Directorate of Distance Education

UNIVERSITY OF JAMMU

JAMMU

SELF LEARNING MATERIAL

M.A. ECONOMICS

SEMESTER-I — **COURSE NO. ECO-101**

UNIT I-IV — **LESSON 1-24**

Prof. Prakash C. Antahal
Course Coordinator

Dr. Neelam Choudhary
Teacher Incharge

http:/www.distanceeducationju.in

Printed and Published on behalf of the Directorate of Distance Education, University of Jammu, Jammu by the Director, DDE, University of Jammu, Jammu.

MICROECONOMICS

COURSE CONTRIBUTOR:

Prof. Falendra Kumar Sudan

Printed at :- **SETHI ART PRINTERS** /20/ 500 No.

Semester I

Foundation Courses

Detailed Syllabus

Course No: ECO- 101 **Title: Microeconomics**
Credits: 6 **Maximum Marks: 100**
Sessional Assessment : 20 marks
External Examination : 80 marks

Duration of Major Examination: 3:00 hrs

Microeconomics

Syllabus for the Examination to be held in December 2019 to December 2021

Preamble: This foundation course on Microeconomics intends to help students see Economics as a coherent whole, and to reinforce its usefulness through the applications to the real world problems and their solutions. Most economists, when approaching a problem, begin by thinking about buyers and sellers, and the markets in which they come together to trade. To understand what Economics is about, the students need to understand demand and supply, consumer and producer behaviour, and the markets and its functioning. To help them do so, this foundation course has identified and stressed on a "three-step process" that economists use in analyzing problems. The three key steps are: characterize the market (decide which market or markets best suit the problem being analyzed, and identify the decision makers - buyers and sellers who interact there, find the equilibrium (describe the conditions necessary for equilibrium in the market, and a method for determining that equilibrium), and determine what happens when things change (explore how events or government policies change the market equilibrium). Therefore, this foundation course is organized around this three-step process, which will help the students learn how to think like economists, and in a very natural way so that they see Economics as a unified whole, rather than as a series of disconnected ideas.

UNIT – I: Theory of Demand and Supply

Scarcity and Individual Choice, Opportunity Cost, Scarcity and Social Choice, Opportunity Cost and Society's Trade Offs, Society's Production Choices, Production Possibilities

Frontier and Opportunity Cost, Operating Inside Production Possibilities Frontier, Resource Allocation: Reasons and Methods, Market and its Characteristics, Law of Demand, Law of Supply, Equilibrium Price and Quantity, Equilibrium with Change in Demand and Supply

Elasticity of Demand, Price Elasticity of Demand, Slope of Demand Curve, Elasticity Approach: Calculating Price Elasticity of Demand, Categorizing Demand, Elasticity and Straight-Line Demand Curves, Elasticity and Total Revenue, Determinants of Elasticity, Time Horizons and Demand Curves, Income Elasticity of Demand, Cross-Price Elasticity of Demand, Price Elasticity of Supply

UNIT – II: Theory of Consumer Behaviour

Consumer Choice: Budget Constraint, Changes in Budget Line, Consumer Preferences, Consumer Decisions: Marginal Utility Approach, Combining Budget Constraint and Preferences, Consumer Decisions with Change in Income and Price, Consumer's Demand Curve, Income and Substitution Effects, Consumers in Markets

Indifference Curve Approach: Indifference Curve and Marginal Rate of Substitution, Properties of Indifference Curve, Indifference Map, Indifference Curve and Consumer Decision Making, Change in Income and Consumer Decisions, Deriving Demand Curve with Indifference Curves, Consumer Decision and Revealed Preference Approach, Consumer Surplus

UNIT – III: Theory of Production and Cost

Production Function, Short-Run versus Long-Run Production Decisions, Production in Short Run, Marginal Returns to Labour, Concept of Cost, Sunk Costs, Explicit versus Implicit Costs, Cost in Short Run: Measuring Short-Run Costs, Shape of Marginal Cost Curve, Relationship between Average and Marginal Costs, Production and Cost in Long Run, Relationship between Long-Run and Short-Run Costs, Explaining the Shape of *LRATC* Curve

Concept of Isoquant and its Properties, Marginal Rate of Technical Substitution, Isocost Lines, Properties of Isocost Lines, Least-Cost Input Combination, Firms Decisions: Goal of Profit Maximization, Firm's Constraints: Demand Curve Facing Firm, Cost Constraint, Profit-Maximizing Output Level: Total Revenue (TR) and Total Cost (TC) Approach, Marginal Revenue (MR) and Marginal Cost (MC) Approach, Profit Maximization Using

Graphs: TR-TC Approach, MR-MC Approach, Dealing with Losses: Short Run and Shutdown Rule, Long Run and Exit Decision

UNIT – IV: Theory of Markets

Perfect Competition: Characteristics and Profit Maximization, Profit-Maximizing Output Level: TR-TC Approach and MC-MR Approach, Measuring Profit or Loss, Firm's Short-Run Supply Curve and Shut Down Price, Competitive Markets: Short-Run and Long-Run Equilibrium, Perfect Competition and Plant Size, Competitive Firm and Changes in Demand, Market Signals and Economy: Change in Demand and Reallocation of Resources, Technological Change in Perfect Competition

Monopoly and its Causes, Monopoly Behaviour: Price or Output Decision, Monopoly: Profit and Loss, Monopoly Markets: Short-Run and Long-Run Equilibrium, Monopoly Vs Perfect Competition, Monopoly Decisions and Changes in Demand and Cost-Saving Technology, Price Discrimination: Conditions, Effects and Types, Price Discrimination and Multiple Prices

Monopolistic Competition: Characteristics, Monopolistic Competition in Short Run and Long Run, Oligopoly and Its Causes, Competition versus Cartel, Non-Cooperative Oligopoly: Cournot Model, Stackelberg Model, and Bertrand Model, Oligopoly versus Other Market Structures: Game Theory Approach, Simple Oligopoly Games, Cooperative Behaviour in Oligopoly, Advertising in Monopolistic Competition and Oligopoly

NOTE FOR PAPER SETTER :

There shall be two types of questios in each Unit four short answer type (each of 250 words) and two medium answer type (each of 500 words). The candidate will have to attempt two short answer type questions and one medium answer type question from each Unit. Each short answer type question shall carry 4 marks and each medium anwer type question carry 12 marks.

Basic Readings:

1. Baumol, W.j. (1982) Economic Theory and Operations Anaylysis, Prentice Hall of India, New Delhi, Delhi.

2. Da Costa G.C (1980) Production, Prices and Distribution, Tata Macgraw Hill, New Delhi

3. Hirshleifer, J and A. Glazer (1997) Price Theory and Application, Prentice Hall of India, New Delhi.

4. Jack Hirshleifer, Amihal Glazer (1997) Price Theory and Application, Prentice Hall of India, New Delhi.

5. Koutsoyiannis, A (1979), Modern Microeconomics, (2nd Edition) Macmillan Press, London.

6. Richard G Lipsey: Colin Harbury, (1992) First Principles of Economics, (2nd Edition), Oxford University Press.

7. Stigler, G.(1996) Theory of Price, (4th Edition), Prentice Hall of India, New Delhi.

8. Varian, H. (2000) Microeconomics Analysis, W.W. Norton, New York

TABLE OF CONTENTS

This lesson will focus on the following:

1. Scarcity and Individual Choice
2. Opportunity Cost
3. Scarcity and Social Choice
4. Opportunity Cost and Society's Trade Offs
5. Society's Production Choices

Introduction

The word *economy* comes from the Greek word *oikonomos*, which means "one who manages a household." At first, this origin might seem peculiar. But in fact, households and economies have much in common. A household faces many decisions. It must decide which members of the household do which tasks and what each member gets in return:

a) Who cooks dinner?

b) Who does the laundry?

c) Who gets the extra dessert at dinner?

d) Who gets to choose?

e) What TV show to watch?

In short, the household must allocate its scarce resources among its various members, taking into account each member's abilities, efforts and desires.

Like a household, a society faces many decisions. A society must find some way to decide what jobs will be done and who will do them. It needs some people to grow food, other people to make clothing, and still others to design computer software. Once society has allocated people (as well as land, buildings, and machines) to various jobs, it must also allocate the output of goods and services they produce. The management of society's resources is important because resources are scarce.

Meaning of Scarcity

Scarcity means that society has limited resources and therefore cannot produce all the goods and services people wish to have. Just as each member of a household cannot get everything he or she wants, each individual in a society cannot attain the highest standard of living to which he or she might aspire.

Meaning of Economics

Economics is the study of how society manages its scarce resources. In most societies, resources are allocated not by an all-powerful dictator but through the combined actions of millions of households and firms. Economists therefore study how people make decisions:

1. How much they work?
2. What they buy?
3. How much they save? and
4. How they invest their savings?

Economists also study how people interact with one another. For instance, they examine how the multitude of buyers and sellers of a good together determine the price at which the good is sold and the quantity that is sold.

Finally, economists analyze forces and trends that affect the economy as a whole, including the growth in average income, the fraction of the population that cannot find work, and the rate at which prices are rising.

Field of Economics

The field of economics is divided into two major parts: microeconomics and macroeconomics.

Microeconomics comes from the Greek word *mikros*, meaning "small." It takes a close-up view of the economy, as if looking through a microscope. Microeconomics is concerned with the behaviour of *individual* actors on the economic scene—households, business firms, and governments.

It looks at the choices they make and how they interact with each other when they come together to trade *specific* goods and services.

What will happen to the cost of movie tickets over the next five years?

How many management-trainee jobs will open up for college graduates?

These are microeconomic questions because they analyze individual *parts* of an economy rather than the *whole*.

Macroeconomics comes from the Greek word *makros*, meaning "large"—takes an *overall* view of the economy. Instead of focusing on the production of carrots or computers, macroeconomics lumps all goods and services together and looks at the economy's *total output*. Instead of focusing on employment of management trainees or manufacturing workers, it considers *total employment* in the economy. Macroeconomics focuses on the big picture.

Positive Economics & Normative Economics

The micro versus macro distinction is based on the level of detail we want to consider. Another useful distinction has to do with our *purpose* in analyzing a problem.

Positive economics explains how the economy works, plain and simple. If someone says, "The decline in home prices during 2017 and 2018 was a major cause of the recent recession," he or she is making a positive economic statement. A statement need not be accurate or even sensible to be classified as positive.

For example, "Government policy has no effect on our standard of living" is a statement that virtually every economist would regard as false. But it is still a positive economic statement. Whether true or not, it's about how the economy works and its accuracy can be tested by looking at the facts—and just the facts.

Normative economics *prescribes solutions* to economic problems. It goes beyond just "the facts" and tells us what we should *do* about them. Normative economics requires us to make judgments about different outcomes and therefore depends on our values.

If an economist says, "We should cut total government spending," he or she is engaging in normative economic analysis. Cutting government spending would benefit some citizens and harm others, so the statement rests on a value judgment. A normative statement—like the one about government spending earlier—cannot be proved or disproved by the facts alone.

Positive and normative economics are intimately related in practice. For one thing, we cannot properly argue about what we should or should not do unless we know certain facts about the world. Every normative analysis is therefore based on an underlying positive analysis. But while a positive analysis can, at least in principle, be conducted without value judgments, a normative analysis is always based, at least in part, on the values of the person conducting it.

Why Economists Disagree about Policy

Suppose the country is suffering from a serious recession—a significant, nationwide decrease in production and employment. Two economists are interviewed on a cable news show. A says, "We should increase government spending on roads, bridges, and other infrastructure. This would directly create jobs and help end the recession." Economist B says, "No, we should cut taxes instead. This will put more money in the hands of households and businesses, leading them to spend more and create jobs that way." Why do they disagree?

It might be based on *positive* economics—different views about how the economy works. Economist A might think that government spending will create more jobs, dollar for dollar, than will tax cuts. Economist B might believe the reverse. Positive differences like these can arise because our knowledge of how the economy works—while always improving—remains imperfect.

But the disagreement might stem from a difference in values—specifically, what each economist believes about government's proper role in the economy. Those toward the left of the political spectrum tend to believe that government should play a larger economic role. They tend to view increases in government spending more favorably. Those toward the right tend to believe that government's role should be smaller. They would prefer tax cuts that result in more private, rather than government, spending. This difference in values can explain why two economists—even if they have the same *positive* views about the outcome of a policy—might disagree about its wisdom.

Policy differences among economists arise from

(1) positive disagreements(about what the outcome of different policies will be), or

(2) differences in values (how those outcomes are evaluated).

Policy disputes among economists are common. But on *some* policy issues, most economists agree. For example, in microeconomics there is wide agreement that certain types of goods and services should be provided by private business firms and that certain others are best provided by government.

Why Study Economics?

- **To Understand the World Better**

Applying the tools of economics can help you understand global and catastrophic events such as wars, famines, epidemics, and depressions. But it can also help you understand much of what happens to you locally and personally—the salary you will earn after you graduate, or the rent you'll pay on your apartment. Economics has the power to help us understand these phenomena because they result, in large part, from the choices we make under conditions of scarcity.

Economics has its limitations, of course. But it is hard to find any aspect of life about which economics does not have *something* important to say. Economics cannot explain why so many Indian like to watch television, but it *can* explain how TV networks decide which programs to offer.

- **To Achieve Social Change**

If you are interested in making the world a better place, economics is indispensable. There is no shortage of serious social problems worthy of our attention— unemployment, hunger, poverty, disease, child abuse, drug addiction, violent crime. Economics can help us understand the origins of these problems, explain why previous efforts to solve them haven't succeeded, and help us to design new, more effective solutions.

- **To Help Prepare for Other Careers**

Economics has long been a popular subject in college for individuals intending to work in business. But it has also been popular among those planning careers in politics, international relations, law, medicine, engineering, psychology, and other professions. This is for good reason: Practitioners in each of these fields often find themselves confronting economic issues. For example, lawyers increasingly face judicial rulings based on the principles of economic efficiency. And doctors will need to understand how new technologies or changes in the structure of health insurance will affect their practices.

- **To Become an Economist**

Economists are hired by banks to assess the risk of investing abroad; by manufacturing companies to help them determine new methods of producing, marketing, and pricing their products; by government agencies to help design policies to fight crime, disease, poverty, and pollution; by international organizations to help create and reform aid programs for less developed countries; by the media to help the public interpret global, national, and local events; and by nonprofit organizations to provide advice on controlling costs and raising funds more effectively.

Explaining Resource Scarcity

We all face the problem of **scarcity**.

At first glance, it may seem that you suffer from an infinite variety of scarcities. There are so many things you might like to have right now—a larger room or apartment, a new car, more clothes . . . the list is endless. But a little reflection suggests that your limited ability to satisfy these desires is based on two other, more basic limitations:

scarce *time* and scarce *spending power. As individuals, we face a scarcity of time and spending power. Given more of either, we could each have more of the goods and services that we desire.*

Scarcity: A situation in which the amount of something available is insufficient to satisfy the desire for it. The scarcity of spending power is no doubt familiar to you. We've all wished for higher incomes so that we could afford to buy more of the things we want.

But the scarcity of time is equally important. So many of the activities we enjoy—seeing movies, taking vacations, making phone calls—require time as well as money. Just as we have limited spending power, we also have a limited number of hours in each day to satisfy our desires. Because of the scarcities of time and spending power, each of us is forced to make *choices*.

We must allocate our scarce *time* to different activities: work, play, education, sleep, shopping, and more. We must allocate our scarce *spending power* among different goods and services: housing, food, furniture, travel, and many others. And each time we choose to buy something or do something, we also choose *not* to buy or do something else.

Economists study the choices we make as individuals, as well as their consequences. When some of the consequences are harmful, economists study what—if anything—the government can or should do about them. For example, in India, as incomes have risen, more and more people have chosen to purchase automobiles. The result is increasing traffic jams in our major cities. The problem is even worse in rapidly developing countries.

In India, for example, recent income growth and migration from rural to urban areas has led to an explosion of driving. Economists have come up with some creative ideas to reduce traffic congestion, while preserving individual choices about driving. A few cities have used these ideas, with some success, and more are considering them.

Explaining Opportunity Cost

What does it cost you to go to the movies? If you answered 150 or 220 rupees because that is the price of a movie ticket, then you are leaving out a lot. Most of us are used to thinking of "cost" as the money we must pay for something. Certainly, the money we pay for goods or services is a *part* of its cost.

But economics takes a broader view of costs. The true cost of any choice we make—buying a car, producing a computer, or even reading a book—is everything we must *give up* when we take that action.

This cost is called the *opportunity cost* of the action, because we give up the opportunity to have other desirable things. *The **opportunity cost** of any choice is what we must forego when we make that choice.*

Opportunity cost is the most accurate and complete concept of cost—the one we should use when making our own decisions or analyzing the decisions of others. Suppose, for example, it's 8 p.m. on a weeknight and you're spending a couple of hours reading this chapter.

We know there are many other things you could be doing: going to a movie, having dinner with friends, playing ping pong, earning some extra money, watching TV. . . . But, assuming you're still reading—and you haven't just run out the door because we've given you better ideas—let's relate this to opportunity cost.

What *is* the opportunity cost of reading this lesson? Is it *all* of those other possibilities we've listed? Not really, because in the time it takes to read this lesson, you'd probably be able to do only *one* of those other activities. You'd no doubt choose whichever one you regarded as best. So, by reading, you sacrifice only the *best* choice among the alternatives that you could be doing instead.

When the alternatives to a choice are mutually exclusive, only the next best choice—the one that would actually be chosen—is used to determine the opportunity cost of the choice.

For many choices, a large part of the opportunity cost is the money sacrificed. If you spend Rs. 150 on a new DVD, you have to part with Rs. 150, which is money you could have spent on something else (whatever the best choice among the alternatives turned out to be). But for other choices, money may be only a small part, or no part, of what is sacrificed. If you walk your dog a few blocks, it will cost you time but not money.

Still, economists often like to attach a monetary value even to the parts of opportunity cost that *don't* involve money. The opportunity cost of a choice can then be expressed as a dollar value, albeit a roughly estimated one. That, in turn, enables us to compare the cost of a choice with its benefits, which we also often express in dollars.

Costs like these—for which rupees are actually paid out—are called **explicit costs**, and they are *part* of the opportunity cost. But there are also the **implicit costs**—sacrifices for which no money changes hands.

The biggest sacrifice in this category is *time*. But what is that time worth?

That depends on what you *would* be doing if you weren't in school. For many students, the alternative would be working full-time at a job. If you are one of these students, attending college requires the sacrifice of the income you *could* have earned at a job—a sacrifice we call *foregone income*.

Summing the explicit and implicit costs gives us a rough estimate of the opportunity cost.

Much of that income is earned in the future, and a rupee gained years from now is worth less than a rupee spent today. Also, *some* of the higher earnings of college graduates result from the **personal characteristics** of people who are likely to attend college, rather than from the education or the degree itself.

But even when we make reasonable adjustments for these facts, attending college appears to be one of the best ***financial*** investments you can make.

Time Is Money

"Time is money." For some people, this maxim applies directly: when they spend time on something, they *actually* give up money—money they *could* have earned during that time. Consider Dinesh, a freelance writer with a backlog of projects on which he can earn Rs. 250 per hour. For each hour Dinesh spends *not* working, he sacrifices Rs. 250.

1. What if Dinesh decides to see a movie?
2. What is the opportunity cost, in rupee terms?

Suppose the ticket costs Rs. 100 and the entire activity takes three hours— including time spent getting there and back. The opportunity cost is the sum of the explicit cost (Rs. 100 for the ticket) and the implicit cost (Rs.750 for three hours of foregone income), making the total opportunity cost Rs. 850.

The idea that a movie "costs" Rs. 85 might seem absurd. But if you think about it, Rs. 850 is a much better estimate than Rs. 100 of what the movie actually costs Dinesh—Rs. 850 is what he sacrifices to see the movie.

The explicit (direct money) cost of a choice may only be a part—and sometimes a small part—of the opportunity cost of a choice.

Circular Flow in Economy

The economy consists of millions of people engaged in many activities—buying, selling, working, hiring, manufacturing, and so on. To understand how the economy works, we must find some way to simplify our thinking about all these activities. In other words, we need a model that explains, in general terms, how the economy is organized and how participants in the economy interact with one another.

Figure 1 presents a visual model of the economy called a **circular-flow diagrammememe**. In this model, the economy is simplified to include only two types of decision makers—firms and households. Firms produce goods and services using inputs, such as labour, land, and capital (buildings and machines).

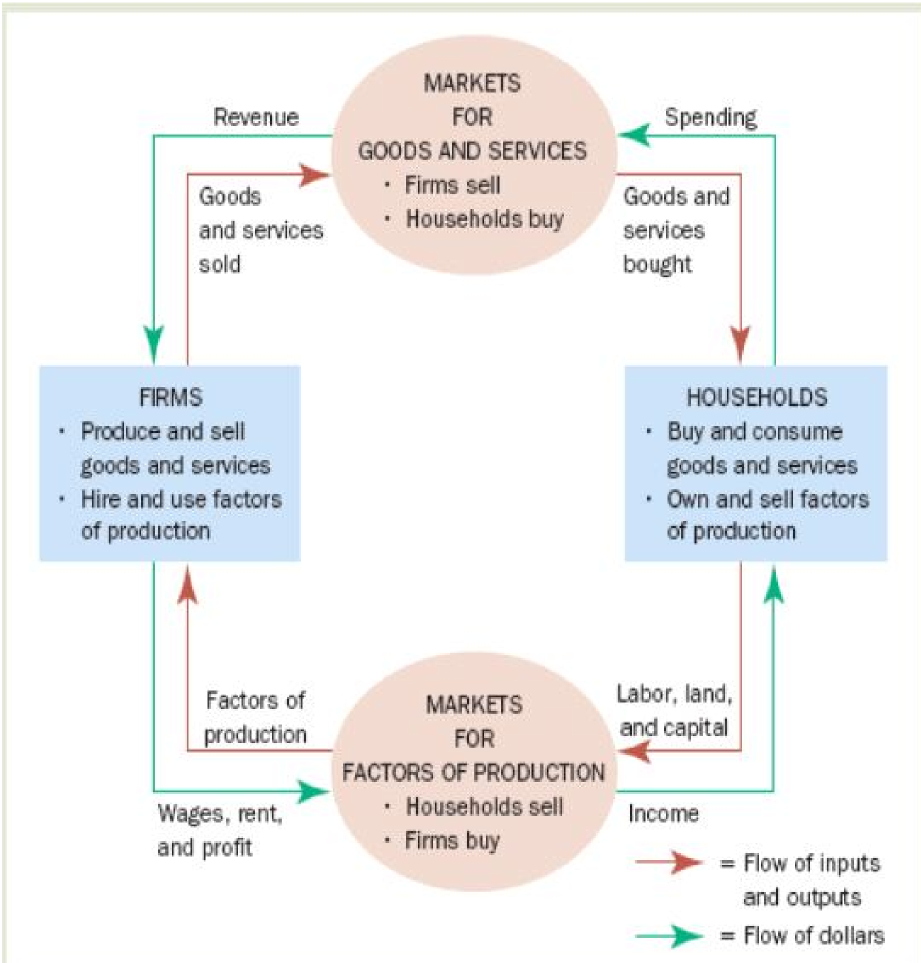
These inputs are called the *factors of production.* Households own the factors of production and consume all the goods and services that the firms produce. Households and firms interact in two types of markets. In the *markets for goods and services,* households are buyers, and firms are sellers. In particular, households buy the output of goods and services that firms produce. In the *markets for the factors of production,* households are sellers, and firms are buyers. In these markets, households provide the inputs that firms use to produce goods and services.

