	13,000	12.5	41.7	54.2	200.0

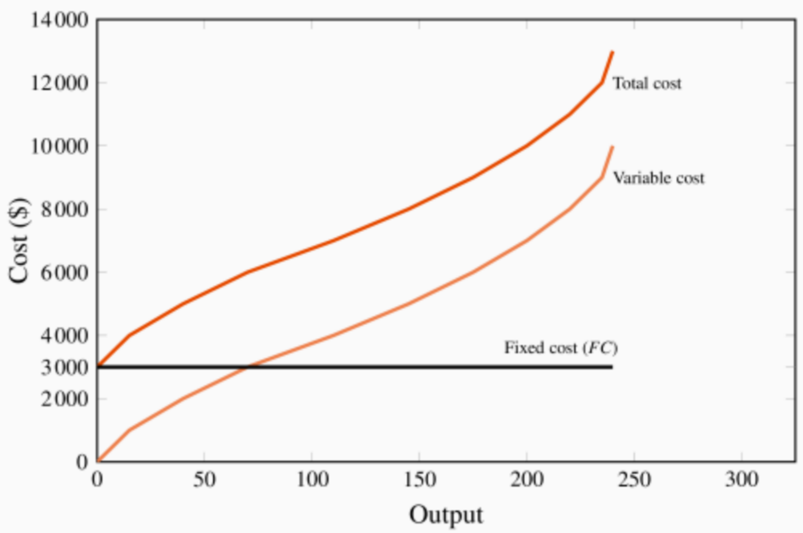
Fixed costs are costs that are independent of the level of output.

Variable costs are related to the output produced.

Total cost is the sum of fixed cost and variable cost.

Total costs are illustrated in Figure 8.3 as the vertical sum of variable and fixed costs. For example, Table 8.2 indicates that the total cost of producing 220 units of output is the sum of $3,000 in fixed costs plus $8,000 in variable costs. Therefore, at the output level 220 on the horizontal axis in Figure 8.3, the sum of the cost components yields a value of $11,000 that forms one point on the total cost curve. Performing a similar calculation for every possible output yields a series of points that together form the complete total cost curve.

Figure 8.3 Total cost curves



Total cost is the vertical sum of the variable and fixed costs.

Average costs are given in the next three columns of Table 8.2. Average cost is the cost per unit of output, and we can define an average cost corresponding to each of the fixed, variable, and total costs defined above. **Average fixed cost** (AFC) is the total fixed cost divided by output; **average variable** cost (AVC) is the total variable cost divided by output; and **average total cost** (ATC) is the total cost divided by output.

AFC = (Fixed cost)/Q = FC/Q
AVC = (Total variable costs)/Q = TVC/Q
ATC = AFC+AVC

Average fixed cost is the total fixed cost per unit of output.

Average variable cost is the total variable cost per unit of output.

Average total cost is the sum of all costs per unit of output.

The Productivity-Cost Relationship

Consider the average variable cost - average product relationship, as developed in column 7 of Table 8.2 (replaced with Figure 2 below); its corresponding variable cost curve is plotted in Figure 8.4. In this example, AVC first decreases and then increases. The intuition behind its shape is straightforward (and realistic) if you have understood why productivity varies in the short run: The variable cost, which represents the cost of labour, is constant per unit of labour, because the wage paid to each worker does not change. However, each worker's productivity varies. Initially, when we hire more workers, they become more productive, perhaps because they have less 'down time' in switching between tasks. This means that the labour costs per snowboard must decline. At some point, however, the law of diminishing returns sets in: As before, each additional worker is paid a constant amount, but as productivity declines the labour cost per snowboard increases.

Source: Lynham, 2018, Section 7.2, CC-BY 4.0 (Original source, Figure 2 only)

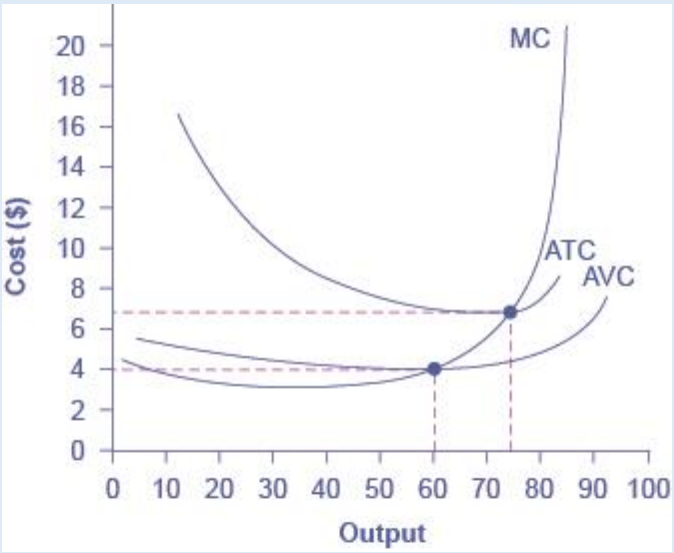


Figure 2

The MC intersects the ATC and AVC at their minimum values

In this numerical example the AP is at a maximum when six units of labour are employed and output is 175. This is also the point where the AVC is at a minimum. This maximum/minimum relationship is also illustrated in Figures 8.2 and 8.4.

Now consider the **marginal cost - marginal product relationship**. The marginal cost (MC) defines the cost of producing one more unit of output. In Table 8.2, the marginal cost of output is given in the final column. It is the additional cost of production divided by the additional number of units produced. For example, in going from 15 units of output to 40, total costs increase from $4,000 to $5,000. The MC is the cost of those additional units divided by the number of additional units. In this range of output, MC is $1,000/25=$40. We could also calculate the MC as the addition to variable costs rather than the addition to total costs, because the addition to each is the same—fixed costs are fixed. Hence:

$$\text{MC} = \frac{\text{Change in total costs}}{\text{Change in output produced}} = \frac{\Delta\text{TC}}{\Delta\text{Q}}$$

$$= \frac{\text{Change in variable costs}}{\text{Change in output produced}} = \frac{\Delta\text{TVC}}{\Delta\text{Q}}$$

Marginal cost of production is the cost of producing each additional unit of output.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

Just as the behaviour of the AVC curve is determined by the AP curve, so too the behaviour of the MC is determined by the MP curve. When the MP of an additional worker exceeds the MP of the previous worker, this implies that the cost of the additional output produced by the last worker hired must be declining. To summarize:

> If the marginal product of labour increases, then the marginal cost of output declines;
> If the marginal product of labour declines, then the marginal cost of output increases.

In our example, the MPL reaches a maximum when the fourth unit of labour is employed (or 110 units of output are produced), and this also is where the MC is at a minimum. This illustrates that the marginal cost reaches a minimum at the output level where the marginal product reaches a maximum.

The average total cost is the sum of the fixed cost per unit of output and the variable cost per unit of output. Typically, fixed costs are the dominant component of total costs at low output levels, but become less dominant at higher output levels. Unlike average variable costs, note that the average fixed cost must always decline with output, because a fixed cost is being spread over more units of output. Hence, when the ATC curve eventually increases, it is because the increasing variable cost component eventually dominates the declining AFC component. In our example, this occurs when output increases from 220 units (8 workers) to 235 (9 workers).

Finally, observe the interrelationship between the MC curve on the one hand and the ATC and AVC on the other. Note from Figure 8.4 that the MC cuts the AVC and the ATC at the minimum point of each of the latter. The logic behind this pattern is analogous to the logic of the relationship between marginal and average product curves: When the cost of an additional unit of output is less than the average, this reduces the average cost; whereas, if the cost of an additional unit of output is above the average, this raises the average cost. This must hold true regardless of whether we relate the MC to the ATC or the AVC.

> When the marginal cost is less than the average cost, the average cost must decline;
> When the marginal cost exceeds the average cost, the average cost must increase.

Notation: We use both the abbreviations ATC and AC to denote average total cost. The term 'average cost' is understood in economics to include both fixed and variable costs.

CHAPTER 8: The Producer Side of Markets- Long Run Production, Costs & Technology

Source: Curtis & Irvine, 2016, Section 8.1, CC-BY-NC-SA 3.0 (Original source, page 2 only)

Economists distinguish between two concepts of efficiency: One is **technological efficiency**; the other is **economic efficiency**. To illustrate the difference, consider the case of auto assembly in Oshawa Megamobile Inc., an auto manufacturer. Megamobile could assemble its vehicles either by using a large number of assembly workers and a plant that has a relatively small amount of machinery, or it could use fewer workers accompanied by more machinery in the form of robots. Each of these processes could be deemed technologically efficient, provided that there is no waste. If the workers without robots are combined with their capital to produce as much as possible, then that production process is technologically efficient. Likewise, in the scenario with robots, if the workers and capital are producing as much as possible, then that process too is efficient in the technological sense.

Technological efficiency means that the maximum output is produced with the given set of inputs.

Economic efficiency is concerned with more than just technological efficiency. Since the entrepreneur's goal is to make profit, she must consider which technologically efficient process best achieves that objective. More broadly, any budget-driven process should focus on being economically efficient, whether in the public or private sector. An economically efficient production structure is the one that produces output at least cost.

Economic efficiency defines a production structure that produces output at least cost.

Auto-assembly plants the world over have moved to using robots during the last two decades. Why? The reason is not that robots were invented 20 years ago; they were invented long before that. The real reason is that, until recently, this technology was not economically efficient. Robots were too expensive; they were not capable of high-precision assembly. But once their cost declined and their accuracy increased they became economically efficient. The development of robots represented technological progress. When this progress reached a critical point, entrepreneurs embraced it.

8.1 Long-Run Production and Costs

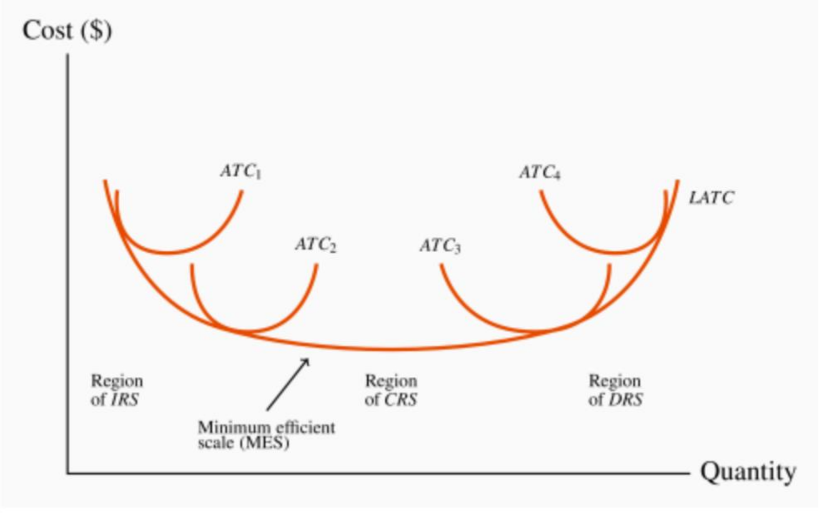
Source: Curtis & Irvine, 2016, Section 8.6, CC-BY-NC-SA 3.0 (Original source, pages 1-5 only, first paragraphs of pages 1 & 4 removed)

Production costs almost always decline when the scale of the operation initially increases. We refer to this phenomenon simply as economies of scale. There are several reasons why scale economies are encountered. One is that production flows can be organized in a more efficient manner when more is being produced. Another is that the opportunity to make greater use of task specialization presents itself; for example, Black Diamond Snowboards may be able to subdivide tasks within the laminating and packaging stations. If scale economies do define the real world, then a bigger plant—one that is geared to produce a higher level of output—should have an average total cost curve that is "lower" than the cost curve corresponding to the smaller scale of operation we considered in the example above.

Average Costs in the Long run

Figure 8.5 illustrates a possible relationship between the ATC curves for four different scales of operation. ATC1 is the average total cost curve associated with a small-sized plant; think of it as the plant built in the entrepreneur's garage. ATC2 is associated with a somewhat larger plant, perhaps one she has put together in a rented industrial or commercial space. The further a cost curve is located to the right of the diagram the larger the production facility it defines, given that output is measured on the horizontal axis. If there are economies associated with a larger scale of operation, then the average costs associated with producing larger outputs in a larger plant should be lower than the average costs associated with lower outputs in a smaller plant, assuming that the plants are producing the output levels they were designed to produce. For this reason, the cost curve ATC2 and the cost curve ATC3 each have a segment that is lower than the lowest segment on ATC1. However, in Figure 8.5 the cost curve ATC4 has moved upwards. What behaviours are implied here?

Figure 8.5 Long-run and short-run average costs



The long-run ATC curve, LATC, is the lower envelope of all short-run ATC curves. It defines the least cost per unit of output when all inputs are variable. Minimum efficient scale is that output level at which the LATC is a minimum, indicating that further increases in the scale of production will not reduce unit costs.

In many production environments, beyond some large scale of operation, it becomes increasingly difficult to reap further cost reductions from specialization, organizational economies, or marketing economies. At such a point, the scale economies are effectively exhausted, and larger plant sizes no longer give rise to lower (short-run) ATC curves. This is reflected in the similarity of the ATC2 and the ATC3 curves. The pattern suggests that we have almost exhausted the possibilities of further scale advantages once we build a plant size corresponding to ATC2. Consider next what is implied by the position of the ATC4 curve relative to the ATC2 and ATC3 curves. The relatively higher position of the ATC4 curve implies that unit costs will be higher in a yet larger plant. Stated differently:

If we increase the scale of this firm to extremely high output levels, we are actually encountering **diseconomies of scale**. Diseconomies of scale imply that unit costs increase as a result of the firm's becoming too large: Perhaps co-ordination difficulties have set in at the very high output levels, or quality-control monitoring costs have risen. These coordination and management difficulties are reflected in increasing unit costs in the long run. Corporations in the modern era are at times broken up into separate operating units and then sold as independent units. Because of the diseconomies of scale, the components of multi-unit corporations may be more valuable independently than when grouped together: When together coordination problems arise; when independent the coordination challenges vanish.

The terms **increasing, constant, and decreasing returns to scale** underlie the concepts of scale economies and diseconomies: Increasing returns to scale (IRS) implies that, when all inputs are increased by a given proportion, output increases more than proportionately. Constant returns to scale (CRS) implies that output increases in direct proportion to an equal proportionate increase in all inputs. Decreasing returns to scale (DRS) implies that an equal proportionate increase in all inputs leads to a less than proportionate increase in output.

Increasing returns to scale implies that, when all inputs are increased by a given proportion, output increases more than proportionately.

Constant returns to scale implies that output increases in direct proportion to an equal proportionate increase in all inputs.

Decreasing returns to scale implies that an equal proportionate increase in all inputs leads to a less than proportionate increase in output.

Increasing returns to scale characterize businesses with large initial costs and relatively low costs of producing each unit of output. Computer chip manufacturers, pharmaceutical manufacturers, even brewers all appear to benefit from scale economies. In the beer market, brewing, bottling and shipping are all low-cost operations relative to the capital cost of setting up a brewery. Consequently, we observe surprisingly few breweries in any brewing company, even in large land-mass economies such as Canada or the US.

Application Box 8.1 Decreasing returns to scale

The CEO of Hewlett Packard announced in October 2012 that the company would reduce its labour force by 29,000 workers (out of a total of 350,000). The problem was that communications within the company were so complex and strained as to increase unit costs. In addition, the company was producing an excessive product variety – 2,100 variants of laser printer!

In addition to the four short-run average total cost curves, Figure 8.5 contains a curve that forms an envelope around the bottom of these short-run average cost curves. This envelope is the **long-run average total cost** (LATC) curve, because it defines average cost as we move from one plant size to another. Remember that in the long run both labour and capital are variable, and as we move from one short-run average cost curve to another, that is exactly what happens—all factors of production are variable. Hence, the collection of short-run cost curves in Figure 8.5 provides the ingredients for a long-run average total cost curve[1].

LATC = (Long-run total costs)/Q = LTC/Q

Long-run average total cost is the lower envelope of all the short-run ATC curves.

The particular range of output on the LATC where it begins to flatten out is called the range of **minimum efficient scale**. This is an important concept in industrial policy, as we shall see in later chapters. At such an output level, the producer has expanded sufficiently to take advantage of virtually all the scale economies available.

Minimum efficient scale defines a threshold size of operation such that scale economies are almost exhausted.

[1] *Note that the long-run average total cost is not the collection of minimum points from each short-run average cost curve. The envelope of the short-run curves will pick up mainly points that are not at the minimum, as you will see if you try to draw the outcome. The intuition behind the definition is this: With increasing returns to scale, it may be better to build a plant size that operates with some spare capacity than to build one that is geared to producing a smaller output level. In building the larger plant, we can take greater advantage of the scale economies, and it may prove less costly to produce in such a plant than to produce with a smaller plant that has less unused capacity and does not exploit the underlying scale economies. Conversely, in the presence of decreasing returns to scale, it may be less costly to produce output in a plant that is used "overtime" than to use a larger plant that suffers from scale diseconomies.*

Source: Lynham, 2018, Section 5.1, CC-BY 4.0 (Original source, Table 1 only)

Economies of Scale Once a firm has determined the least costly production technology, it can consider the optimal scale of production, or quantity of output to produce. Many industries experience economies of scale. **Economies of scale** refers to the situation where, as the quantity of output goes up, the cost per unit goes down. This is the idea behind “warehouse stores” like Costco or Walmart. In everyday language: a larger factory can produce at a lower average cost than a smaller factory.

Figure 1 illustrates the idea of economies of scale, showing the average cost of producing an alarm clock falling as the quantity of output rises. For a small-sized factory like S, with an output level of 1,000, the average cost of production is $12 per alarm clock. For a medium-sized factory like M, with an output level of 2,000, the average cost of production falls to $8 per alarm clock. For a large factory like L, with an output of 5,000, the average cost of production declines still further to $4 per alarm clock.

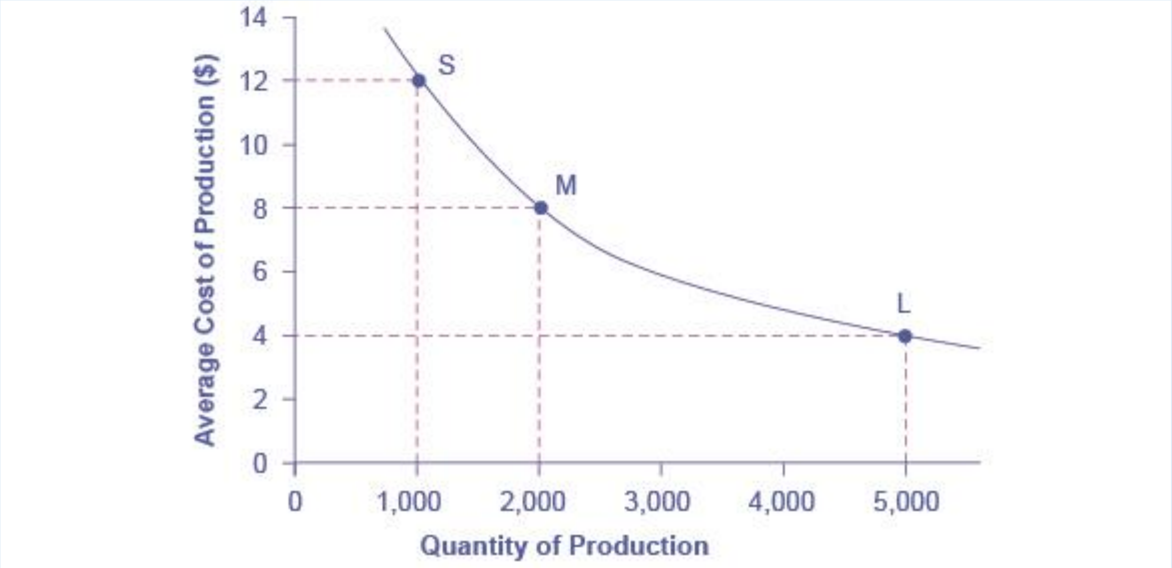


Figure 1. *Economies of Scale. A small factory like S produces 1,000 alarm clocks at an average cost of $12 per clock. A medium factory like M produces 2,000 alarm clocks at a cost of $8 per clock. A large factory like L produces 5,000 alarm clocks at a cost of $4 per clock. Economies of scale exist because the larger scale of production leads to lower average costs.*

8.2 Technological Change in the Very Long Run

Source: Curtis & Irvine, 2016, Section 8.6, CC-BY-NC-SA 3.0 (Original source, footnote removed)

Technological change represents innovation that can reduce the cost of production or bring new products on line. As stated earlier, the very long run is a period that is sufficiently long for new technology to evolve and be implemented.

Technological change represents innovation that can reduce the cost of production or bring new products on line.

Technological change has had an enormous impact on economic life for several centuries. It is not something that is defined in terms of the recent telecommunications revolution. The industrial revolution began in eighteenth century Britain. It was accompanied by a less well-recognized, but equally important, agricultural revolution. The improvement in cultivation technology, and ensuing higher yields, freed up enough labour to populate the factories that were the core of the industrial revolution2. The development and spread of mechanical power dominated the nineteenth century, and the mass production line of Henry Ford heralded in the twentieth century. The modern communications revolution has reduced costs, just like its predecessors. But it has also greatly sped up **globalization**, the increasing integration of national markets.

Globalization is the tendency for international markets to be ever more integrated.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

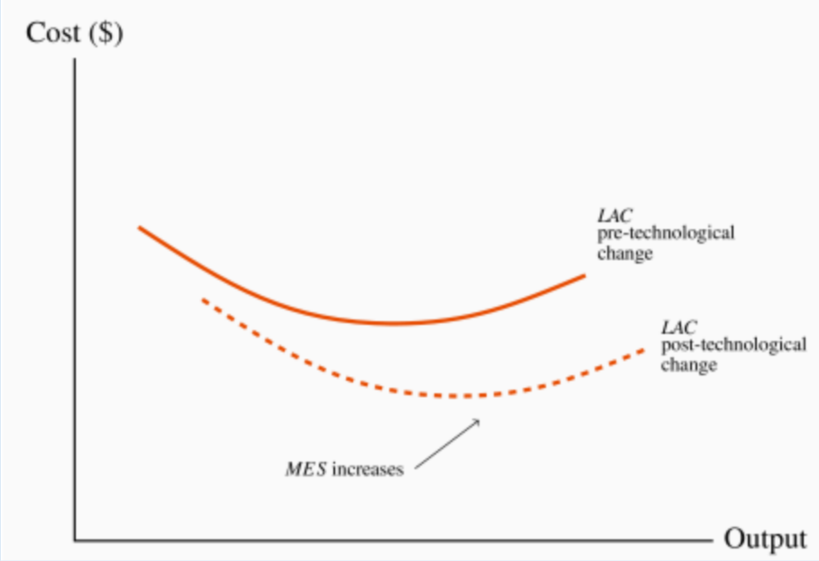
Globalization has several drivers, the most important of which are lower transportation and communication costs, reduced barriers to trade and capital mobility, and the spread of new technologies that facilitate cost and quality control. New technology and better communications have been critical in both increasing the minimum efficient scale of operation and reducing diseconomies of scale; they facilitate the efficient management of large companies.

The continued reduction in trade barriers in the post-World War II era has also meant that the effective marketplace has become the globe rather than the national economy for many products. Companies like *Apple, Microsoft*, and *Dell* are visible worldwide. Globalization has been accompanied by the collapse of the Soviet Union, the adoption of an outward looking philosophy on the part of China, and an increasing role for the market place in India. These developments together have facilitated the outsourcing of much of the West's manufacturing to lower-wage economies.