The circular-flow diagramme offers a simple way of organizing the economic transactions that occur between households and firms in the economy. The two loops of the circular-flow diagramme are distinct but related. The inner loop represents the flows of inputs and outputs. The households sell the use of their labour, land, and capital to the firms in the markets for the factors of production.

The firms then use these factors to produce goods and services, which in turn are sold to households in the markets for goods and services. The outer loop of the diagramme represents the corresponding flow of dollars. The households spend money to buy goods and services from the firms. The firms use some of the revenue from these sales to pay for the factors of production, such as the wages of their workers. What's left is the profit of the firm owners, who themselves are members of households.

The circular-flow diagramme in Figure 1 is one simple model of the economy. It dispenses with details that, for some purposes, are significant. A more complex and realistic circular-flow model would include, for instance, the roles of government and international trade.

Figure 1: The Circular Flow



This diagramme is a schematic representation of the organization of the economy. Decisions are made by households and firms. Households and firms interact in the markets for goods and services (where households are buyers and firms are sellers) and in the markets for the factors of production (where firms are buyers and households are sellers). The outer set of arrows shows the flow of dollars, and the inner set of arrows shows the corresponding flow of inputs and outputs.

Scarcity and Social Choice

Now let's think about scarcity and choice from *society*'s point of view.

1. What are the goals of our society?
2. What is holding us back from accomplishing all of these goals in a way that would satisfy everyone?

You already know the answer: scarcity.

In society's case, the problem is a scarcity of **resources**—the things we use to make goods and services that help us achieve our goals.

Four Resources

Resources are the most basic elements used to make goods and services. We can classify resources into four categories:

• **Labour**—the time human beings spend producing goods and services.

• **Capital**—any long-lasting tool, that is itself produced, and helps us make other goods and services.

More specifically, **physical capital** consists of things like machinery and equipment, factory buildings, computers, and even hand tools like hammers and screwdrivers. These are all long-lasting *physical* tools that we produce to help us make other goods and services.

Another type of capital is **human capital**—the skills and knowledge possessed by workers. These satisfy our definition of capital: They are *produced* (through education and training), they help us produce *other* things, and they last for many years, typically through an individual's working life.

Note the word *long-lasting* in the definition. If something is used up quickly in the production process—like the flour a baker uses to make bread—it is generally *not* considered capital. A good rule of thumb is that capital should last at least a year, although most types of capital last considerably longer.

The **capital stock** is the total amount of capital at a nation's disposal at any point in time. It consists of all the capital—physical and human—created in previous periods that is still productively useful.

• **Land**—the physical space on which production takes place, as well as useful materials—*natural resources*—found under it or on it, such as crude oil, iron, coal, or fertile soil.

• **Entrepreneurship**—the ability (and the willingness to *use* it) to combine the *other* resources into a productive enterprise. An entrepreneur may be an *innovator* who comes up with an original idea for a business or a *risk taker* who provides her own funds or time to nurture a project with uncertain rewards. Anything *produced* in the economy comes, ultimately, from some combinations of the four resources.

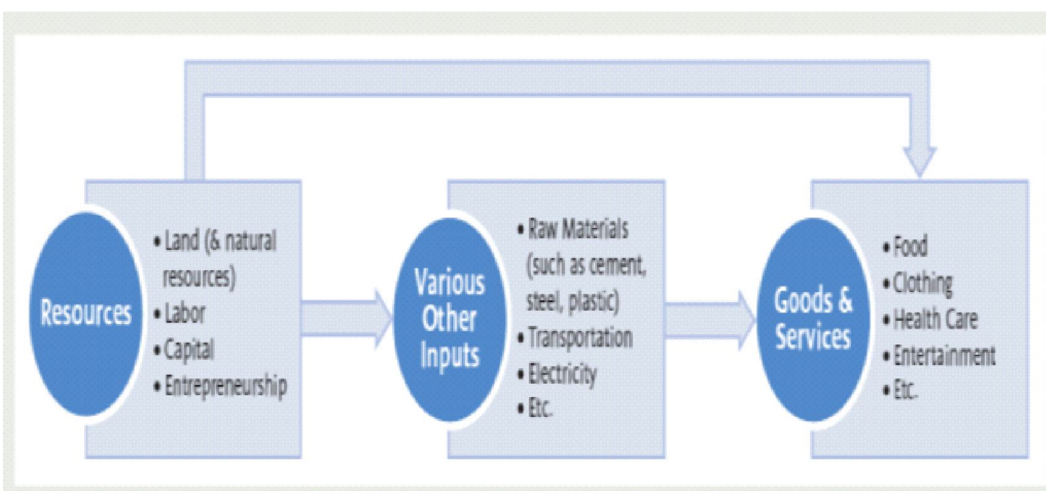
Resources versus inputs

The term *resources* is often confused with another, more general term—**inputs.** An input is *anything* used to make a good or service. Inputs include not only resources but also many other things made from them (cement, rolled steel, electricity), which are, in turn, used to make goods and services. *Resources*, by contrast, are the *special* inputs that fall into one of four categories: labour, land, capital, and entrepreneurship. They are the ultimate source of everything that is produced.

Opportunity Cost and Society's Tradeoff

For an individual, opportunity cost arises from the scarcity of time or money. But for society as a whole, opportunity cost arises from the scarcity of *resources*. Our desire for goods is limitless, but we have limited resources to produce them. Therefore, *virtually all production carries an opportunity cost: To produce more of one thing, society must shift resources away from producing something else.*

Figure 2: Resources and Production



All goods and services come ultimately from the four resources. Resources are used directly by firms that produce goods and services. They are also used indirectly, to make the other inputs firms use to produce goods and services.

For example, we'd all like better health for our citizens.

What would be needed to achieve this goal?

Perhaps more frequent medical checkups for more people and greater access to top-flight medicine when necessary. These, in turn, would require more and better-trained doctors, more hospital buildings and labouratories, and more high-tech medical equipment. In order for us to produce these goods and services, we would have to pull resources—land, labour, capital, and entrepreneurship—out of producing other things that we also enjoy. The opportunity cost of improved health care, then, consists of those other goods and services we would have to do without.

Society's Production Choices

Let's consider a specific choice that faces every society.

How much of its resources to allocate toward national defence versus how much to use for civilian production. To make this choice more concrete, we'll make a simplifying assumption: In the economy we're studying, there is one kind of military good (tanks) and one kind of civilian good (wheat).

Table 1 lists some possible combinations of yearly tank production and yearly wheat production this society could manage, given its available resources and the currently available production technology.

For example, the first row of the table (choice A) tells us what would happen if all available resources were devoted to wheat production and no resources at all to producing tanks. The resulting quantity of wheat—1 million bushels per year—is the most this society could possibly produce. In the second row (choice B), society moves enough resources into tank production to make 1,000 tanks per year. This leaves fewer resources for wheat production, which now declines to 950,000 bushels per year.

Table 1: Production of Tanks and Wheat

Choice	Tank Production (number per year)	Wheat Production (bushels per year)
A	0	1,000,000
B	1,000	950,000
C	2,000	850,000
D	3,000	700,000
E	4,000	400,000
F	5,000	0

As we continue down the table, moving to choices C, D, and E, tank production increases by increments of 1,000. The last column shows us the maximum quantity of wheat that can be produced for each given quantity of tanks. Finally, look at the last row (choice F). It shows us that when society throws all of its resources into tank production (with none for wheat), tank production is 5,000 and wheat production is zero.

The table gives us a quantitative measure of opportunity cost for this society. For example, suppose this society currently produces 1,000 tanks per year, along with 950,000 bushels of wheat (choice B). What would be the opportunity cost of producing another 1,000 tanks? Moving down to choice C, we see that producing another 1,000 tanks (for a total of 2,000) would require wheat production to drop from 950,000 to 850,000 bushels, a decrease of 100,000 bushels. Thus, the opportunity cost of 1,000 more tanks is 100,000 bushels of wheat. The opportunity cost of having more of one good is measured in the units of the other good that must be sacrificed.

One of the most fundamental concepts in economics is *opportunity cost*. The opportunity cost of any choice is what we give up when we make that choice. *Economics* is the study of choice under conditions of scarcity. As individuals—and as a society—we have unlimited desires for goods and services.

Unfortunately, our ability to satisfy those desires is limited, so we must usually sacrifice something for any choice we make. The correct measure of the cost of a choice is not just the money price we pay, but the *opportunity cost*: what we must give up when we make a choice.

At the individual level, opportunity cost arises from the scarcity of time or money. For society as a whole, it arises from the scarcity of *resources*: *land, labour, capital,* and *entrepreneurship*. To produce and enjoy more of one thing, society must shift resources away from producing something else. Therefore, we must choose which desires to satisfy and how to satisfy them.

Economics provides the tools that explain those choices. The field of economics is divided into two major areas. *Microeconomics* studies the behaviour of individual households, firms, and governments as they interact in specific markets. *Macroeconomics*, by contrast, concerns itself with the behaviour of the entire economy. It considers variables such as total output, total employment, and the overall price level.

Economics makes heavy use of *models*—abstract representations of reality—to help us understand how the economy operates. All models are simplifications, but a good model will have just enough detail for the purpose at hand. The *simplifying* assumptions in a model just make it easier to use. The *critical* assumptions are the ones that affect the model's conclusions.

Questions

1. Discuss whether each statement is a purely positive statement, or also contains normative elements and/or value judgments:

    a. An increase in the personal income tax will slow the growth rate of the economy.

    b. The goal of any country's economic policy should be to increase the well-being of its poorest, most vulnerable citizens.

    c. The best way to reduce the national poverty rate is to increase the federal minimum wage.

    d. The 1970s were a disastrous decade for the Indian economy. Income inequality increased to its highest level since before Independence.

2. For each of the following, state whether economists would consider it a *resource*, and if they would, identify which of the four types of resources the item is.

   a. A computer used by a CBI agent to track the whereabouts of suspected criminals.

   b. The office building in which the CBI agent works.

   c. The time that a CBI agent spends on a case.

   d. A farmer's tractor.

   e. The farmer's knowledge of how to operate the tractor.

   f. Crude oil.

   g. A package of frozen vegetables.

   h. A food scientist's knowledge of how to commercially freeze vegetables.

   i. The ability to bring together resources to start a frozen food company.

   j. Plastic bags used by a frozen food company to hold its product.

3. Suppose that you are considering what to do with an upcoming weekend. Here are your options, from least to most preferred:

   (1) study for upcoming midterms;

   (2) fly to Sri Nagar for a quick ski trip;

   (3) go into seclusion in your room and try to improve your score on a computer game. What is the opportunity cost of a decision to play the computer game all weekend?

This lesson will focus on the following:

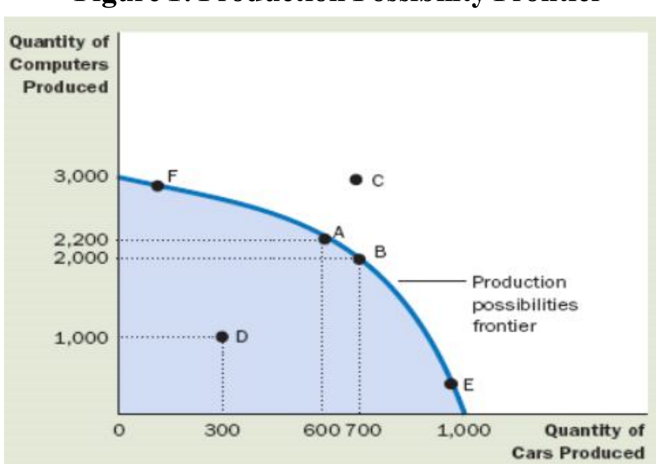
1. Production Possibilities Frontier and Opportunity Cost
2. Operating Inside Production Possibilities Frontier
3. Resource Allocation: Reasons and Methods

Production Possibilities Frontier

Most economic models, unlike the circular-flow diagramme, are built using the tools of mathematics. Here we use one of the simplest such models, called the production possibilities frontier, to illustrate some basic economic ideas. Although real economies produce thousands of goods and services, let's assume an economy that produces only two goods—cars and computers. Together, the car industry and the computer industry use all of the economy's factors of production.

The **production possibilities frontier** is a graph that shows the various combinations of output—in this case, cars and computers—that the economy can possibly produce given the available factors of production and the available production technology that firms use to turn these factors into output.

Figure 1: Production Possibility Frontier



The production possibilities frontier shows the combinations of output—in this case, cars and computers—that the economy can possibly produce. The economy can produce any combination on or inside the frontier. Points outside the frontier are not feasible given the economy's resources good without producing less of the other. Point D represents an *inefficient* outcome.

For some reason, perhaps widespread unemployment, the economy is producing less than it could from the resources it has available: It is producing only 300 cars and 1,000 computers. If the source of the inefficiency is eliminated, the economy can increase its production of both goods. For example, if the economy moves from point D to point A, its production of cars increases from 300 to 600, and its production of computers increases from 1,000 to 2,200.

People face trade-offs. The production possibilities frontier shows one trade-off that society faces. Once we have reached the efficient points on the frontier, the only way of getting more of one good is to get less of the other. When the economy moves from point A to point B, for instance, society produces 100 more cars but at the expense of producing 200 fewer computers.

There is another trade-off to understand *the Principles of Economics.*

The cost of something is what you give up to get it.

This is called the *opportunity cost.*

The production possibilities frontier shows the opportunity cost of one good as measured in terms of the other good. When society moves from point A to point B, it gives up 200 computers to get 100 additional cars. That is, at point A, the opportunity cost of 100 cars is 200 computers. Put another way, the opportunity cost of each car is two computers. Notice that the opportunity cost of a car equals the slope of the production possibilities frontier.

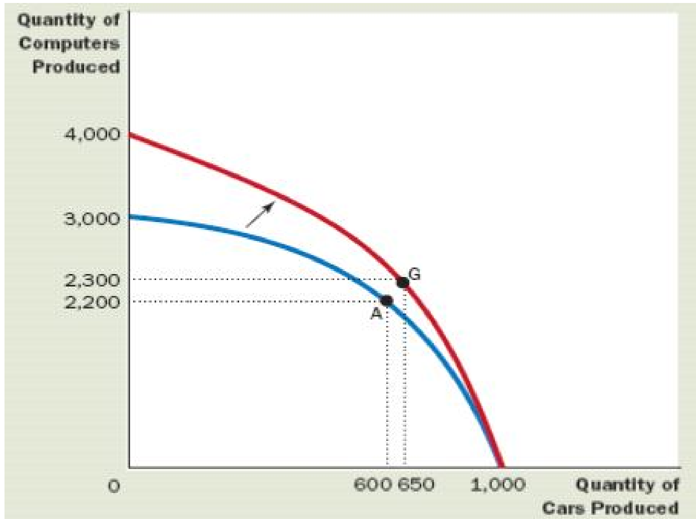
The opportunity cost of a car in terms of the number of computers is not constant in this economy but depends on how many cars and computers the economy is producing. This is reflected in the shape of the production possibilities frontier. Because the production possibilities frontier in Figure 1 is bowed outward, the opportunity cost of a car is highest when the economy is producing many cars and fewer computers, such as at point E, where the frontier is steep. When the economy is producing few cars and many computers, such as at point F, the frontier is flatter, and the opportunity cost of a car is lower.

Economists believe that production possibilities frontiers often have this bowed shape. When the economy is using most of its resources to make computers, such as at point F, the resources best suited to car production, such as skilled autoworkers, are being used in the computer industry. Because these workers probably aren't very good at making computers, the economy won't have to lose much computer production to increase car production by one unit. The opportunity cost of a car in terms of computers is small, and the frontier is relatively flat.

By contrast, when the economy is using most of its resources to make cars, such as at point E, the resources best suited to making cars are already in the car industry. Producing an additional car means moving some of the best computer technicians out of the computer industry and making them autoworkers. As a result, producing an additional car will mean a substantial loss of computer output.

The opportunity cost of a car is high, and the frontier is steep. The production possibilities frontier shows the trade-off between the outputs of different goods at a given time, but the trade-off can change over time. For example, suppose a technological advance in the computer industry raises the number of computers that a worker can produce per week. This advance expands society's set of opportunities. For any given number of cars, the economy can make more computers. If the economy does not produce any computers, it can still produce 1,000 cars, so one endpoint of the frontier stays the same. But the rest of the production possibilities frontier shifts outward, as in Figure 2.

Figure 2: Production Possibilities Frontier Depicting Economic Growth



This figure illustrates economic growth. Society can move production from a point on the old frontier to a point on the new frontier. Which point it chooses depends on its preferences for the two goods. In this example, society moves from point A to point G, enjoying more computers (2,300 instead of 2,200) and more cars (650 instead of 600).

The production possibilities frontier simplifies a complex economy to highlight some basic but powerful ideas:

- scarcity,
- efficiency,
- trade-offs,
- opportunity cost, and
- economic growth

A Shift in the Production Possibilities Frontier

A technological advance in the computer industry enables the economy to produce more computers for any given number of cars. As a result, the production possibilities frontier shifts outward. If the economy moves from point A to point G, then the production of both cars and computers increases.

Figure 2 shows this economy's production possibilities frontier. If the economy uses all its resources in the car industry, it produces 1,000 cars and no computers. If it uses all its resources in the computer industry, it produces 3,000 computers and no cars. The two endpoints of the production possibilities frontier represent these extreme possibilities. More likely, the economy divides its resources between the two industries, and this yields other points on the production possibilities frontier.

For example, it can produce 600 cars and 2,200 computers, shown in the figure by point A. Or, by moving some of the factors of production to the car industry from the computer industry, the economy can produce 700 cars and 2,000 computers, represented by point B.

Because resources are scarce, not every conceivable outcome is feasible. For example, no matter how resources are allocated between the two industries, the economy cannot produce the amount of cars and computers represented by point C. Given the technology available for manufacturing cars and computers, the economy does not have enough of the factors of production to support that level of output.

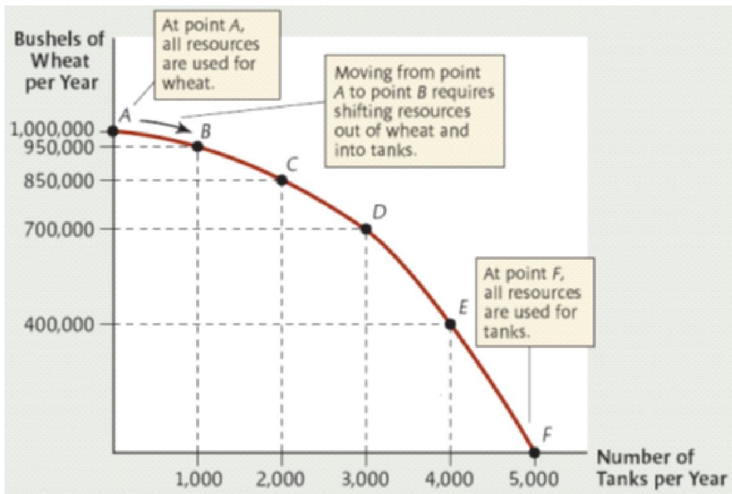
With the resources it has, the economy can produce at any point on or inside the production possibilities frontier, but it cannot produce at points outside the frontier. An outcome is said to be *efficient* if the economy is getting all it can from the scarce resources it has available.

Points on (rather than inside) the production possibilities frontier represent efficient levels of production. When the economy is producing at such a point, say point A, there is no way to produce more of one

We can see opportunity cost even more clearly in Figure 3, where the data in Table 1 has been plotted on a graph. In the figure, tank production is measured along the horizontal axis, and wheat production along the vertical axis. Each of the six points labeled *A* through *F* corresponds to one of society's choices in the table. For example, point *B* represents 1,000 tanks and 950,000 bushels of wheat.

When we connect these points with a smooth line, we get a curve called society's **production possibilities frontier (PPF).** Specifically, this PPF tells us the maximum quantity of wheat that can be produced for each quantity of tanks produced. Alternatively, it tells us the maximum number of tanks that can be produced for **Production possibilities frontier (PPF)** A curve showing all combinations of two goods that can be produced with the resources and technology currently available.

Figure 3: Production Possibilities Frontier and Opportunity Cost



Positions outside the frontier are unattainable with the technology and resources at the economy's disposal. Society's choices are limited to points on or inside the PPF. Now recall our earlier example of moving from choice B to choice C in the table.

When tank production increased from 1,000 to 2,000, wheat production decreased from 950,000 to 850,000. In the graph, this change would be represented by a movement along the PPF from point *B* to point *C*. We're moving rightward (1,000 more tanks) and also downward (100,000 fewer bushels of wheat). Thus, the opportunity cost of 1,000 more tanks can be viewed as the vertical drop along the PPF as we move from point *B* to point *C*.

Increasing Opportunity Cost

Suppose we have arrived at point *C* and society then decides to produce still more tanks. Once again, resources must be shifted into tank production to make an additional 1,000 of them, moving from point *C* to point *D*. This time, however, there is an even *greater opportunity cost*: Production of wheat falls from 850,000 to 700,000 bushels, a sacrifice of 150,000 bushels. The opportunity cost of 1,000 more tanks has risen. Graphically, the vertical drop along the curve is greater for the same move rightward.

As we continue to increase tank production by increments of 1,000—moving from point *C* to point *D* to point *E* to point *F*—the opportunity cost of producing an additional 1,000 tanks keeps rising, until the last 1,000 tanks costs us 400,000 bushels of wheat. (You can also see this in the table, by running down the numbers in the right column. Each time tank production rises by 1,000, wheat production falls by more and more.)

The behaviour of opportunity cost described here—the more tanks we produce, the greater the opportunity cost of producing still more—applies to a wide range of choices facing society. It can be generalized as the *law of increasing opportunity cost*. *According to the law of increasing opportunity cost, the more of something we produce, the greater the opportunity cost of producing even more of it.*

The law of increasing opportunity cost causes the PPF to have a concave (upside down bowl) shape, becoming steeper as we move rightward and downward. That's because the slope of the PPF—the change in the quantity of wheat divided by the change in the quantity of tanks—can be interpreted as the change in wheat *per additional tank*.

For example, moving from point *C* to point *D*, we give up 150,000 bushels of wheat to get 1,000 more tanks, or *150 bushels of wheat per tank.* Thus, the slope of the PPF between points *C* and *D* is approximately 150. (We say approximately because the PPF is curved, so its slope changes slightly as we move along the interval from *C* to *D*.) If we remove the minus sign from this slope and consider just its absolute value, it tells us the opportunity cost of *one more tank*. Now—as we've seen—this opportunity cost increases as we move rightward.

Therefore, the absolute value of the PPF's slope must rise as well. The PPF gets steeper and steeper, giving us the concave shape we see in Figure 3.

Reason for Increasing Opportunity Cost

Why does opportunity cost increase as we move along a PPF?

Most resources—by their very nature—are better suited to some purposes than to others. If the economy were operating at point *A*, for example, we'd be using *all* of our resources for wheat, even those that are much better suited to make tanks. People who would be better at factory work than farming would nevertheless be pressed into working on farms. And we'd be growing wheat on all the land available, even land that would be fine for a tank factory but awful for growing crops.

Now, as we begin to move rightward along the PPF, say from *A* to *B*, we would shift resources out of wheat production and into tank production. But we would *first* shift those resources *best suited* to tank production—and least suited for wheat. When these resources are shifted, an additional thousand tanks causes only a small drop in wheat production. This is why, at first, the PPF is very flat: a small vertical drop for the rightward movement.

As we continue moving rightward, however, we are forced to shift resources away from wheat—resources that are less and less suited to tanks and more and more suited to wheat. As a result, the PPF becomes steeper.

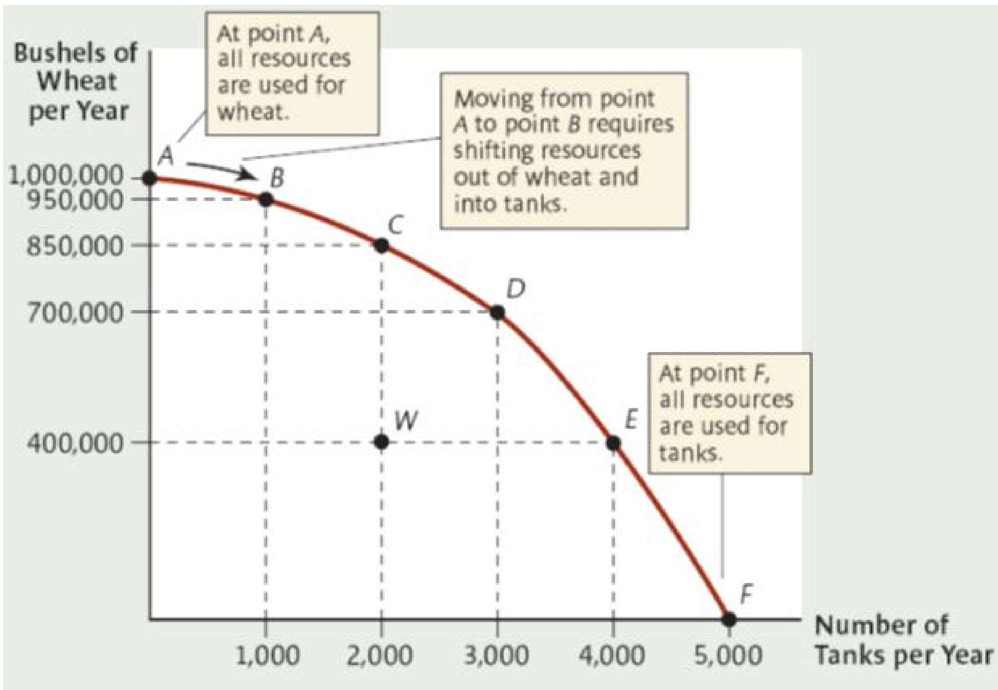
The principle of increasing opportunity cost applies to most of society's production choices, not just that between wheat and tanks. If we look at society's choice between food and oil, we would find that some land is better suited to growing food and other land is better suited to drilling for oil. As we continue to produce more oil, we would find ourselves drilling on land that is less and less suited to producing oil, but better and better for producing food.

The opportunity cost of producing additional oil will therefore increase. The same principle applies if we want to produce more health care, more education, more automobiles, or more computers: The more of something we produce, the greater the opportunity cost of producing still more.

Operating Inside the Production Possibilities Frontier

What if an economy is not living up to its productive potential, but is instead operating *inside* its PPF? For example, in Figure 4, suppose we are currently operating at point *W*, where we are producing 2,000 tanks and 400,000 bushels of wheat. Then we could move from point *W* to point *E* and produce 2,000 more tanks, with no sacrifice of wheat. Or, starting at point *W*, we could move to point *C* (more wheat with no sacrifice of tanks), or to a point like *D* (more of *both* wheat and tanks). But why would an economy ever operate inside its PPF?

Figure 4: Operating Inside the Production Possibilities Frontier



Productive Inefficiency

One possibility is that resources are not being used in the most productive way. Suppose, for example, that many people who could be outstanding wheat farmers are instead making tanks, and many who would be great at tank production are instead stuck on farms. Then switching people from one job to the other could enable us to have more of *both* tanks *and* wheat.

That is, because of the mismatch of workers and jobs, we would be *inside* the PPF at a point like *W*. Creating better job matches would then move us to a point *on* the PPF (such as point *E*).

Economists use the phrase *productive inefficiency* to describe this type of situation that puts us inside our PPF.

*A firm, an industry, or an entire economy is **productively inefficient** if it could produce more of at least one good without pulling resources from the production of any other good.*

The phrase *productive efficiency* means the absence of any productive *in*efficiency. Although no firm, industry, or economy is ever 100 percent productively efficient, cases of gross inefficiency are not as common as you might think. Business firms have strong incentives to identify and eliminate productive inefficiency, since any waste of resources increases their costs and decreases their profit. When one firm discovers a way to eliminate waste, others quickly follow.

For example, empty seats on an airline flight represent productive inefficiency. Since the plane is making the trip anyway, filling the empty seat would enable the airline to serve more people with the flight (produce more transportation services) without using any additional resources (other than the trivial resources of in-flight snacks).

Therefore, more people could fly without sacrificing any other good or service. When American Airlines developed a computer model in the late 1980s to fill its empty seats by altering schedules and fares, the other airlines followed its example very rapidly. And when—in the late 1990s—Priceline.com enabled airlines to auction off empty seats on the Internet, several airlines jumped at the chance and others quickly followed. As a result, a case of productive inefficiency in the airline industry—and therefore in the economy—was eliminated.

Starbucks provides another example. In 2000, the company analyzed how it makes drinks and eliminated several productively inefficient practices that it hadn't previously noticed. For example, it ended the practice of requiring signatures for small credit card purchases. It also began using larger scoops so that iced drinks could be made with one dip into the ice machine instead of two. These and other changes freed up labour time and enabled the company to make more drinks and serve more customers without using any additional resources. Economists, logistics experts, and engineers are continually identifying and designing policies to eliminate cases of productive inefficiency.

But many instances still remain. Why?

One economist has estimated that this simple change would save a total of 250 million hours per year (for those who currently fill out their own returns), and $2 billion per year (for those who pay accountants). With resources freed up by this change, we could produce and enjoy more of all the things that we value.

When political obstacles prevent us from eliminating inefficiency, we are back to where we started: producing more of one thing requires taking resources away from something else we value, rather than getting "free" resources from greater efficiency. Productive inefficiency does create a theoretical possibility for a free lunch. But in practice, it does not offer as many hearty meals as you might think.

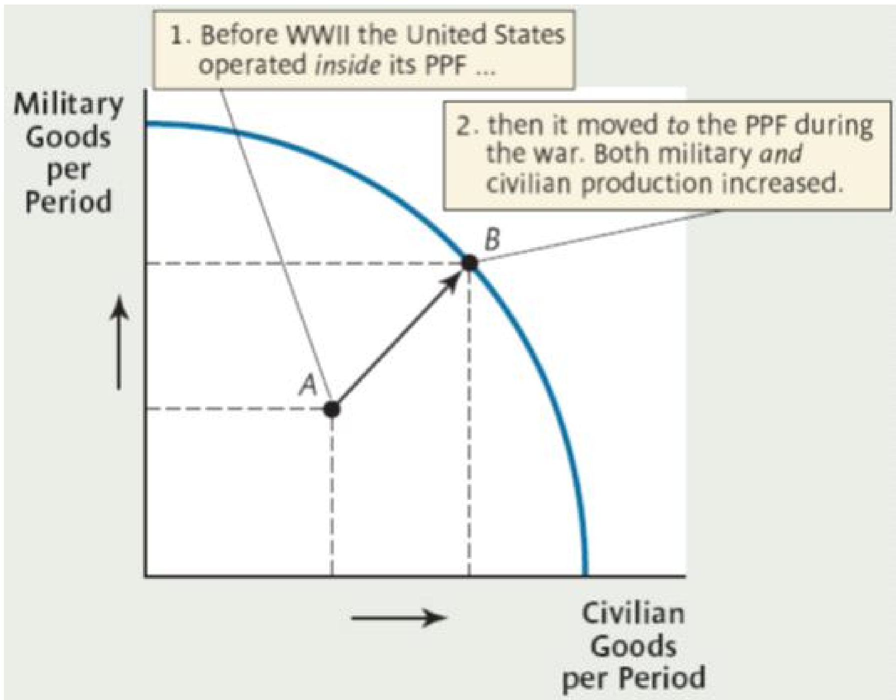
Recessions

Another reason an economy might operate inside its PPF is a *recession*—a slowdown in overall economic activity. During recessions, many resources are idle. For one thing, there is widespread *unemployment*—people *want* to work but are unable to find jobs. In addition, factories shut down, so we are not using all of our available capital. An end to the recession would move the economy from a point *inside* its PPF to a point *on* its PPF—using idle resources to produce more goods and services without sacrificing anything.

In Figure 5, this moved our economy from a point like *A*, *inside* the PPF, to a point like *B*, *on* the frontier. Military production like tanks increased, but so did the production of civilian goods such as wheat. Although there were shortages of some consumer goods, the overall result was a rise in total production and an increase in the material well-being of the average Indian citizen.

No government would ever *choose* war as a purely economic policy to end a downturn because other, economically superior policies could accomplish the same goal. But do these other methods of promoting recovery give us a free lunch—more of some things without any sacrifice? Not really.

Figure 5: Production and Unemployment



Economic Growth

1. What if the PPF itself were to change?
2. Couldn't we then produce more of everything?

This is exactly what happens when an economy's productive capacity grows. One way that productivity capacity grows is by an increase in available resources. Historically, the resource that has contributed most to rising living standards is capital. More physical capital (factory buildings, tractors, and medical equipment) or more human capital (skilled doctors, engineers, and construction workers) can enable us to produce more of *any* goods and services that use these tools.

The other major source of economic growth is *technological change*—the discovery of new ways to produce more from a given quantity of resources. The development of the Internet, for example, enabled people to find information in a few minutes that used to require hours of searching through printed documents. As a result, a variety of professionals—teachers, writers, government officials, attorneys, and physicians—can produce more of their services with the same amount of labour hours.

Figure 6 shows three examples of how economic growth can change the PPF. Panel (a) illustrates a change that initially affects only wheat production—say, the acquisition of more tractors (usable in wheat farming but not in tank-making) or the discovery of a new, higher-yielding technique for growing wheat. If we used *all* of our resources to produce wheat, we could now produce more of it than before.

For that reason, the vertical intercept of the PPF rises from point *A* to a point like *A*", where the economy could produce a maximum of 1,200,000 bushels per year. But the horizontal intercept of the PPF remains at point *F*, because the changes we're considering apply only to wheat. If we were to use all of our resources in tank production, we'd be able to produce the same number of tanks as before. The final effect is to stretch the PPF upward along the vertical axis.

Suppose we were originally operating at point *D* on the old PPF. Then, with our new PPF, we could choose to produce more wheat and the same number of tanks (point *H*). Or we could produce more of *both* goods (point *J*). We could even choose to produce more tanks and the same amount of wheat as before. (See if you can identify this point on the new PPF.) But wait.

How can having more tractors or a new type of seed—changes that directly affect only the wheat industry—enable us to produce more tanks? The answer is: after the change in the PPF, society can choose to shift some resources out of wheat farming and have the same amount of wheat as before at point *D* on the original PPF. The shifted resources can be used to increase tank production.

Panel (b) illustrates the opposite type of change in the PPF—from a technological change in producing tanks, or an increase in resources usable only in the tank industry. The *horizontal* intercept of the PPF increases, while the vertical intercept remains unchanged.

Can you explain why? As before, we could choose to produce more tanks, more wheat, or more of both. (See if you can identify points on the new PPF in panel (b) to illustrate all three cases.)

Finally, panel (c) illustrates the case where technological change occurs in both the wheat and the tank industries, or there is an increase in resources (such as workers or capital) that could be used in either. Both the horizontal and the vertical intercepts of the PPF increase. Society can choose to locate anywhere along the new PPF, producing more tanks, more wheat, or more of both.

Figure 6: Economic Growth and Production Possibilities Frontier

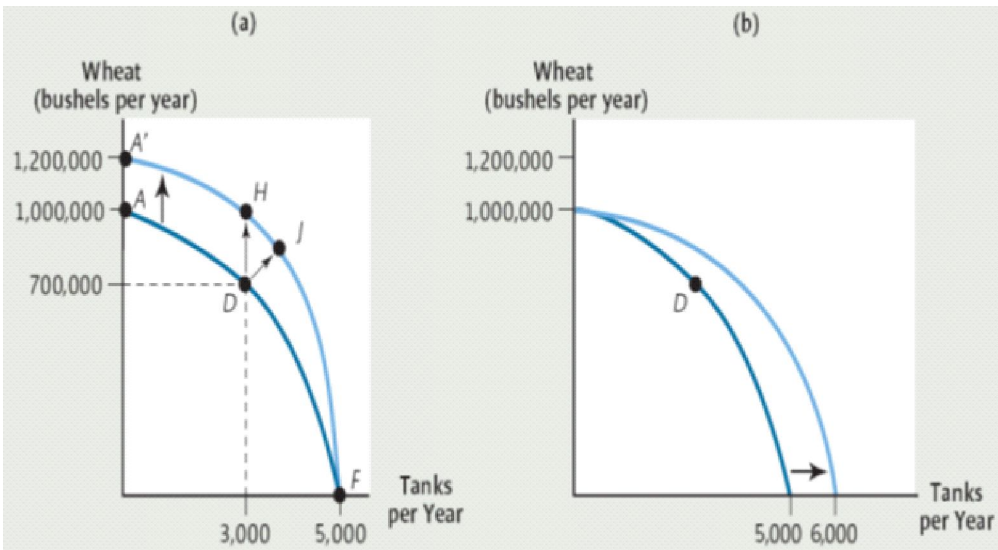
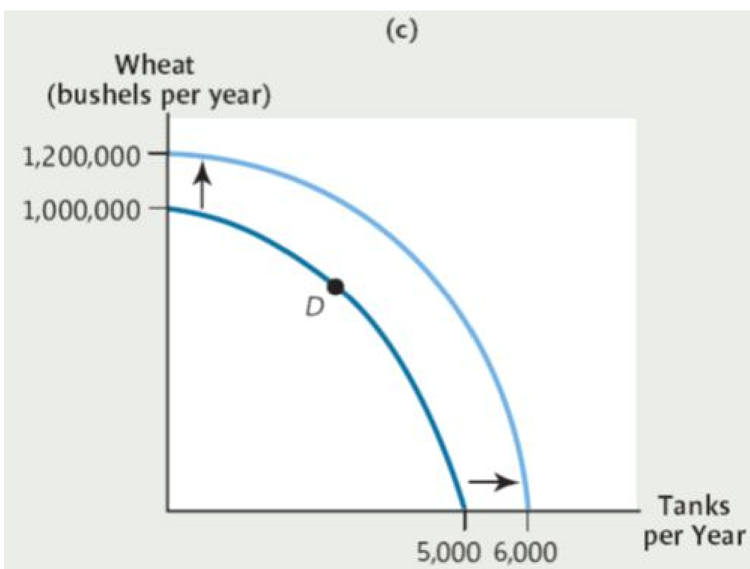


Figure 6: Economic Growth and Production Possibilities Frontier



All three panels show economic growth from an increase in resources or a technological change. In panel (a), the additional resources or technological advance directly affect only wheat production. However, society can choose to have more wheat and more tanks if it desires, such as at point J. *In panel (b), the additional resources or technological advance directly affect only tank production. But once again, society can choose to have more of both goods. In panel (c), the additional resources or technological advance directly affect production of both goods.*

Panels (a) and (b) can be generalized to an important principle about economic growth:

A technological change or an increase in resources, even when the direct impact is to increase production of just one type of good, allows us to choose greater production of all types of goods.

This conclusion certainly *seems* like a free lunch. But is it?

Yes . . . and no. True, comparing the new PPF to the old, it looks like we can have more of something—in fact, more of everything—without any sacrifice. But Figure 6 tells only part of the story. It leaves out the sacrifice that creates the change in the PPF in the first place.

Consumption vs. Growth

Suppose we want more capital. First, note that capital plays two roles in the economy. On the one hand, capital is a *resource* that we use to produce goods and services. On the other hand, capital is *itself* a good and is produced . . . using resources! A tractor, for example, is produced using land, labour, and *other* capital (a tractor factory and all of the manufacturing equipment inside the factory).

Each year, we must choose how much of our available resources to devote to producing capital, as opposed to other things. On the plus side, the more capital we produce this year, the more capital we'll have in the future to produce other things. (Remember: capital, once produced, is a long-lasting tool.)

But there's a tradeoff:

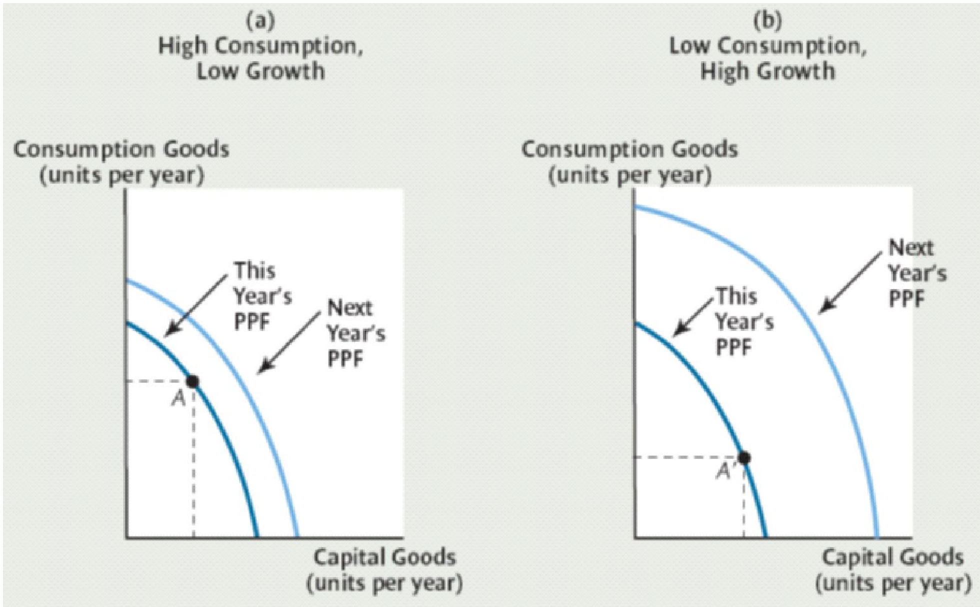
Any resources used to produce capital this year are *not* being used to produce *consumer goods*—food, health care, and other things we can enjoy right now. Figure 7 illustrates this tradeoff. In each panel, the total quantity of capital goods is measured on the horizontal axis, and consumption goods on the vertical axis. In each panel, the darker curve is *this year's* PPF.

In the left panel, point *A* shows one choice we could make this year: relatively high production of consumer goods and low production of capital goods. We'd have a relatively high standard of living this year (lots of consumer goods), but we'd be adding little to our total stock of capital during the year. As a result, *next year's* PPF—the lighter curve—will be shifted outward somewhat (because we'll have a bit more capital next year than we had this year), but not by much.

The right panel shows a different choice for the same economy. If we are situated at point *A?˝* on this year's PPF, we sacrifice more consumption goods now and produce more capital goods than at point *A* in the left panel. Living standards are lower this year. But next year, when we have considerably more capital, the PPF will have shifted outward even more. We can then choose a point on next year's PPF with greater production of consumer goods than we could have had if we had chosen point *A*. So, choosing point *A?˝* rather than *A* can lead to a greater rise in living standards next year, but requires greater sacrifice of consumer goods this year. A similar tradeoff exists when technological change drives growth.

New technologies don't just "happen"—resources must be used *now* for research and development. These resources could have been used to produce to other things that we'd enjoy today. For example, doctors who work at developing new drugs in pharmaceutical companies could instead be providing health care to patients right now. We could show this using the same PPFs as in Figure 5. But on the horizontal axis, instead of "capital goods," we'd have some measure of "research and development activities."

Figure 7: How Current Production Affects Economic Growth



In panel (a), production is tilted toward current consumption goods, with relatively few resources devoted to production of capital goods. As a result, in the future, there will not be much of an increase in capital, so the PPF will not shift out much in the future.

In panel (b), production is tilted more toward capital goods, with a greater sacrifice of current consumption. As a result, there will be a greater increase in capital, so the PPF will shift out more in the future.

And we'd come to the same conclusion about technological change that we came to earlier about having more capital:

In order to produce more goods and services in the future, we must shift resources toward R&D and capital production, and away from producing things we'd enjoy right now.

We must conclude that although economic growth—at first glance—*appears* to be a free lunch, someone ends up paying the check. In this case, the bill is paid by the part of society who will have to make do with less in the present.

Economic Systems

As you read these words—perhaps sitting at home or in the library—you are experiencing a very private moment. It is just you and this book; the rest of the world might as well not exist. Or so it seems. . . . Actually, even in this supposedly private moment, you are connected to others in ways you may not have thought about. In order for you to be reading this book, the authors had to write it.

Someone had to edit it, to help make sure that all necessary material was covered and explained as clearly as possible. Someone else had to prepare the graphics. Others had to run the printing presses and the binding machines, and still others had to pack the book, ship it, unpack it, put it on a store shelf, and then sell it to you. And there's more. People had to manufacture all kinds of goods: paper and ink, the boxes used for shipping, the computers used to keep track of inventory, and so on. It is no exaggeration to say that thousands of people were involved in putting this book in your hands.