But new technology not only helps existing companies grow large; it also enables new ones to start up. It is now cheaper for small producers to manage their inventories and maintain contact with their own suppliers. Contrary to what is often claimed, new technology has not led to a greater concentration of the economy's output in the hands of a small number of big producers; the competitive forces that the new technology has unleashed are too strong.

The impact of technology on typical cost curves is illustrated in Figure 8.7. It both reduces the whole cost structure of production, and at the same time increases the minimum efficient scale.

Figure 8.7 Technological change and LAC



Technological change reduces the unit production cost for any output produced and may also increase the minimum efficient scale (MES) *threshold.*

CHAPTER 9: Perfect Competition

Source: Curtis & Irvine, 2016, Sections 9.1-9.2, CC-BY-NC-SA 3.0

9.1 THE PERFECT COMPETITION PARADIGM

A competitive market is one that encompasses a very large number of suppliers, each producing a similar or identical product. Each supplier produces an output that forms a small part of the total market, and the sum of all of these individual outputs represents the production of that sector of the economy. Florists, barber shops, corner stores and dry cleaners all fit this description.

At the other extreme, a market that has just a single supplier is a monopolist. For example, the National Hockey League is the sole supplier of top-quality professional hockey games in North America; Hydro Quebec is a monopoly electricity distributor in Quebec; Via Rail is the only supplier of passenger rail services between Windsor, Ontario and the city of Quebec.

We use the word 'paradigm' in the title to this section: It implies that we will develop a model of supply behaviour for a market in which there are many small suppliers, producing essentially the same product, competing with one-another to meet the demands of consumers.

The structures that we call perfect competition and monopoly are extremes in the market place. Most sectors of the economy lie somewhere between these limiting cases. For example, the market for internet services usually contains several providers in any area – some provide using a fibre cable, others by satellite. The market for smart-phones is dominated by two major players – *Apple and Samsung*. Hence, while these markets that have a limited number of suppliers are competitive in that they freely and fiercely compete for the buyer's expenditure, these are not **perfectly competitive** markets, because they do not have a very large number of suppliers.

In all of the models we develop in this chapter we will assume that the objective of firms is to maximize profit – the difference between revenues and costs.

A **perfectly competitive** industry is one in which many suppliers, producing an identical product, face many buyers, and no one participant can influence the market.

Profit maximization is the goal of competitive suppliers – they seek to maximize the difference between revenues and costs.

The presence of so many sellers in perfect competition means that each firm recognizes its own small size in relation to the total market, and that its actions have no perceptible impact on the market price for the good or service being traded. Each firm is therefore a *price taker*—in contrast to a monopolist, who is a *price setter*.

The same 'smallness' characteristic was assumed when we examined the demands of individuals earlier. Each buyer takes the price as given. He or she is not big enough to be able to influence the price. In contrast, when international airlines purchase or lease aircraft from *Boeing* or *Airbus*, they negotiate over the price and other conditions of supply. The market models underlying these types of transactions are examined [later].

Hence, when we describe a market as being perfectly competitive we do not mean that other market types are not competitive; all market structure are competitive in the sense that the suppliers wish to make profit, and they produce as efficiently as possible in order to meet that goal.

9.2 MARKET CHARACTRISTICS

The key attributes of a perfectly competitive market are the following:

1. There must be *many firms*, each one small and powerless relative to the entire industry.
2. The product *must be standardized*. Barber shops offer a standard product, but a Lexus differs from a Ford. Barbers tend to be price takers, but Lexus does not charge the same price as Ford, and is a price setter.
3. Buyers are assumed to have *full information* about the product and its pricing. For example, buyers know that the products of different suppliers really are the same in quality.
4. There are *many buyers*.
5. There is *free entry and exit* of firms.

In terms of the demand curve that suppliers face, these market characteristics imply that the demand curve facing the perfectly competitive firm is horizontal, or infinitely elastic, as we defined in Chapter 4. In contrast, the demand curve facing the whole industry is downward sloping. The demand curve facing a firm is represented in Figure 9.1. It implies that the supplier can sell any output he chooses at the going price P_0. But what quantity should he choose, or what quantity will maximize his profit? The profit-maximizing choice is his target, and the MC curve plays a key role in this decision.

9.3 SUPPLY DECISIONS IN THE SHORT RUN

Source: Curtis & Irvine, 2016, Sections 9.3, CC-BY-NC-SA 3.0 (Original source, page 8 removed)

The concept **of marginal revenue** is key to analyzing the supply decision of an individual firm. We have used marginal analysis at several points to date. In consumer theory, we saw how consumers balance the utility per dollar at the margin in allocating their budget. Marginal revenue is the additional revenue accruing to the firm from the sale of one more unit of output.

Marginal revenue is the additional revenue accruing to the firm resulting from the sale of one more unit of output.

In perfect competition, a firm's marginal revenue (MR) is the price of the good. Since the price is constant for the individual supplier, each additional unit sold at the price P brings in the same additional revenue. Therefore, P=MR. For example, whether a dry cleaning business launders 10 shirts or 100 shirts per day, the price charged to customers is the same. This equality holds in no other market structure, as we shall see in the following chapters.

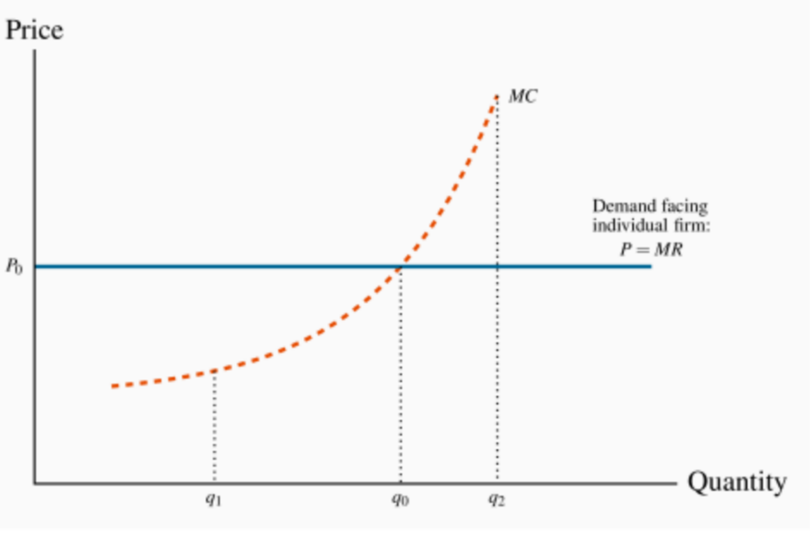
Supply in the Short Run

Recall how we defined the short run in the previous chapter: Each firm's plant size is fixed in the short run, so too is the number of firms in an industry. In the long run, each individual firm can change its scale of operation, and at the same time new firms can enter or existing firms can leave the industry.

Perfectly competitive suppliers face the choice of how much to produce at the going market price: That is, the amount that will maximize their profit. We abstract for the moment on how the price in the marketplace is determined. We shall see later in this chapter that it emerges as the value corresponding to the intersection of the supply and demand curves for the whole market – as described in Chapter 3.

The firm's MC curve is critical in defining the optimal amount to supply at any price. In Figure 9.1, MC is the firm's marginal cost curve in the short run. At the price P_0 the optimal amount to supply is q_0, the amount determined by the intersection of the MC and the demand. To see why, imagine that the producer chose to supply the quantity q_1. Such an output would leave the opportunity for further profit untapped. By producing one additional unit beyond q_1, the supplier would get P0 in additional revenue and incur an additional cost that is less than P_0 in producing this unit. In fact, on every unit between q_1 and q_0 he can make a profit, because the MR exceeds the associated cost, MC. By the same argument, it makes no sense to increase output beyond q_0, to q_2 for example, because the cost of such additional units of output, MC, exceeds the revenue from them. *The MC therefore defines an optimal supply response.*

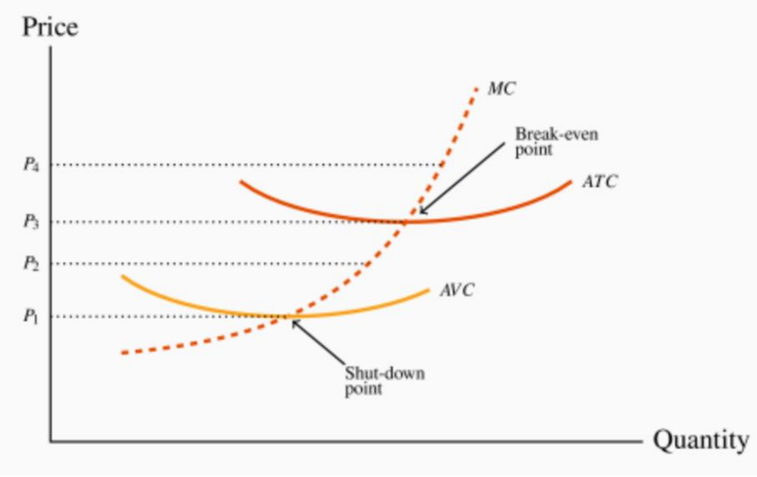
Figure 9.1 The competitive firm's optimal output



Here, q_0 *represents the optimal supply decision when the price is* P_0*. At output* q_1 *the cost of additional units is less than the revenue from such units and therefore it is profitable to increase output beyond* q_1*. Conversely, at* q_2*the* MC *of production exceeds the revenue obtained, and so output should be reduced.*

While the choice of the output q0 is the best choice for the producer, Figure 9.1 does not tell us anything about profit. To determine profit we need information on costs. Accordingly, in Figure 9.2 the firm's AVC and ATC curves have been added to Figure 9.1. As explained in the previous chapter, the ATC curve includes both fixed and variable cost components, and the MC curve cuts the AVC and the ATC at their minima.

Figure 9.2 Short-run supply for the competitive firm

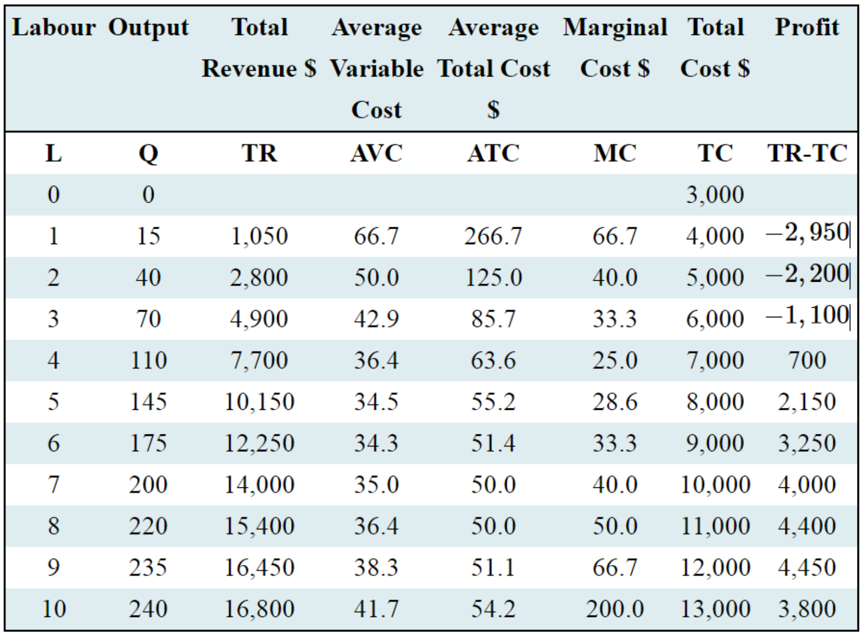


A price below P_1 *does not cover variable costs, so the firm should shut down. Between prices* P_1 *and* P_3*, the producer can cover variable, but not total, costs and therefore should produce in the short run if fixed costs are 'sunk'. In the long run the firm must close if the price does not reach* P_3*. Profits are made if the price exceeds* P_3*. The short-run supply curve is the quantity supplied at each price. It is therefore the* MC *curve above* P_1.

First, note that any price below P3, which corresponds to the minimum of the ATC curve, yields no profit, since it does not enable the producer to cover all of his costs. This price is therefore called the **break-even price**. Second, any price below P1, which corresponds to the minimum of the AVC, does not even enable the producer to cover variable costs. What about a price such as P2, that lies between these? The answer is that, if the supplier has already incurred some fixed costs, he should continue to produce, provided he can cover his variable cost. But in the long run he must cover all of his costs, fixed and variable. Therefore, if the price falls below P1, he should shut down, even in the short run. This price is therefore called the **shut-down price**. If a price at least equal to P3 cannot be sustained in the long run, he should leave the industry. But at a price such as P2 he can cover variable costs and therefore should continue to produce in the short run. His optimal output at P2 is defined by the intersection of the P2 line with the MC curve. The firm's short-run supply curve is, therefore, that portion of the MC curve above the minimum of the AVC.

To illustrate this more concretely, consider again the example of our snowboard producer, and imagine that he is producing in a perfectly competitive marketplace. How should he behave in response to different prices? Table 9.1 reproduces the data from Table 8.2.

Table 9.1 Profit maximization in the short run



Labour	Output	Total Revenue $	Average Variable Cost	Average Total Cost $	Marginal Cost $	Total Cost $	Profit
L	**Q**	**TR**	**AVC**	**ATC**	**MC**	**TC**	**TR-TC**
0	0					3,000	
1	15	1,050	66.7	266.7	66.7	4,000	−2,950
2	40	2,800	50.0	125.0	40.0	5,000	−2,200
3	70	4,900	42.9	85.7	33.3	6,000	−1,100
4	110	7,700	36.4	63.6	25.0	7,000	700
5	145	10,150	34.5	55.2	28.6	8,000	2,150
6	175	12,250	34.3	51.4	33.3	9,000	3,250
7	200	14,000	35.0	50.0	40.0	10,000	4,000
8	220	15,400	36.4	50.0	50.0	11,000	4,400
9	235	16,450	38.3	51.1	66.7	12,000	4,450
10	240	16,800	41.7	54.2	200.0	13,000	3,800

Output Price=$70; Wage=$1,000; Fixed Cost=$3,000. The shut-down point occurs at a price of $34.3, *where the* AVC *attains a minimum. Hence no production, even in the short run, takes place unless the price exceeds this value. The break-even level of output occurs at a price of* $50, *where the* ATC *attains a minimum.*

The **shut-down price** corresponds to the minimum value of the AVC curve.

The **break-even price** corresponds to the minimum of the ATC curve.

The firm's **short-run supply curve** is that portion of the MC curve above the minimum of the AVC.

Suppose that the price is $70. How many boards should he produce? The answer is defined by the behaviour of the MC curve. For any output less than or equal to 235, the MC is less than the price. For example, at L=9 and Q=235, the MC is $66.7. At this output level, he makes a profit on the marginal unit produced, because the MC is less than the revenue he gets ($70) from selling it.

But, at outputs above this, he registers a loss on the marginal units because the MC exceeds the revenue. For example, at L=10 and Q=240, the MC is $200. Clearly, 235 snowboards is the optimum. To produce more would generate a loss on each additional unit, because the additional cost would exceed the additional revenue. Furthermore, to produce fewer snowboards would mean not availing of the potential for profit on additional boards.

His profit is based on the difference between revenue per unit and cost per unit at this output: (P−ATC). Since the ATC for the 235 units produced by the nine workers is $51.1, his profit margin is $70−$51.1=$18.9 per board, and total profit is therefore 235×$18.9=$4,441.5.

Let us establish two other key outputs and prices for the producer. First, the shut-down point is the minimum of his AVC curve. Table 9.1 indicates that the price must be at least $34.3 for him to be willing to supply any output, since that is the value of the AVC at its minimum. Second, the minimum of his ATC is at $50. Accordingly, provided the price exceeds $50, he will cover both variable and fixed costs and make a maximum profit when he chooses an output where P=MC, above P=$50. It follows that the short-run supply curve for Black Diamond Snowboards is the segment of the MC curve in Figure 8.4 above the AVC curve.

Given that we have developed the individual firm's supply curve, the next task is to develop the industry supply curve.

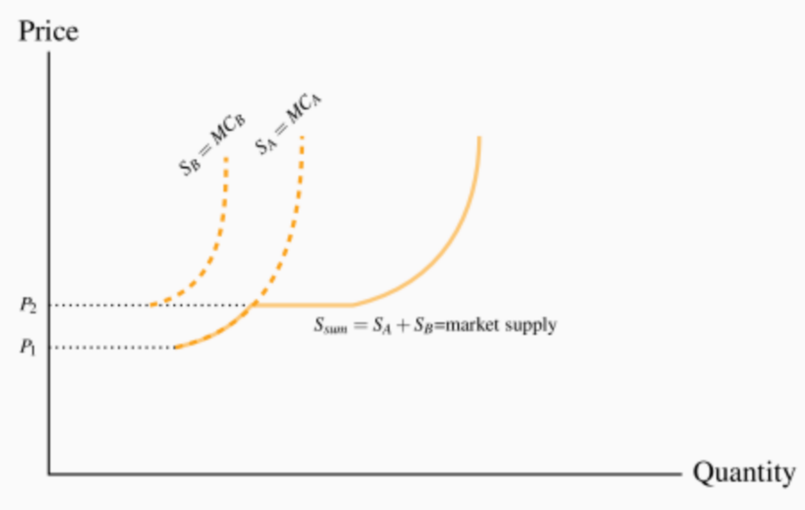
Industry Supply in the Short Run

In Chapter 3 it was demonstrated that individual demands can be aggregated into an industry demand by summing them horizontally. The industry supply is obtained in exactly the same manner—by summing the firms' supply quantities across all firms in the industry.

To illustrate, imagine we have many firms, possibly operating at different scales of output and therefore having different short-run MC curves. The MC curves of two of these firms are illustrated in Figure 9.3. The MC of A is below the MC of B; therefore, B likely has a smaller scale of plant than A. Consider first the supply decisions in the price range P1 to P2. At any price between these limits, only firm A will supply output – firm B does not cover its AVC in this price range. Therefore, the joint contribution to industry supply of firms A and B is given by the MC curve of firm A. But once a price of P2 is attained, firm B is now willing to supply. The Ssum schedule is the horizontal addition of their supply quantities. Adding the supplies of every firm in the industry in this way yields the industry supply.

Industry supply (short run) in perfect competition is the horizontal sum of all firms' supply curves.

Figure 9.3 Deriving industry supply

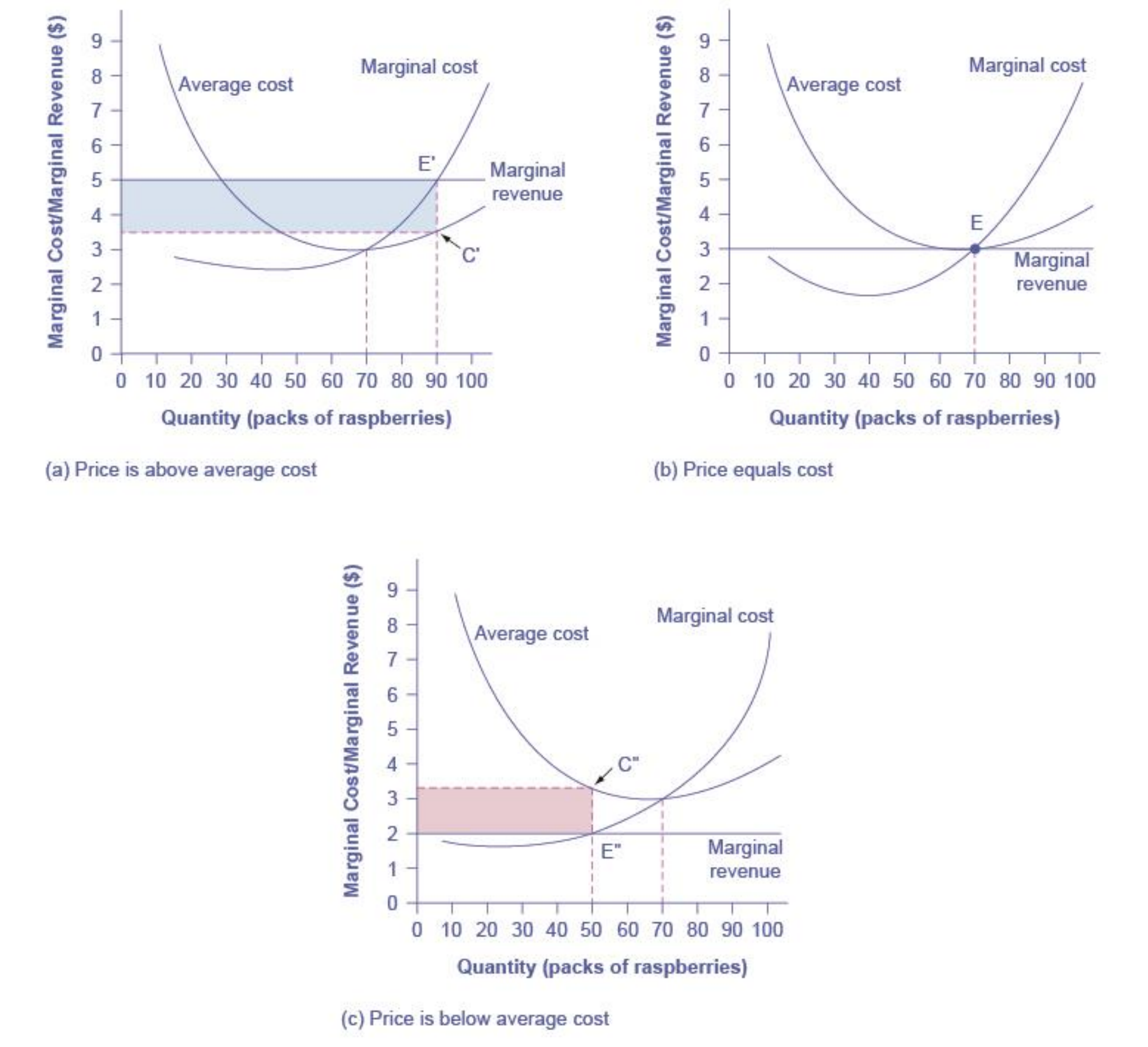


The market supply curve S *is the sum of each firm's supply or* MC *curve above the shut-down price.* D *is the sum of individual demands. The market equilibrium price and quantity are defined by* PE *and* QE.

Profits and Losses with the Average Cost Curve

Source: Lynham, 2018, Section 8.2, CC-BY 4.0 (Original source, paragraph 15, Figure 4 & Figure 6 only)

Does maximizing profit (producing where MR = MC) imply an actual economic profit? The answer depends on the relationship between price and average total cost. If the price that a firm charges is higher than its average cost of production for that quantity produced, then the firm will earn profits. Conversely, if the price that a firm charges is lower than its average cost of production, the firm will suffer losses. You might think that, in this situation, the farmer may want to shut down immediately. Remember, however, that the firm has already paid for fixed costs, such as equipment, so it may continue to produce and incur a loss. Figure 4 illustrates three situations: (a) where price intersects marginal cost at a level above the average cost curve, (b) where price intersects marginal cost at a level equal to the average cost curve, and (c) where price intersects marginal cost at a level below the average cost curve.