Take a walk in your town or city, and you will see even more evidence of our economic interdependence: People are collecting garbage, helping school children cross the street, transporting furniture across town, constructing buildings, repairing roads, painting houses.

Everyone is producing goods and services for *other people*. Why is it that so much of what we consume is produced by other people? Why are we all so heavily dependent on each other for our material well-being?

Resource Allocation

More than ten thousand years ago, the Neolithic age began and human society switched from hunting and gathering to farming and simple manufacturing. At the same time, human want grew beyond mere food and shelter to the infinite variety of things that can be *made*. Ever since, all societies have been confronted with three important questions:

1. ***Which* goods and services should be produced with society's resources?**

Should we produce more consumer goods for enjoyment now or more capital goods to increase future production? Should we produce more health care, and if so, what should we produce less of? In other words, where on its production possibilities frontier should the economy operate?

2. ***How*** **should they be produced?**

Most goods and services can be produced in a variety of different ways, with each method using more of some resources and less of others. For example, there are many ways to dig a ditch. We could use *no capital at all* and have dozens of workers digging with their bare hands. We could use *a small amount of capital* by giving each worker a shovel and thereby use less labour, since each worker would now be more productive. Or we could use *even more capital*—a power trencher—and dig the ditch with just one or two workers. In every economic system, there must always be some mechanism that determines *how* goods and services will be produced from the many options available.

3. ***Who*** **should get them?**

This is where economics interacts most strongly with politics. There are so many ways to divide ourselves into groups: men and women, rich and poor, skilled and unskilled, workers and owners, families and single people, young and old . . . the list is almost endless. How should the products of our economy be distributed among these different groups and among individuals within each group?

Determining *who* gets the economy's output is always the most controversial aspect of resource allocation. Over the last half-century, our society has become more sensitized to the way goods and services are distributed, and we increasingly ask whether that distribution is fair.

Three Methods of Resource Allocation

Throughout history, every society has relied primarily on one of three mechanisms for allocating resources. In a **traditional economy,** resources are allocated according to the long-lived practices of the past. Tradition was the dominant method of

Traditional economy An economy in which resources are allocated according to long-lived practices from the past.

In a **command economy,** resources are allocated mostly by explicit instructions from some higher authority. Because the government must plan these instructions in advance, command economies are also called **centrally planned economies.** But command economies are disappearing fast.

Until about three decades ago, examples would have included the former Soviet Union, Poland, Romania, Bulgaria, Albania, China, and many others. Beginning in the late 1980s, all of these nations began abandoning central planning. The only examples left today are Cuba and North Korea, and even these economies—though still dominated by central planning—occasionally take steps away from it.

The third method of allocating resources—and the one with which you are no doubt most familiar—is "the market." In a **market economy,** instead of following long held traditions or commands from above, people are largely free to do what they want with the resources at their disposal. In the end, resources are allocated as a result of individual decision making.

1. *Which* goods and services are produced?
2. The ones that producers *choose* to produce.
3. *How* are they produced?
4. However producers *choose* to produce them.
5. *Who* gets these goods and services?
6. Anyone who *chooses* to buy them.

Of course, in a **market economy**, freedom of choice is constrained by the resources one controls. And in this respect, we do not all start in the same place in the economic race. Some of us have inherited great intelligence, talent, or beauty; and some, such as the children of successful professionals, are born into a world of helpful personal contacts.

Others, unfortunately, will inherit none of these advantages. In a market system, those who control more resources will have more choices available to them than those who control fewer resources. Nevertheless, given these different starting points, individual choice plays the major role in allocating resources in a market economy.

But wait . . . isn't there a problem here? People acting according to their own desires, without command or tradition to control them? This sounds like a recipe for chaos! How, in such a free-for-all, could resources possibly be *allocated*? The answer is contained in two words: *markets* and *prices*.

Nature of Markets

The market economy gets its name from something that nearly always happens when people are free to do what they want with the resources they possess. Inevitably, people decide to specialize in the production of one or a few things— often organizing themselves into business firms—and then sellers and buyers *come together to trade*. A **market** is a collection of buyers and sellers who have the potential to trade with one another.

In some cases, the market is *global;* that is, the market consists of buyers and sellers who are spread across the globe. The market for oil is an example of a global market, since buyers in any country can buy from sellers in any country. In other cases, the market is local. Markets for restaurant meals, haircuts, and taxi service are examples of local markets.

Command or centrally planned economy An economic system in which resources are allocated according to explicit instructions from a central authority.

Market economy An economic system in which resources are allocated through individual decision making.

Market A group of buyers and sellers with the potential to trade with each other.

Markets play a major role in allocating resources by forcing individual decision makers to consider very carefully their decisions about buying and selling. They do so because of an important feature of every market: the *price* at which a good is bought and sold.

Importance of Prices

A **price** is *the amount of money a buyer must pay to a seller for a good or service*. Price is not always the same as *cost*. In economics, as you've learned in this chapter, cost means *opportunity cost*—the *total* sacrifice needed to buy the good. While the price of a good is a *part* of its opportunity cost, it is not the only cost.

For example, the price does not include the value of the time sacrificed to buy something. Buying a new jacket will require you to spend time traveling to and from the store, trying on different styles and sizes, and waiting in line at the cash register. Still, in most cases, the price of a good is a significant part of its opportunity cost.

For large purchases such as a home or automobile, the price will be *most* of the opportunity cost. And this is why prices are so important to the overall working of the economy: They confront individual decision makers with the costs of their choices.

Consider the example of purchasing a car. Because you must pay the price, you know that buying a new car will require you to cut back on purchases of other things. In this way, the opportunity cost to *society* of making another car is converted to an opportunity cost *for you*. If you value a new car more highly than the other things you must sacrifice for it, you will buy it. If not, you won't buy it. Why is it so important that people face the opportunity costs of their actions?

The following thought experiment can answer this question.

A Thought Experiment: Free Cars

Imagine that the government passes a new law: When anyone buys a new car, the government will reimburse that person for it immediately. The consequences would be easy to predict.

First, on the day the law was passed, everyone would rush out to buy new cars. Why not, if cars are free? The entire stock of existing automobiles would be gone within days—maybe even hours. Many people who didn't value cars much at all, and who seldom used them, would find themselves owning several—one for each day of the week or to match the different colors in their wardrobe. Others who weren't able to act in time—including some who desperately needed a new car for their work or to run their households—would be unable to find one at all.

Over time, automobile companies would drastically increase production to meet the surge in demand for cars. So much of our available labour, capital, land, and entrepreneurial talent would be diverted to making cars that we'd have to sacrifice huge quantities of all other goods and services. Thus, we'd end up *paying* for those additional cars in the end, by having less education, less medical care, perhaps even less food—all to support the widespread, frivolous use of cars. Almost everyone would conclude that society had been made worse off with the new "free-car" policy.

By eliminating a price for automobiles, and severing the connection between society's opportunity cost of producing cars and individuals' decisions to have them, we would have created quite a mess for ourselves.

When resources are allocated by the market, and people must pay for their purchases, they are forced to consider the opportunity cost to society of their individual actions. In this way, markets are able to create a sustainable allocation of resources.

Price The amount of money that must be paid to a seller to obtain a good or service.

Resource Allocation in a Market Economy

India has always been considered the leading example of a market economy. Each day, millions of distinct items are produced and sold in markets. Our grocery stores are always stocked with broccoli and tomato soup, and the drugstore always has Kleenex and aspirin—all due to the choices of individual producers and consumers. The goods that are traded, the way they are traded, and the price at which they trade are determined by the traders themselves.

But even in the market economy, there are numerous cases of resource allocation *outside* the market. For example, many economic decisions are made within families, which do *not* operate like little market economies. Instead, many decisions are based on tradition. For example, even when children get an allowance or have other earnings, they don't have to pay for goods consumed within the home. Other decisions are based on command ("No TV until you finish your homework!").

In the broader economy, there are many examples of resource allocation by command. Various levels of government collect, in total, about one-third of our incomes as taxes. We are *told* how much tax we must pay, and those who don't comply suffer serious penalties, including imprisonment. Government—rather than individual decision makers—spends the tax revenue. In this way, the government plays a major role in allocating resources—especially in determining which goods are produced and who gets them.

There are also other ways, aside from strict commands, that the government limits our market freedoms. Regulations designed to protect the environment, maintain safe workplaces, and ensure the safety of our food supply are just a few examples of government-imposed constraints on our individual choice.

Because for each example we can find where resources are allocated by tradition or command, or where government restrictions seriously limit some market freedom, we can find hundreds of examples where individuals make choices according to their own desires. The things we buy, the jobs at which we work, the homes in which we live— in almost all cases, these result from market choices. The market, though not pure, is certainly the dominant method of resource allocation.

The complete label for the economic system is *market capitalism*. While the market describes how resources are *allocated,* capitalism refers to one way that resources are *owned.* Under **capitalism,** most resources are owned by private citizens, who are mostly free to sell or rent them to others as they wish. The alternative mode of ownership is **socialism,** a system in which most resources are owned by the state, as in the former Soviet Union.

Just as the U.S. is a leading example of resource allocation by the market, it is also a leading example of capitalism. True, there are examples of state ownership of resources (national parks, government buildings, state highways systems, and more). But most of our nation's land, labour, and capital are privately owned and managed, and can be sold or rented in markets as the owners wish.

Understanding the Market

The market is simultaneously the most simple and the most complex way to allocate resources. For individual buyers and sellers, the market is simple. There are no traditions or commands to be memorized and obeyed. Instead, we enter the markets we *wish* to trade in, and we respond to prices there as we *wish* to, unconcerned about the overall process of resource allocation.

Capitalism A type of economic system in which most resources are owned privately.

Socialism A type of economic system in which most resources are owned by the state.

But from the economist's point of view, the market is quite complex. Resources are allocated indirectly, as a *byproduct* of individual decision making, rather than through easily identified traditions or commands. As a result, it often takes some skillful economic detective work to determine just how individuals are behaving and how resources are being allocated as a consequence.

Conclusion

The *production possibilities frontier* (PPF) is a simple model to illustrate the opportunity cost of society's choices. When we are *productively efficient* (operating *on* the PPF), producing more of one thing requires producing less of something else. The *law of increasing opportunity cost* tells us that the more of something we produce, the greater the opportunity cost of producing still more. Even when we are operating inside the PPF—say because of productive inefficiency or a recession—it is not necessarily easy or costless to move to the PPF and avoid opportunity cost.

In a world of scarce resources, each society must have an *economic system*: its way of organizing economic activity. All economic systems feature *specialization,* in which each person and firm concentrates on a limited number of productive activities, and exchange, through which we obtain most of what we desire by trading with others. Specialization and exchange enable us to enjoy higher living standards than would be possible under self-sufficiency.

In a *market economy,* resources are allocated primarily through markets. Prices play an important role in markets by forcing decision makers to take account of society's opportunity cost when they make choices.

Questions

1. Name one economic interaction that isn't covered by the simplified circular-flow diagramme.
2. Draw and explain a production possibilities frontier for an economy that produces milk and cookies. What happens to this frontier if disease kills half of the economy's cows?
3. Use a production possibilities frontier to describe the idea of "efficiency."
4. What are the two subfields into which economics is divided? Explain what each subfield studies.
5. What is the difference between a positive and a normative statement? Give an example of each.
6. Why do economists sometimes offer conflicting advice to policymakers?

This lesson will focus on the following:

1. Market and its Characteristics
2. Law of Demand
3. Law of Supply
4. Equilibrium Price and Quantity
5. Equilibrium with Change in Demand and Supply

Supply and demand

Supply **and** ***demand*** are the two words economists use most often—and for good reason. Supply and demand are the forces that make market economies work. They determine the quantity of each good produced and the price at which it is sold. If you want to know how any event or policy will affect the economy, you must think first about how it will affect supply and demand.

This lesson introduces the theory of supply and demand. It considers how buyers and sellers behave and how they interact with one another. It shows how supply and demand determine prices in a market economy and how prices, in turn, allocate the economy's scarce resources.

The terms *supply* and *demand* refer to the behaviour of people as they interact with one another in competitive markets. Before discussing how buyers and sellers behave, let's first consider more fully what we mean by the terms *market* and *competition.*

What is a Market?

A **market** is a group of buyers and sellers of a particular good or service. The buyers as a group determine the demand for the product, and the sellers as a group determine the supply of the product.

Markets take many forms. Sometimes markets are highly organized, such as the markets for many agricultural commodities. In these markets, buyers and sellers meet at a specific time and place, where an auctioneer helps set prices and arrange sales.

More often, markets are less organized. For example, consider the market for ice cream in a particular town. Buyers of ice cream do not meet together at any one time. The sellers of ice cream are in different locations and offer somewhat different products. There is no auctioneer calling out the price of ice cream. Each seller posts a price for an ice-cream cone, and each buyer decides how much ice cream to buy at each store. Nonetheless, these consumers and producers of ice cream are closely connected. The ice-cream buyers are choosing from the various ice-cream sellers to satisfy their hunger, and the ice-cream sellers are all trying to appeal to the same ice-cream buyers to make their businesses successful. Even though it is not organized, the group of ice-cream buyers and ice-cream sellers forms a market.

What is Competition?

The market for ice cream, like most markets in the economy, is highly competitive. Each buyer knows that there are several sellers from which to choose, and each seller is aware that his or her product is similar to that offered by other sellers. As a result, the price of ice cream and the quantity of ice cream sold are not determined by any single buyer or seller. Rather, price and quantity are determined by all buyers and sellers as they interact in the marketplace.

Economists use the term **competitive market** to describe a market in which there are so many buyers and so many sellers that each has a negligible impact on the market price. Each seller of ice cream has limited control over the price because other sellers are offering similar products. A seller has little reason to charge less than the going price, and if he or she charges more, buyers will make their purchases elsewhere. Similarly, no single buyer of ice cream can influence the price of ice cream because each buyer purchases only a small amount.

In this lesson, we assume that markets are *perfectly competitive.* To reach this highest form of competition, a market must have two characteristics:

(1) the goods offered for sale are all exactly the same, and

(2) the buyers and sellers are so numerous that no single buyer or seller has any influence over the market price.

Because buyers and sellers in perfectly competitive markets must accept the price the market determines, they are said to be *price takers.* At the market price, buyers can buy all they want, and sellers can sell all they want.

There are some markets in which the assumption of perfect competition applies perfectly. In the wheat market, for example, there are thousands of farmers who sell wheat and millions of consumers who use wheat and wheat products. Because no single buyer or seller can influence the price of wheat, each takes the price as given.

Not all goods and services, however, are sold in perfectly competitive markets. Some markets have only one seller, and this seller sets the price. Such a seller is called a *monopoly.* Your local cable television company, for instance, may be a monopoly. Residents of your town probably have only one cable company from which to buy this service. Still other markets fall between the extremes of perfect competition and monopoly.

Despite the diversity of market types we find in the world, assuming perfect competition is a useful simplification and, therefore, a natural place to start. Perfectly competitive markets are the easiest to analyze because everyone participating in the market takes the price as given by market conditions. Moreover, because some degree of competition is present in most markets, many of the lessons that we learn by studying supply and demand under perfect competition apply in more complicated markets as well.

Characteristics of a Market

The first step in analyzing a market is to figure out *which* market we are analyzing. This might seem easy. But we can choose to define a market in different ways, depending on our purpose.

Broad versus Narrow Definition of Market

Suppose we want to study the personal computer industry in India.

Should we define the market very broadly (“the market for computers”), or very narrowly (“the market for ultra-light laptops”), or something in between (“the market for laptops”)?

The right choice depends on the problem we’re trying to analyze.

For example, if we're interested in why computers *in general* have come down in price over the past decade, we'd treat all types of computers as if they were the same good. Economists call this process **aggregation**—combining a group of distinct things into a single whole. But suppose we're asking a different question: Why do laptops always cost more than desktops with similar computing power? Then we'd aggregate all laptops together as one good, and all desktops as another, and look at each of these more narrowly defined markets.

We can also choose to define the *geography* of a market more broadly or more narrowly, depending on our purpose. We'd analyze the *national* market for gasoline if we're explaining general nationwide trends in gas prices. But we'd define it more locally to explain, say, why gas prices are rising more rapidly in Delhi than in other areas of the country.

In economics, ***markets*** *can be defined broadly or narrowly, depending on our purpose.*

How markets are defined is one of the most important differences between *macro*economics and *micro*economics.

In macroeconomics, goods and services are aggregated to the highest levels. Macro models even lump all consumer goods— breakfast cereals, cell phones, blue jeans, and so forth— into the single category "consumption goods" and view them as if they are traded in a single, national "market for consumption goods."

Defining markets this broadly allows macroeconomists to take an overall view of the economy without getting bogged down in the details.

In microeconomics, by contrast, markets are defined more narrowly. Instead of asking how much we'll spend on *consumer goods,* a micro economist might ask how much we'll spend on *health care* or *video games*.

Even in microeconomics, there is always some aggregation, but not as much as in macroeconomics.

Product and Resource Markets

The simple **circular flow** model of the economy, illustrates two different types of markets and how they relate to each other. The upper half illustrates **product markets,** where goods and services are bought and sold. The blue arrows show the flow of products from the business firms who supply them to the households who buy them. The green arrows show the associated flow of dollars from the households who spend the dollars, to the business firms who receive these dollars as revenue. (In the real world, businesses also sell products to the government and to other businesses, but this simple version leaves out these details.)

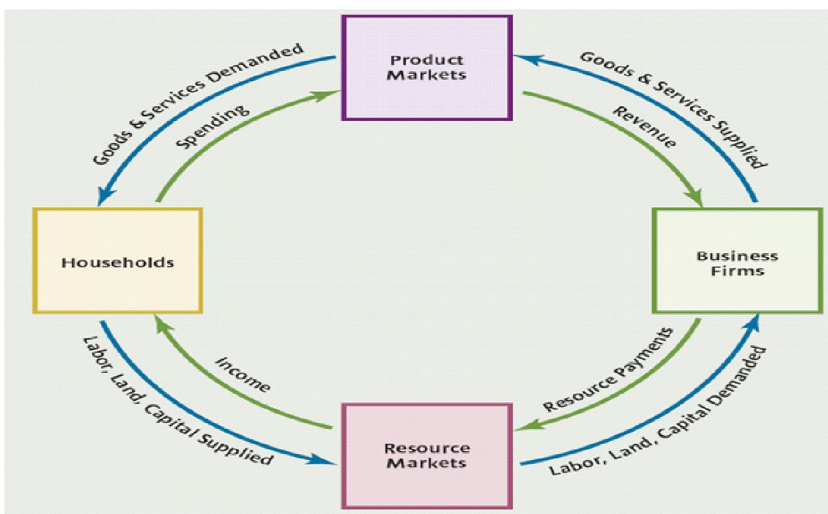
Aggregation The process of combining distinct things into a single whole.

Markets A group of buyers and seller with the potential to trade with each other

Circular flow A simple model that shows how goods, resources, and dollar payments flow between households and firms.

Product markets Markets in which firms sell goods and services to households

Figure 1: Circular Flow Model and Markets



The outer loop of the diagramme shows the flows of goods and resources, and the markets in which they are traded. Households sell resources to firms in resource markets. Business firms use the resources to produce goods and services, which they sell to households in product markets. The inner loop shows money flows. The resource payments made by firms become income to households. Households use the income to purchases goods and services from firms.

The lower half depicts a different set of markets: **resource markets,** where labour, land, and capital are bought and sold. Here, the roles of households and firms are reversed. The blue arrows show resources flowing from the households (who own and supply them) to the business firms (who demand them). The associated flow of dollars is indicated by the green arrows: Business firms pay for the resources they use, and households receive these payments as income.

Competition in Markets

A final issue in defining a market is how prices are determined. In some markets, individual buyers or sellers have an important influence over the price. For example, in the national market for cornflakes, Kellogg's—an individual *seller*—simply sets its price every few months. It can raise the price and sell fewer boxes of cereal or lower the price and sell more. In a small-town, a major *buyer* of antiques may be

Resource markets: Markets in which households that own resources sell them to firms able to negotiate special discount prices with the local antique shops. These are examples of *imperfectly competitive* markets.

In ***imperfectly competitive markets,*** *individual buyers or sellers can influence the price of the product.*

But now think about the Indian market for wheat. Can an individual seller have any impact on the market price? Not really. On any given day there is a going price for wheat—say, Rs. 5.80 per bushel. If a farmer tries to charge more than that—say, Rs. 5.85 per bushel—he won't sell any wheat at all! His customers will instead go to one of his many competitors and buy the identical product from them for less. Each wheat farmer must take the price of wheat as a "given." The same is true of a single wheat *buyer:* If he tries to negotiate a lower price from a seller, he'd be laughed off the farm. "Why should I sell my wheat to you for Rs. 5.75 per bushel, when there are others who will pay me Rs. 5.80?" Accordingly, each buyer must take the market price as a given.

The market for wheat is an example of a *perfectly competitive market.*

In ***perfectly competitive markets*** *(or just* ***competitive markets****), each buyer and seller takes the market price as a given.*

What makes some markets imperfectly competitive and others perfectly competitive? You'll learn the complete answer, along with more formal definitions, when you are further into your study of *microeconomics*. But here's a hint: In perfectly competitive markets, there are many small buyers and sellers, each is a small part of the market, and the product is standardized, like wheat. Imperfectly competitive markets, by contrast, have just a few large buyers or sellers, or else the product of each seller is unique in some way. Understanding competition in markets is important in this chapter for one simple reason:

The supply and demand model is designed to show how prices are determined in perfectly competitive markets.

Competition in the Real World

Markets that are truly perfectly competitive—where no buyer or seller has *any* influence over the price—are rare.

Does that mean we can only use supply and demand in those rare cases, such as the market for wheat?

Not at all. Supply and demand is useful for many real-world markets, even when the competition is somewhat imperfect.

Consider the market for laptop computers. Laptops made by Lenovo, Hewlett Packard, Toshiba, Apple, and other producers differ in important ways: memory, speed, operating system, and more. And even within a smaller group—say, Windows laptops with the same memory and speed—there are still differences. The keyboards feel different, the reputations for reliability and service are different, and more. For this reason, each producer can charge a different price, even for very similar laptops. Because there is no single market price that all producers take as given, the market is not *strictly* perfectly competitive.

Imperfectly competitive market A market in which a single buyer or seller has the power to influence the price of the product.

Perfectly competitive market (informal definition) A market in which no buyer or seller has the power to influence the price.

But laptops made by different firms, while not identical, are not *that* different. So the freedom to set price is limited. For example, if other similar laptops are selling for between Rs. 90000 and Rs. 1,00,000, Lenovo cannot charge Rs. 1,40,000; it would lose almost all of its customers to competitors. While there is no single market price, each producer views the *range* of prices it can charge as given.

Thus, the laptop market—while not perfectly competitive—is still somewhat competitive. And in cases like these, supply and demand can help us see how the price *range* is determined, and what makes that range rise and fall.

More generally, *while few markets are* strictly *perfectly competitive, most markets have enough competition for supply and demand to explain broad movements in prices.*

Law of Demand

It's tempting to think of the "demand" for a product psychologically—a pure "want" or "desire." But that kind of thinking can lead us astray. For example, you *want* all kinds of things: a bigger apartment, a better car, nicer clothes, more and better vacations. The list is endless. But you don't always *buy* them. Why not?

Because in addition to your wants which you'd very much like to satisfy you also face *constraints*. First, you have to *pay*. Second, your spending power is limited, so every decision to buy one thing is also a decision *not* to buy something else (or a decision to save less, and have less buying power in the future). As a result, every purchase confronts you with an opportunity cost. Your "wants," together with the real-world constraints that you face, determine what you will choose to buy in any market. Hence, the following definition:

*The **quantity demanded** of a good or service is the number of units that all buyers in a market would choose to buy over a given time period, given the constraints that they face.*

Since this definition plays a key role in any supply and demand analysis, it's worth taking a closer look at it.

***Quantity Demanded Implies a* Choice.** Quantity demanded doesn't tell us the amount of a good that households feel they "need" or "desire" in order to be happy. Instead, it tells us how much households would choose to buy *when they take into account the opportunity cost* of their decisions. The opportunity cost arises from the

Quantity demanded The quantity of a good that all buyers in a market would choose to buy during a period of time, given their constraints.

Quantity Demanded Is **Hypothetical.** Will households actually be *able* to purchase the amount they want to purchase? As you'll soon see, usually yes. But there are special situations—analyzed in microeconomics—in which households are frustrated in buying all that they would like to buy. Quantity demanded makes no assumptions about the availability of the good. Instead, it's the answer to a hypothetical question: How much would households buy, given the constraints that they face, if the units they wanted to buy were available?

Quantity Demanded Depends on **Price.** The price of the good is just one variable among many that influences quantity demanded. But because the price is a key variable that our model will ultimately determine, we try to keep that variable front and- center in our thinking. This is why for the next few pages we'll assume that all other influences on demand are held constant, so we can explore the relationship between price and quantity demanded.

Explaining the Law of Demand

How does a change in price affect quantity demanded?

You probably know the answer to this already: When something is more expensive, people tend to buy less of it. This common observation applies to air travel, magazines, guitars, and virtually everything else that people buy. For all of these goods and services, price and quantity are *negatively related:* that is, when price rises, quantity demanded falls; when price falls, quantity demanded rises. This negative relationship is observed so regularly in markets that economists call it the *law of demand*.

The ***law of demand*** *states that when the price of a good rises and everything else remains the same, the quantity of the good demanded will fall.*

The law of demand tells us what would happen *if* all the other influences on buyers' choices remained unchanged, and only one influence—the price of the good—changed.

This is an example of a common practice in economics. In the real world, many variables change *simultaneously*. But to understand changes in the economy, we must first understand the effect of each variable *separately*. So we conduct a series of mental experiments in which we ask: "What would happen if this one influence— and only this one—were to change?" The law of demand is the result of one such mental experiment, in which we imagine that the price of the good changes, but all other influences on quantity demanded remain constant.

Law of demand As the price of a good increases, the quantity demanded decreases.
Ceteris paribus Latin for "all else remaining the same."

Demand Schedule and Demand Curve

To make our discussion more concrete, let's look at a specific market: the market for real maple syrup in the United States. In this market, we'll view the buyers as U.S. households, whereas the sellers (to be considered later) are maple syrup producers in the United States or Canada.

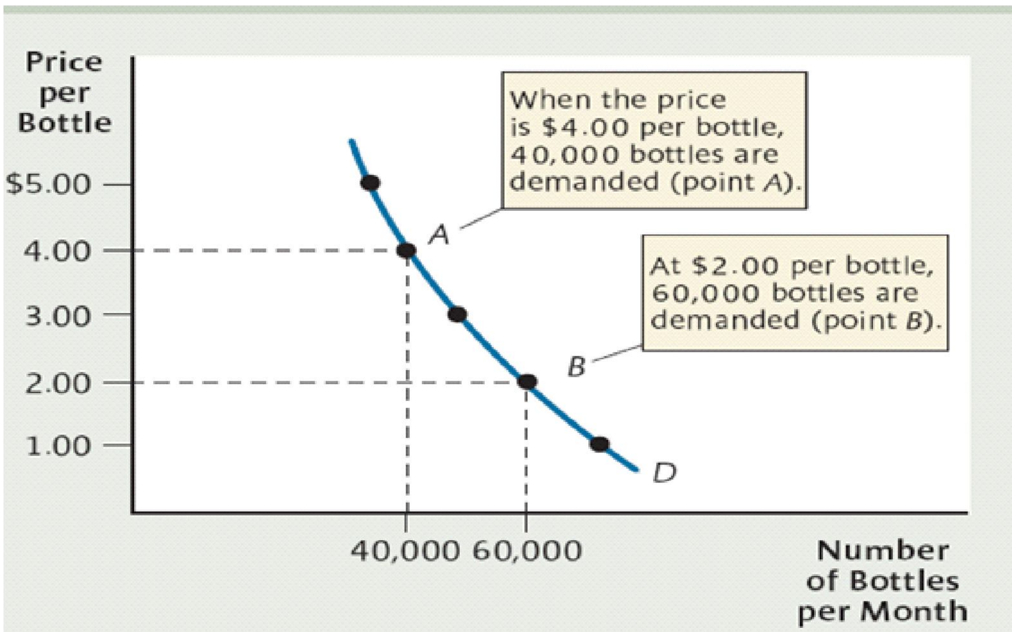
Table 1: Demand Schedule for Maple Syrup in the United States

Price (per bottle)	Quantity Demanded (bottles per month)
$1.00	75,000
$2.00	60,000
$3.00	50,000
$4.00	40,000
$5.00	35,000

Table 1 shows a hypothetical **demand schedule** for maple syrup in this market. This is *a list of different quantities demanded at different prices, with all other variables that affect the demand decision assumed constant*. For example, the demand schedule tells us that when the price of maple syrup is $2.00 per bottle, the quantity demanded will be 60,000 bottles per month. Notice that the demand schedule obeys the law of demand: As the price of maple syrup increases, *ceteris paribus,* the quantity demanded falls.

Now look at Figure 2. It shows a diagramme that will appear again and again in your study of economics. In the figure, each price-and-quantity combination in Table 1 is represented by a point. For example, point *A* represents the price $4.00 and quantity 40,000, while point *B* represents the pair $2.00 and 60,000. When we connect all of these points with a line, we obtain the famous *demand curve,* labeled with a *D* in the figure.

Figure 2: Demand Curve



*The **demand curve** shows the relationship between the price of a good and the quantity demanded in the market, holding constant all other variables that influence demand. Each point on the curve shows the total quantity that buyers would choose to buy at a specific price.*

Notice that the demand curve in Figure 2—like virtually all demand curves—*slopes downward*. This is just a graphical representation of the law of demand.

Shifts versus Movement along the Demand Curve

Markets are affected by a variety of events. Some events will cause us to *move along* the demand curve; others will cause the entire demand curve to *shift*. It is crucial to distinguish between these two very different types of effects.

Let's go back to Figure 2. There, you can see that when the price of maple syrup rises from $2.00 to $4.00 per bottle, the number of bottles demanded falls from 60,000 to 40,000. This is a movement *along* the demand curve, from point *B* to point *A*. In general, *a change in the price of a good causes a movement* along *the demand curve.*

In Figure 2, a *fall* in price would cause us to move *rightward* along the demand curve (from point *A* to point *B*), and a *rise* in price would cause us to move *leftward* along the demand curve (from *B* to *A*).

Remember, though, that when we draw a demand curve, we assume all other variables that might influence demand are *held constant* at some particular value.

For example, the demand curve in Figure 2 might have been drawn to give us quantity demanded at each price when average household income in the United States remains constant at, say, $40,000 per year.

But suppose average income increases to $50,000. With more income, we'd expect households to buy more of *most* things, including maple syrup. This is illustrated in Table 2. At the original income level, households would choose to buy 60,000 bottles of maple syrup at $2.00 per bottle. But after income rises, they would choose to buy more at that price—80,000 bottles. After income rises, households would choose to buy more than before. In other words, the rise in income *changes the entire relationship between price and quantity demanded*. We now have a *new* demand curve.

Table 2: Increase in Demand for Maple Syrup in the United States

Price (per bottle)	Original Quantity Demanded (average income = $40,000)	New Quantity Demanded (average income = $50,000)
$1.00	75,000	95,000
$2.00	60,000	80,000
$3.00	50,000	70,000
$4.00	40,000	60,000
$5.00	35,000	55,000

Demand schedule A list showing the quantities of a good that consumers would choose to purchase at different prices, with all other variables held constant.

Figure 3 plots the new demand curve from the quantities in the third column of Table 2. The new demand curve lies to the *right* of the old curve. For example, at a price of \$2.00, quantity demanded increases from 60,000 bottles on the old curve (point *B*) to 80,000 bottles on the *new* demand curve (point *C*). As you can see, the rise in household income has *shifted* the demand curve to the right.

More generally, *a change in any variable that affects demand—except for the good's price— Causes the demand curve to shift.*

When buyers would choose to buy a greater quantity at any price, the demand curve shifts *rightward*. If they would decide to buy a smaller quantity at any price, the demand curve shifts *leftward*.

Shift in Demand Curve

Because the market demand curve holds other things constant, it need not be stable over time. If something happens to alter the quantity demanded at any given price, the demand curve shifts.

Figure 3 illustrates shifts in demand. Any change that increases the quantity demanded at every price, shifts the demand curve to the right and is called an *increase in demand*. Any change that reduces the quantity demanded at every price shifts the demand curve to the left and is called a *decrease in demand.*

There are many variables that can shift the demand curve. Here are the most important.

"Change in Quantity Demanded" versus "Change in Demand"

The term *quantity demanded* means a *particular amount* that buyers would choose to buy at a specific price, represented by a single point on a demand curve.

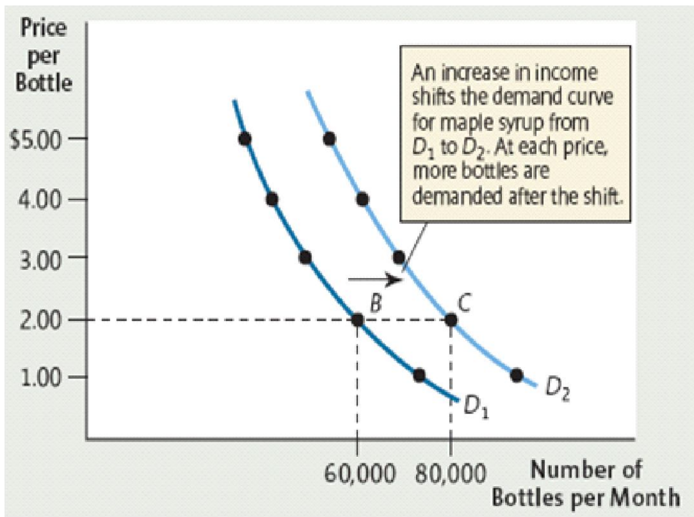
Demand, by contrast, means the *entire relationship* between price and quantity demanded, represented by the entire demand curve.

For this reason, when a change in the price of a good moves us *along* a demand curve, we call it a **change in quantity demanded.** For example, in Figure 2, the movement from point *A* to point *B* is an *increase* in quantity demanded. This is a change from one number (40,000 bottles) to another (60,000 bottles).

Change in quantity demanded

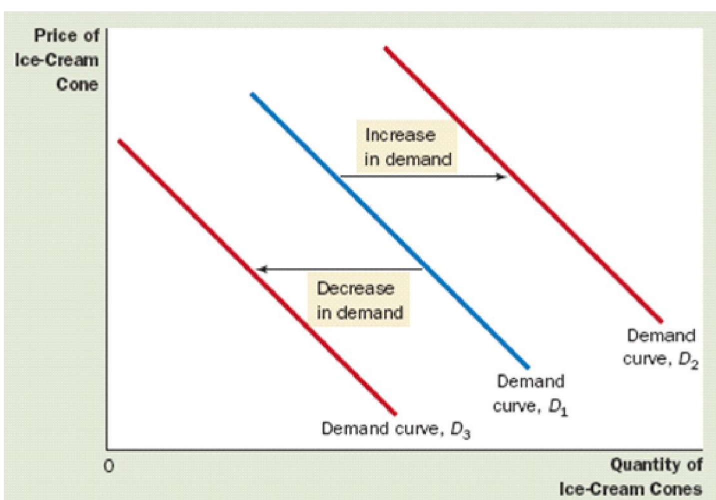
A movement along a demand curve in response to a change in price.

Figure 3: A Shift of the Demand Curve



Any change that raises the quantity that buyers wish to purchase at any given price shifts the demand curve to the right. Any change that lowers the quantity that buyers wish to purchase at any given price shifts the demand curve to the left.

Figure 4: Change in Demand



When something *other* than the price changes, causing the entire demand curve to shift, we call it a **change in demand.** In Figure 4, for example, the shift in the curve would be called an *increase in demand*.

Factors Shifting Demand Curve

Let's take a closer look at what might cause a change in demand (a shift of the demand curve). Keep in mind that for now, we're exploring *one factor at a time*, always keeping *all other determinants of demand constant*.

Income

In Figure 4, an increase in **income** shifted the demand for maple syrup to the right. In fact, a rise in income increases demand for *most* goods.

We call these **normal goods.** Housing, automobiles, health club memberships, and real maple syrup are all examples of normal goods. But not all goods are normal.

For some goods—called **inferior goods**—a rise in income would *decrease* demand—shifting the demand curve *leftward*.

Regular-grade ground chuck is a good example. It's a cheap source of protein, but not as high in quality as sirloin. With higher income, households could more easily afford better types of meat—ground sirloin or steak, for example. As a result, higher incomes for buyers might cause the demand for ground chuck to *decrease*. For similar reasons, we might expect that Greyhound bus tickets (in contrast to airline tickets) and single-ply paper towels (in contrast to two-ply) are inferior goods. *A rise in income will* increase *the demand for a* normal *good, and* decrease *the demand for an* inferior *good.*

Wealth

Your **wealth** at any point in time is the total value of everything you *own* (cash, bank accounts, stocks, bonds, real estate or any other valuable property) minus the total dollar amount you *owe* (home mortgage, credit card debt, auto loan, student loan, and so on). Although income and wealth are different, they have similar effects on demand. Increases in wealth among buyers—because of an increase in the value of their stocks or bonds, for example—gives them more funds with which to purchase goods and services. As you might expect, *an increase in wealth will* increase *demand (shift the curve rightward)* *for a normal good, and* decrease *demand (shift the curve leftward) for an inferior good.*

Prices of Related Goods

A **substitute** is a good that can be used in place of another good and that fulfills more or less the same purpose. For example, many people use real maple syrup to sweeten their pancakes, but they could use a number of other things instead: honey, sugar, jam, or *artificial* maple syrup. Each of these can be considered a substitute for real maple syrup. When the price of a substitute rises, people will choose to buy *more* maple syrup. For example, when the price of jam rises, some jam users will switch to maple syrup, and the demand for maple syrup will increase. In general, *a rise in the price of a substitute increases the demand for a good, shifting the demand curve to the right.*

Change in demand A shift of a demand curve in response to a change in some variable other than price.

Income The amount that a person or firm earns over a particular period.

Normal good A good that people demand more of as their income rises.

Inferior good A good that people demand less of as their income rises.

Wealth The total value of everything a person or firm owns, at a point in time, minus the total amount owed.

Substitute A good that can be used in place of some other good and that fulfills more or less the same purpose.

Of course, if the price of a substitute falls, we have the opposite result: Demand for the original good decreases, shifting its demand curve to the left.

A **complement** is the opposite of a substitute: It's used *together with* the good we are interested in. Pancake mix is a complement to maple syrup, since these two goods are used frequently in combination. If the price of pancake mix rises, some consumers will switch to other breakfasts—bacon and eggs, for example—that *don't* include maple syrup. The demand for maple syrup will decrease. *A rise in the price of a complement decreases the demand for a good, shifting the demand curve to the left.*

Population

As the population increases in an area, the number of buyers will ordinarily increase as well, and the demand for a good will increase. The growth of the U.S. population over the last 50 years has been an important reason (but not the only reason) for rightward shifts in the demand curves for food, housing, automobiles, and many other goods and services.

Expected Price

If buyers expect the price of maple syrup to rise next month, they may choose to purchase more *now* to stock up before the price hike. If people expect the price to drop, they may postpone buying, hoping to take advantage of the lower price later. *In many markets, an expectation that price will rise in the future shifts the* current *demand curve rightward, while an expectation that price will fall shifts the current demand curve leftward.*

Expected price changes for goods are especially important for goods that can be purchased and stored until needed later. Expected price changes are also important in the markets for financial assets such as stocks and bonds and in the market for housing, as you'll see in the next chapter.

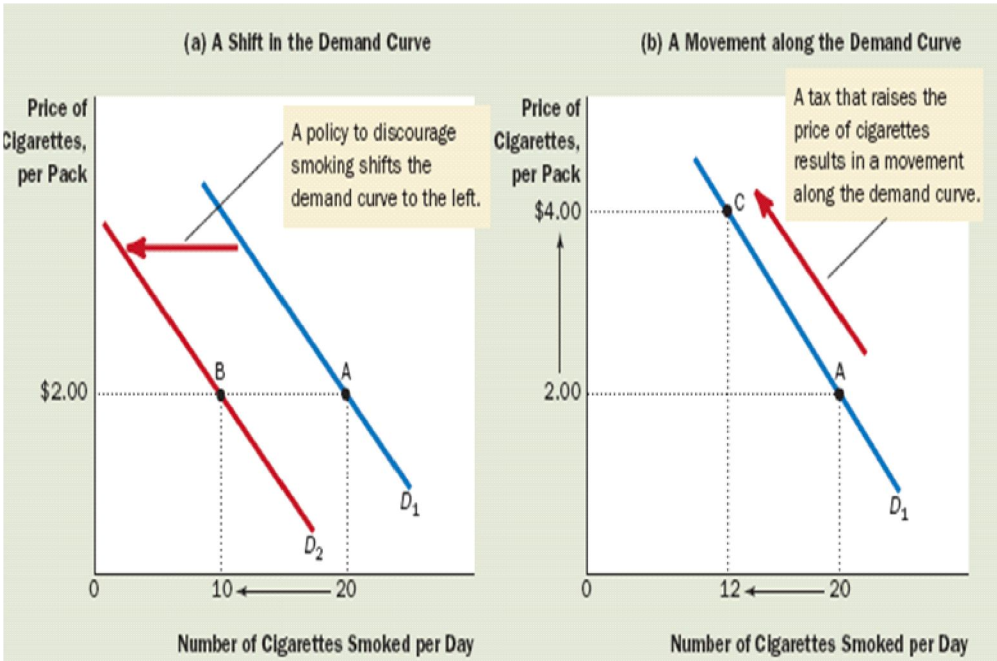
Tastes

Not everyone likes maple syrup. And among those who do, some *really* like it, and some like it just a little. Buyers' basic attitudes toward a good are based on their *tastes* or *preferences.* Economists are sometimes interested in where these tastes come from or what makes them change. But for the most part, economics deals with the *consequences* of a change in tastes, whatever the reason for its occurrence.

When tastes change *toward* a good (people favor it more), demand increases, and the demand curve shifts to the right. When tastes change *away* from a good, demand decreases, and the demand curve shifts to the left. An example of this is the change in tastes away from cigarettes over the past several decades. The cause may have been an aging population, a greater concern about health among people of *all* ages, or successful antismoking advertising. But regardless of the cause, the effect has been to decrease the demand for cigarettes, shifting the demand curve to the left.

Complement A good that is used together with some other good

Figure 5: Shifts in the Demand Curve versus Movements along the Demand Curve



If warnings on cigarette packages convince smokers to smoke less, the demand curve for cigarettes shifts to the left. In panel (a), the demand curve shifts from *D*1 to *D*2. At a price of $2.00 per pack, the quantity demanded falls from 20 to 10 cigarettes per day, as reflected by the shift from point A to point B.

Other Shift Variables

Many other things, besides those we've discussed, can shift the demand curve. For example, if the government began to offer subsidies to households who buy maple syrup, demand would shift rightward. Also, if business firms (rather than just households) are among the buyers, then changes in the demand for their own products will influence their demand for maple syrup.

Market Demand versus Individual Demand Curve

The **quantity demanded** of any good is the amount of the good that buyers are willing and able to purchase. As we will see, many things determine the quantity demanded of any good, but when analyzing how markets work, one determinant plays a central role—the price of the good. If the price of ice cream rose to $20 per scoop, you would buy less ice cream. You might buy frozen yogurt instead. If the price of ice cream fell to $0.20 per scoop, you would buy more. This relationship between price and quantity demanded is true for most goods in the economy and, in fact, is so pervasive that economists call it the **law of demand**: Other things equal, when the price of a good rises, the quantity demanded of the good falls, and when the price falls, the quantity demanded rises.

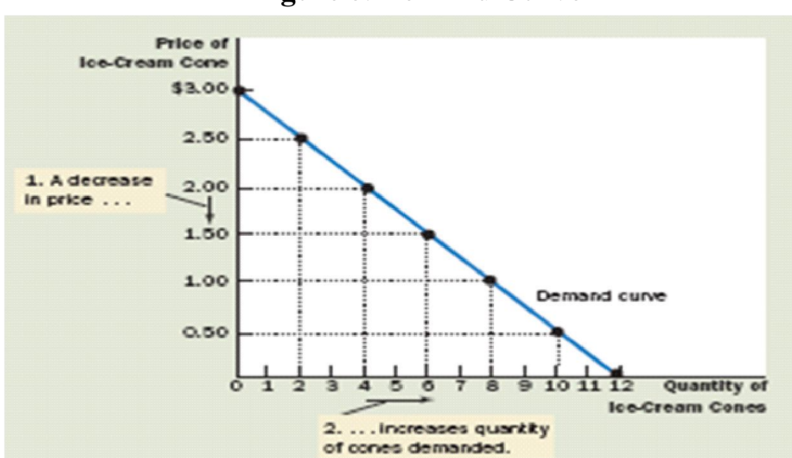
The table 3 shows how many ice-cream cones Catherine buys each month at different prices of ice cream. If ice cream is free, Catherine eats 12 cones per month. At $0.50 per cone, Catherine buys 10 cones each month. As the price rises further, she buys fewer and fewer cones. When the price reaches $3.00, Catherine doesn't buy any ice cream at all.