(a) Price is above average cost

(b) Price equals cost

(c) Price is below average cost

Figure 4. *Price and Average Cost at the Raspberry Farm. In (a), price intersects marginal cost above the average cost curve. Since price is greater than average cost, the firm is making a profit. In (b), price intersects marginal cost at the minimum point of the average cost curve. Since price is equal to average cost, the firm is breaking even. In (c), price intersects marginal cost below the average cost curve. Since price is less than average cost, the firm is making a loss.*

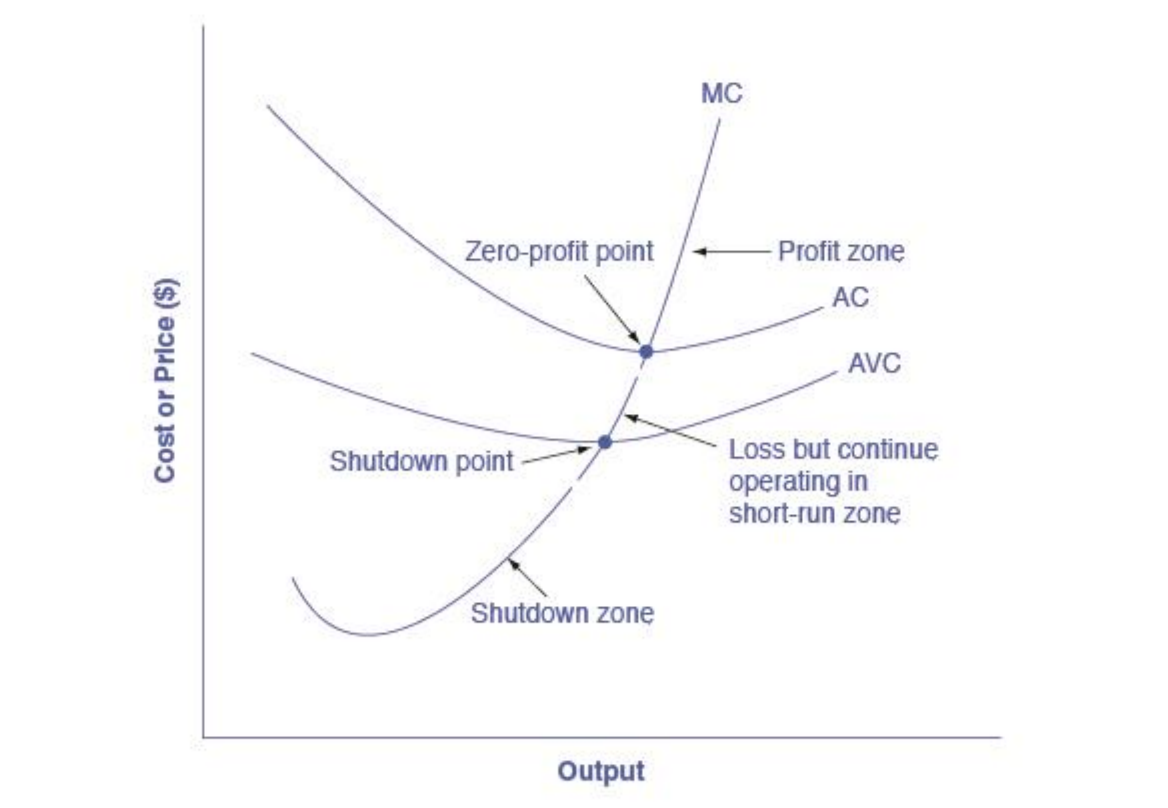


Figure 6. *Profit, Loss, Shutdown. The marginal cost curve can be divided into three zones, based on where it is crossed by the average cost and average variable cost curves. The point where MC crosses AC is called the zero-profit point. If the firm is operating at a level of output where the market price is at a level higher than the zero-profit point, then price will be greater than average cost and the firm is earning profits. If the price is exactly at the zero-profit point, then the firm is making zero profits. If price falls in the zone between the shutdown point and the zero-profit point, then the firm is making losses but will continue to operate in the short run, since it is covering its variable costs. However, if price falls below the price at the shutdown point, then the firm will shut down immediately, since it is not even covering its variable costs.*

9.4 DECISIONS IN THE LONG RUN

Source: Curtis & Irvine, 2016, Section 9.3, CC-BY-NC-SA 3.0 (Original source, page 2 and Application Box 9.2 removed)

Dynamics: Entry and Exit

We have now described the market and firm-level equilibrium in the short run. However, this equilibrium may be only temporary; whether it can be sustained or not depends upon whether profits (or losses) are being incurred, or whether all participant firms are making what are termed normal profits. Such profits are considered an essential part of a firm's operation. They reflect the opportunity cost of the resources used in production. Firms do not operate if they cannot make a minimal, or normal, profit level. Above such profits are **economic profits** (also called **supernormal profits**), and these are what entice entry into the industry.

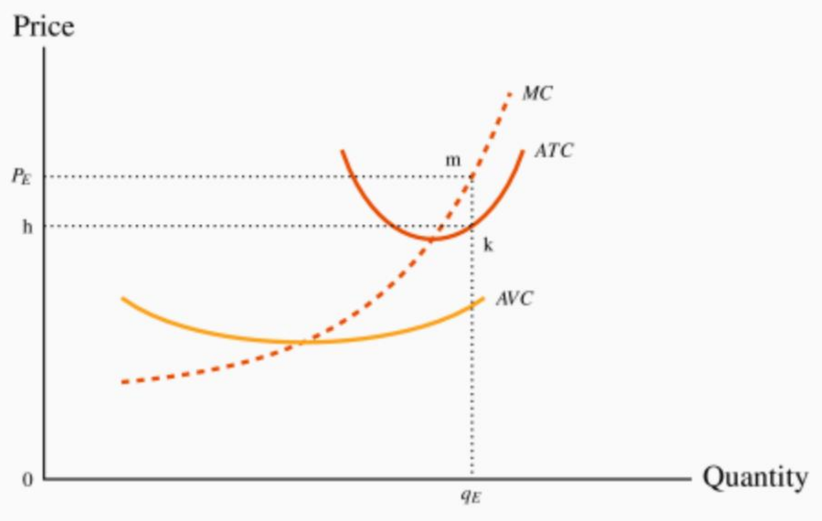
Recall from Chapter 7 that accounting and economic profits are different. The economist includes opportunity costs in determining profit, whereas the accountant considers actual revenues and costs. In the example developed in Section 7.2 the entrepreneur recorded accounting profit, but not economic profit. Suppose now that the numbers were slightly different, and are as defined in Table 9.2: Felicity invests $250,000 in her business in the form of capital, as before. But she now has gross revenues of $165,000 and incurs a cost of $90,000 to buy the clothing wholesale that she then sells retail. She pays herself a salary of $35,000. If these numbers represent her balance sheet, then she records an accounting profit of $40,000.

Table 9.2 Economic profits

Sales		$165,000
Materials costs		$90,000
Wage costs		$35,000
Accounting profit		**$40,000**
Capital invested	$250,000	
Implicit return on capital at 4%		$10,000
Additional implicit wage costs		$20,000
Total implicit costs		**$30,000**
Economic profit		**$10,000**

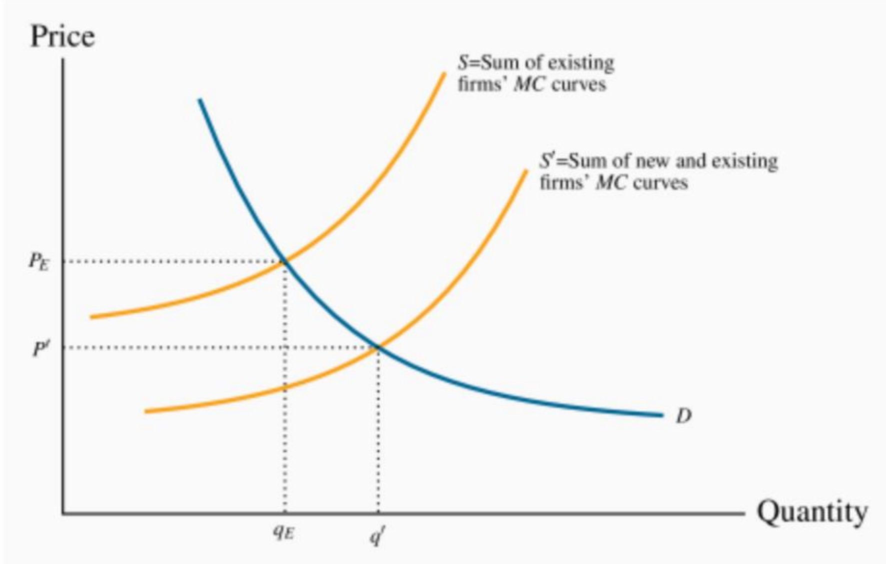
Let us return to our graphical analysis, and begin by supposing that the market equilibrium described in Figure 9.4 results in profits being made by some firms. Such an outcome is described in Figure 9.5, where the price exceeds the ATC. At the price PE, a profit-making firm supplies the quantity qE, as determined by its MC curve. On average, the cost of producing each unit of output, qE, is defined by the point on the ATC at that output level, point k. Profit per unit is thus given by the value (m−k) – the difference between revenue per unit and cost per unit. Total (economic) profit is therefore the area PEmkh, which is quantity times profit per unit.

Figure 9.5 Short-run profits for the firm



At the price P_E*, determined by the intersection of market demand and market supply, an individual firm produces the amount* Q_E*. The* ATC *of this output is k and therefore profit per unit is mk. Total profit is therefore*
$P_E mkh = 0q_E \times mk = TR - TC$.

Figure 9.6 Entry of firms due to economic profits



If economic profits result from the price P_E *new firms enter the industry. This entry increases the market supply to* S' *and the equilibrium price falls to* P'*. Entry continues as long as economic profits are present. Eventually the price is driven to a level where only normal profits are made, and entry ceases.*

While qE represents an equilibrium for the firm, it is only a short-run, or temporary, equilibrium for the industry. The assumption of free entry and exit implies that the presence of economic profits will induce new entrepreneurs to enter and start producing. The impact of this dynamic is illustrated in Figure 9.6. An increased number of firms shifts supply rightwards to become S′, thereby increasing the amount supplied at any price. The impact on price of this supply shift is evident: With an unchanged demand, the equilibrium price must fall.

How far will the price fall, and how many new firms will enter this profitable industry? As long as economic profits exist new firms will enter and the resulting increase in supply will continue to drive the price downwards. But, once the price has been driven down to the minimum of the ATC of a representative firm, there is no longer an incentive for new entrepreneurs to enter. Therefore, the **long-run industry equilibrium** is where the market price equals the minimum point of a firm's ATC curve. This generates normal profits, and there is no incentive for firms to enter or exit.

A **long-run equilibrium** in a competitive industry requires a price equal to the minimum point of a firm's ATC. At this point, only normal profits exist, and there is no incentive for firms to enter or exit.

In developing this dynamic, we began with a situation in which economic profits were present. However, we could have equally started from a position of losses. With a market price between the minimum of the AVC and the minimum of the ATC in Figure 9.5, revenues per unit would exceed variable costs but not total costs per unit. When firms cannot cover their ATC in the long run, they will cease production. Such closures must reduce aggregate supply; consequently the market supply curve contracts, rather than expands as it did in Figure 9.6. The reduced supply drives up the price of the good. This process continues as long as firms are making losses. A final industry equilibrium is attained only when the price reaches a level where firms can make a normal profit. Again, this will be at the minimum of the typical firm's ATC.

Accordingly, the long-run equilibrium is the same, regardless of whether we begin from a position in which firms are incurring losses, or where they are making profits.

CHAPTER 10: Monopoly, Cartels, and Price Discrimination

Source: Curtis & Irvine, 2016, Section 10.2, CC-BY-NC-SA 3.0 (Original source, page 4/Figure 10.4 removed)

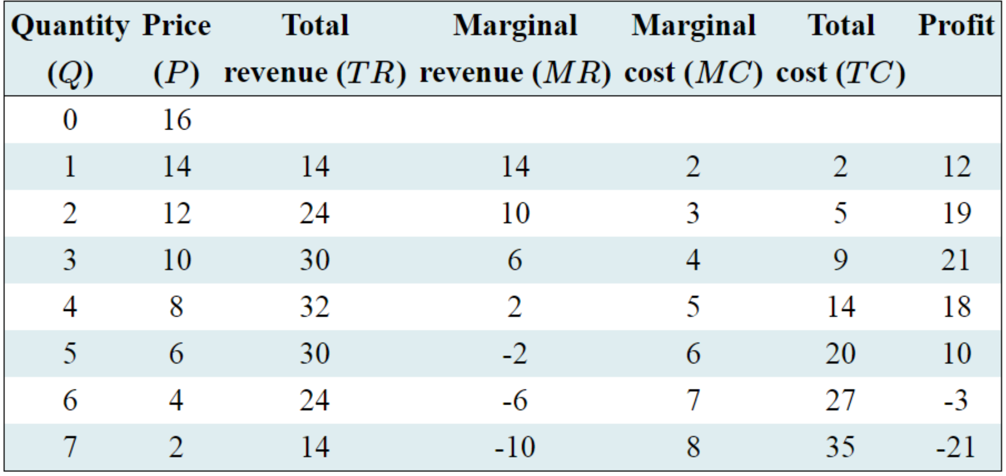
10.1 MONOPOLY

Profit Maximization

We established in the previous chapter that, in deciding upon a profit-maximizing output, any firm should produce up to the point where the additional cost equals the additional revenue from a unit of output. What distinguishes the supply decision for a monopolist from the supply decision of the perfect competitor is that the monopolist faces a downward sloping demand. A monopolist is the sole supplier and therefore must meet the full market demand. This means that if more output is produced, the price must fall. We will illustrate the choice of a profit maximizing output using first a marginal-cost/marginal-revenue approach; then a supply/demand approach.

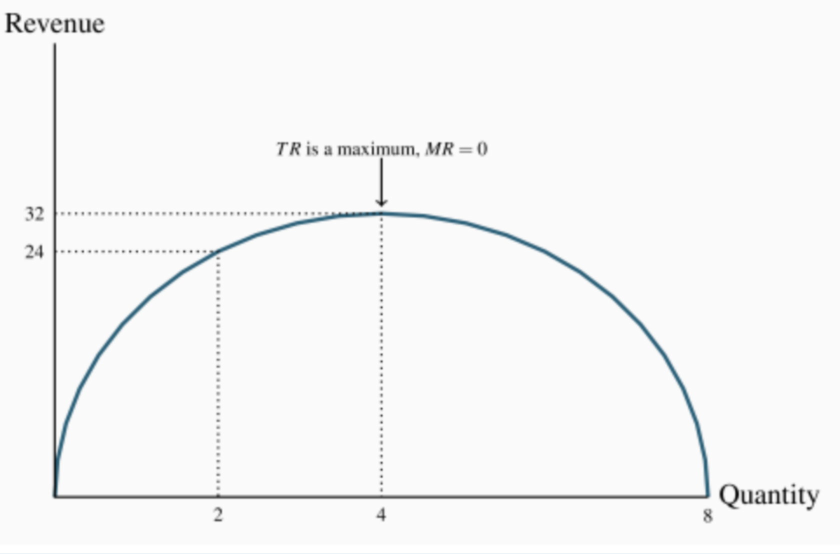
Marginal Revenue and Marginal Cost Table 10.1 displays price and quantity values for a demand curve in columns 1 and 2. Column three contains the sales revenue generated at each output. It is the product of price and quantity. Since the price denotes the revenue per unit, it is sometimes referred to as average revenue. The total revenue (TR) reaches a maximum at $32, where 4 units of output are produced. A greater output necessitates a lower price on every unit sold, and in this case revenue falls if the fifth unit is brought to the market. Even though the fifth unit sells for a positive price, the price on the other 4 units is now lower and the net effect is to reduce total revenue. This pattern reflects what we examined in Chapter 4: As price is lowered from the highest possible value of $14 (where 1 unit is demanded) and the corresponding quantity increases, revenue rises, peaks, and ultimately falls as output increases. In Chapter 4 we explained that this maximum revenue point occurs where the price elasticity is unity (-1), at the midpoint of a linear demand curve.

Table 10.1 A profit maximizing monopolist



Quantity (Q)	Price (P)	Total revenue (TR)	Marginal revenue (MR)	Marginal cost (MC)	Total cost (TC)	Profit
0	16					
1	14	14	14	2	2	12
2	12	24	10	3	5	19
3	10	30	6	4	9	21
4	8	32	2	5	14	18
5	6	30	-2	6	20	10
6	4	24	-6	7	27	-3
7	2	14	-10	8	35	-21

Figure 10.3 Total revenue and marginal revenue



When the quantity sold increases total revenue/expenditure initially increases also. At a certain point, further sales require a price that not only increases quantity, but reduces revenue on units already being sold to such a degree that TR *declines – where the demand elasticity equals* –1 *(the mid point of a linear demand curve). Here the midpoint occurs at* Q=4. *Where the* TR *is a maximum the* MR=0.

Related to the total revenue function is the **marginal revenue** function. It is the addition to total revenue due to the sale of one more unit of the commodity.

Marginal revenue is the change in total revenue due to selling one more unit of the good.

Average revenue is the price per unit sold.

The MR in this example is defined in the fourth column of Table 10.1. When the quantity sold increases from 1 unit to 2 units total revenue increases from $14 to $24. Therefore the marginal revenue associated with the second unit of output is $10. When a third unit is sold TR increases to $30 and therefore the MR of the third unit is $6. As output increases the MR declines and eventually becomes negative – at the point where the TR is a maximum: If TR begins to decline then the additional revenue is by definition negative.

The Optimal Output This producer has a marginal cost structure given in the fifth column of the table, and this too is plotted in Figure 10.4. Our profit maximizing rule from Chapter 8 states that it is optimal to produce a greater output as long as the additional revenue exceeds the additional cost of production on the next unit of output. In perfectly competitive markets the additional revenue is given by the fixed price for the individual producer, whereas for the monopolist the additional revenue is the marginal revenue. Consequently as long as MR exceeds MC for the next unit a greater output is profitable, but once MC exceeds MR the production of additional units should cease.

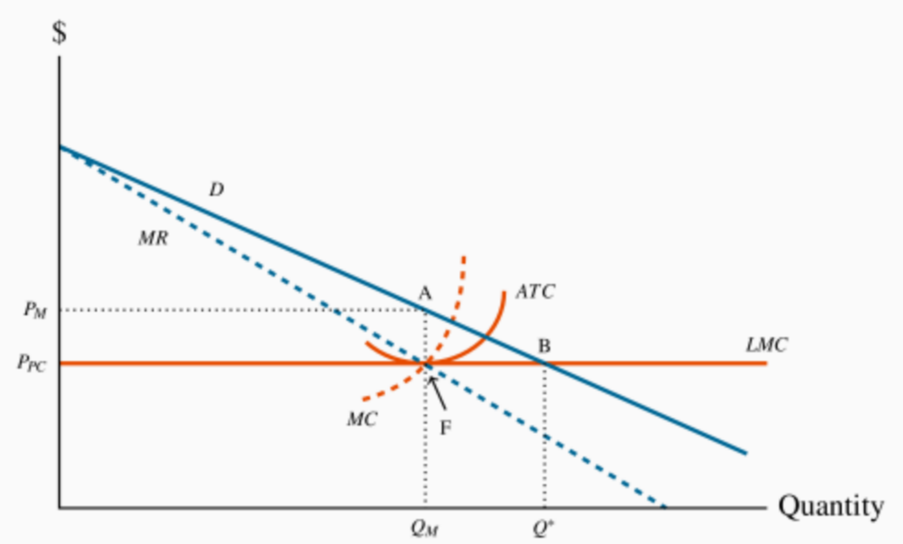
From Table 10.1 and Figure 10.4 [removed] it is clear that the optimal output is at 3 units. The third unit itself yields a profit of 2$, the difference between MR ($6) and MC ($4). A fourth unit however would reduce profit by $3, because the MR ($2) is less than the MC ($5). What price should the producer charge? The price, as always, is given by the demand function. At a quantity sold of 3 units, the corresponding price is $10, yielding total revenue of $30.

Profit is the difference between total revenue and total cost. In Chapter 8 we computed total cost as the average cost times the number of units produced. It can also be computed as the sum of costs associated with each unit produced: The first unit costs $2, the second $3 and the third $4. The total cost of producing 3 units is the sum of these dollar values: $9=$2+$3+$4. The profit-maximizing output therefore yields a profit of $21 ($30−$9).

Source: Curtis & Irvine, 2016, Section 10.4, CC-BY-NC-SA 3.0 (Original source, last paragraph & footnote removed)

Output Inefficiency A characteristic of perfect competition is that it secures an efficient allocation of resources when there are no externalities in the market: Resources are used up to the point where their marginal cost equals their marginal value – as measured by the price that consumers are willing to pay. But a monopoly structure does not yield this output. Consider Figure 10.10.

Figure 10.10 Monopoly output inefficiency



A monopolist maximizes profit at QM*. Here the value of marginal output exceeds cost. If output expands to* Q* *a gain arises equal to the area ABF. This is the deadweight loss associated with the output* QM *rather than* Q**. If the monopolist's long-run* MC *is equivalent to a competitive industry's supply curve, then the deadweight loss is the cost of having a monopoly rather than a perfectly competitive market.*

The monopolist's profit-maximizing output QM is where MC equals MR. This output is inefficient for the reason that we developed in Chapter 5: If output is increased beyond QM the additional benefit exceeds the additional cost of producing it. The additional benefit is measured by the willingness of buyers to pay – the market demand curve. The additional cost is the long-run MC curve under the assumption of constant returns to scale. Using the terminology from Chapter 5, there is a deadweight loss equal to the area ABF. This is termed **allocative inefficiency.**

Allocative inefficiency arises when resources are not appropriately allocated and result in deadweight losses.

Perfect competition versus monopoly The area ABF can also be considered as the efficiency loss associated with having a monopoly rather than a perfectly competitive market structure. In perfect competition the supply curve is horizontal. This is achieved by having firms enter and exit when more or less must be produced. Accordingly, if the perfectly competitive industry's supply curve approximates the monopolist's long-run marginal cost curve, we can say that if the monopoly were turned into a competitive industry, output would increase from QM to Q∗. The deadweight loss is one measure of the superiority of the perfectly competitive structure over the monopoly structure.

Note that this critique of monopoly is not initially focused upon profit. While monopoly profits are what frequently irk the public, we have focused upon resource allocation inefficiencies. But in a real sense the two are related: Monopoly inefficiencies arise through output being restricted, and it is this output reduction – achieved by maintaining a higher than competitive price – that gives rise to those profits. Nonetheless, there is more than just a shift in purchasing power from the buyer to the seller. Deadweight losses arise because output is at a level lower than the point where the MC equals the value placed on the good; thus the economy is sacrificing the possibility of creating additional surplus.

How Monopolies Form: Barriers to Entry

Source: Lynham, 2018, Section 9.1, CC-BY 4.0 (Original source, paragraphs 1-10 only)

Because of the lack of competition, monopolies tend to earn significant economic profits. These profits should attract vigorous competition as described in Perfect Competition, and yet, because of one particular characteristic of monopoly, they do not. **Barriers to entry** are the legal, technological, or market forces that discourage or prevent potential competitors from entering a market. Barriers to entry can range from the simple and easily surmountable, such as the cost of renting retail space, to the extremely restrictive. For example, there are a finite number of radio frequencies available for broadcasting. Once the rights to all of them have been purchased, no new competitors can enter the market.

In some cases, barriers to entry may lead to monopoly. In other cases, they may limit competition to a few firms. Barriers may block entry even if the firm or firms currently in the market are earning profits. Thus, in markets with significant barriers to entry, it is not true that abnormally high profits will attract new firms, and that this entry of new firms will eventually cause the price to decline so that surviving firms earn only a normal level of profit in the long run.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

There are two types of monopoly, based on the types of barriers to entry they exploit. One is **natural monopoly**, where the barriers to entry are something other than legal prohibition. The other is **legal monopoly**, where laws prohibit (or severely limit) competition.

Natural Monopoly Economies of scale can combine with the size of the market to limit competition. (This theme was introduced in Cost and Industry Structure). Figure 1 presents a long-run average cost curve for the airplane manufacturing industry. It shows economies of scale up to an output of 8,000 planes per year and a price of P0, then constant returns to scale from 8,000 to 20,000 planes per year, and diseconomies of scale at a quantity of production greater than 20,000 planes per year.