This table is a **demand schedule**, a table that shows the relationship between the price of a good and the quantity demanded, holding constant everything else that influences how much consumers of the good want to buy.

Table 3: Demand Schedule

Price of Ice-Cream Cone	Quantity of Cones Demanded
$0.00	12 cones
0.50	10
1.00	8
1.50	6
2.00	4
2.50	2
3.00	0

Figure 6: Demand Curve



The Figure 6 uses the numbers from the table 3 to illustrate the law of demand. By convention, the price of ice cream is on the vertical axis, and the quantity of ice cream demanded is on the horizontal axis. The downward-sloping line relating price and quantity demanded is called the **demand curve**.

Quantity demanded : the amount of a good that buyers are willing and able to purchase

Law of demand: the claim that, other things equal, the quantity demanded of a good falls when the price of the good rises

Demand schedule: a table that shows the relationship between the price of a good and the quantity demanded

The demand curve in Figure 6 shows an individual's demand for a product. To analyze how markets work, we need to determine the *market demand*, the sum of all the individual demands for a particular good or service.

Table 4: Market Demand Curve as Sum of Individual Demand Curves

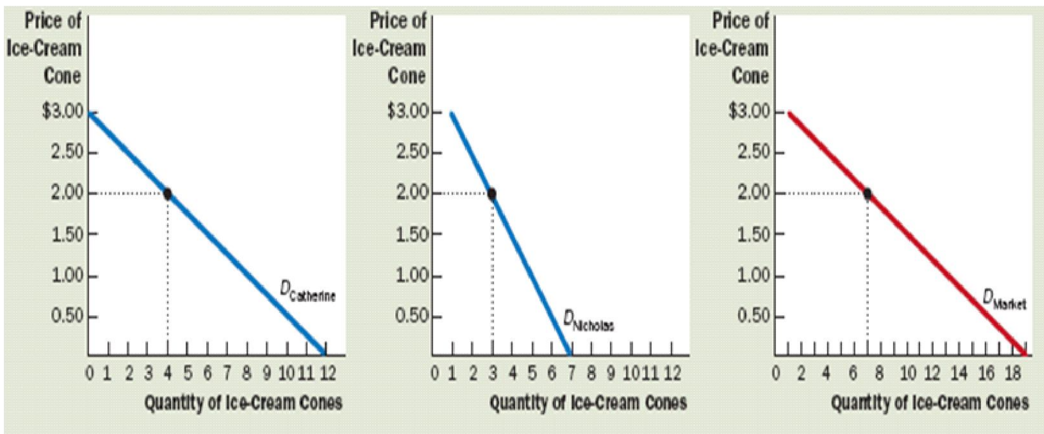
Price of Ice-Cream Cone	Catherine		Nicholas		Market
$0.00	12	+	7	=	19 cones
0.50	10		6		16
1.00	8		5		13
1.50	6		4		10
2.00	4		3		7
2.50	2		2		4
3.00	0		1		1

The table 4 shows the demand schedules for ice cream of the two individuals in this market—Catherine and Nicholas. At any price, Catherine's demand schedule tells us how much ice cream she buys, and Nicholas's demand schedule tells us how much ice cream he buys. The market demand at each price is the sum of the two individual demands.

The Figure 7 shows the demand curves that correspond to these demand schedules. Notice that we sum the individual demand curves horizontally to obtain the market demand curve. That is, to find the total quantity demanded at any price, we add the individual quantities, which are found on the horizontal axis of the individual demand curves.

Because we are interested in analyzing how markets function, we work most often with the market demand curve. The market demand curve shows how the total quantity demanded of a good varies as the price of the good varies, while all the other factors that affect how much consumers want to buy are held constant.

Figure 7: Market Demand as the Sum of Individual Demands

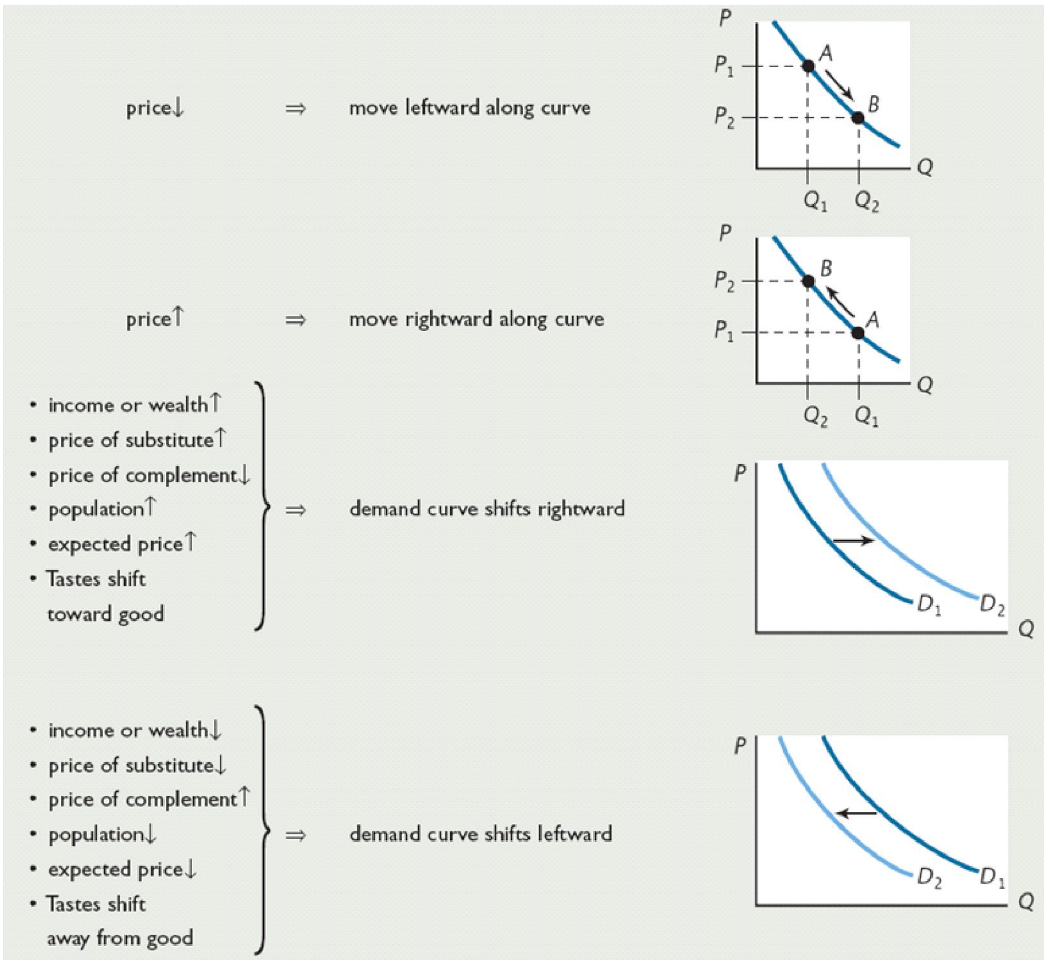


The quantity demanded in a market is the sum of the quantities demanded by all the buyers at each price. Thus, the market demand curve is found by adding horizontally the individual demand curves. At a price of $2.00, Catherine demands 4 ice-cream cones, and Nicholas demands 3 ice-cream cones. The quantity demanded in the market at this price is 7 cones.

Demand : Summary

Figure 8 summarizes the key variables we've discussed that affect the demand side of the market and how their effects are represented with a demand curve. Notice the important distinction between events that move us *along* the curve (changes in price) and events that *shift* the curve.

Figure 8: Demand Curve—A Summary



Law of Supply

When most people hear the word *supply,* their first thought is that it's the amount of something "available," as if this amount were fixed in stone. For example, someone might say, "We can only drill so much oil from the ground," or "There are only so many apartments for rent in this town." And yet, the world's known oil reserves—as well as yearly production of oil—have increased dramatically over the last half century, as oil companies have found it worth their while to look harder for oil.

Similarly, in most towns and cities, short buildings have been replaced with tall ones, and the number of apartments has increased. Supply, like demand, can change, and the amount of a good supplied in a market depends on the *choices* made by those who produce it.

What governs these choices?

We assume that those who supply goods and services have a goal: to earn the highest profit possible. But they also face constraints.

First, in a competitive market, the price they can charge for their product is a *given*—the market price.

Second, firms have to pay the *costs* of producing and selling their product. These costs will depend on the production process they use, the prices they must pay for their inputs, and more. Business firms' desire for profit, together with the real-world constraints that they face, determines how much they will choose to sell in any market.

Hence, the following definition:

Quantity supplied *is the number of units of a good that* all *sellers in the market would choose to sell over some time period, given the constraints that they face.*

Let's briefly go over the notion of quantity supplied to clarify what it means and doesn't mean.

Quantity Supplied Implies a Choice

Quantity supplied doesn't tell us the amount of, say, maple syrup that sellers would like to sell *if* they could charge a thousand dollars for each bottle, and *if* they could produce it at zero cost. Instead, it's the quantity that firms *choose* to sell—the quantity that gives them the highest profit given the constraints they face.

Quantity Supplied Is Hypothetical

Will firms actually be *able* to sell the amount they want to sell at the going price? You'll soon see that they usually can. But the definition of quantity supplied makes no assumptions about firms' ability to sell the good. Quantity supplied answers the hypothetical question: How much *would* suppliers sell, given their constraints, if they were able to sell all that they wanted to.

***Quantity Supplied Depends on* Price**

The price of the good is just one variable among many that influences quantity supplied.

Quantity supplied The specific amount of a good that all sellers in a market would choose to sell over some time period, given their constraints.

But—as with demand—we want to keep that variable foremost in our thinking. This is why for the next couple of pages we'll assume that all other influences on supply are held constant, so we can explore the relationship between price and quantity supplied.

Explaining Law of Supply

How does a change in price affect quantity supplied?

When a seller can get a higher price for a good, producing and selling it become more profitable. Producers will devote more resources toward its production—perhaps even pulling resources from other goods they produce—so they can sell more of the good in question.

For example, a rise in the price of laptop (but not desktop) computers will encourage computer makers to shift resources out of the production of other things (such as desktop computers) and toward the production of laptops.

In general, price and quantity supplied are *positively related:* When the price of a good rises, the quantity supplied will rise as well. This relationship between price and quantity supplied is called the law of supply, the counterpart to the law of demand we discussed earlier.

*The **law of supply** states that when the price of a good rises, and everything else remains the same, the quantity of the good supplied will rise.*

Once again, notice the very important words "everything else remains the same"— *ceteris paribus*. Although many other variables influence the quantity of a good supplied, the law of supply tells us what would happen if all of them remained unchanged and only one—the price of the good—changed.

Supply Schedule and Supply Curve

Let's continue with our example of the market for maple syrup. Who are the suppliers in this market? The market quantity supplied is the amount of syrup all of these producers together would offer for sale at each price for maple syrup.

Table 1 shows the **supply schedule** for maple syrup—a *list of different quantities supplied at different prices, with all other variables held constant*.

As you can see, the supply schedule obeys the law of supply: As the price of maple syrup rises, the quantity supplied rises along with it. But how can this be? After all, maple trees must be about 40 years old before they can be tapped for syrup, so any rise in quantity supplied now or in the near future cannot come from an increase in planting.

What, then, causes quantity supplied to rise as price rises?

With higher prices, firms will find it profitable to tap existing trees more intensively. Evaporating and bottling can be done more carefully, so that less maple syrup is spilled and more is available for shipping. Or the product can be diverted from other areas and shipped. For example, if the price of maple syrup rises, producers would shift deliveries away so they could sell more.

Law of supply As the price of a good increases, the quantity supplied increases.

Supply schedule A list showing the quantities of a good or service that firms would choose to produce and sell at different prices, with all other variables held constant.

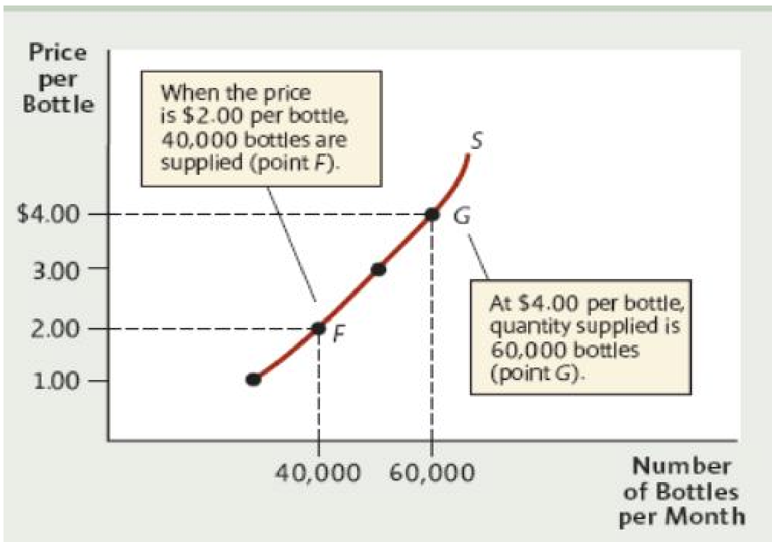
Table 1: Supply Schedule

Price (per bottle)	Quantity Supplied (bottles per month)
$1.00	25,000
$2.00	40,000
$3.00	50,000
$4.00	60,000
$5.00	65,000

Now look at Figure 1, which shows a very important curve—the counterpart to the demand curve. In Figure 1, each point represents a price-quantity pair taken from Table 1. For example, point *F* in the figure corresponds to a price of $2.00 per bottle and a quantity of

40,000 bottles per month, while point *G* represents the price-quantity pair $4.00 and 60,000 bottles. Connecting all of these points with a solid line gives us the *supply curve* for maple syrup, labeled with an *S* in the figure.

Figure 2: Supply Curve



*The **supply curve** shows the relationship between the price of a good and the quantity supplied in the market, holding constant the values of all other variables that affect supply. Each point on the curve shows the quantity that sellers would choose to sell at a specific price.*

Notice that the supply curve in Figure 1—like all supply curves for goods and services—is *upward sloping*. This is the graphical representation of the law of supply.

Supply curve A graph of a supply schedule, showing the quantity of a good or service supplied at various prices, with all other variables held constant.

Shifts versus Movements along Supply Curve

As with the demand curve, it's important to distinguish those events that will cause us to *move along* a given supply curve for the good, and those that will cause the entire supply curve to *shift*.

If you look once again at Figure 5, you'll see that if the price of maple syrup rises from \$2.00 to \$4.00 per bottle, the number of bottles supplied rises from 40,000 to 60,000. This is a movement *along* the supply curve, from point *F* to point *G*. In general, *a change in the price of a good causes a movement* along *the supply curve.*

In the figure, a *rise* in price would cause us to move *rightward* along the supply curve (from point *F* to point *G*) and a *fall* in price would move us *leftward* along the curve (from point *G* to point *F*).

But remember that when we draw a supply curve, we assume that all other variables that might influence supply are *held constant* at some particular values. For example, the supply curve in Figure 1 might tell us the quantity supplied at each price when the cost of an important input—transportation from the farm to the point of sale—remains constant.

But suppose the cost of transportation drops. Then, at any given price for maple syrup, firms would find it more profitable to produce and sell it. This is illustrated in Table 1. With the original transportation cost, and a selling price of \$4.00 per bottle, firms would choose to sell 60,000 bottles. But after transportation cost falls, they would choose to produce and sell more 80,000 bottles in our example—assuming they could still charge \$4.00 per bottle.

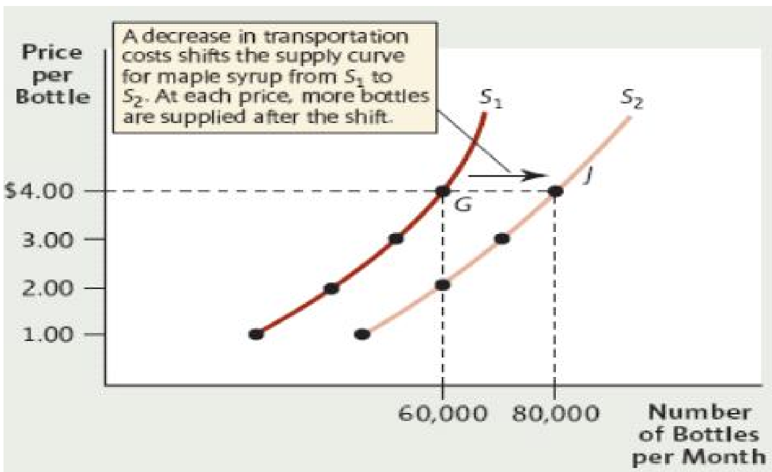
A similar change would occur for any other price of maple syrup we might imagine: After transportation costs fall, firms would choose to sell more than before. In other words, *the entire relationship between price and quantity supplied has changed,* so we have a *new* supply curve.

Figure 2 plots the new supply curve from the quantities in the third column of Table 2. The new supply curve lies to the *right* of the old one. For example, at a price of \$4.00, quantity supplied increases from 60,000 bottles on the old curve (point *G*) to 80,000 bottles on the *new* supply curve (point *J*). The drop in the transportation costs has *shifted* the supply curve to the right.

Table 2: Increase in Supply

Price (per bottle)	Original Quantity Supplied	Quantity Supplied After Decrease in Transportation Cost
$1.00	25,000	45,000
$2.00	40,000	60,000
$3.00	50,000	70,000
$4.00	60,000	80,000
$5.00	65,000	90,000

Figure 2: Shift in Supply Curve



In general, *a change in any variable that affects supply—except for the good's price—causes the supply curve to shift.*

If sellers want to sell a greater quantity at any price, the supply curve shifts *rightward.* If sellers would prefer to sell a smaller quantity at any price, the supply curve shifts *leftward*.

Change in Quantity Supplied versus Change in Supply

As we stressed in our discussion of the demand side of the market, be careful about language when thinking about supply. The term *quantity supplied* means a *particular amount* that sellers would choose to sell at a *particular* price, represented by a single point on the supply curve. The term *supply,* however, means the *entire relationship* between price and quantity supplied, as represented by the entire supply curve.

For this reason, when the price of the good changes, and we move *along* the supply curve, we have a **change in quantity supplied.** For example, in Figure 1, the movement from point *F* to point *G* is an *increase* in quantity supplied.

When something *other* than the price changes, causing the entire supply curve to shift, we call it a **change in supply.** The shift in Figure 2, for example, would be called an *increase in supply*.

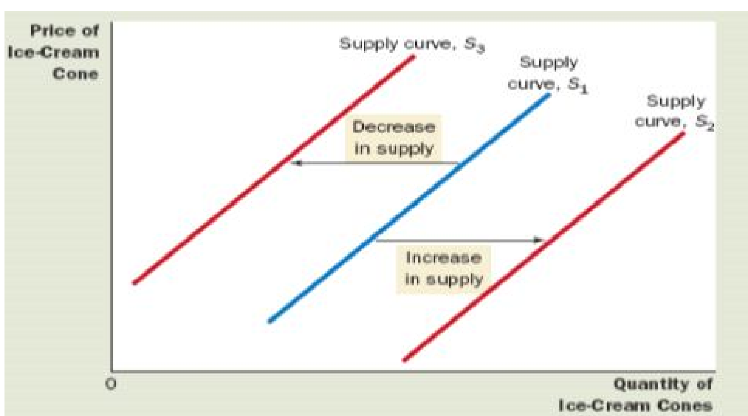
Factors shifting Supply Curve

Let's look at some of the *causes* of a change in supply (a shift of the supply curve). As always, we're considering *one* variable at a time, keeping all other determinants of supply constant.

Change in quantity supplied A movement along a supply curve in response to a change in price.

Change in supply A shift of a supply curve in response to a change in some variable other than price.

Figure 3: Shifts in the Supply Curve



Any change that raises the quantity that sellers wish to produce at any given price shifts the supply curve to the right. Any change that lowers the quantity that sellers wish to produce at any given price shifts the supply curve to the left.

Input Prices

In Figure 2, we saw that a drop in transportation costs shifted the supply curve for maple syrup to the right. But producers of maple syrup use a variety of other inputs besides transportation: land, maple trees, sap pans, evaporators, labour, glass bottles, and more. A lower price for any of these means a lower cost of producing and selling maple syrup, making it more profitable. As a result, we would expect producers to shift resources into maple syrup production, causing an increase in supply.

In general, *a fall in the price of an input causes an increase in supply, shifting the supply curve to the right.*

Similarly, a rise in the price of an input causes a decrease in supply, shifting the supply curve to the left. If, for example, the wages of maple syrup workers rose, the supply curve in Figure 2 would shift to the left.

Price of Alternatives

Many firms can switch their production rather easily among several different goods or services, each of which requires more or less the same inputs. For example, a dermatology practice can rather easily switch its specialty from acne treatments for the young to wrinkle treatments for the elderly. An automobile producer can—without too much adjustment—switch to producing light trucks. And a maple syrup producer could dry its maple syrup and produce maple *sugar* instead. Or it could even cut down its maple trees and sell maple wood as lumber. These other goods that firms *could* produce are called **alternate goods** and their prices influence the supply curve.

For example, if the price of maple *sugar* rose, then at any given price for maple *syrup,* producers would shift some production from syrup to sugar. This would be a decrease in the supply of maple syrup.

When the price for an alternative rises—either an alternate good or the same good in an alternate market—the supply curve shifts leftward.

Similarly, a decrease in the price of an alternate good (or a lower price in an alternate market) will shift the supply curve rightward.

Technology

A *technological advance* in production occurs whenever a firm can produce a given level of output in a new and cheaper way than before. Examples would include a new, more efficient tap that draws more maple syrup from each tree, or a new bottling method that reduces spillage. Advances like these would reduce the cost of producing maple syrup, making it more profitable, and producers would want to make and sell more of it at any price.

In general, *cost-saving technological advances increase the supply of a good, shifting the supply curve to the right.*

Alternate goods Other goods that firms in a market could produce instead of the good in question.

Alternate market A market other than the one being analyzed in which the same good could be sold.

Number of Firms

A change in the number of firms in a market will change the quantity that all sellers together would want to sell at any given price. For example, if—over time—more people decided to open up maple syrup farms because it was a profitable business, the supply of maple syrup would increase. And if maple syrup farms began closing down, their number would be reduced and supply would decrease.