Now consider the market demand curve in the diagram, which intersects the long-run average cost (LRAC) curve at an output level of 6,000 planes per year and at a price P1, which is higher than P0. In this situation, the market has room for only one producer. If a second firm attempts to enter the market at a smaller size, say by producing a quantity of 4,000 planes, then its average costs will be higher than the existing firm, and it will be unable to compete. If the second firm attempts to enter the market at a larger size, like 8,000 planes per year, then it could produce at a lower average cost—but it could not sell all 8,000 planes that it produced because of insufficient demand in the market.

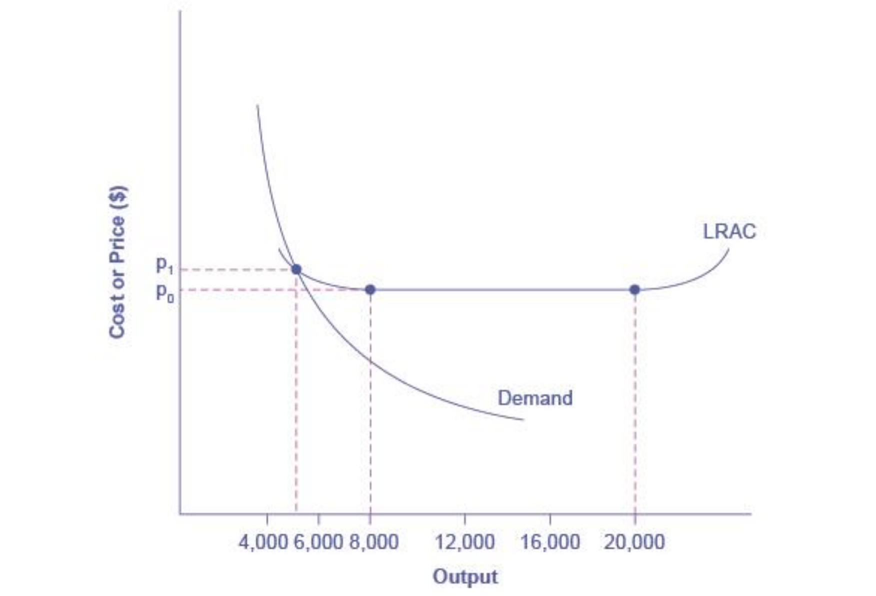


Figure 1. *Economies of Scale and Natural Monopoly. In this market, the demand curve intersects the long-run average cost (LRAC) curve at its downward-sloping part. A natural monopoly occurs when the quantity demanded is less than the minimum quantity it takes to be at the bottom of the long-run average cost curve.*

The graph represents a natural monopoly as evidenced by the demand curve intersecting with the downward-sloping part of the LRAC curve.

Figure 1. Economies of Scale and Natural Monopoly. In this market, the demand curve intersects the long-run average cost (LRAC) curve at its downward-sloping part. A natural monopoly occurs when the quantity demanded is less than the minimum quantity it takes to be at the bottom of the long-run average cost curve.

This situation, when economies of scale are large relative to the quantity demanded in the market, is called a natural monopoly. Natural monopolies often arise in industries where the marginal cost of adding an additional customer is very low, once the fixed costs of the overall system are in place. Once the main water pipes are laid through a neighborhood, the marginal cost of providing water service to another home is fairly low. Once electricity lines are installed through a neighborhood, the marginal cost of providing additional electrical service to one more home is very low. It would be costly and duplicative for a second water company to enter the market and invest in a whole second set of main water pipes, or for a second electricity company to enter the market and invest in a whole new set of electrical wires. These industries offer an example where, because of economies of scale, one producer can serve the entire market more efficiently than a number of smaller producers that would need to make duplicate physical capital investments.

A natural monopoly can also arise in smaller local markets for products that are difficult to transport. For example, cement production exhibits economies of scale, and the quantity of cement demanded in a local area may not be much larger than what a single plant can produce. Moreover, the costs of transporting cement over land are high, and so a cement plant in an area without access to water transportation may be a natural monopoly.

Control of a Physical Resource Another type of natural monopoly occurs when a company has control of a scarce physical resource. In the U.S. economy, one historical example of this pattern occurred when ALCOA—the Aluminum Company of America—controlled most of the supply of bauxite, a key mineral used in making aluminum. Back in the 1930s, when ALCOA controlled most of the bauxite, other firms were simply unable to produce enough aluminum to compete.

As another example, the majority of global diamond production is controlled by DeBeers, a multi-national company that has mining and production operations in South Africa, Botswana, Namibia, and Canada. It also has exploration activities on four continents, while directing a worldwide distribution network of rough cut diamonds. Though in recent years they have experienced growing competition, their impact on the rough diamond market is still considerable.

Legal Monopoly For some products, the government erects barriers to entry by prohibiting or limiting competition. Under U.S. law, no organization but the U.S. Postal Service is legally allowed to deliver first-class mail. Many states or cities have laws or regulations that allow households a choice of only one electric company, one water company, and one company to pick up the garbage. Most legal monopolies are considered utilities—products necessary for everyday life—that are socially beneficial to have. As a consequence, the government allows producers to become regulated monopolies, to insure that an appropriate amount of these products is provided to consumers. Additionally, legal monopolies are often subject to economies of scale, so it makes sense to allow only one provider.

10.2 CARTELS: ACTING LIKE A MONOPOLIST

Source: Curtis & Irvine, 2016, Section 10.4, CC-BY-NC-SA 3.0 (Original source, Application Box 10.1 removed)

A **cartel** is a group of suppliers that colludes to operate like a monopolist. The cartel formed by the members of the Organization of Oil Exporting Countries (OPEC) is an example of a cartel that was successful in achieving its objectives for a long period. This cartel first flexed its muscles in 1973, by increasing the world price of oil from $3 per barrel to $10 per barrel. The result was to transfer billions of dollars from the energy-importing nations in Europe and North America to OPEC members – the demand for oil is relatively inelastic, hence an increase in price increases total expenditures.

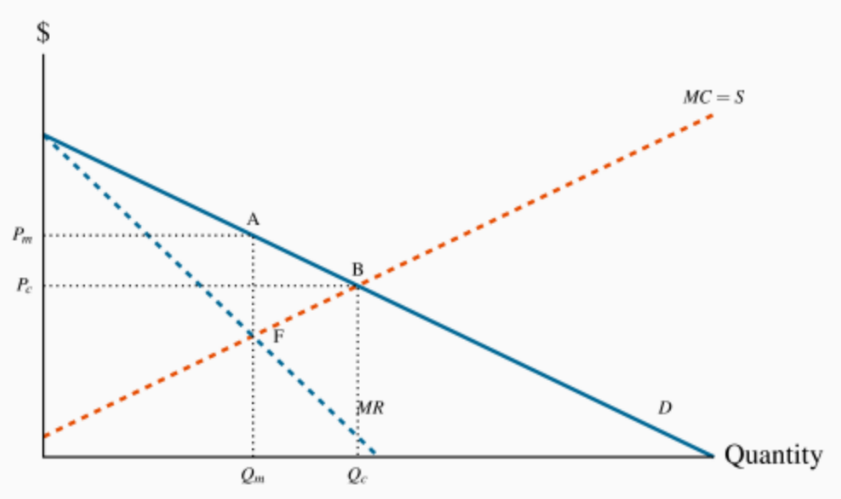
A **cartel** is a group of suppliers that colludes to operate like a monopolist.

A second renowned cartel is managed by De Beers, which controls a large part of the world's diamond supply. A third is Major League Baseball in the US. Lesser-known cartels, but in some cases very effective, are those formed by the holders of taxi licenses in many cities throughout the world, and by agricultural marketing boards in many developed economies. These cartels may have thousands of members. By limiting entry, through requiring a production 'quota' or a license (taxi medallion), the incumbents can charge a higher price than if entry to the industry were free.

Some cartels are sustained through violence, and frequently wars break out between competing cartels or groups who want to sustain their market power. Drug gangs frequently fight for hegemony over distribution. The recent drug wars in Mexico between rival cartels have seen tens of thousands on individuals killed as of 2015.

To illustrate the dynamics of cartels consider Figure 10.14. Several producers, with given production capacities, come together and agree to restrict output with a view to increasing price and therefore profit. This may be done with the agreement of the government, or it may be done secretively, and possibly against the law. Each firm has a MC curve, and the industry supply is defined as the sum of these marginal cost curves, as illustrated in Figure 9.3. The resulting cartel is effectively one in which there is a single supplier with many different plants – a multi-plant monopolist. To maximize profits this organization will choose an output level Qm where the MR equals the MC. In contrast, if these firms act competitively the output chosen will be Qc. The competitive output yields no supernormal profit, whereas the monopoly/cartel output does.

Figure 10.14 Cartelizing a competitive industry



A cartel is formed when individual suppliers come together and act like a monopolist in order to increase profit. If MC *is the joint supply curve of the cartel, profits are maximized at the output* Q_m*, where* MC=MR*. In contrast, if these firms operate competitively output increases to* Q_c.

The cartel results in a deadweight loss equal to the area ABF, just as in the standard monopoly model.

Cartel instability Cartels tend to be unstable in the long run for a number of reasons.

In the first instance, the degree of instability depends on the authority that the governing body of the cartel can exercise over its members, and upon the degree of information it has on the operations of its members. If a cartel is simply an arrangement among producers to limit output, each individual member of the cartel has an incentive to increase its output, because the monopoly price that the cartel attempts to sustain exceeds the cost of producing a marginal unit of output. In Figure 10.14 each firm has a MC of output equal to $F when the group collectively produces the output Qm. Yet any firm that brings output to market, *beyond its agreed production limit*, at the price Pm will make a profit of AF on that additional output – *provided the other members of the cartel agree to restrict their output*. Since each firm faces the same incentive to increase output, it is difficult to restrain all members from doing so.

Individual members are more likely to abide by the cartel rules if the organization can sanction them for breaking the supply-restriction agreement. Alternatively, if the actions of individual members are not observable by the organization, then the incentive to break ranks may be too strong for the cartel to sustain its monopoly power.

Cartels within individual economies are almost universally illegal. Yet at the international level there exists no governing authority to limit such behaviour. In practice, governments are unwilling to see their own citizens and consumers being 'gouged', but are relatively unconcerned if their national or multinational corporations are willing and successful in gouging the consumers of other economies! We will see in Chapter 14 that Canada's

Competition Act forbids the formation of cartels, as it forbids many other anti-competitive practices.

In the second instance, cartels may be undermined eventually by the emergence of new products and new technologies. OPEC has lost much of its power in the modern era because of technological developments in oil recovery. Canada's 'tar sands' yield oil, as a result of technological developments that enabled producers to separate the oil from the earth it is mixed with. Fracking technologies are the latest means of extracting oil that is discovered in small pockets and encased in rock. These technologies do not enable oil to be produced as cheaply as in traditional oil wells, but they limit the degree to which cartels can raise prices.

10.3 PRICE DISCRIMINATION

Source: Curtis & Irvine, 2016, Section 10.5, CC-BY-NC-SA 3.0 (Original source, page 4 removed)

A common characteristic in the pricing of many goods is that different individuals pay different prices for goods or services that are essentially the same. Examples abound: Seniors get a reduced rate for coffee in Burger King; hair salons charge women more than they charge men; bank charges are frequently waived for juniors. **Price discrimination** involves charging different prices to different consumers in order to increase profit.

Price discrimination involves charging different prices to different consumers in order to increase profit.

A strict definition of discrimination involves different prices for identical products. We all know of a school friend who has been willing to take the midnight flight to make it home at school break at a price he can afford. In contrast, the business executive prefers the seven a.m. flight to arrive for a nine a.m. business meeting in the same city at several times the price. These are very mild forms of price discrimination, since a midnight flight (or a midday flight) is not a perfect substitute for an early morning flight. Price discrimination is practiced because buyers are willing to pay different amounts for a good or service, and the supplier may have a means of profiting from this. Consider the following example.

Family Flicks is the local movie theatre. It has two distinct groups of customers – those of prime age form one group; youth and seniors form the other. Family Flicks has done its market research and determined that each group accounts for 50 percent of the total market of 100 potential viewers per screening. It has also established that the prime-age group members are willing to pay $12 to see a movie, while the seniors and youth are willing to pay just $5. How should the tickets be priced?

Family Flicks has no variable costs, only fixed costs. It must pay a $100 royalty to the movie maker each time it shows the current movie, and must pay a cashier and usher $20 each. Total costs are therefore $140, regardless of how many people show up – short-run MC is zero. On the pricing front, as illustrated in Table 10.3 below, if Family Flicks charges $12 per ticket it will attract 50 viewers, generate $600 in revenue and therefore make a profit of $460.

Table 10.3 Price discrimination

	P=**$5**	P=**$12**	**Twin price**
No. of customers	100	50	
Total revenue	$500	$600	$850
Total costs	$140	$140	$140
Profit	$360	$460	$710

In contrast, if it charges $5 it can fill the theatre, because each of the prime-age individuals is willing to pay more than $5, but the seniors and youth are now offered a price they too are willing to pay. However, the total revenue is now only $500 (100×$5=$500), and profits are reduced to $360. It therefore decides to charge the high price and leave the theatre half-empty, because this strategy maximizes its profit.

Suppose finally that the theatre is able to segregate its customers. It can ask the young and senior customers for identification upon entry, and in this way charge them a lower price, *while still maintaining the higher price to the prime-age customers*. If it can execute such a plan Family Flicks can now generate $850 in revenue – $600 from the prime-age group and $250 from the youth and seniors groups. Profit soars to $710.

There are two important conditions for this scheme to work:

1. The seller must be able to *segregate the market* at a reasonable cost. In the movie case this is achieved by asking for identification.

2. The second condition is that *resale must be impossible or impractical*. For example, we rule out the opportunity for young buyers to resell their tickets to the prime-age individuals. Sellers have many ways of achieving this – they can require immediate entry to the movie theatre upon ticket purchase, they can stamp the customer's hand, they can demand the showing of ID with the ticket when entering the theatre area.

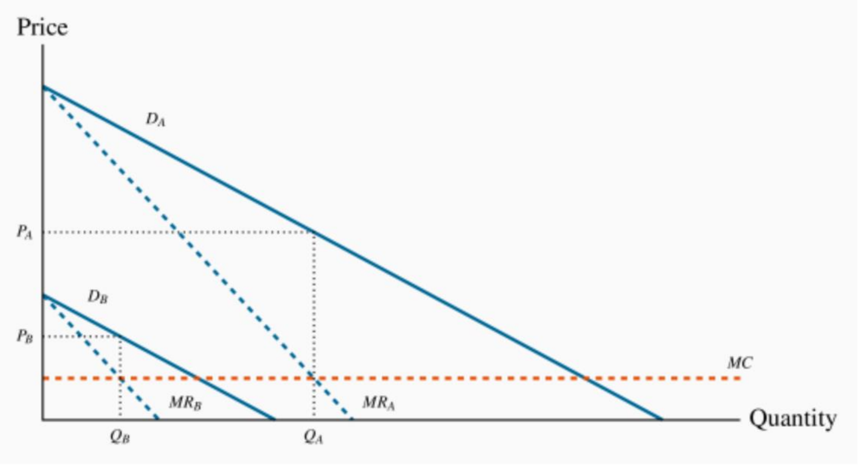
Frequently we think of sellers who offer price reductions to specific groups as being generous. For example, hotels may levy only a nominal fee for the presence of a child, once the parents have paid a suitable rate for the room or suite in which a family stays. The hotel knows that if it charges too much for the child, it may lose the whole family as a paying unit. The coffee shop offering cheap coffee to seniors is interested in getting a price that will cover its variable cost and so contribute to its profit. It is unlikely to be motivated by philanthropy, or to be concerned with the financial circumstances of seniors.

In our Family Flicks example, the profit maximizing monopolist that did not, or could not, price discriminate *left 50 customers unsupplied who were willing to pay $5 for a good that had a zero* MC. This is a deadweight loss of $250 because 50 seniors and youth valued a commodity at $5 that had a zero MC. Their demand was not met because, in the absence of an ability to discriminate between consumer groups, Family Flicks made more profit by satisfying the demand of the prime-age group alone. But in this example, by segregating its customers, the firm's profit maximization behaviour resulted in the DWL being eliminated, because it supplied the product to those additional 50 individuals. In this instance *price discrimination improves welfare*, because more of a good is supplied in a situation where market valuation exceeds marginal cost.

In the preceding example we simplified the demand side of the market by assuming that every individual in a given group was willing to pay the same price – either $12 or $5. More realistically each group can be defined by a downward-sloping demand curve, reflecting the variety of prices that buyers in a given market segment are willing to pay. It is valuable to extend the analysis to include this reality. For example, a supplier may face different demands from her domestic and foreign buyers, and if she can segment these markets she can price discriminate effectively.

Consider Figure 10.12 where two segmented demands are displayed, DA and DB, with their associated marginal revenue curves, MRA and MRB. We will assume that marginal costs are constant for the moment. It should be clear by this point that the profit maximizing solution for the monopoly supplier is to supply an amount to each market where the MC equals the MR in each market: Since the buyers in one market cannot resell to buyers in the other, the monopolist considers these as two different markets and therefore maximizes profit by applying the standard rule. She will maximize profit in market A by supplying the quantity QA and in market B by supplying QB. The prices at which these quantities can be sold are PA and PB. These prices, unsurprisingly, are different – the objective of segmenting markets is to increase profit by treating the markets as distinct.

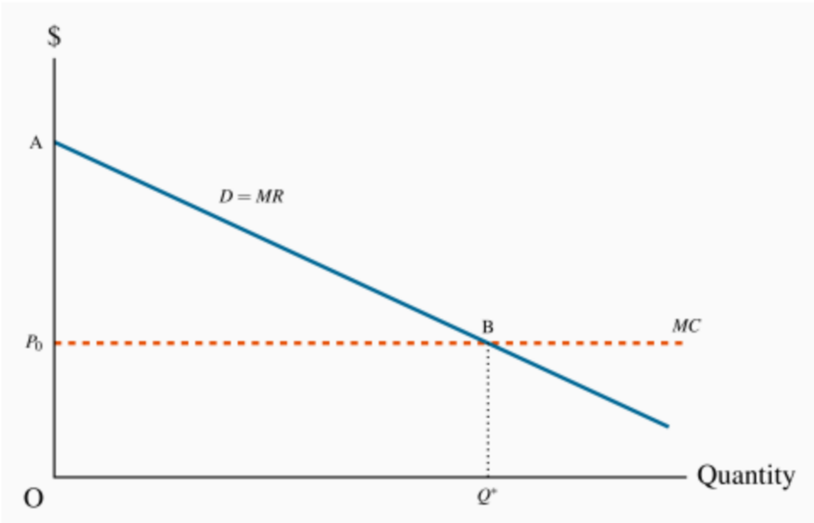
Figure 10.12 Pricing in segregated markets



With two separate markets defined by DA *and* DB*, and their associated* MR*curves* MRA *and* MRB*, a profit maximizing strategy is to produce where* MC=MRA=MRB*, and discriminate between the two markets by charging prices* PA *and* PB.

The preceding examples involved two separable groups of customers and are very real. This kind of group segregation is sometimes called *third degree price discrimination*. But it may be possible to segregate customers into several groups rather than just two. In the limit, if we could charge a different price to every consumer in a market, or for every unit sold, the revenue accruing to the monopolist would be the area under the demand curve up to the output sold. Though primarily of theoretical interest, this is illustrated in Figure 10.13. It is termed *perfect price discrimination*, and *sometimes first degree price discrimination*. Such discrimination is not so unrealistic: A tax accountant may charge different customers a different price for providing the same service; home renovators may try to charge as much as any client appears willing to pay.

Figure 10.13 Perfect price discrimination



A monopolist who can sell each unit at a different price maximizes profit by producing Q**. With each consumer paying a different price the demand curve becomes the* MR *curve. The result is that the monopoly DWL is eliminated because the efficient output is produced, and the monopolist appropriates all the consumer surplus. Total revenue for the perfect price discriminator is OAB*Q**.*

Second degree price discrimination is based on a different concept of buyer identifiability. In the cases we have developed above, the seller is able to distinguish the buyers by *observing* a vital characteristic that signals their type. It is also possible that, while individuals might have defining traits which influence their demands, such traits might not be detectable by the supplier. Nonetheless, it is frequently possible for the supplier to offer different pricing options (corresponding to different uses of a product) that buyers would choose from, with the result that her profit would be greater than under a uniform price with no variation in the use of the service. Different cell phone 'plans', or different internet plans that users can choose from are examples of this second-degree discrimination.

CHAPTER 11: Between the Extremes and Strategic Behaviour

11.1 IMPERFECT COMPETITION

The Four-Firm Concentration Ratio

Source: Lynham, 2018, Section 11.1, CC-BY 4.0 (Original source, paragraphs 7-9 only)

Regulators have struggled for decades to measure the degree of monopoly power in an industry. An early tool was the **concentration ratio**, which measures what share of the total sales in the industry are accounted for by the largest firms, typically the top four to eight firms. For an explanation of how high market concentrations can create inefficiencies in an economy, refer to Monopoly.

Say that the market for replacing broken automobile windshields in a certain city has 18 firms with the market shares shown in Table 1, where the **market share** is each firm's proportion of total sales in that market. The four-firm concentration ratio is calculated by adding the market shares of the four largest firms: in this case, 16 + 10 + 8 + 6 = 40. This concentration ratio would not be considered especially high, because the largest four firms have less than half the market.

If the market shares in the market for replacing automobile windshields are:

Smooth as Glass Repair Company	16% of the market
The Auto Glass Doctor Company	10% of the market
Your Car Shield Company	8% of the market
Seven firms that each have 6% of the market	42% of the market, combined
Eight firms that each have 3% of the market	24% of the market, combined

Then the four-firm concentration ratio is 16 + 10 + 8 + 6 = 40.

Table 1. Calculating Concentration Ratios from Market Shares

The concentration ratio approach can help to clarify some of the fuzziness over deciding when a merger might affect competition. For instance, if two of the smallest firms in the hypothetical market for repairing automobile windshields merged, the four-firm concentration ratio would not change—which implies that there is not much worry that the degree of competition in the market has notably diminished. However, if the top two

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

firms merged, then the four-firm concentration ratio would become 46 (that is, 26 + 8 + 6 + 6). While this concentration ratio is modestly higher, the four-firm concentration ratio would still be less than half, so such a proposed merger might barely raise an eyebrow among antitrust regulators.

Monopolistic Competition and Oligopoly

Source: Lynham, 2018, Chapter 10 Introduction, CC-BY 4.0 (Original source, 'The Temptation to Defy the Law' box removed)

Perfect competition and monopoly are at opposite ends of the competition spectrum. A perfectly competitive market has many firms selling identical products, who all act as price takers in the face of the competition. If you recall, **price takers** are firms that have no market power. They simply have to take the market price as given.

Monopoly arises when a single firm sells a product for which there are no close substitutes. Microsoft, for instance, has been considered a monopoly because of its domination of the operating systems market.

What about the vast majority of real world firms and organizations that fall between these extremes, firms that could be described as **imperfectly competitive**? What determines their behavior? They have more influence over the price they charge than perfectly competitive firms, but not as much as a monopoly would. What will they do?

One type of imperfectly competitive market is called **monopolistic competition**. Monopolistically competitive markets feature a large number of competing firms, but the products that they sell are not identical. Consider, as an example, the Mall of America in Minnesota, the largest shopping mall in the United States. In 2010, the Mall of America had 24 stores that sold women's "ready-to-wear" clothing (like Ann Taylor and Urban Outfitters), another 50 stores that sold clothing for both men and women (like Banana Republic, J. Crew, and Nordstrom's), plus 14 more stores that sold women's specialty clothing (like Motherhood Maternity and Victoria's Secret). Most of the markets that consumers encounter at the retail level are monopolistically competitive.

The other type of imperfectly competitive market is **oligopoly**. Oligopolistic markets are those dominated by a small number of firms. Commercial aircraft provides a good example: Boeing and Airbus each produce slightly less than 50% of the large commercial aircraft in the world. Another example is the U.S. soft drink industry, which is dominated by Coca-Cola and Pepsi. Oligopolies are characterized by high barriers to entry with firms choosing output, pricing, and other decisions strategically based on the decisions of the other firms in the market. In this chapter, we first explore how monopolistically competitive firms will choose their profit-maximizing level of output. We will then discuss oligopolistic firms, which face two conflicting temptations: to collaborate as if they were a single monopoly, or to individually compete to gain profits by expanding output levels and cutting prices. Oligopolistic markets and firms can also take on elements of monopoly and of perfect competition.

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

11.2 MONOPOLISTIC COMPETITION

Source: Curtis & Irvine, 2016, Section 11.3, CC-BY-NC-SA 3.0

Monopolistic competition presumes a large number of quite small producers or suppliers, each of whom may have a slightly **differentiated product**. The competition element of this name signifies that there are many participants, while the monopoly component signifies that each supplier faces a downward-sloping demand. In concrete terms, your local coffee shop that serves "fair trade" coffee has a product that differs slightly from that of neighbouring shops that sell the traditional product. They coexist in the same sector, and probably charge different prices: The fair trade supplier likely charges a higher price, but knows nonetheless that too large a difference between her price and the prices of her competitors will see some of her clientele migrate to those lower-priced establishments. That is to say, she faces a downward-sloping demand curve.

The competition part of the name also indicates that there is *free entry and exit*. There are no barriers to entry. As a consequence, we know at the outset that only normal profits will exist in a long-run equilibrium. Economic profits will be competed away by entry, just as losses will erode due to exit.

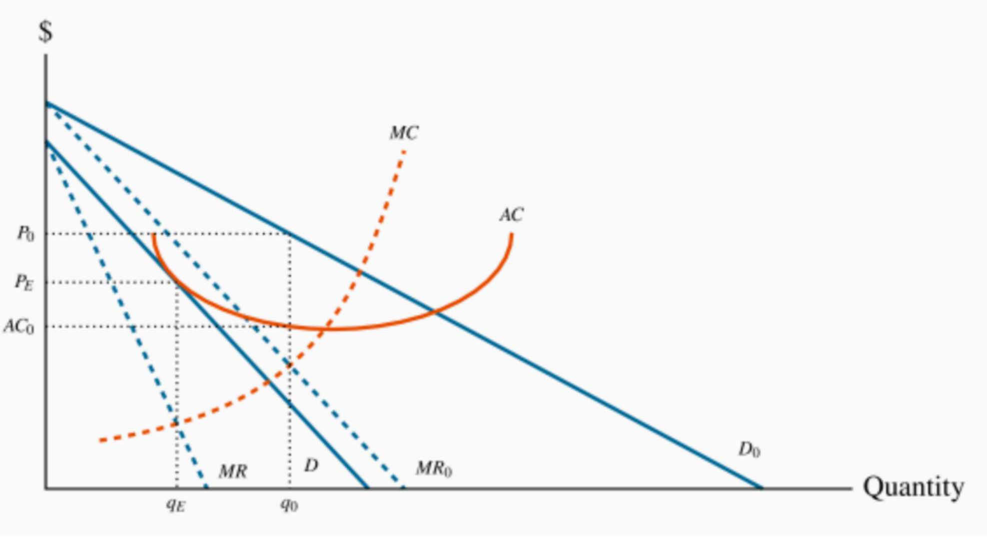
As a general rule then, each firm can influence its market share to some extent by changing its price. Its demand curve is not horizontal because different firms' products are only limited substitutes. A lower price level may draw some new customers away from competitors, but convenience or taste will prevent most patrons from deserting their local businesses. In concrete terms: A pasta special at the local Italian restaurant that reduces the price below the corresponding price at the competing local Thai restaurant will indeed draw clients away from the latter, but the foods are sufficiently different that only some customers will leave the Thai restaurant. The differentiated menus mean that many customers will continue to pay the higher price.

A **differentiated product** is one that differs slightly from other products in the same market.

Given that there are very many firms, the theory also envisages limits to scale economies. Firms are small and, with many competitors, individual firms do not compete strategically with *particular* rivals. Because the various products offered are slightly differentiated, we avoid graphics with a *market* demand, because this would imply that a uniform product is being considered. At the same time the market is a well-defined concept—it might be composed of all those restaurants within a reasonable distance, for example, even though each one is slightly different from the others. The market share of each firm depends on the price that it charges *and* on the number of competing firms. For a given number of suppliers, a shift in industry demand also shifts the demand facing each firm. Likewise, the presence of more firms in the industry reduces the demand facing each one.

The market equilibrium is illustrated in Figure 11.2. Here D0 is the initial demand facing a representative firm, and MR0 is the corresponding marginal revenue curve. Profit is maximized where MC=MR, and the price P_0 is obtained from the demand curve corresponding to the output q_0. Total profit is the product of output times the difference between price and average cost, which equals $q_0 \times (P_0 - AC_0)$.

Figure 11.2 Equilibrium for a monopolistic competitor



Profits exist at the initial equilibrium (q_0,P_0)*. Hence, new firms enter and reduce the share of the total market faced by each firm, thereby shifting back their demand curve. A final equilibrium is reached where economic profits are eliminated: At* $AC=P_E$ *and* MR=MC.

With free entry, such profits attract new firms. The increased number of firms reduces the share of the market that any one firm can claim. That is, the firm's demand curve shifts inwards when entry occurs. As long as (economic) profits exist, this process continues. For entry to cease, average cost must equal price. A final equilibrium is illustrated by the combination (PE,qE), where the demand has shifted inward to D.

At this long-run equilibrium, two conditions must hold: First, the optimal pricing rule must be satisfied—that is MC = MR; second it must be the case that only normal profits are made at the final equilibrium. Economic profits are competed away as a result of free entry. And these two conditions must exist at the optimal output. Graphically this implies that ATC must equal price at the output where MC = MR. In turn this implies that the ATC is tangent to the demand curve where P = ATC. While this could be proven mathematically, it is easy to intuit why this tangency must exist: If ATC merely intersected the demand curve at the output where MC = MR, we could find some other output where the demand price would be above ATC, suggesting that profits could be made at such an output. Clearly that could not represent an equilibrium.

The **monopolistically competitive equilibrium** in the long run requires the firm's demand curve to be tangent to the ATC curve at the output where MR = MC.

Source: “Macroeconomic principles, 2012, [section], CC-BY-NC-SA 3.0

Excess Capacity: The Price of Variety The long-run equilibrium solution in monopolistic competition always produces zero economic profit at a point to the left of the minimum of the average total cost curve. That is because the zero profit solution occurs at the point where the downward-sloping demand curve is tangent to the average total cost curve, and thus the average total cost curve is itself downward-sloping. By expanding output, the firm could lower average total cost. The firm thus produces less than the output at which it would minimize average total cost. A firm that operates to the left of the lowest point on its average total cost curve has **excess capacity.**

11.3 OLIGOPOLY AND GAME THEORY

Source: Curtis & Irvine, 2016, Section 11.4, CC-BY-NC-SA 3.0 (Original source, pages 4-5 removed)

The *players* in the game try to maximize their own *payoffs*. In an oligopoly, the firms are the players and their payoffs are their profits. Each player must choose a **strategy**, which is a plan describing how a player moves or acts in different situations.

> A **strategy** is a game plan describing how a player acts, or moves, in each possible situation.

How do we arrive at an equilibrium in these games? Let us begin by defining a commonly used concept of equilibrium. A **Nash equilibrium** is one in which each player chooses the best strategy, given the strategies chosen by the other players and there is no incentive to move or change choice.

> A **Nash equilibrium** is one in which each player chooses the best strategy, given the strategies chosen by the other player, and there is no incentive for any player to move.

In such an equilibrium, no player wants to change strategy, since the other players' strategies were already figured into determining each player's own best strategy. This concept and theory are attributable to the Princeton mathematician John Nash, who was popularized by the Hollywood movie version of his life *A Beautiful Mind*.

In most games, each player's best strategy depends on the strategies chosen by their opponents. Sometimes, though not always, a player's best strategy is independent of those chosen by rivals. Such a strategy is called a **dominant strategy**.

> A **dominant strategy** is a player's best strategy, independent of the strategies adopted by rivals.

We now illustrate these concepts with the help of two different games. These games differ in their outcomes and strategies. Table 11.2 contains the domestic happiness game[1]. Will and Kate are attempting to live in harmony, and their happiness depends upon each of them carrying out domestic chores such as shopping, cleaning and cooking. The first element in each pair defines Will's outcome, the second Kate's outcome. If both contribute to domestic life they each receive a happiness or utility level of 5 units. If one contributes and the other does not the happiness levels are 2 for the contributor and 6 for the non-contributor, or 'free-rider'. If neither contributes happiness levels are 3 each. When each follows the same strategy the payoffs are on the diagonal, when they follow different strategies the payoffs are on the off-diagonal. Since the elements of the table define the payoffs resulting from various choices, this type of matrix is called a **payoff matrix**.

A payoff matrix defines the rewards to each player resulting from particular choices.

So how is the game likely to unfold? In response to Will's choice of a contribute strategy, Kate's utility maximizing choice involves lazing: She gets 6 units by not contributing as opposed to 5 by contributing. Instead, if Will decides to be lazy what is in Kate's best interest? Clearly it is to be lazy also because that strategy yields 3 units of happiness compared to 2 units if she contributes. In sum, Kate's best strategy is to be lazy, regardless of Will's behaviour. So the strategy of not contributing is a *dominant strategy*.

Will also has a dominant strategy – identical to Kate's. This is not surprising since the payoffs are symmetric in the table. Hence, since each has a dominant strategy of not contributing the Nash equilibrium is in the bottom right cell, where each receives a payoff of 3 units. Interestingly this equilibrium is not the one that yields maximum combined happiness.

Table 11.2 A game with dominant strategies

		Kate's choice	
		Contribute	Laze
Will's choice	Contribute	5,5	2,6
	Laze	6,2	3,3

The first element in each cell denotes the payoff or utility to Will; the second element the utility to Kate.

The Prisoner's Dilemma

Source: Lynham, 2018, Chapter 10 Introduction, CC-BY 4.0 (Original source, paragraphs 9-16 only)

Because of the complexity of oligopoly, which is the result of mutual interdependence among firms, there is no single, generally-accepted theory of how oligopolies behave, in the same way that we have theories for all the other market structures. Instead, economists use **game theory**, a branch of mathematics that analyzes situations in which players must make decisions and then receive payoffs based on what other players decide to do. Game theory has found widespread applications in the social sciences, as well as in business, law, and military strategy.

The **prisoner's dilemma** is a scenario in which the gains from cooperation are larger than the rewards from pursuing self-interest. It applies well to oligopoly. The story behind the prisoner's dilemma goes like this:

> Two co-conspiratorial criminals are arrested. When they are taken to the police station, they refuse to say anything and are put in separate interrogation rooms.

> Eventually, a police officer enters the room where Prisoner A is being held and says: “You know what? Your partner in the other room is confessing. So your partner is going to get a light prison sentence of just one year, and because you’re remaining silent, the judge is going to stick you with eight years in prison. Why don’t you get smart? If you confess, too, we’ll cut your jail time down to five years, and your partner will get five years, also.” Over in the next room, another police officer is giving exactly the same speech to Prisoner B. What the police officers do not say is that if both prisoners remain silent, the evidence against them is not especially strong, and the prisoners will end up with only two years in jail each.

The game theory situation facing the two prisoners is shown in Table 3. To understand the dilemma, first consider the choices from Prisoner A’s point of view. If A believes that B will confess, then A ought to confess, too, so as to not get stuck with the eight years in prison. But if A believes that B will not confess, then A will be tempted to act selfishly and confess, so as to serve only one year. The key point is that A has an incentive to confess regardless of what choice B makes! B faces the same set of choices, and thus will have an incentive to confess regardless of what choice A makes. Confess is considered the dominant strategy or the strategy an individual (or firm) will pursue regardless of the other individual’s (or firm’s) decision. The result is that if prisoners pursue their own self-interest, both are likely to confess, and end up doing a total of 10 years of jail time between them.

		Prisoner B	
		Remain Silent (cooperate with other prisoner)	Confess (do not cooperate with other prisoner)
Prisoner A	Remain Silent (cooperate with other prisoner)	A gets 2 years, B gets 2 years	A gets 8 years, B gets 1 year
	Confess (do not cooperate with other prisoner)	A gets 1 year, B gets 8 years	A gets 5 years B gets 5 years

Table 3. The Prisoner’s Dilemma Problem

The game is called a dilemma because if the two prisoners had cooperated by both remaining silent, they would only have had to serve a total of four years of jail time between them. If the two prisoners can work out some way of cooperating so that neither one will confess, they will both be better off than if they each follow their own individual self-interest, which in this case leads straight into longer jail terms.

The Oligopoly Version of the Prisoner’s Dilemma

The members of an oligopoly can face a prisoner’s dilemma, also. If each of the oligopolists cooperates in holding down output, then high monopoly profits are possible. Each oligopolist, however, must worry that while it is holding down output, other firms are taking advantage of the high price by raising output and earning higher profits. Table 4 shows the prisoner’s dilemma for a two-firm oligopoly—known as a **duopoly**. If Firms A and B both agree to hold down output, they are acting together as a monopoly and will each earn $1,000 in profits. However, both firms’ dominant strategy is to increase output, in which case each will earn $400 in profits.

		Firm B	
		Hold Down Output (cooperate with other firm)	Increase Output (do not cooperate with other firm)
Firm A	Hold Down Output (cooperate with other firm)	A gets $1,000, B gets $1,000	A gets $200, B gets $1,500
	Increase Output (do not cooperate with other firm)	A gets $1,500, B gets $200	A gets $400, B gets $400

Table 4. A Prisoner’s Dilemma for Oligopolists

Can the two firms trust each other? Consider the situation of Firm A:

- If A thinks that B will cheat on their agreement and increase output, then A will increase output, too, because for A the profit of $400 when both firms increase output (the bottom right-hand choice in Table 4) is better than a profit of only $200 if A keeps output low and B raises output (the upper right-hand choice in the table).
- If A thinks that B will cooperate by holding down output, then A may seize the opportunity to earn higher profits by raising output. After all, if B is going to hold down output, then A can earn $1,500 in profits by expanding output (the bottom left-hand choice in the table) compared with only $1,000 by holding down output as well (the upper left-hand choice in the table).

Thus, firm A will reason that it makes sense to expand output if B holds down output and that it also makes sense to expand output if B raises output. Again, B faces a parallel set of decisions.

The result of this prisoner's dilemma is often that even though A and B could make the highest combined profits by cooperating in producing a lower level of output and acting like a monopolist, the two firms may well end up in a situation where they each increase **output** and earn only $400 each in **profits**.

The Collusion Model

Source: "Boundless economics", OER2Go, n.d. [section], CC-BY-SA 4.0 International (Original source, paragraphs 3, 5, 6 & Figure 11.3 removed)

There is no single model of profit-maximizing oligopoly behavior that corresponds to economists' models of perfect competition, monopoly, and monopolistic competition. Uncertainty about the interaction of rival firms makes specification of a single model of oligopoly impossible. Instead, economists have devised a variety of models that deal with the uncertain nature of rivals' responses in different ways. In this section we review one type of oligopoly model, the collusion model. After examining this traditional approach to the analysis of oligopoly behavior, we shall turn to another method of examining oligopolistic interaction: game theory.

Firms in any industry could achieve the maximum profit attainable if they all agreed to select the monopoly price and output and to share the profits. One approach to the analysis of oligopoly is to assume that firms in the industry collude, selecting the monopoly solution.

In the simplest form of collusion, overt collusion, firms openly agree on price, output, and other decisions aimed at achieving monopoly profits. Firms that coordinate their activities through overt collusion and by forming collusive coordinating mechanisms make up a cartel.

An alternative to overt collusion is tacit collusion, an unwritten, unspoken understanding through which firms agree to limit their competition. Firms may, for example, begin following the price leadership of a particular firm, raising or lowering their prices when the leader makes such a change. The price leader may be the largest firm in the industry, or it may be a firm that has been particularly good at assessing changes in demand or cost. At various times, tacit collusion has been alleged to occur in a wide range of industries, including steel, cars, and breakfast cereals.

It is difficult to know how common tacit collusion is. The fact that one firm changes its price shortly after another one does cannot prove that a tacit conspiracy exists. After all, we expect to see the prices of all firms in a perfectly competitive industry moving together in response to changes in demand or production costs.

Intended Entry Barriers

Source: Curtis & Irvine, 2016, Section 11.6, CC-BY-NC-SA 3.0 (Original source, page 2 only, 'predatory pricing' definition box added from section 14.5)

1. Patent law is one form of protection for incumbent firms. Research and development is required for the development of many products in the modern era. Pharmaceuticals are an example. If innovations were not protected, firms and individuals would not be incentivized to devote their energies and resources to developing new drugs. Society would be poorer as a result. Patent protection is obviously a legal form of protection.

2. *Advertizing* is a second form of entry deterrence. In this instance firms attempt to market their product as being distinctive and even enviable. For example, *Coca-Cola* and *PepsiCo* invest hundreds of millions annually to project their products in this light. They sponsor sports, artistic and cultural events. Entry into the cola business is not impossible, but brand image is so strong for these firms that potential competitors would have a very low probability of entering this sector profitably. Likewise, in the 'energy-drinks' market, *Red Bull* spends hundreds of millions of dollars per annum to project its brand as being just as unique and desirable as Pepsi or Coca-Cola.

3. *Predatory pricing* is an illegal form of entry deterrence. It involves an incumbent charging an artificially low price for its product in the event of entry of a new competitor. This is done with a view to making it impossible for the entrant to earn a profit. Given that incumbents have generally greater resources than entrants, they can survive a battle of losses for a more prolonged period, thus ultimately driving out the entrant.
Network externalities arise when the existing number of buyers itself influences the total demand for a product. *Facebook* is now a classic example. It has many more members than *MySpace* or *Google+*, and hence finds it easier to attract new users. An individual contemplating joining a social network has an incentive to join one where she has many existing 'friends'.

Predatory pricing is a practice that is aimed at driving out competition by artificially reducing the price of one product sold by a supplier.

4. *Transition costs* can be erected by firms who do not wish to lose their customer base. Cell-phone plans are a good example. Contract-termination costs are one obstacle to moving to a new supplier. Some carriers grant special low rates to users communicating with other users within the same network, or offer special rates for a block of users (perhaps within a family).

CHAPTER 12: Economic Efficiency and Government Policy

12.1 EFFICIENCY IN PERFECTLY COMPETITIVE MARKETS

Source: Lynham, 2018, SECTION 8.4, CC-BY 4.0 (Original source, paragraphs 5, 7-10 removed)

When profit-maximizing firms in perfectly competitive markets combine with utility-maximizing consumers, something remarkable happens: the resulting quantities of outputs of goods and services demonstrate both productive and allocative efficiency (terms that were first introduced in (Choice in a World of Scarcity) .

Productive efficiency means producing without waste, so that the choice is on the production possibility frontier. In the long run in a perfectly competitive market, because of the process of entry and exit, the price in the market is equal to the minimum of the long-run average cost curve. In other words, goods are being produced and sold at the lowest possible average cost.

Allocative efficiency means that among the points on the production possibility frontier, the point that is chosen is socially preferred—at least in a particular and specific sense. In a perfectly competitive market, price will be equal to the marginal cost of production. Think about the price that is paid for a good as a measure of the social benefit received for that good; after all, willingness to pay conveys what the good is worth to a buyer. Then think about the marginal cost of producing the good as representing not just the cost for the firm, but more broadly as the social cost of producing that good. When perfectly competitive firms follow the rule that profits are maximized by producing at the quantity where price is equal to marginal cost, they are thus ensuring that the social benefits received from producing a good are in line with the social costs of production.

To explore what is meant by **allocative efficiency**, it is useful to walk through an example. Begin by assuming that the market for wholesale flowers is perfectly competitive, and so P = MC. Now, consider what it would mean if firms in that market produced a lesser quantity of flowers. At a lesser quantity, marginal costs will not yet have increased as much, so that price will exceed marginal cost; that is, P > MC. In that situation, the benefit to society as a whole of producing additional goods, as measured by the willingness of consumers to pay for marginal units of a good, would be higher than the cost of the inputs of labor and physical capital needed to produce the marginal good. In other words, the gains to society as a whole from producing additional marginal units will be greater than the costs.

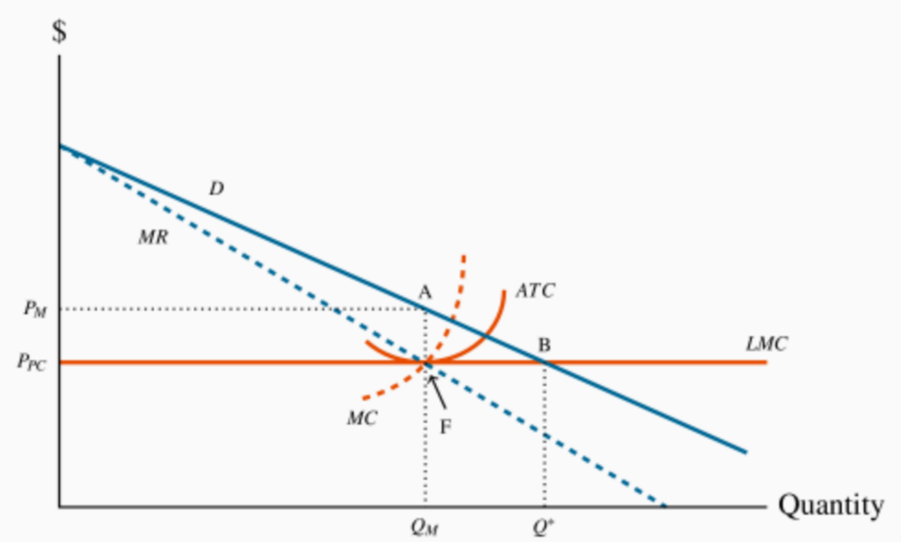
When perfectly competitive firms maximize their profits by producing the quantity where P = MC, they also assure that the benefits to consumers of what they are buying, as measured by the price they are willing to pay, is equal to the costs to society of producing the marginal units, as measured by the marginal costs the firm must pay—and thus that allocative efficiency holds.

Recall: The Allocative Inefficiency of a Monopoly

Source: Curtis & Irvine, 2016, Section 10.4, CC-BY-NC-SA 3.0 (Original source, last paragraph & footnote removed)

Output Inefficiency A characteristic of perfect competition is that it secures an efficient allocation of resources when there are no externalities in the market: Resources are used up to the point where their marginal cost equals their marginal value – as measured by the price that consumers are willing to pay. But a monopoly structure does not yield this output. Consider Figure 10.10.