Expected Price

If it is expected that the market price of maple syrup, for example, to rise next month. Supplier wants to postpone selling some of your maple syrup until the price is higher and your profit greater. Therefore, at any given price *now,* you might slow down production, or just slow down sales by warehousing more of what you produce. If other firms have similar expectations of a price hike, they'll do the same.

Thus, an expectation of a *future* price hike will decrease supply *in the present*. Suppose instead you expect the market price to *drop* next month. Then—at any given price—you'd want to sell more *now,* by stepping up production and even selling out of your inventories. So an expected future drop in the price would cause an increase in supply in the present.

Expected price is especially important when suppliers can hold inventories of goods for later sale, or when they can easily shift production from one time period to another.

In many markets, an expectation of a future price rise *shifts the current supply curve* leftward. *Similarly, an expectation of a* future price drop *shifts the current supply curve* rightward.

Changes in Weather and Other Natural Events

Weather conditions are an especially important determinant of the supply of agricultural goods.

Favorable weather *increases crop yields, and causes a* rightward *shift of the supply curve for that crop.* Unfavorable weather *destroys crops and shrinks yields, and shifts the supply curve* leftward.

In addition to bad weather, natural disasters such as fires, hurricanes, and earthquakes can destroy or disrupt the productive capacity of *all* firms in a region. If many sellers of a particular good are located in the affected area, the supply curve for that good will shift leftward.

Other Shift Variables

Many other things, besides those listed earlier, can shift the supply curve. For example, a government tax imposed on maple syrup producers would raise the cost of making and selling maple syrup. To suppliers, this would have the same effect as a higher price for transportation: it would shift the supply curve leftward. We'll discuss other shift variables for supply as they become relevant in later chapters.

Market versus Individual Supply Curve

The **quantity supplied** of any good or service is the amount that sellers are willing and able to sell. There are many determinants of quantity supplied, but once again, price plays a special role in our analysis. When the price of ice cream is high, selling ice cream is profitable, and so the quantity supplied is large. Sellers of ice cream work long hours, buy many ice-cream machines, and hire many workers.

By contrast, when the price of ice cream is low, the business is less profitable, and so sellers produce less ice cream. At a low price, some sellers may even choose to shut down, and their quantity supplied falls to zero. This relationship between price and quantity supplied is called the **law of supply**: Other things equal, when the price of a good rises, the quantity supplied of the good also rises, and when the price falls, the quantity supplied falls as well.

The table 3 shows the quantity of ice-cream cones supplied each month by Ben, an ice-cream seller, at various prices of ice cream. At a price below $1.00, Ben does not supply any ice cream at all. As the price rises, he supplies a greater and greater quantity. This is the **supply schedule**, a table that shows the relationship between the price of a good and the quantity supplied, holding constant everything else that influences how much producers of the good want to sell.

Table 3: Supply Schedule

Price of Ice-Cream Cone	Quantity of Cones Supplied
$0.00	0 cones
0.50	0
1.00	1
1.50	2
2.00	3
2.50	4
3.00	5

The Figure 4 uses the numbers from the table to illustrate the law of supply. The curve relating price and quantity supplied is called the **supply curve**. The supply curve slopes upward because, other things equal, a higher price means a greater quantity supplied.

Figure 4: Supply Curve

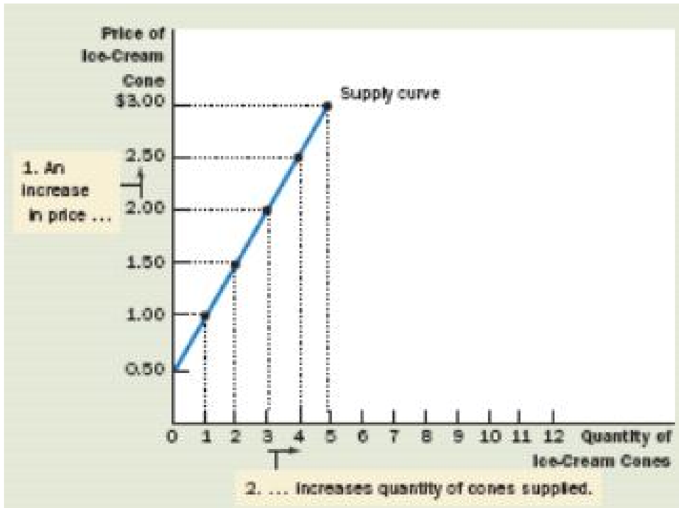


Table 4: Market Supplies as the Sum of Individual Supplies

Price of Ice-Cream Cone	Ben		Jerry		Market
$0.00	0	+	0	=	0 cones
0.50	0		0		0
1.00	1		0		1
1.50	2		2		4
2.00	3		4		7
2.50	4		6		10
3.00	5		8		13

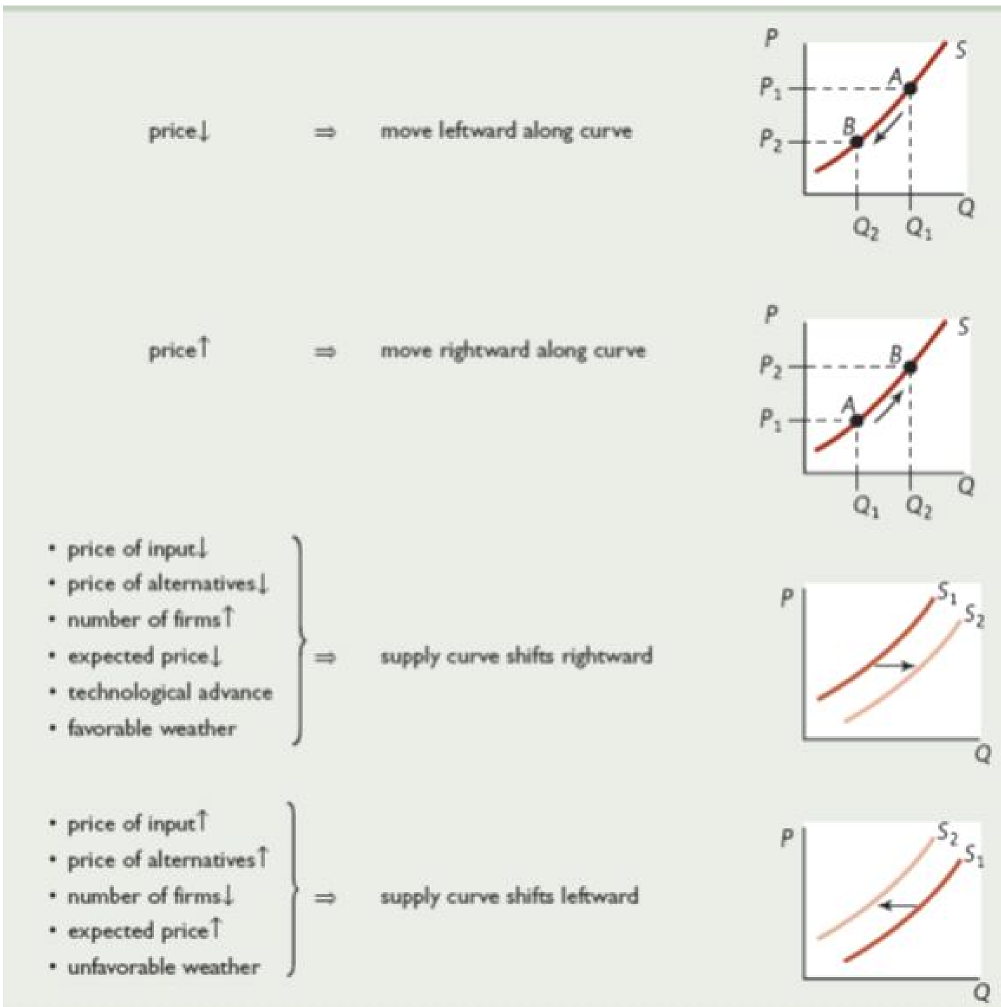
Figure 5: Markey Supply Curve

The quantity supplied in a market is the sum of the quantities supplied by all the sellers at each price. Thus, the market supply curve is found by adding horizontally the individual supply curves. At a price of $2.00, Ben supplies 3 ice-cream cones, and Jerry supplies 4 ice-cream cones. The quantity supplied in the market at this price is 7 cones.

Supply: Summary

Figure 6 summarizes the various factors we've discussed that affect the supply side of the market, and how we illustrate them using a supply curve. As with demand, notice which events move us along the supply curve (changes in price) and which shift the curve.

Figure 6: Supply Curve—A Summary



Putting Supply and Demand Together

What happens when buyers and sellers, each having the desire and the ability to trade, come together in a market?

The two sides of the market certainly have different agendas. Buyers would like to pay the lowest possible price, while sellers would like to charge the highest possible price.

Is there chaos when they meet, with buyers and sellers endlessly chasing after each other or endlessly bargaining for advantage, so that trade never takes place?

A casual look at the real world suggests not. In most markets, most of the time, there is order and stability in the encounters between buyers and sellers. In most cases, prices do not fluctuate wildly from moment to moment but seem to hover around a stable value. Even when this stability is short-lived—lasting only a day, an hour, or even a minute in some markets—for this short-time the market seems to be at rest. Whenever we study a market, therefore, we look for this state of rest—a price and quantity at which the market will settle, at least for a while.

Economists use the word *equilibrium* when referring to a state of rest. When a market is in equilibrium, both the price of the good and the quantity bought and sold have settled into a state of rest.

More formally, *the* ***equilibrium price*** *and* ***equilibrium quantity*** *are values for price and quantity in the market that, once achieved, will remain constant—unless and until the supply curve or the demand curve shifts.*

Equilibrium Price and Quantity

Look at Table 1, which combines the supply and demand schedules for maple syrup from previous Tables. We'll use Table 1 to find the equilibrium price in this market through the process of elimination.

Prices below the Equilibrium Price

Let's first ask what would happen if the price were less than $3.00 per bottle—say, $1.00. At this price, Table 1 tells us that buyers would want to buy 75,000 bottles each month, while sellers would offer to sell only 25,000. There would be an **excess demand** of 50,000 bottles.

What would happen in this case?

Buyers would compete with each other to get more maple syrup than was available, and would offer to pay a higher price rather than do without. The price would then rise. The same would occur if the price were $2.00, or any other price below $3.00. We conclude that any price less than $3.00 cannot be an equilibrium price.

If the price starts below $3.00, it would start rising—*not* because the supply curve or the demand curve had shifted, but from natural forces within the market itself. This directly contradicts our definition of equilibrium price.

Figure 1 illustrates the same process by putting the supply and demand curves together on the same graph. As you can see, at a price of $1.00, quantity supplied of 25,000 bottles is found at point *H* on the supply curve, while quantity demanded is at point *J* on the demand curve. The horizontal difference between the two curves at $1.00 is a graphical representation of the excess demand at that price.

At this point, we should ask another question: If the price were initially $1.00, would it ever *stop* rising?

Yes. Since excess demand is the reason for the price to rise, the process will stop when the excess demand is gone. And as you can see in Figure 1, the rise in price *shrinks* the excess demand in two ways.

First, as price rises, buyers demand a smaller quantity—a leftward movement along the demand curve.

Second, sellers increase supply to a larger quantity—a rightward movement along the supply curve. Finally, when the price reaches $3.00 per bottle, the excess demand is gone and the price stops rising.

This logic tells us that $3.00 is an *equilibrium* price in this market—a value that won't change as long as the supply and demand curves stay put. But is it the *only* equilibrium price?

Prices above the Equilibrium Price

We've shown that any price *below* $3.00 is not equilibrium, but what about a price *greater* than $3.00?

Let's see. Suppose the price of maple syrup was, say, $5.00 per bottle. Look again at Table 1 and you'll find that, at this price, quantity supplied would be 65,000 bottles per month, while quantity demanded would be only 35,000 bottles. There is an **excess supply** of 30,000 bottles. Sellers would compete with each other to sell more maple syrup than buyers wanted to buy, and the price would fall. Thus, $5.00 cannot be the equilibrium price.

Equilibrium price The market price that, once achieved, remains constant until either the demand curve or supply curve shifts.

Equilibrium quantity The market quantity bought and sold per period that, once achieved, remains constant until either the demand curve or supply curve shifts.

Excess demand At a given price, the amount by which quantity demanded exceeds quantity supplied.

Table 1: Finding the Market Equilibrium

Price (per bottle)	Quantity Demanded (bottles per month)	Quantity Supplied (bottles per month)	Excess Demand or Supply?	Consequence
$1.00	75,000	25,000	Excess Demand	Price will Rise
$2.00	60,000	40,000	Excess Demand	Price will Rise
$3.00	**50,000**	**50,000**	**Neither**	**No Change in price**
$4.00	40,000	60,000	Excess Supply	Price will Fall
$5.00	35,000	65,000	Excess Supply	Price will Fall

Figure 1: Excess Demand Causes Price to Rise

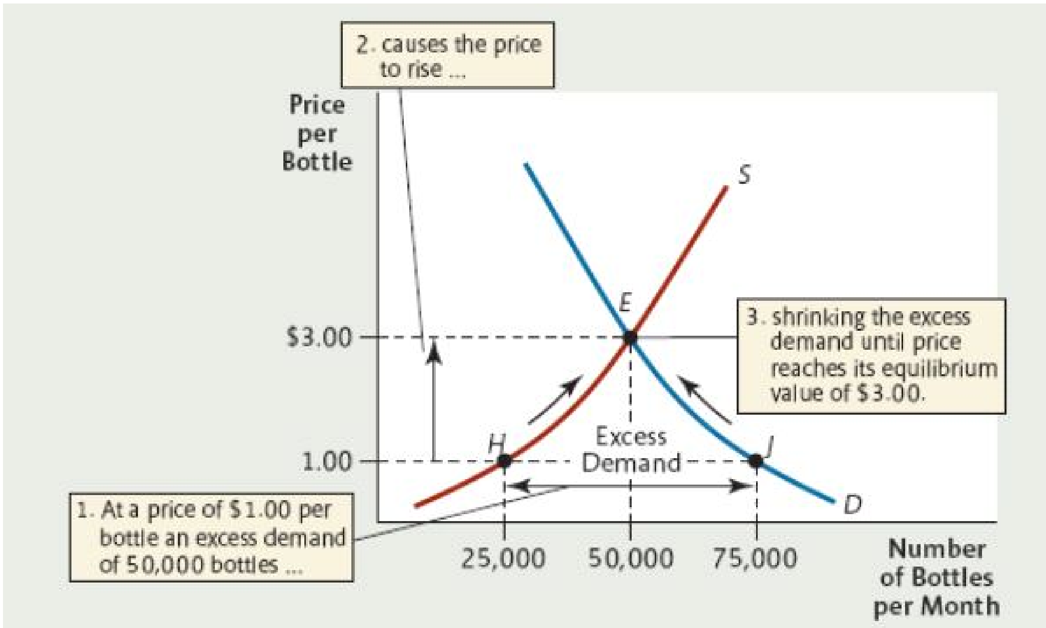
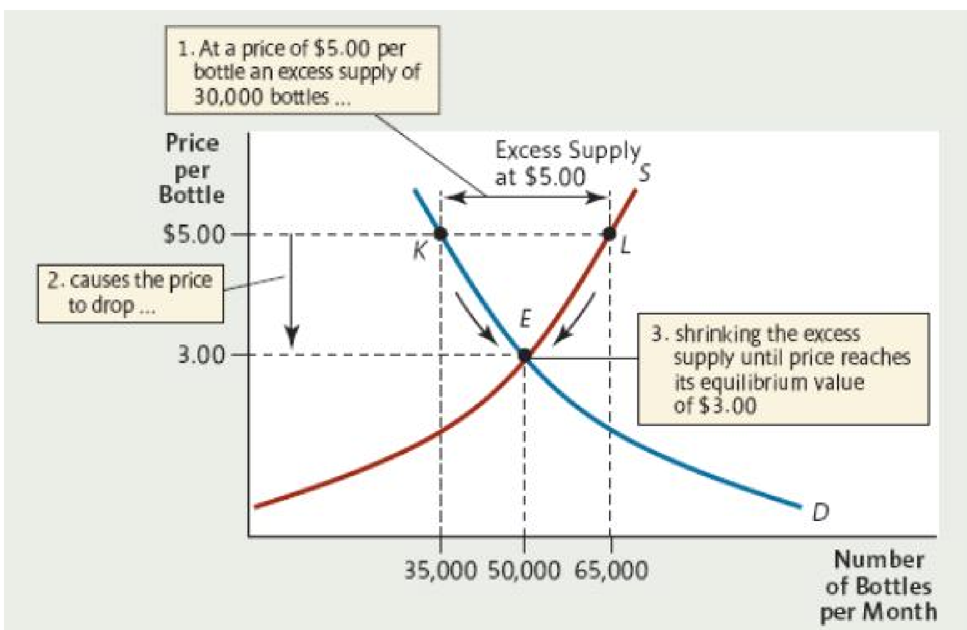


Figure 2 provides a graphical view of the market in this situation.

Figure 2: Excess Supply Causes Price to Fall



With a price of $5.00, the excess supply is the horizontal distance between points *K* (on the demand curve) and *L* (on the supply curve).

In the figure, the resulting drop in price would move us along both the supply curve (leftward) and the demand curve (rightward). As these movements continued, the excess supply of maple syrup would shrink until it disappeared, once again, at a price of $3.00 per bottle.

Our conclusion: If the price happens to be above $3.00, it will fall to $3.00 and then stop changing.

You can see that $3.00 is the equilibrium price—and the *only* equilibrium price— in this market. Moreover, at this price, sellers would want to sell 50,000 bottles—the same quantity that households would want to buy. So, when price comes to rest at $3.00, quantity comes to rest at 50,000 per month—the *equilibrium quantity*.

Equilibrium on a Graph

No doubt, you have noticed that $3.00 happens to be the price at which the supply and demand curves cross. This leads us to an easy, graphical technique for locating our equilibrium:

To find the equilibrium in a competitive market, draw the supply and demand curves. Market equilibrium occurs where the two curves cross. At this crossing point, the equilibrium price is found on the vertical axis, and the equilibrium quantity on the horizontal axis.

Excess supply At a given price, the amount by which quantity supplied exceeds quantity demanded.

Notice that in equilibrium, the market is operating on *both* the supply curve *and* the demand curve so that—at a price of $3.00—quantity demanded and quantity supplied is equal. There are no unsatisfied buyers unable to find goods they want to purchase, nor are there any frustrated sellers unable to sell goods they want to sell. Indeed, this is why $3.00 is the equilibrium price. It's the only price that creates consistency between what buyers choose to buy and sellers choose to sell. But we don't expect a market to stay at any particular equilibrium forever.

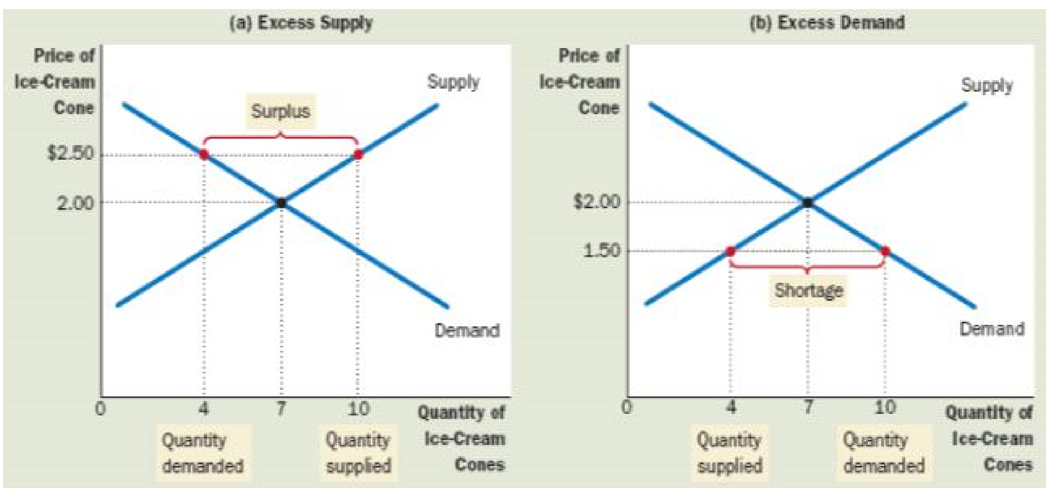
What Happens When Things Change?

Remember that in order to draw the supply and demand curves in the first place, we had to assume particular values for all the other variables—besides price—that affect demand and supply. If one of these variables changes, then either the supply curve or the demand curve will shift, and our equilibrium will change as well. Let's look at some examples.

Income Rises, Causing an Increase in Demand

In Figure 3, point *E* shows an initial equilibrium in the market for maple syrup, with an equilibrium price of $3.00 per bottle, and equilibrium quantity of 50,000 bottles per month. Suppose that the incomes of buyers rise because the economy recovers rapidly from a recession. We know that income is one of the

Figure 3: Market Not in Equilibrium



In panel (a), there is a surplus. Because the market price of $2.50 is above the equilibrium price, the quantity supplied (10 cones) exceeds the quantity demanded (4 cones). Suppliers try to increase sales by cutting the price of a cone, and this moves the price toward its equilibrium level. In panel (b), there is a shortage. Because the market price of $1.50 is below the equilibrium price, the quantity demanded (10 cones) exceeds the quantity supplied (4 cones). With too many buyers chasing too few goods, suppliers can take advantage of the shortage by raising the price. Hence, in both cases, the price adjustment moves the market toward the equilibrium of supply and demand.