Figure 10.10 Monopoly output inefficiency



A monopolist maximizes profit at QM*. Here the value of marginal output exceeds cost. If output expands to* Q∗ *a gain arises equal to the area ABF. This is the deadweight loss associated with the output* QM *rather than* Q∗*. If the monopolist's long-run* MC *is equivalent to a competitive industry's supply curve, then the deadweight loss is the cost of having a monopoly rather than a perfectly competitive market.*

The monopolist's profit-maximizing output QM is where MC equals MR. This output is inefficient for the reason that we developed in Chapter 5: If output is increased beyond QM the additional benefit exceeds the additional cost of producing it. The additional benefit is measured by the willingness of buyers to pay – the market demand curve. The additional cost is the long-run MC curve under the assumption of constant returns to scale. Using the terminology from Chapter 5, there is a deadweight loss equal to the area ABF. This is termed **allocative inefficiency.**

Allocative inefficiency arises when resources are not appropriately allocated and result in deadweight losses.

Perfect competition versus monopoly The area ABF can also be considered as the efficiency loss associated with having a monopoly rather than a perfectly competitive market structure. In perfect competition the supply curve is horizontal. This is achieved by having firms enter and exit when more or less must be produced. Accordingly, if the perfectly competitive industry's supply curve approximates the monopolist's long-run marginal cost curve, we can say that if the monopoly were turned into a competitive industry, output would increase from QM to Q∗. The deadweight loss is one measure of the superiority of the perfectly competitive structure over the monopoly structure.

Note that this critique of monopoly is not initially focused upon profit. While monopoly profits are what frequently irk the public, we have focused upon resource allocation inefficiencies. But in a real sense the two are related: Monopoly inefficiencies arise through output being restricted, and it is this output reduction – achieved by maintaining a higher than competitive price – that gives rise to those profits. Nonetheless, there is more than just a shift in purchasing power from the buyer to the seller. Deadweight losses arise because output is at a level lower than the point where the MC equals the value placed on the good; thus the economy is sacrificing the possibility of creating additional surplus.

12.2 REGULATION AND COMPETITION POLICY

Source: Curtis & Irvine, 2016, Section 14.5, CC-BY-NC-SA 3.0 (Original source, page 1 only)

Goals and Objectives

The goals of competition policy are relatively uniform across developed economies: The promotion of domestic competition; the development of new ideas, new products and new enterprises; the promotion of efficiency in the resource-allocation sense; the development of manufacturing and service industries that can compete internationally.

In addition to these economic objectives, governments and citizens frown upon monopolies or monopoly practices if they lead to an undue *concentration of political power*. Such power can lead to a concentration of wealth and influence in the hands of an elite.

Canada's regulatory body is the *Competition Bureau*, whose activity is governed primarily by the *Competition Act* of 1986. This act replaced the *Combines Investigation Act*. The *Competition Tribunal* acts as an adjudication body, and is composed of judges and non- judicial members. This tribunal can issue orders on the maintenance of competition in the marketplace. Canada has had anti-combines legislation since 1889, and the act of 1986 is the most recent form of such legislation and policy. The Competition Act does not forbid monopolies, but it does rule as unlawful the *abuse* of monopoly power. Canada's competition legislation is aimed at anti-competitive practices, and a full description of its activities is to be found on its website at www.competitionbureau.gc.ca. Let us examine some of these proscribed policies.

The Choices in Regulating a Natural Monopoly

Source: Lynham, 2018, Section 11.3, CC-BY 4.0 (Original source, paragraphs 3-8 only, Table 5 removed, subtitles added)

So what then is the appropriate competition policy for a natural monopoly? Figure 1 illustrates the case of natural monopoly, with a market demand curve that cuts through the downward-sloping portion of the **average cost curve**. Points A, B, C, and F illustrate four of the main choices for regulation. Table 5 [removed] outlines the regulatory choices for dealing with a natural monopoly.

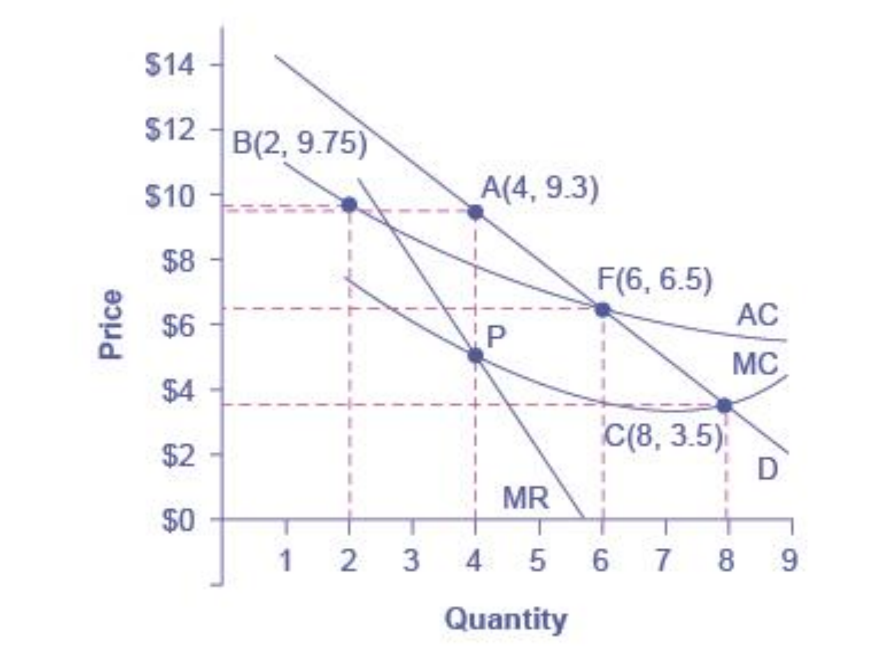


Figure 1. *Regulatory Choices in Dealing with Natural Monopoly. A natural monopoly will maximize profits by producing at the quantity where marginal revenue (MR) equals marginal costs (MC) and by then looking to the market demand curve to see what price to charge for this quantity. This monopoly will produce at point A, with a quantity of 4 and a price of 9.3. If antitrust regulators split this company exactly in half, then each half would produce at point B, with average costs of 9.75 and output of 2. The regulators might require the firm to produce where marginal cost crosses the market demand curve at point C. However, if the firm is required to produce at a quantity of 8 and sell at a price of 3.5, the firm will suffer from losses. The most likely choice is point F, where the firm is required to produce a quantity of 6 and charge a price of 6.5.*

The first possibility is to leave the natural monopoly alone. In this case, the monopoly will follow its normal approach to maximizing profits. It determines the quantity where MR = MC, which happens at point P at a quantity of 4. The firm then looks to point A on the demand curve to find that it can charge a price of 9.3 for that profit-maximizing quantity. Since the price is above the average cost curve, the natural monopoly would earn economic profits.

A second outcome arises if antitrust authorities decide to divide the company, so that the new firms can compete. As a simple example, imagine that the company is cut in half.

Thus, instead of one large firm producing a quantity of 4, two half-size firms each produce a quantity of 2. Because of the declining average cost curve (AC), the average cost of production for each of the half-size companies each producing 2, as shown at point B, would be 9.75, while the average cost of production for a larger firm producing 4 would only be 7.75. Thus, the economy would become less productively efficient, since the good is being produced at a higher average cost. In a situation with a downward-sloping average cost curve, two smaller firms will always have higher average costs of production than one larger firm for any quantity of total output. In addition, the antitrust authorities must worry that splitting the natural monopoly into pieces may be only the start of their problems. If one of the two firms grows larger than the other, it will have lower average costs and may be able to drive its competitor out of the market. Alternatively, two firms in a market may discover subtle ways of coordinating their behavior and keeping prices high. Either way, the result will not be the greater competition that was desired.

Marginal-Cost Pricing A third alternative is that regulators may decide to set prices and quantities produced for this industry. The regulators will try to choose a point along the market demand curve that benefits both consumers and the broader social interest. Point C illustrates one tempting choice: the regulator requires that the firm produce the quantity of output where marginal cost crosses the demand curve at an output of 8, and charge the price of 3.5, which is equal to **marginal cost** at that point. This rule is appealing because it requires price to be set equal to marginal cost, which is what would occur in a perfectly competitive market, and it would assure consumers a higher quantity and lower price than at the monopoly choice A. In fact, efficient allocation of resources would occur at point C, since the value to the consumers of the last unit bought and sold in this market is equal to the marginal cost of producing it.

Attempting to bring about point C through force of regulation, however, runs into a severe difficulty. At point C, with an output of 8, a price of 3.5 is below the average cost of production, which is 5.7, and so if the firm charges a price of 3.5, it will be suffering losses. Unless the regulators or the government offer the firm an ongoing public subsidy (and there are numerous political problems with that option), the firm will lose money and go out of business.

Average-Cost Pricing Perhaps the most plausible option for the regulator is point F; that is, to set the price where AC crosses the demand curve at an output of 6 and a price of 6.5. This plan makes some sense at an intuitive level: let the natural monopoly charge enough to cover its average costs and earn a normal rate of profit, so that it can continue operating, but prevent the firm from raising prices and earning abnormally high monopoly profits, as it would at the monopoly choice A. Of course, determining this level of output and price with the political pressures, time constraints, and limited information of the real world is much harder than identifying the point on a graph. For more on the problems that can arise from a centrally determined price, see the discussion of price floors and price ceilings in Demand and Supply.

Source: Curtis & Irvine, 2016, Section 14.1, CC-BY-NC-SA 3.0 (Original source, paragraph 1 only)

Market Failure

Markets are fine institutions when all of the conditions for their efficient operation are in place. In Chapter 5 we explored the meaning of efficient resource allocation, by developing the concepts of consumer and producer surpluses. But, while we have emphasized the benefits of efficient resource allocation in a market economy, there are many situations where markets deliver inefficient outcomes. Several problems beset the operation of markets. The principal sources of *market failure* are: *Externalities, public goods, asymmetric information,* and the *concentration of power*. In addition markets may produce outcomes that are *unfavourable* to certain groups – perhaps those on low incomes. The circumstances described here lead to what is termed **market failure**.

Market failure defines outcomes in which the allocation of resources is not efficient.

CHAPTER 13: LABOUR MARKETS

Source: Curtis & Irvine, 2016, Section 13.1, CC-BY-NC-SA 3.0

13.1 HUMAN CAPITAL AND INCOME DISTRIBUTION

Human capital, HK, is the stock of knowledge and ability accumulated by a worker that determines future productivity and earnings. It depends on many different attributes – education, experience, intelligence, interpersonal skills etc. While individuals look upon human capital as a determinant of their own earnings, it also impacts the productivity of the economy at large, and is therefore a vital force in determining long-run growth. Canada has been investing heavily in human capital in recent decades, and this suggests that future productivity and earnings will benefit accordingly.

Human capital is the stock of knowledge and ability accumulated by a worker that determines future productivity and earning.

Several features of Canada's recent human capital accumulation are noteworthy. First, Canada's enrolment rate in post-secondary education now exceeds the US rate, and that of virtually every economy in the world. Second is the fact that the number of women in third-level institutions exceeds the number of men. Almost 60% of university students are women. Third, international testing of high-school students sees Canadian students performing well, which indicates that the quality of the Canadian educational system appears to be high. These are positive aspects of a system that frequently comes under criticism. Nonetheless, the distribution of income that emerges from market forces in Canada has become more unequal.

Let us now try to understand individuals' motivation for embarking on the accumulation of human capital, and in the process see why different groups earn different amounts. We start by analyzing the role of education and then turn to on-the-job training.

Source: Curtis & Irvine, 2016, Section 13.2, CC-BY-NC-SA 3.0 (Original source, page 3, paragraph 1 only)

The education premium Individuals with different education levels earn different wages. The **education premium** is the difference in earnings between the more and less highly educated. Quantitatively, Professors Kelly Foley and David Green have recently proposed that the completion of a college or trade certification adds about 15% to one's income, relative to an individual who has completed high school. A Bachelor's degree brings a premium of 20-25%, and a graduate degree several percentage points more[1]. The failure to complete high school penalizes individuals to the extent of about 10%. These are average numbers, and they vary depending upon the province of residence, time period and gender. Nonetheless the findings underline that more human capital is associated with higher earnings. The earnings premium depends upon both the supply and demand of high HK individuals. *Ceteris paribus*, if high-skill workers are heavily in demand by employers, then the premium should be greater than if lower-skill workers are more in demand.

Education premium: the difference in earnings between the more and less highly educated.

Source: Curtis & Irvine, 2016, Section 13.1, CC-BY-NC-SA 3.0

On-the-job training earning on the job is central to the age-earnings profiles of the better educated, and is less important for those with lower levels of education. **On-the-**

job training improves human capital through work experience. If on-the-job training increases worker productivity, who should pay for this learning – firms or workers? To understand who should pay, we distinguish between two kinds of skills: **Firm-specific skills** that raise a worker's productivity in a particular firm, and **general skills** that enhance productivity in many jobs or firms.

Firm-specific HK could involve knowing how particular components of a somewhat unique production structure functions, whereas general human capital might involve an understanding of engineering or architectural principles that can be applied universally. As for who should pay for the accumulation of skills: An employer should be willing to undertake most of the cost of firm-specific skills, because they are of less value to the worker should she go elsewhere. Firms offering general or transferable training try to pass the cost on to the workers, perhaps by offering a wage-earnings profile that starts very low, but that rises over time. Low-wage apprenticeships are examples. Hence, whether an employee is a medical doctor in residence, a plumber in an apprenticeship or a young lawyer in a law partnership, she 'pays' for the accumulation of her portable HK by facing a low wage when young. Workers are willing to accept such an earnings profile because their projected future earnings will compensate for lower initial earning.

On-the-job training improves human capital through work experience.

Firm-specific skills raise a worker's productivity in a particular firm.

General skills enhance productivity in many jobs or firms.

Source: Curtis & Irvine, 2016, Section 13.6, CC-BY-NC-SA 3.0

Discrimination Wage differences are a natural response to differences in human capital. But we frequently observe wage differences that might be discriminatory. For example, women on average earn less than men with similar qualifications; older workers may be paid less than those in their prime years; immigrants may be paid less than native-born Canadians, and ethnic minorities may be paid less than traditional white workers. The term **discrimination** describes an earnings differential that is attributable to a trait other than human capital.

If two individuals have the same HK, in the broadest sense of having the same capability to perform a particular task, then a wage premium paid to one represents discrimination. Correctly measured then, the discrimination premium between individuals from these various groups is the differential in earnings after correcting for HK differences. Thousands of studies have been undertaken on discrimination, and most conclude that discrimination abounds. Women, particularly those who have children, are paid less than men, and frequently face a 'glass ceiling' – a limit on their promotion possibilities within organizations.

Discrimination implies an earnings differential that is attributable to a trait other than human capital.

In contrast, women no longer face discrimination in university and college admissions, and form a much higher percentage of the student population than men in many of the higher paying professions such as medicine and law. Immigrants to Canada also suffer from a wage deficit. This is especially true for the most recent cohorts of working migrants who now come predominantly, not from Europe, as was once the case, but from China, South Asia, Africa and the Caribbean. For similarly-measured HK as Canadian-born individuals, these migrants frequently have an initial wage deficit of 30%, and require a period of more than twenty years to catch-up.

Monopsony

Source: Curtis & Irvine, 2016, Section 12.1, CC-BY-NC-SA 3.0 (Original source, page 7, paragraph 1 only)

Some firms may have to pay a higher wage in order to employ more workers. Think of Hydro Quebec building a dam in Northern Quebec. Not every hydraulic engineer would be equally happy working there as in Montreal. Some engineers may demand only a small wage premium to work in the North, but others will demand a high premium. If so, Hydro Quebec must pay a higher wage to attract more workers – it faces an upward sloping supply of labour curve. Hydro Quebec is the sole buyer in this particular market and is called a **monopsonist** – a single buyer. Our general optimizing principle governing the employment of labour still holds, even if we have different names for the various functions: Hire any factor of production up to the point where the cost of an additional unit equals the value generated for the firm by that extra worker. The essential difference here is that when a firm faces an upward sloping labour supply it will have to pay more to attract additional workers and *also pay more to its existing workers*. This will impact the firm's willingness to hire additional workers.

A **monopsonist** is the sole buyer of a good or service and faces an upward-sloping supply curve.

13.2 LABOUR UNIONS

Source: Lynham, 2018, Section 15.1, CC-BY 4.0 (Original source, paragraphs 1-2, 7-11 only)

A **labor union** is an organization of workers that negotiates with employers over wages and working conditions. A labor union seeks to change the balance of power between employers and workers by requiring employers to deal with workers collectively, rather than as individuals. Thus, negotiations between unions and firms are sometimes called **collective bargaining**.

The subject of labor unions can be controversial. Supporters of labor unions view them as the workers' primary line of defense against efforts by profit-seeking firms to hold down wages and benefits. Critics of labor unions view them as having a tendency to grab as much as they can in the short term, even if it means injuring workers in the long run by driving firms into bankruptcy or by blocking the new technologies and production

Download for free at http://cnx.org/contents/ea2f225e-6063-41ca-bcd8-36482e15ef65@11.9.

methods that lead to economic growth. We will start with some facts about union membership in the United States.

Higher Wages for Union Workers

Why might union workers receive higher pay? What are the limits on how much higher pay they can receive? To analyze these questions, let's consider a situation where all firms in an industry must negotiate with a single union, and no **firm** is allowed to hire nonunion labor. If no labor union existed in this market, then equilibrium (E) in the labor market would occur at the intersection of the demand for labor (D) and the supply of labor (S) in Figure 2. The union can, however, threaten that, unless firms agree to the wages they demand, the workers will strike. As a result, the labor union manages to achieve, through negotiations with the firms, a union wage of Wu for its members, above what the equilibrium wage would otherwise have been.

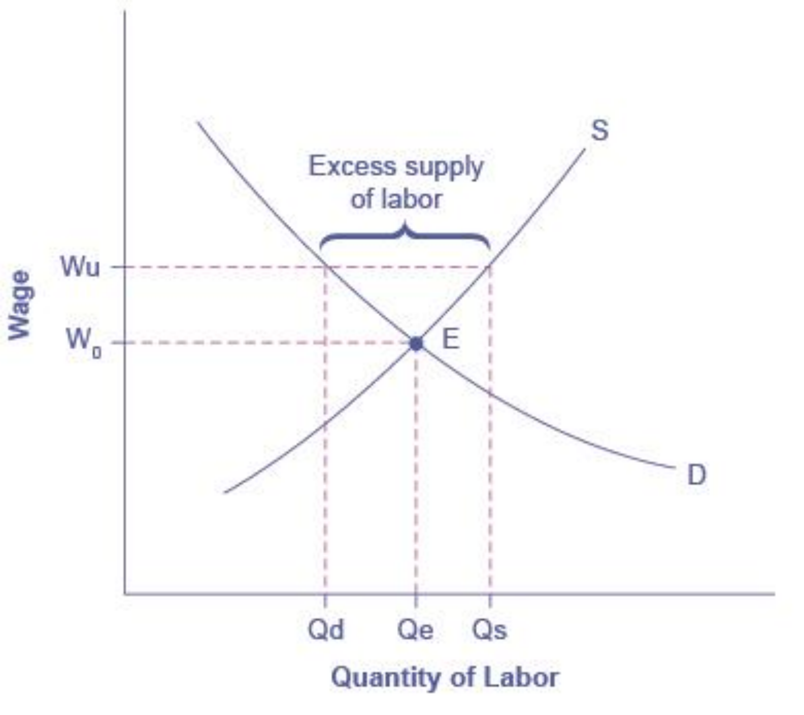


Figure 2. *Union Wage Negotiations. Without a union, the equilibrium at E would have involved the wage We and the quantity of labor Qe. However, the union is able to use its bargaining power to raise the wage to Wu. The result is an excess supply of labor for union jobs. That is, a quantity of labor supplied, Qs is greater than firms' quantity demanded for labor, Qd.*

This labor market situation resembles what a **monopoly firm** does in selling a product, but in this case a union is a monopoly selling labor to firms. At the higher union wage Wu, the firms in this industry will hire less labor than they would have hired in equilibrium. Moreover, an excess supply of workers want union jobs, but firms will not be hiring for such jobs.

From the union point of view, workers who receive higher wages are better off. However, notice that the quantity of workers (Qd) hired at the union wage Wu is smaller than the quantity Qe that would have been hired at the original equilibrium wage. A sensible union must recognize that when it pushes up the wage, it also reduces the incentive of firms to hire. This situation does not necessarily mean that union workers are fired. Instead, it may be that when union workers move on to other jobs or retire, they are not always replaced. Or perhaps when a firm expands production, it expands employment somewhat less with a higher union wage than it would have done with the lower equilibrium wage. Or perhaps a firm decides to purchase inputs from nonunion producers, rather than producing them with its own highly paid unionized workers. Or perhaps the firm moves or opens a new facility in a state or country where unions are less powerful.

From the firm's point of view, the key question is whether the higher wage of union workers is matched by higher productivity. If so, then the firm can afford to pay the higher union wages and, indeed, the demand curve for "unionized" labor could actually shift to the right. This could reduce the job losses as the equilibrium employment level shifts to the right and the difference between the equilibrium and the union wages will have been reduced. If worker unionization does not increase productivity, then the higher union wage will cause lower profits or losses for the firm.

Union workers might have higher productivity than nonunion workers for a number of reasons. First, higher wages may elicit higher productivity. Second, union workers tend to stay longer at a given job, a trend that reduces the employer's costs for training and hiring and results in workers with more years of experience. Many unions also offer job training and apprenticeship programs.

13.3 THE INCOME DISTRIBUTION

Source: Curtis & Irvine, 2016, Section 13.7, CC-BY-NC-SA 3.0 (Original source, pages 1-3 only)

How does all of our preceding discussion play out when it comes to the income distribution? That is, when we examine the incomes of all individuals or households in the economy, how equally or unequally are they distributed?

The study of inequality is a critical part of economic analysis. It recognizes that income differences that are in some sense 'too large' are not good for society. Inordinately large differences can reflect poverty and foster social exclusion and crime. Economic growth that is concentrated in the hands of the few can increase social tensions, and these can have economic as well as social or psychological costs. Crime is one reflection of the divide between 'haves' and 'have-nots'. It is economically costly; but so too is child poverty. Impoverished children rarely achieve their social or economic potential and this is a loss both to the individual and the economy at large.

In this section we will first describe a subset of the basic statistical tools that economists use to measure inequality. Second, we will examine how income inequality has evolved in recent decades. We shall see that, while the picture is complex, market income inequality has indeed increased in Canada. Third, we shall investigate some of the proposed reasons for the observed increase in inequality. Finally we will examine if the government offsets the inequality that arises from the marketplace through its taxation and redistribution policies.

It is to be emphasized that income inequality is just one proximate measure of the distribution of wellbeing. The extent of poverty is another such measure. Income is not synonymous with happiness but, that being said, income inequality can be computed reliably, and it provides a good measure of households' control over economic resources.

Theory and Measurement

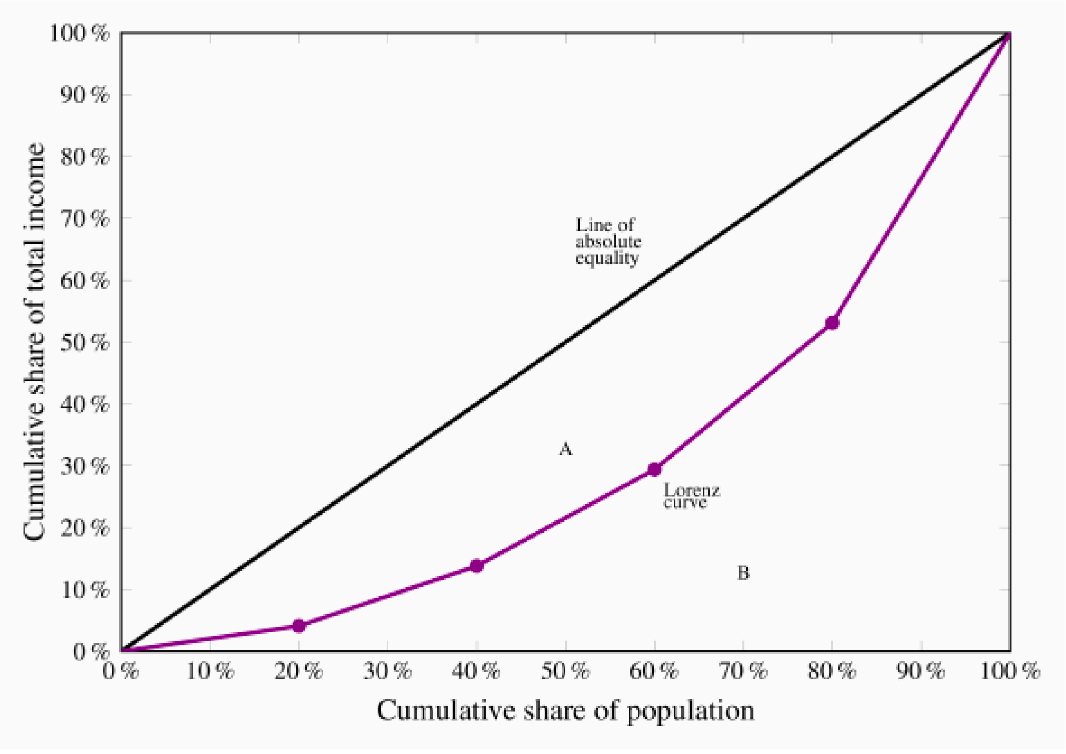
Let us rank the market incomes of all households in the economy from poor to rich, and categorize this ordering into different quantiles or groups. With five such quantiles the shares are called *quintiles*. The richest group forms the highest quintile, while the poorest group forms the lowest quintile. Such a representation is given in Table 13.2. The first numerical column displays the income in each quintile as a percentage of total income. If we wanted a finer breakdown, we could opt for decile (ten), or even vintile (twenty) shares, rather than quintile shares. These data can be graphed in a variety of ways. Since the data are in share, or percentage, form, we can compare, in a meaningful manner, distributions from economies that have different average income levels.

Table 13.2 Quintile shares of total family income in Canada, 2006

	Quintile share of total income	Cumulative share
First quintile	4.1	4.1
Second quintile	9.6	13.7
Third quintile	15.3	29.0
Fourth quintile	23.8	52.8
Fifth quintile	47.2	100.0
Total	100	

Source: Statistics Canada, CANSIM Matrix 2020405. These combinations are represented by the circles in the figure.

Figure 13.3 Gini index and Lorenz curve



The more equal are the income shares, the closer is the Lorenz curve to the diagonal line of equality. The Gini index is the ratio of the area A to the area (A+B). The Lorenz curve plots the cumulative percentage of total income against the cumulative percentage of the population.

An interesting way of presenting these data graphically is to plot the cumulative share of income against the cumulative share of the population. This is given in the final column, and also presented graphically in Figure 13.3. The bottom quintile has 4.1% of total income. The bottom two quintiles together have 13.7%(4.1%+9.6%), and so forth. By joining the coordinate pairs represented by the circles, a **Lorenz curve** is obtained. Relative to the diagonal line it is a measure of how unequally incomes are distributed: If everyone had the same income, each 20% of the population would have 20% of total income and by joining the points for such a distribution we would get a straight diagonal line joining the corners of the box. In consequence, if the Lorenz curve is further from the line of equality the distribution is less equal than if the Lorenz curve is close to the line of equality.

Lorenz curve describes the cumulative percentage of the income distribution going to different quantiles of the population.

This suggests that the area A relative to the area (A + B) forms a measure of inequality in the income distribution. This fraction obviously lies between zero and one, and it is called the **Gini index**. A larger value of the Gini index indicates that inequality is greater. We will not delve into the mathematical formula underlying the Gini, but for this set of numbers its value is 0.4.

Gini index: a measure of how far the Lorenz curve lies from the line of equality. Its maximum value is one; its minimum value is zero.

The Gini index is what is termed *summary index* of inequality – it encompasses a lot of information in one number. There exist very many other such summary statistics.

It is important to recognize that very different Gini index values emerge for a given economy by using different income definitions of the variable going into the calculations. For example, the quintile shares of the *earnings of individuals* rather than the *incomes of households* could be very different. Similarly, the shares of income *post tax* and *post transfers* will differ from their shares on a *pre-tax, pre-transfer* basis.

Government Policies to Reduce Inequality

Source: Lynham, 2018, Section 14.5, CC-BY 4.0 (Original source, paragraphs 1-10 only)

No society should expect or desire complete equality of income at a given point in time, for a number of reasons. First, most workers receive relatively low earnings in their first few jobs, higher earnings as they reach middle age, and then lower earnings after retirement. Thus, a society with people of varying ages will have a certain amount of income inequality. Second, people's preferences and desires differ. Some are willing to work long hours to have income for large houses, fast cars and computers, luxury vacations, and the ability to support children and grandchildren.

These factors all imply that a snapshot of inequality in a given year does not provide an accurate picture of how people's incomes rise and fall over time. Even if some degree of economic inequality is expected at any point in time, how much inequality should there be? There is also the difference between income and wealth, as the following Clear It Up feature explains.

Even if they cannot answer the question of how much inequality is too much, economists can still play an important role in spelling out policy options and tradeoffs. If a society decides to reduce the level of economic inequality, it has three main sets of tools: redistribution from those with high incomes to those with low incomes; trying to assure that a ladder of opportunity is widely available; and a tax on inheritance.

Redistribution means taking income from those with higher incomes and providing income to those with lower incomes. Earlier in this chapter, we considered some of the key government policies that provide support for the poor: the welfare program TANF, the earned income tax credit, SNAP, and Medicaid. If a reduction in inequality is desired, these programs could receive additional funding.

The programs are paid for through the federal income tax, which is a **progressive tax system** designed in such a way that the rich pay a higher percent in income taxes than the poor. Data from household income tax returns in 2009 shows that the top 1% of households had an average income of $1,219,700 per year in pre-tax income and paid an average federal tax rate of 28.9%. The **effective income tax**, which is total taxes paid divided by total income (all sources of income such as wages, profits, interest, rental income, and government transfers such as veterans' benefits), was much lower. The effective tax paid by the top 1% of householders was 20.4%, while the bottom two quintiles actually paid negative effective income taxes, because of provisions like the earned income tax credit. News stories occasionally report on a high-income person who has managed to pay very little in taxes, but while such individual cases exist, according to the Congressional Budget Office, the typical pattern is that people with higher incomes pay a higher average share of their income in federal income taxes.

Of course, the fact that some degree of redistribution occurs now through the federal income tax and government antipoverty programs does not settle the questions of how much redistribution is appropriate, and whether more redistribution should occur.

The Ladder of Opportunity Economic inequality is perhaps most troubling when it is not the result of effort or talent, but instead is determined by the circumstances under which a child grows up. One child attends a well-run grade school and high school and heads on to college, while parents help out by supporting education and other interests, paying for college, a first car, and a first house, and offering work connections that lead to internships and jobs. Another child attends a poorly run grade school, barely makes it through a low-quality high school, does not go to college, and lacks family and peer support. These two children may be similar in their underlying talents and in the effort they put forth, but their economic outcomes are likely to be quite different.

Public policy can attempt to build a ladder of opportunities so that, even though all children will never come from identical families and attend identical schools, each child has a reasonable opportunity to attain an economic niche in society based on their interests, desires, talents, and efforts. Some of those initiatives include those shown in Table 13.

Children	College Level	Adults
• Improved day care	• Widespread loans and grants for those in financial need	• Opportunities for retraining and acquiring new skills
• Enrichment programs for preschoolers	• Public support for a range of institutions from two-year community colleges to large research universities	• Prohibiting discrimination in job markets and housing on the basis of race, gender, age, and disability
• Improved public schools	–	–

Children	College Level	Adults
• After school and community activities	–	–
• Internships and apprenticeships	–	–

Table 13. Public Policy Initiatives

The United States has often been called a land of opportunity. Although the general idea of a ladder of opportunity for all citizens continues to exert a powerful attraction, specifics are often quite controversial. Society can experiment with a wide variety of proposals for building a ladder of opportunity, especially for those who otherwise seem likely to start their lives in a disadvantaged position. Such policy experiments need to be carried out in a spirit of open-mindedness, because some will succeed while others will not show positive results or will cost too much to enact on a widespread basis.

Inheritance Taxes There is always a debate about inheritance taxes. It goes like this: On the one hand, why should people who have worked hard all their lives and saved up a substantial nest egg not be able to give their money and possessions to their children and grandchildren? In particular, it would seem un-American if children were unable to inherit a family business or a family home. On the other hand, many Americans are far more comfortable with inequality resulting from high-income people who earned their money by starting innovative new companies than they are with inequality resulting from high-income people who have inherited money from rich parents.

The United States does have an **estate tax**—that is, a tax imposed on the value of an inheritance—which suggests a willingness to limit how much wealth can be passed on as an inheritance. However, according to the Center on Budget and Policy Priorities, in 2015 the estate tax applied only to those leaving inheritances of more than $5.43 million and thus applies to only a tiny percentage of those with high levels of wealth.

CHAPTER 14: Market Failures

14.1 EXTERNALITIES

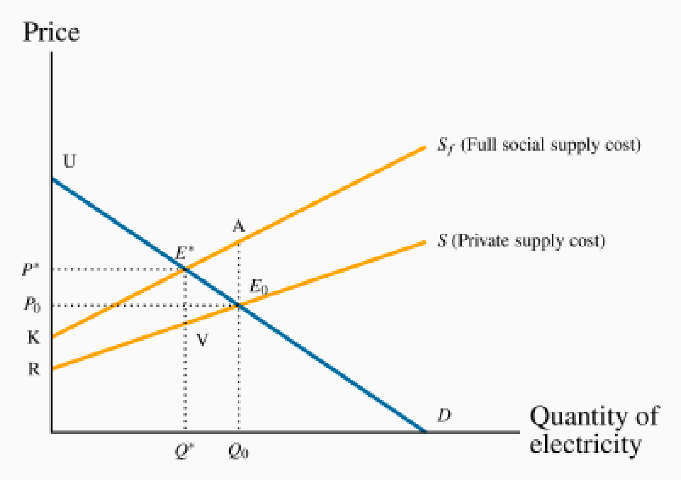
Source: Curtis & Irvine, 2016, Section 5.5, CC-BY-NC-SA 3.0

The consumer and producer surplus concepts we have developed are extremely powerful tools of analysis, but the world is not always quite as straightforward as simple models indicate. For example, many suppliers generate pollutants that adversely affect the health of the population, or damage the environment, or both. The term externality is used to denote such impacts. Externalities impact individuals who are not participants in the market in question, and the effects of the externalities may not be captured in the market price. For example, electricity-generating plants that use coal reduce air quality, which, in turn, adversely impacts individuals who suffer from asthma or other lung ailments. While this is an example of a negative externality, externalities can also be positive.

An externality is a benefit or cost falling on people other than those involved in the activity's market. It can create a difference between private costs or values and social costs or values.

We will now show why markets characterized by externalities are not efficient, and also show how these externalities might be corrected or reduced. The essence of an externality is that it creates a divergence between private costs/benefits and social costs/benefits. If a steel producer pollutes the air, and the steel buyer pays only the costs incurred by the producer, then the buyer is not paying the full "social" cost of the product. The problem is illustrated in Figure 5.5.

Figure 5.5 Negative externalities and inefficiency



A negative externality is associated with this good. S reflects private costs, whereas S_f reflects the full social cost. The socially optimal output is Q^*, not the market outcome Q_0. Beyond Q^* the real cost exceeds the demand value; therefore Q_0 is not an efficient output. A tax that increases P to P^* and reduces output is one solution to the externality.

Negative Externalities

In Figure 5.5, the supply curve S represents the cost to the supplier, whereas Sf (the full cost) reflects, in addition, the cost of bad air to the population. Of course, we are assuming that this external cost is ascertainable, in order to be able to characterize Sf accurately. Note also that this illustration assumes that, as power output increases, the external cost per unit rises, because the difference between the two supply curves increases with output. This implies that low levels of pollution do less damage per unit: Perhaps the population has a natural tolerance for low levels, but higher levels cannot be tolerated easily and so the cost per unit is greater.

Despite the externality, an efficient level of production can still be defined. It is given by $Q\ast$, not $Q0$. To see why, consider the impact of reducing output by one unit from $Q0$. At $Q0$ the willingness of buyers to pay for the marginal unit supplied is $E0$. The (private) supply cost is also $E0$. But from a societal standpoint there is a pollution/health cost of $AE0$ associated with that unit of production. The full cost, as represented by Sf, exceeds the buyer's valuation. Accordingly, if the last unit of output produced is cut, society gains by the amount $AE0$, because the cut in output reduces the excess of true cost over value.

Applying this logic to each unit of output between $Q0$ and $Q\ast$, it is evident that society can increase its well-being by the dollar amount equal to the area $E\ast AE0$, as a result of reducing production.

Next, consider the consequences of reducing output further from $Q\ast$. Note that some pollution is being created here, and environmentalists frequently advocate that pollution should be reduced to zero. However, an efficient outcome may not involve a zero level of pollution! If the production of power were reduced below $Q\ast$, the loss in value to buyers, as a result of not being able to purchase the good, would exceed the full cost of its production.

If the government decreed that, instead of producing $Q\ast$, no pollution would be tolerated, then society would forgo the possibility of earning the total real surplus equal to the area $UE\ast K$. Economists do not advocate such a zero-pollution policy; rather, we advocate a policy that permits a "tolerable" pollution level – one that still results in net benefits to society. In this particular example, the total cost of the tolerated pollution equals the area between the private and full supply functions, $KE\ast VR$.

As a matter of policy, how is this market influenced to produce the amount $Q\ast$ rather than $Q0$? One option would be for the government to intervene directly with production quotas for each firm. An alternative would be to impose a corrective tax on the good whose production causes the externality: With an appropriate increase in the price, consumers will demand a reduced quantity. In Figure 5.5 a tax equal to the dollar value $VE\ast$ would shift the supply curve upward by that amount and result in the quantity $Q\ast$ being traded.

A **corrective tax** seeks to direct the market towards a more efficient output.

We are now venturing into the field of environmental policy, and this is explored in the following section. The key conclusion of the foregoing analysis is that an efficient working of the market continues to have meaning in the presence of externalities. An efficient output level still maximizes economic surplus where surplus is correctly defined.

Positive Externalities

Externalities of the *positive* kind enable individuals or producers to get a type of 'free ride' on the efforts of others. Real world examples abound: When a large segment of the population is inoculated against disease, the remaining individuals benefit on account of the reduced probability of transmission.

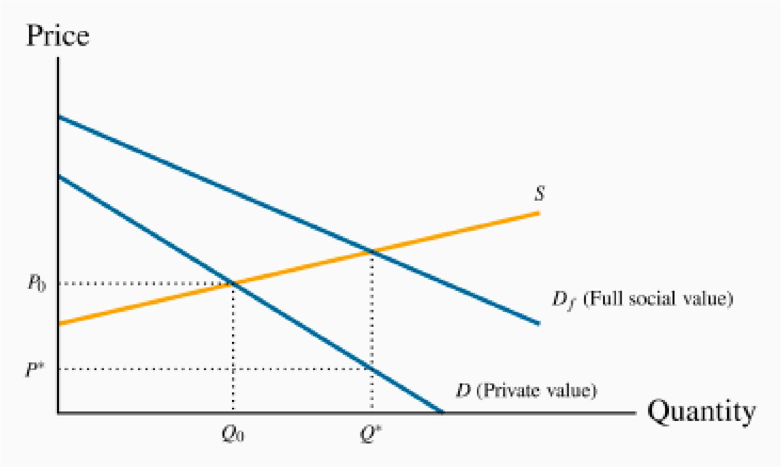
A less well recognized example is the benefit derived by many producers world-wide from research and development (R&D) undertaken in advanced economies and in universities and research institutes. The result is that society at large, including the corporate sector, gain from this enhanced understanding of science, the environment, or social behaviours.

The free market may not cope any better with these positive externalities than it does with negative externalities, and government intervention may be beneficial. Furthermore, firms that invest heavily in research and development would not undertake such investment if competitors could have a complete free ride and appropriate the fruits. This is why patent laws exist, as we shall see later in discussing Canada's competition policy. These laws prevent competitors from copying the product development of firms that invest in R&D. If such protection were not in place, firms would not allocate sufficient resources to R&D, which is a real engine of economic growth. In essence, the economy's research-directed resources would not be appropriately rewarded, and thus too little research would take place.

While patent protection is one form of corrective action, subsidies are another. We illustrated above that an appropriately formulated tax on a good that creates negative externalities can reduce demand for that good, and thereby reduce pollution. A subsidy can be thought of as a negative tax, and can stimulate the supply of goods and services that have positive externalities. Consider the example in Figure 5.6.

Individuals have a demand for flu shots given by D. This reflects their private valuation – their personal willingness to pay. But the social value of flu shots is greater. When a given number of individuals are inoculated, the probability that others will be infected falls. Additionally, with higher rates of inoculation, the health system will incur fewer costs in treating the infected. Therefore, the value to society of any quantity of flu shots is greater than the sum of the values that individuals place on them.

Figure 5.6 Positive externalities – the market for flu shots



The value to society of vaccinations exceeds the value to individuals: The greater the number of individuals vaccinated, the lower is the probability of others contracting the virus. D_f *reflects this additional value. Consequently, the social optimum is* Q^* *which exceeds* Q_0.

Let Df reflects the full social value of any quantity of flu shots. If S is the supply curve, the socially optimal, efficient, market outcome is Q∗. How can we influence the market to move from Q0 to Q∗? One solution is a subsidy that would reduce the price from P0 to P∗. Rather than shifting the supply curve upwards, as a tax does, the subsidy would shift the supply *downward*, sufficiently to intersect D at the output Q∗. In some real world examples, the value of the positive externality is so great that the government may decide to drive the price to zero, and thereby provide the inoculation at a zero price. For example, children typically get their MMR shots (measles, mumps, and rubella) free of charge. Graphically, this case would have the intersection between S and Df occurring to the right of the D intercept on the horizontal axis.

14.2 MARKET FAILURES

Source: Curtis & Irvine, 2016, Section 14.1, CC-BY-NC-SA 3.0 (Original source, last paragraph removed)

Markets are fine institutions when all of the conditions for their efficient operation are in place. In Chapter 5 we explored the meaning of efficient resource allocation, by developing the concepts of consumer and producer surpluses. But, while we have emphasized the benefits of efficient resource allocation in a market economy, there are many situations where markets deliver inefficient outcomes. Several problems beset the operation of markets. The principal sources of *market failure* are: *Externalities, public goods, asymmetric information,* and the *concentration of power*. In addition markets may produce outcomes that are *unfavourable* to certain groups – perhaps those on low incomes. The circumstances described here lead to what is termed **market failure**.

Market failure defines outcomes in which the allocation of resources is not efficient.

Externalities

A negative externality is one resulting, perhaps, from the polluting activity of a producer, or the emission of greenhouse gases into the atmosphere. A positive externality is one where the activity of one individual confers a benefit on others. An example here is where individuals choose to get immunized against a particular illness. As more people become immune, the lower is the probability that the illness can propagate itself through hosts, and therefore the greater the benefits to those not immunized.

Solutions to these market failures come in several forms: Government taxes and subsidies, or quota systems that place limits on the production of products generating externalities. Such solutions were explored in Chapter 5. Taxes on gasoline discourage its use and therefore reduce the emission of poisons into the atmosphere. Taxes on cigarettes and alcohol lower the consumption of goods that may place an additional demand on our publicly-funded health system. The provision of free, or low-cost, immunization against specific diseases to children benefits the whole population.

These measures attempt to *compensate for the absence of a market* in certain activities. Producers may not wish to pay for the right to emit pollutants, and consequently if the government steps in to counter such an externality, the government is effectively implementing a solution to the missing market.

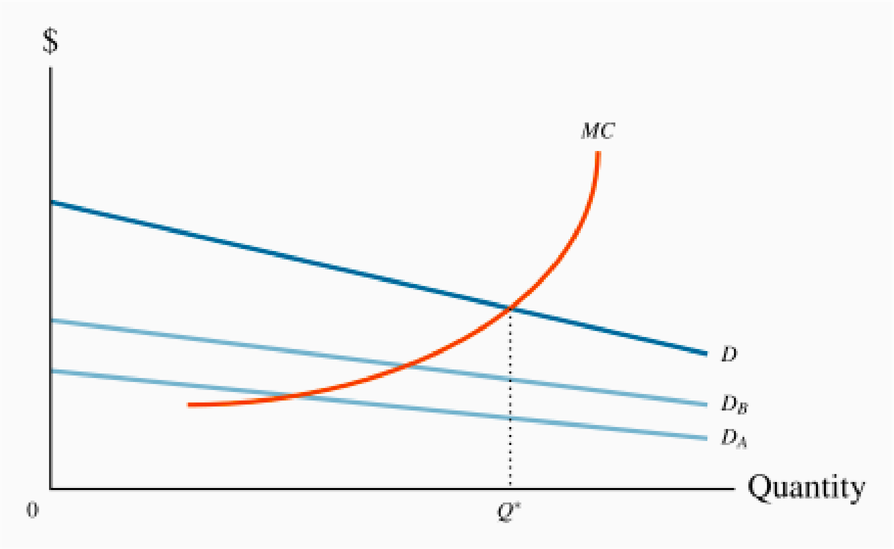
Public Goods

Public goods are sometimes called collective consumption goods, on account of their non-rivalrous and non-excludability characteristics. For example, if the government meteorological office provides daily forecasts over the nation's airwaves, it is no more expensive to supply that information to one million than to one hundred individuals in the same region. Its provision to one is not rivalrous with its provision to others – in contrast to private goods that cannot be 'consumed' simultaneously by more than one individual. In addition, it may be difficult to exclude certain individuals from receiving the information.

Public goods are non-rivalrous, in that they can be consumed simultaneously by more than one individual; additionally they may have a non-excludability characteristic.

Examples of such goods and services abound: Highways (up to their congestion point), street lighting, information on trans-fats and tobacco, or public defence provision. Such goods pose a problem for private markets: If it is difficult to exclude individuals from their consumption, then potential private suppliers will likely be deterred from supplying them because the suppliers cannot generate revenue from *free-riders*. Governments therefore normally supply such goods and services. But how much should governments supply? An answer is provided with the help of Figure 14.1.

Figure 14.1 Optimal provision of a public good



The total demand for the public good D is the vertical sum of the individual demands D_A and D_B. The optimal provision is where the MC equals the aggregate marginal valuation, as defined by the demand curve D. At the optimum Q^*, each individual is supplied the same amount of the public good.

This is a supply-demand diagram with a difference. The supply side is conventional, with the MC of production representing the supply curve. An efficient use of the economy's resources, we already know, dictates that an amount should be produced so that the cost at the margin equals the benefit to consumers at the margin. In contrast to the total market demand for private goods, which is obtained by summing individual demands horizontally, the demand for public goods is obtained by summing individual demands *vertically*.

Figure 14.1 depicts an economy with just two individuals whose demands for street lighting are given by DA and DB. These demands reveal the value each individual places on the various output levels of the public good, measured on the x-axis. However, since each individual can consume the public good *simultaneously*, the aggregate value of any output produced is the *sum of each individual valuation*. The valuation in the market of any quantity produced is therefore the vertical sum of the individual demands. D is the vertical sum of DA and DB, and the optimal output is Q*. At this equilibrium each individual consumes the same quantity of street lighting, and the MC of the last unit supplied equals the value placed upon it by society – both individuals. Note that this 'optimal' supply depends upon the income distribution, as we have stated several times to date. A different distribution of income may give rise to different demands DA and DB, and therefore a different 'optimal' output.

Efficient supply of public goods is where the marginal cost equals the sum of individual marginal valuations, and each individual consumes the same quantity.

Application Box 14.1 Are Wikipedia, Google and MOOCs public goods?
Wikipedia is one of the largest on-line sources of free information in the world. It is an encyclopedia that functions in multiple languages and that furnishes information on millions of topics. It is freely accessible, and is maintained and expanded by its users. Google is the most frequently used search engine on the World Wide Web. It provides information to millions of users simultaneously on every subject imaginable, free of charge. MOOCs are 'monster open online courses' offered by numerous universities, frequently for no charge to the student. Are these services public goods in the sense we have described?

Very few goods and services are pure public goods, some have the major characteristics of public goods nonetheless. In this general sense, Google, Wikipedia and MOOCs have public good characteristics. Wikipedia is funded by philanthropic contributions, and its users expand its range by posting information on its servers. Google is funded from advertising revenue. MOOCs are funded by university budgets.

A pure public good is available to additional users at zero marginal cost. This condition is essentially met by these services since their server capacity rarely reaches its limit. Nonetheless, they are constantly adding server capacity, and in that sense cannot furnish their services to an unlimited number of additional users at no additional cost.

Knowledge is perhaps the ultimate public good; Wikipedia, Google and MOOCs all disseminate knowledge, knowledge which has been developed through the millennia by philosophers, scientists, artists, teachers, research laboratories and universities.

A challenge in providing the optimal amount of government-supplied public goods is to know the value that users may place upon them – how can the demand curves DA and DB, be ascertained, for example, in Figure 14.1? In contrast to markets for private goods, where consumer demands are essentially revealed through the process of purchase, the demands for public goods may have to be uncovered by means of surveys that are designed so as to elicit the true valuations that users place upon different amounts of a public good. A second challenge relates to the pricing and funding of public goods: For example, should highway lighting be funded from general tax revenue, or should drivers pay for it? These are complexities that are beyond our scope of our current inquiry.

Asymmetric Information

Markets for information abound in the modern economy. Governments frequently supply information on account of its public good characteristics. But the problem of *asymmetric information* poses additional challenges. **Asymmetric information** is where at least one party in an economic relationship has less than full information. This situation characterizes many interactions: Bosses do not always know how hard their subordinates work; life-insurance companies do not have perfect information on the lifestyle and health of their clients.

Asymmetric information is where at least one party in an economic relationship has less than full information and has a different amount of information from another party.

Asymmetric information can lead to two kinds of problems. The first is **adverse selection**. For example, can the life-insurance company be sure that it is not insuring only the lives of people who are high risk and likely to die young? If primarily high-risk people buy such insurance then the insurance company must set its premiums accordingly: The company is getting an adverse selection rather than a random selection of clients. Frequently governments decide to run universal compulsory-membership insurance plans (auto or health are examples in Canada) precisely because they may not wish to charge higher rates to higher-risk individuals.

Adverse selection occurs when incomplete or asymmetric information describes an economic relationship.

A related problem is **moral hazard**. If an individual does not face the full consequences of his actions, his behaviour may be influenced: If the boss cannot observe the worker's effort level, the worker may shirk. Or, if a homeowner has a fully insured home he may be less security conscious than an owner who does not.

In Chapter 7 we described how US mortgage providers lent large sums to borrowers with uncertain incomes in the early years of the new millennium. The lenders were being rewarded on the basis of the amount lent, not the safety of the loan. Nor were the lenders responsible for loans that were not repaid. This 'sub-prime mortgage crisis' was certainly a case of moral hazard.

Moral hazard may characterize behaviour where the costs of certain activities are not incurred by those undertaking them.

Solutions to these problems do not always involve the government, but in critical situations do. For example, the government requires most professional societies and orders to ensure that their members are trained, accredited and capable. Whether for a medical doctor, a plumber or an engineer, a license or certificate of competence is a signal that the work and advice of these professionals is *bona fide*. Equally, the government sets *standards* so that individuals do not have to incur the cost of ascertaining the quality of their purchases – bicycle helmets must satisfy specific crash norms; so too must air-bags in automobiles.

These situations differ from those where solutions to the information problem can be dealt with reasonably well in the market place. For example, life insurance companies can frequently establish the past medical history of its clients, and thus form an estimate of what the client's future health will be.

Concentration of Power

Monopolistic and imperfectly-competitive market structures can give rise to inefficient outcomes, in the sense that the value placed on the last unit of output does not equal the cost at the margin. In monopoly structures this arises because the supplier uses his market power in order to maximize profits.

What can governments do about such power concentrations? Every developed economy has a body similar to Canada's *Competition Bureau*. Such regulatory bodies are charged with seeing that the interests of the consumer, and the economy more broadly, are represented in the market place. Interventions, regulatory procedures and efforts to prevent the abuse of market power come in a variety of forms.

CHAPTER 15: Trade

15.1 TRADE BARRIERS

Source: Curtis & Irvine, 2016, Section 15.5, CC-BY-NC-SA 3.0 (Original source, pages 6 & 7 removed)

Despite the many good arguments favoring free or relatively free trade, we observe numerous trade barriers. These barriers come in several forms. A **tariff** is a tax on an imported product that is designed to limit trade in addition to generating tax revenue. It is a barrier to trade. There also exist **quotas**, which are quantitative restrictions on imports; other **non-tariff barriers**, such as product content requirements; and **subsidies**. By raising the domestic price of imports, a tariff helps domestic producers but hurts domestic consumers. Quotas and other **non-tariff barriers** have similar impacts.

A **tariff** is a tax on an imported product that is designed to limit trade in addition to generating tax revenue. It is a barrier to trade.

A **quota** is a quantitative limit on an imported product.

A **trade subsidy** to a domestic manufacturer reduces the domestic cost and limits imports.

Non-tariff barriers, such as product content requirements, limit the gains from trade.

Application Box 15.2 Tariffs – the national policy of J.A. MacDonald

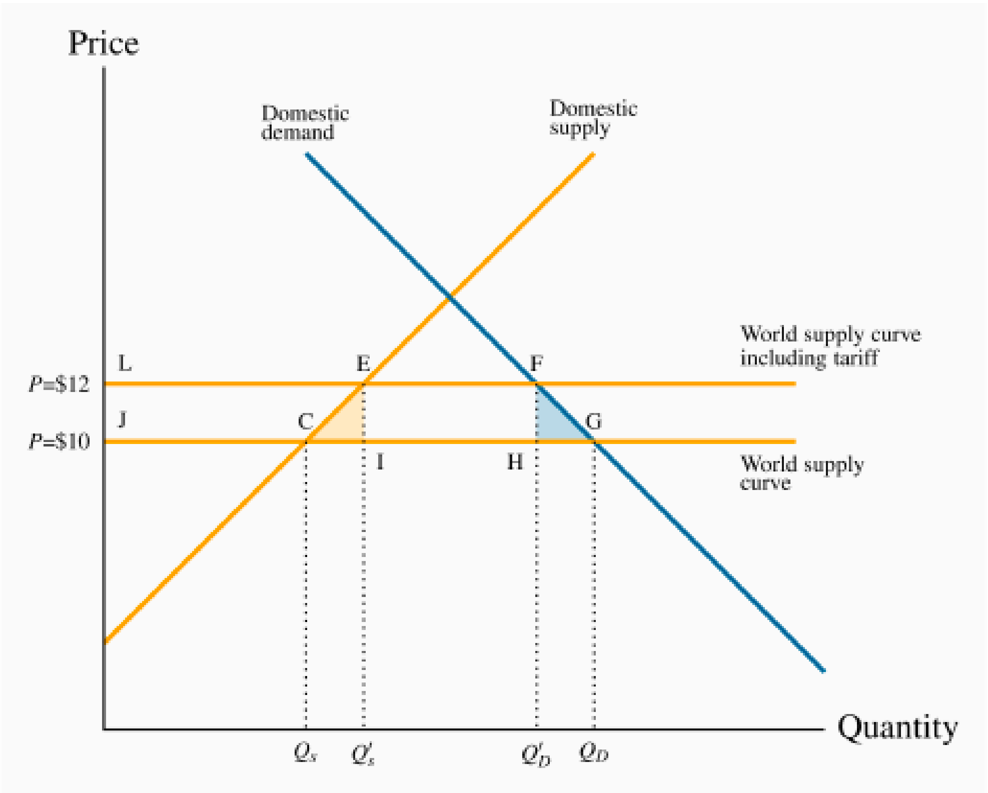
In Canada, tariffs were the main source of government revenues, both before and after Confederation in 1867 and up to World War I. They provided 'incidental protection' for domestic manufacturing. After the 1878 federal election, tariffs were an important part of the National Policy introduced by the government of Sir John A. MacDonald. The broad objective was to create a Canadian nation based on east-west trade and growth.

This National Policy had several dimensions. Initially, to support domestic manufacturing, it increased tariff protection on foreign manufactured goods, but lowered tariffs on raw materials and intermediate goods used in local manufacturing activity. The profitability of domestic manufacturing improved. But on a broader scale, tariff protection, railway promotion, Western settlement, harbour development, and transport subsidies to support the export of Canadian products were intended to support national economic development. Although reciprocity agreements with the United States removed duties on commodities for a time, tariff protection for manufactures was maintained until the GATT negotiations of the post-World War II era.

Tariffs

Figure 15.4 describes how tariffs operate. We can think of this as the Canadian wine market—a market that is heavily taxed in Canada. The world price of Cabernet Sauvignon is $10 per bottle, and this is shown by the horizontal world supply curve at that price. It is horizontal because our domestic market accounts for only a small part of the world demand for wine. International producers can supply us with any amount we wish to buy at the world price. The Canadian demand for this wine is given by the demand curve D, and Canadian suppliers have a supply curve given by S (Canadian Cabernet is assumed to be of the same quality as the imported variety in this example). At a price of $10, Canadian consumers wish to buy QDlitres, and domestic producers wish to supply QS litres. The gap between domestic supply QS and domestic demand QD is filled by imports. This is the *free trade equilibrium*.

Figure 15.4 Tariffs and trade



At a world price of $10 the domestic quantity demanded is Q_D*. Of this amount* Q_S *is supplied by domestic producers and the remainder by foreign producers. A tariff increases the world price to $12. This reduces demand to* Q'_D*; the domestic component of supply increases to* Q'_S*. Of the total loss in consumer surplus (LFGJ), tariff revenue equals EFHI, increased surplus for domestic suppliers equals LECJ, and the deadweight loss is therefore the sum of the triangular areas CEI and HFG.*

If the government now imposes a 20 percent tariff on imported wines, foreign wine sells for $12 a bottle, inclusive of the tariff. The tariff raises the domestic 'tariff-inclusive' price above the world price, and this shifts the supply of this wine upwards. By raising wine prices in the domestic market, the tariff protects domestic producers by raising the domestic price at which imports become competitive. Those domestic suppliers who were previously not quite competitive at a global price of $10 are now competitive. The total quantity demanded falls from QD to Q′D at the new equilibrium F. Domestic producers supply the amount Q′S and imports fall to the amount (Q′D−Q′S). Reduced imports are partly displaced by domestic producers who can supply at prices between $10 and $12. Hence, imports fall both because total consumption falls and because domestic suppliers can displace some imports under the protective tariff.

Since the tariff is a type of tax, its impact in the market depends upon the elasticities of supply and demand, (as illustrated in Chapters 4 and 5). The more elastic is the demand curve, the more a given tariff reduces imports. In contrast, if it is inelastic the quantity of imports declines less.

Costs and benefits of a tariff The costs of a tariff come from the higher price to consumers, but this is partly offset by the tariff revenue that goes to the government. This tariff revenue is a benefit and can be redistributed to consumers or spent on goods from which consumers derive a benefit. But there are also efficiency costs associated with tariffs—deadweight losses, as we call them. These are the real costs of the tariff, and they arise because the marginal cost of production does not equal the marginal benefit to the consumer. Let us see how these concepts apply with the help of Figure 15.4.

Consumer surplus is the area under the demand curve and above the equilibrium market price. It represents the total amount consumers would have been willing to pay for the product but did not have to pay at the equilibrium price. It is a measure of consumer welfare. The tariff raises the market price and reduces this consumer surplus by the amount LFGJ. This area measures by how much domestic consumers are worse off as a result of the price increase caused by the tariff. But this is not the net loss for the whole domestic economy, because the government obtains some tax revenue and domestic producers get more revenue and profit.

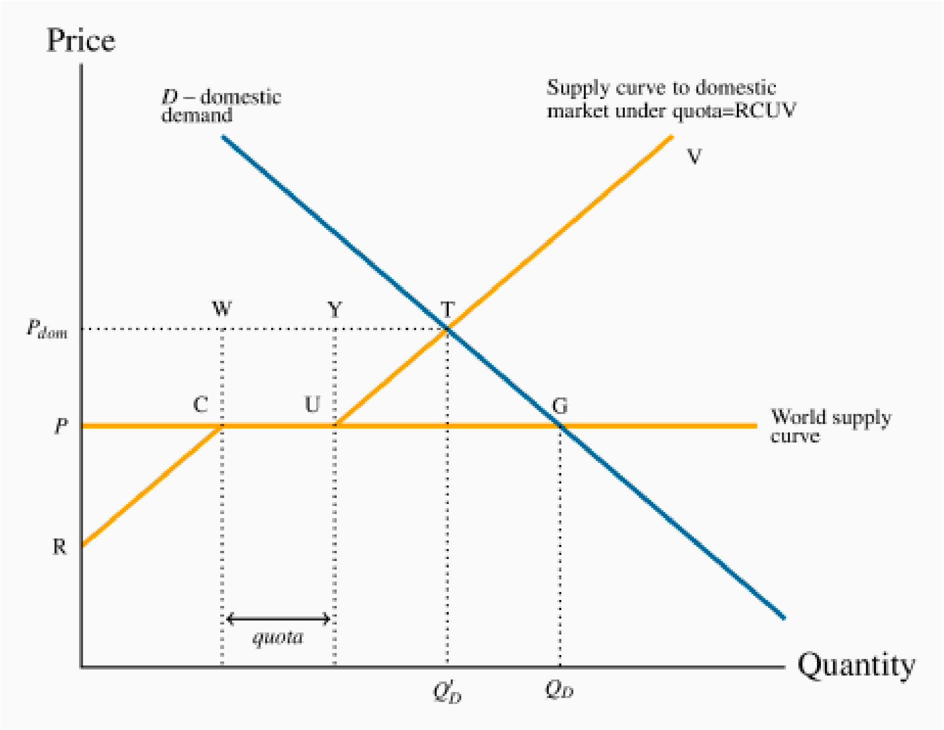
Government revenue accrues from the domestic sales of imports. On imports of (Q′D−Q′S), tax revenue is EFHI. Then, domestic producers obtain an additional profit of LECJ—the excess of additional revenue over their cost per additional bottle. If we are not concerned about who gains and who loses, then there is a net loss to the domestic economy equal to the areas CEI and HFG.

The area HFG is the standard measure of deadweight loss. At the quantity Q′D, the cost of an additional bottle is less than the value placed on it by consumers; and, by not having those additional bottles supplied, consumers forgo a potential gain. The area A tells us that when supply by domestic higher-cost producers is increased, and supply of lower-cost foreign producers is reduced, the corresponding resources are not being used efficiently. The sum of the areas CEI and HFG is therefore the total deadweight loss of the tariff.

Quotas

A quota is a limit placed upon the amount of a good that can be imported. Consider Figure 15.6, where again there is a domestic supply curve coupled with a world price of P. Rather than imposing a tariff, the government imposes a quota that restricts imports to a physical amount denoted by the distance *quota* on the quantity axis. The supply curve facing domestic consumers then has several segments to it. First it has the segment RC, reflecting the fact that domestic suppliers are competitive with world suppliers up to the amount C. Beyond this output, world suppliers can supply at a price of P, whereas domestic suppliers cannot compete at this price. Therefore the supply curve becomes horizontal, but only *up to the amount permitted under the quota*—the quantity CU corresponding to *quota*. Beyond this amount, international supply is not permitted and therefore additional amounts are supplied by the (higher cost) domestic suppliers. Hence the supply curve to domestic buyers becomes the supply curve from the domestic suppliers once again.

Figure 15.6 Quotas and trade



At the world price P, plus a quota, the supply curve becomes RCUV. This has three segments: (i) domestic suppliers who can supply below P; (ii) quota; and (iii) domestic suppliers who can only supply at a price above P. The quota equilibrium is at T, with price P_{dom} and quantity Q'_D; the free-trade equilibrium is at G. Of the amount Q'_D, quota is supplied by foreign suppliers and the remainder by domestic suppliers. The quota increases the price in the domestic market.

The resulting supply curve yields an equilibrium quantity Q′D. There are several features to note about this equilibrium. First, the quota pushes the domestic price above the world price (Pdom is greater than P) because low-cost international suppliers are partially supplanted by higher-cost domestic suppliers. Second, if the quota is chosen 'appropriately', the same domestic market price could exist under the quota as under the tariff in Figure 15.4. Third, in contrast to the tariff case, the government obtains no tax revenue from the quotas. Fourth, there are inefficiencies associated with the equilibrium at Q′D. These inefficiencies arise because the lower-cost international suppliers are not permitted to supply the amount they would be willing to supply at the quota-induced market equilibrium. In other words, more efficient producers are being squeezed out of the market by quotas that make space for less-efficient producers.

Application Box 15.3 Cheese quota in Canada

In 1978 the federal government set a cheese import quota for Canada at just over 20,000 tonnes. This quota was implemented initially to protect the interests of domestic suppliers. Despite a strong growth in population and income in the intervening decades, the import quota has remained unchanged. The result is a price for cheese that is considerably higher than it would otherwise be. The quotas are owned by individuals and companies who have the right to import cheese. The quotas are also traded among importers, at a price. Importers wishing to import cheese beyond their available quota pay a tariff of about 250 percent. So, while the consumer is the undoubted loser in this game, who gains?

First the suppliers gain, as illustrated in Figure 15.6. Canadian consumers are required to pay high-cost domestic producers who displace lower-cost producers from overseas. Second, the holders of the quotas gain. With the increase in demand for cheese that comes with higher incomes, the domestic cheese price increases over time and this in turn makes an individual quota more valuable.

Valid trade barriers: Infant industries and dumping?

Source: Curtis & Irvine, 2016, Section 15.6, CC-BY-NC-SA 3.0 (Original source, page 3 only)

An argument that carries both intellectual and emotional appeal to voters is the 'infant industry' argument. The argument goes as follows: New ventures and sectors of the economy may require time before that can compete internationally. Scale economies may be involved, for example, and time may be required for producers to expand their scale of operation, at which time costs will have fallen to international (i.e. competitive) levels. In addition, learning-by-doing may be critical in more high-tech sectors and, once again, with the passage of time costs should decline for this reason also.

The problem with this stance is that these 'infants' have insufficient incentive to 'grow up' and become competitive. A protection measure that is initially intended to be temporary can become permanent because of the potential job losses associated with a cessation of the protection to an industry that fails to become internationally competitive. Furthermore, employees and managers in protected sectors have insufficient incentive to make their production competitive if they realize that their government will always be there to protect them.

In contrast to the infant industry argument, economists are more favourable to restrictions that are aimed at preventing 'dumping'. **Dumping** is a predatory practice, based on artificial costs aimed at driving out domestic producers.

Dumping is a predatory practice, based on artificial costs aimed at driving out domestic producers.

Dumping may occur either because foreign suppliers choose to sell at artificially low prices (prices below their marginal cost for example), or because of surpluses in foreign markets resulting from oversupply. For example, if, as a result of price support in its own market, a foreign government induced oversupply in butter and it chose to sell such butter on world markets at a price well below the going ('competitive') world supply price, such a sale would constitute dumping. Alternatively, an established foreign supplier might choose to enter our domestic market by selling its products at artificially low prices, with a view to driving domestic competition out of the domestic market. Having driven out the domestic competition it would then be in a position to raise prices. This is predatory pricing as explored in the last chapter. Such behaviour differs from a permanently lower price on the part of foreign suppliers. This latter may be welcomed as a gain from trade, whereas the former may generate no gains and serve only to displace domestic labour and capital.

15.2 INSTITUTIONS GOVERNING TRADE

Source: Curtis & Irvine, 2016, Section 15.7, CC-BY-NC-SA 3.0

In the nineteenth century, world trade grew rapidly, in part because the leading trading nation at the time—the United Kingdom—pursued a vigorous policy of free trade. In contrast, US tariffs averaged about 50 percent, although they had fallen to around 30 percent by the early 1920s. As the industrial economies went into the Great Depression of the late 1920s and 1930s, there was pressure to protect domestic jobs by keeping out imports. Tariffs in the United States returned to around 50 percent, and the United Kingdom abandoned the policy of free trade that had been pursued for nearly a century. The combination of world recession and increasing tariffs led to a disastrous slump in the volume of world trade, further exacerbated by World War II.

The WTO and GATT

After World War II, there was a collective determination to see world trade restored. Bodies such as the International Monetary Fund and the World Bank were set up, and many countries signed the General Agreement on Tariffs and Trade (GATT), a commitment to reduce tariffs successively and dismantle trade restrictions.

Under successive rounds of GATT, tariffs fell steadily. By 1960, United States tariffs were only one-fifth their level at the outbreak of the War. In the United Kingdom, the system of wartime quotas on imports had been dismantled by the mid-1950s, after which tariffs were reduced by nearly half in the ensuing 25 years. Europe as a whole moved toward an enlarged European Union in which tariffs between member countries have been abolished. By the late 1980s, Canada's tariffs had been reduced to about one-quarter of their immediate post-World War II level.

The GATT Secretariat, now called the World Trade Organization (WTO), aims both to dismantle existing protection that reduces efficiency and to extend trade liberalization to more and more countries. Tariff levels throughout the world are now as low as they have ever been, and trade liberalization has been an engine of growth for many economies. The consequence has been a substantial growth in world trade.

NAFTA, the EU and the TPP

In North America, recent policy has led to a free trade area that covers the flow of trade between Canada, the United States, and Mexico. The Canada/United States free trade agreement (FTA) of 1989 expanded in 1994 to include Mexico in the North American Free Trade Agreement (NAFTA). The objective in both cases was to institute free trade between these countries in most goods and services. This meant the elimination of tariffs over a period of years and the reduction or removal of non-tariff barriers to trade, with a few exceptions in specific products and cultural industries. A critical component of the Agreement was the establishment of a dispute-resolution mechanism, under which disputes would be resolved by a panel of 'judges' nominated from the member economies. Evidence of the success of these agreements is reflected in the fact that Canadian exports have grown to more than 30 percent of GDP, and trade with the United States accounts for the lion's share of Canadian trade flows. As of 2015, most of the Pacific Rim economies were engaged in negotiating a new trade pact – the Trans-Pacific Partnership.

The European Union was formed after World War II, with the prime objective of bringing about a greater degree of *political* integration in Europe. Two world wars had laid waste to their economies and social fabric. Closer economic ties and greater trade were seen as the means of achieving this integration. The Union was called the "Common Market" for much of its existence. The Union originally had six member states, but as of 2015 the number is 28, with several other candidate countries in the process of application, most notably Turkey. The European Union (EU) has a secretariat and parliament in Bruxelles. You can find more about the EU at http://europa.eu.

Another Useful Link: Economic Integration: Free Trade Areas, Trade Creation, and Trade Diversion