Figure 4: A Shift in Both Supply and Demand

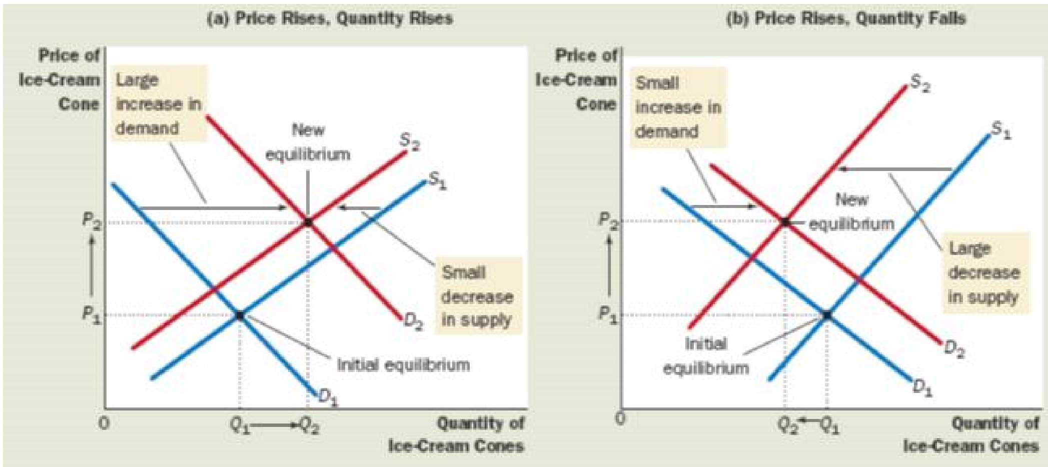
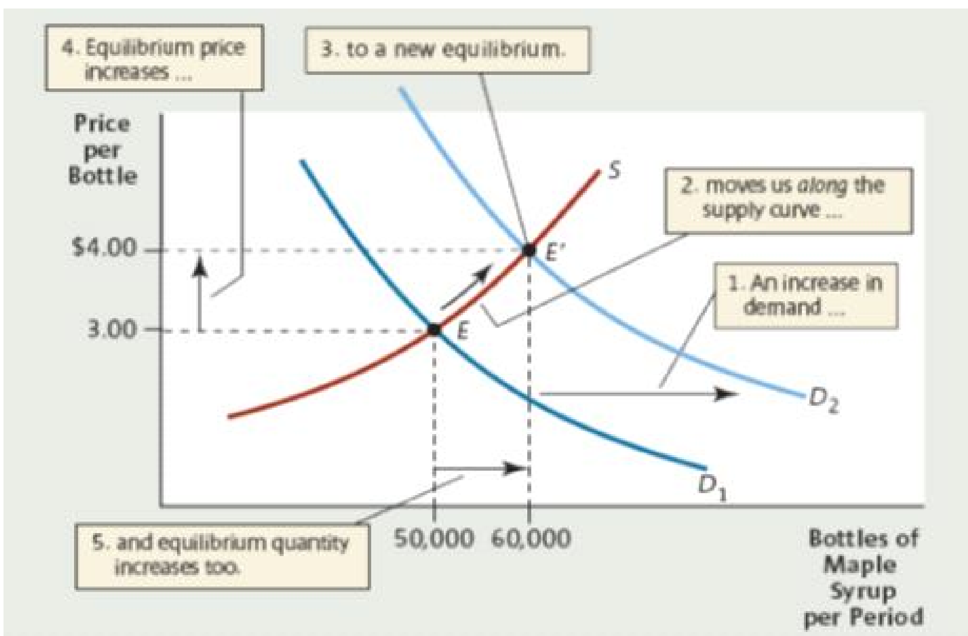


Figure 5: A Shift in Demand and a New Equilibrium



We also can reason that maple syrup is a *normal good,* so the rise in income will cause the demand curve to shift rightward. What happens then? The old price—$3.00—is no longer the equilibrium price.

How do we know?

Because if the price *did* remain at \$3.00 after the demand curve shifted, there would be an excess demand that would drive the price upward. The new equilibrium—at point E'—is the new intersection point of the curves *after* the shift in the demand curve. Comparing the original equilibrium at point E with the new one at point E', we find that the shift in demand has caused the equilibrium price to rise (from \$3.00 to \$4.00) and the equilibrium quantity to rise as well (from 50,000 to 60,000 bottles per month).

Notice, too, that in moving from point E to point E', we move *along* the supply curve. That is, a shift of the demand curve has caused a movement along the supply curve in Figure 5.

Why is this?

The demand shift causes the *price* to rise, and a rise in price always causes a movement *along* the supply curve. But the supply curve itself does not shift because none of the variables that affect sellers—other than the price of the good—has changed.

In this example, income rose. But *any* event that shifted the demand curve rightward would have the same effect on price and quantity. For example, if tastes changed in favor of maple syrup, or a substitute good like jam rose in price, or a complementary good like pancake mix became cheaper, the demand curve for maple syrup would shift rightward, just as it did in Figure 10. So, we can summarize our findings as follows:

A rightward shift in the demand curve causes a rightward movement along *the supply curve. Equilibrium price and equilibrium quantity both rise.*

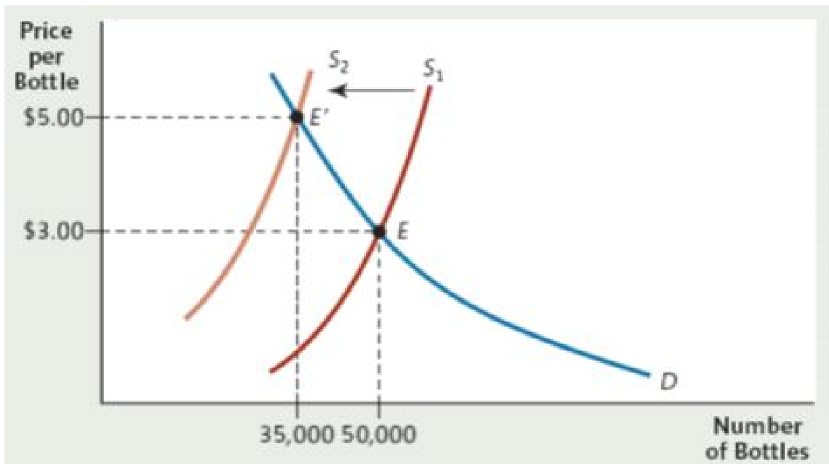
Decrease in Supply

Bad weather can affect supply for most agricultural goods, including maple syrup. Weather can be shift-variable for the supply curve. Look at Figure 6. Initially, the supply curve for maple syrup is $S1$, with the market in equilibrium at Point E. When bad weather hits, the supply curve shifts leftward—say, to $S2$. The result: a rise in the equilibrium price of maple syrup (from \$3.00 to \$5.00 in the figure) and a fall in the equilibrium quantity (from 50,000 to 35,000 bottles).

Any event that shifts the supply curve leftward would have similar effects. For example, if the wages of maple syrup workers increase, or some maple syrup producers go out of business and sell their farms to housing developers, the supply curve for maple syrup would shift leftward, just as in Figure 6.

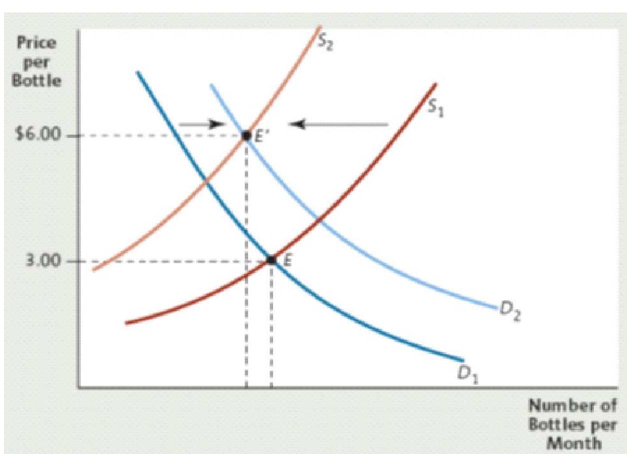
More generally, *A leftward shift of the supply curve causes a leftward movement* along *the demand curve. Equilibrium price rises, but equilibrium quantity falls.*

Figure 6: A Shift of Supply and a New Equilibrium



An ice storm causes supply to decrease from S*1 to* S*2. At the old equilibrium price of $3.00, there is now an excess demand. As a result, the price increases until excess demand is eliminated at point* E *¨. In the new equilibrium, quantity demanded again equals quantity supplied. The price is higher, and fewer bottles are produced and sold.*

Figure 7: A Shift in Both Curves and a New Equilibrium



An increase in income shifts the demand curve rightward from D*1 to* D*2. At the same time, bad weather shifts the supply curve leftward from* S*1 to* S*2. The equilibrium moves from point* E *to point* E *″. While the price must rise after these shifts, quantity could rise or fall or remain the same, depending on the relative sizes of the shifts. In the figure, quantity happens to fall.*

Higher Income and Bad Weather

So far, we've considered example in which just one curve shifts due to a change in a single variable that influences *either* demand or supply.

But what would happen if two changes affected the market simultaneously?

Then both curves would shift.

Figure 7 shows what happens when we take the two factors we've just explored separately (a rise in income and bad weather) and combine them together. The rise in income causes the demand curve to shift rightward, from *D*1 to *D*2. The bad weather causes the supply curve to shift leftward, from *S*1 to *S*2. The result of all this is a change in equilibrium from point *E* to point *E* ″, where the new demand curve *D*2 intersects the new supply curve *S*2.

Notice that the equilibrium price rises from $3.00 to $6.00 in our example. This should come as no surprise. A rightward shift in the demand curve, with no other change, causes price to rise. And a leftward shift in the supply curve, with no other change, causes price to rise. So when we combine the two shifts together, the price must rise. In fact, the increase in the price will be greater than would be caused by either shift alone.

But what about equilibrium quantity?

Here, the two shifts work in *opposite* directions. The rightward shift in demand works to increase quantity, while the leftward shift in supply works to decrease quantity. We can't say what will happen to equilibrium quantity until we know which shift is greater and thus has the greater influence. Quantity could rise, fall, or remain unchanged.

In Figure 7, it just so happens that the supply curve shifts more than the demand curve, so equilibrium quantity falls. But you can easily prove to yourself that the other outcomes are possible. First, draw a graph where the demand curves shifts rightward by more than the supply curve shifts leftward. In your graph, you'll see that equilibrium quantity rises. Then, draw one where both curves shift (in opposite directions) by equal amounts, and you'll see that equilibrium quantity remains unchanged.

We can also imagine other combinations of shifts. A rightward or leftward shift in either curve can be combined with a rightward or leftward shift in the other.

Table 2 lists all the possible combinations. It also shows what happens to equilibrium price and quantity in each case, and when the result is ambiguous (a question mark). For example, the top left entry tells us that when both the supply and demand curves shift rightward, the equilibrium *quantity* will always rise, but the equilibrium price could rise, fall, or remain unchanged, depending on the relative *size* of the shifts.

Table 2: Effect of Simultaneous Shifts in Supply and Demand

	Increase in Demand (Rightward Shift)	No Change in Demand	Decrease in Demand (Leftward Shift)
• Increase in Supply (Rightward Shift)	P↑ Q↑	P↓ Q↑	P↓ Q?
• No change in Supply	P↑ Q↑	No change in P or Q	P↓ Q↓
• Decrease in Supply (Leftward Shift)	P↑ Q?	P↑ Q↓	P? Q↓

The Three-Step Process

Step 1—Characterize the Market:

Decide which market or markets best suit the problem being analyzed, and identify the decision makers (buyers and sellers) who interact there.

In economics, we make sense of the very complex, real-world economy by viewing it as a collection of *markets.* Each of these markets involves a group of *decision makers*—buyers and sellers—who have the potential to trade with each other.

Step 2—Find the Equilibrium: *Describe the conditions necessary for equilibrium in the market, and a method for determining that equilibrium.*

Once we've defined a market, and put buyers and sellers together, we look for the at which the market will come to rest—the equilibrium. In this chapter, we used supply and demand to find the equilibrium price and quantity in a perfectly competitive market, but this is just one example of how economists apply Step 2.

Step 3—What Happens When Things Change: *Explore how events or government policies change the market equilibrium.*

Once you've found the equilibrium, the next step is to ask how different events will *change* it. In this chapter, for example, we explored how rising income or bad weather (or both together) would affect the equilibrium price and quantity for maple syrup.

Conclusion

Economists use the model of supply and demand to analyze competitive markets. In a competitive market, there are many buyers and sellers, each of whom has little or no influence on the market price.

• The demand curve shows how the quantity of a good demanded depends on the price. According to the law of demand, as the price of a good falls, the quantity demanded rises. Therefore, the demand curve slopes downward.

• In addition to price, other determinants of how much consumers want to buy include income, the prices of substitutes and complements, tastes, expectations, and the number of buyers. If one of these factors changes, the demand curve shifts. The supply curve shows how the quantity of a good supplied depends on the price. According to the law of supply, as the price of a good rises, the quantity supplied rises. Therefore, the supply curve slopes upward.

• In addition to price, other determinants of how much producers want to sell include input prices, technology, expectations, and the number of sellers. If one of these factors changes, the supply curve shifts.

• The intersection of the supply and demand curves determines the market equilibrium. At the equilibrium price, the quantity demanded equals the quantity supplied.

• The behaviour of buyers and sellers naturally drives markets toward their equilibrium. When the market price is above the equilibrium price, there is a surplus of the good, which causes the market price to fall. When the market price is below the equilibrium price, there is a shortage, which causes the market price to rise.

• To analyze how any event influences a market, we use the supply-and-demand diagramme to examine how the event affects the equilibrium price and quantity. To do this, we follow three steps. First, we decide whether the event shifts the supply curve or the demand curve (or both). Second, we decide in which direction the curve shifts. Third, we compare the new equilibrium with the initial equilibrium.

• In market economies, prices are the signals that guide economic decisions and thereby allocate scarce resources. For every good in the economy, the price ensures that supply and demand are in balance. The equilibrium price then determines how much of the good buyers choose to consume and how much sellers choose to produce.

In a market economy, prices are determined through the interaction of buyers and sellers in *markets. Perfectly competitive* markets have many buyers and sellers, and none of them individually can affect the market price. If an individual, buyer, or seller has the power to influence the price of a product, the market is *imperfectly competitive.*

The model of *supply and demand* explains how prices are determined in perfectly competitive markets. The *quantity demanded* of any good is the total amount buyers would choose to purchase given the constraints that they face. The *law of demand* states that quantity demanded is negatively related to price; it tells us that the *demand curve* slopes downward. The demand curve is drawn for given levels of income, wealth, tastes, prices of substitute and complementary goods, population, and expected future price. If any of those factors changes, the demand curve will shift. A change in price, however, moves us *along* the demand curve.

The *quantity supplied* of a good is the total amount sellers would choose to produce and sell given the constraints that they face. According to the *law of supply,* supply curves slope upward. The supply curve will shift if there is a change in the price of an input, the price of an alternate good, the price in an alternate market, the number of firms, expectations of future prices, or (for some goods) a change in weather. A change in the price of the good, by contrast, moves us *along* the supply curve.

Equilibrium price and quantity in a market are found where the supply and demand curves intersect. If either or both of these curves shift, price and quantity will change as the market moves to a new equilibrium.

Questions

1. Consider the following statement: "In late 2018, as at other times in history, oil prices came down at the same time as the quantity of oil produced fell. Therefore, one way for us to bring down oil prices is to slow down oil production." True or false? Explain.

2. Discuss, and illustrate with a graph, how each of the following events will affect the market for coffee:

   a. Blight on coffee plants kills off much of the Brazilian crop.

   b. The price of tea declines.

   c. Coffee workers organize themselves into a union and gain higher wages.

   d. Coffee is shown to cause cancer in labouratory rats.

   e. Coffee prices are expected to rise rapidly in the near future.

3. Draw supply and demand diagrammes for market *A* for each of the following. Then use your diagrammes to illustrate the impact of the following events. In each case, determine what happens to price and quantity in each market.

   a. *A* and *B* are substitutes, and the price of good *B* rises.

   b. *A* and *B* satisfy the same kinds of desires, and there is a shift in tastes away from *A* and toward *B*.

   c. *A* is a normal good, and incomes in the community increase.

   d. There is a technological advance in the production of good *A*.

   e. *B* is an input used to produce good *A*, and the price of *B* rises.

13. When we observe an increase in both price and quantity, we know that the demand curve must have shifted rightward. However, we cannot rule a shift in the supply

curve as well. Prove this by drawing a supply and demand graph for each of the following cases:

a. Demand curve shifts rightward, supply curve shifts leftward, equilibrium price and quantity both rise.

b. Demand and supply curves both shift rightward, equilibrium price and quantity both rise.

c. Evaluate the following statement: "During the oil price spike from 2017 to mid-2018, we know the supply curve could not have shifted leftward, because quantity supplied rose." True or False? Explain.

d. "During the oil price spike from 2017 to mid-2018, the supply curve may have shifted leftward (say, because a rise in expected price), but the demand curve must have shifted rightward as well." True of False? Explain.

1. What is a market? Briefly describe a type of market that is not perfectly competitive.

2. What are the demand schedule and the demand curve and how are they related? Why does the demand curve slope downward?

3. Does a change in consumers' tastes lead to a movement along the demand curve or a shift in the demand curve? Does a change in price lead to a movement along the demand curve or a shift in the demand curve?

4. Popy's income declines, and as a result, he buys more spinach. Is spinach an inferior or demand curve for spinach?

5. What are the supply schedule and the supply curve and how are they related? Why does the supply curve slope upward?

6. Does a change in producers' technology lead to a movement along the supply curve or a shift in the supply curve? Does a change in price lead to a movement along the supply curve or a shift in the supply curve?

7. Define the equilibrium of a market. Describe the forces that move a market toward its equilibrium.

This lesson will focus on the following:

1. Elasticity of Demand
2. Price Elasticity of Demand
3. Slope of Demand Curve
4. Elasticity Approach
5. Calculating Price Elasticity of Demand
6. Categorizing Demand
7. Elasticity and Straight-Line Demand Curves

Elasticity of Demand

In this lesson, you will learn about *elasticity:* measures of the sensitivity of one variable to another. As you'll see, economists use a variety of different types of elasticity to make predictions and to recommend policy changes.

Elasticity is a measure of how much buyers and sellers respond to changes in market conditions. When studying how some event or policy affects a market, we can discuss not only the direction of the effects but their magnitude as well.

Consumers usually buy more of a good when its price is lower, when their incomes are higher, when the prices of substitutes for the good are higher, or when the prices of complements of the good are lower. Discussion of demand was qualitative, not quantitative. That is, we discussed the direction in which quantity demanded moves but not the size of the change. To measure how much consumers respond to changes in these variables, economists use the concept of **elasticity**.

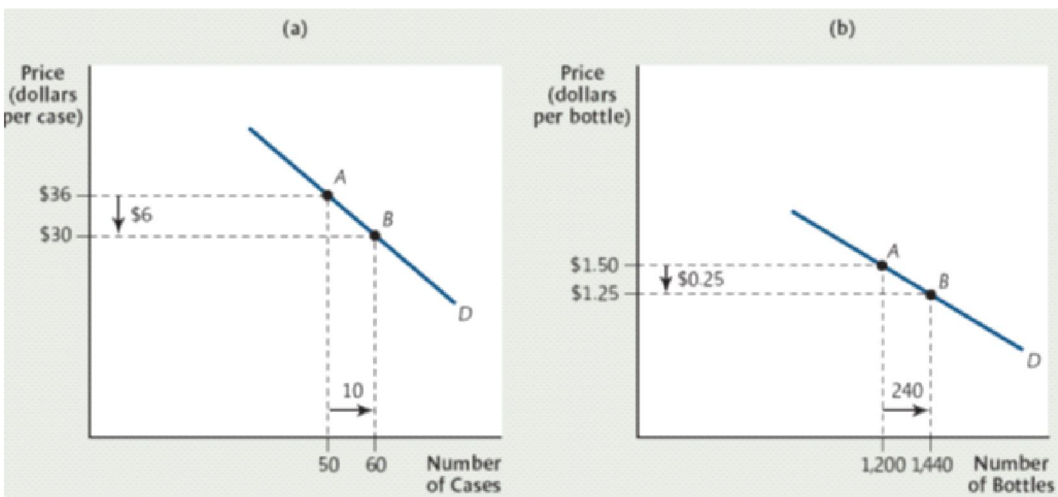
Price Elasticity of Demand

At the most general level, *elasticity* measures the sensitivity of one market variable to another. One of the most important elasticity is the *price elasticity of demand,* which measures the sensitivity of quantity demanded to the price of the good itself.

But how should we measure this sensitivity?

One obvious candidate is the *slope* or *steepness* of the demand curve. After all, for any given rise in price, the flatter the demand curve, the greater will be the decrease in quantity demanded along the curve. So it seems that the flatter the demand curve (the smaller the absolute value of its slope), the greater the sensitivity of quantity demanded to price. But that reasoning is only partially correct, and it can get us into trouble.

Figure 1: Same Buying Behaviour, Different Slopes



In each panel, the movement from A *to* B *represents the same buying behaviour in the market for bottled water. In panel (a), the unit of measurement is* cases. *When price drops by $6 per case, quantity demanded rises by 10 cases, so the slope of the demand curve is - $6/10 = - 0.6. In panel (b), the unit of measurement is* bottles. *The same price decrease ($6 per case) translates to 25 cents per bottle, and the same quantity increase (10 cases) translates to 240 bottles. Using bottles, the slope is - $0.25/240 = - .001. Although the demand behaviour is the same, the slopes are different. This is one reason why slope is a poor measure of the price sensitivity of demand.*

For example, 20-ounce bottles of water are sold both individually and by the case (24 bottles per case). Suppose we use "cases," and find that the demand curve in a market is like that shown in the left panel of Figure 1. In that market, a \$6 drop in price per case causes people to buy 10 more cases per day (moving us from point *A* to point *B*). Because we graph price on the vertical axis and number of cases on the horizontal, the *slope* of the demand curve would be $\Delta P/\Delta Q_D$ = - \$6/10 = - 0.6.

But now, for the same market, let's change our unit of measurement from "number of cases" to "number of bottles." \$6 less per *case* translates to 25 cents less per *bottle*. And a 10-case increase in quantity demanded translates to 240 bottles. The right panel of Figure 1 shows the demand curve for "bottles" (with the scale of the axes adjusted so we can see larger quantities and smaller prices more easily). As we move from point *A* to *B*, the slope of the demand curve is now $\Delta P/\Delta Q_D$ = - 0.25/240 = - .001.

Buyers respond the same way in both examples, yet—due only to an arbitrary change in units—the demand curves have very different slopes. Clearly, we can't rely on the slope as our measure of price sensitivity.

A second problem is that the slope of the demand curve doesn't tell us anything about the *significance* of a change in price or quantity—whether it is a relatively small or a relatively large change. A price drop of \$0.05, for example, is a tiny, hardly noticeable change for a good with a current price of \$500. But it's a relatively huge change if the current price is \$0.08. Our measure of price sensitivity should take this into account.

Elasticity Approach

The elasticity approach solves both of these problems by comparing the *percentage change* in quantity demanded with the *percentage change* in price.

More specifically:

The ***price elasticity of demand*** (E_D) *for a good is the percentage change in quantity demanded divided by the percentage change in price:*

$E_{D\,=}$ % Change in Quantity Demanded ÷ % Change in Price

For example, if the price of newspapers falls by 2 percent, and this causes the quantity demanded to rise by 6 percent, then $E_D = 6\%/2\% = 3.0$. We would say "the price elasticity of demand for newspapers is 3.0."

Of course, when price *falls* by 2 percent, that's a change of *negative* 2 percent, while quantity demanded changes by $_+$6 percent.

So technically speaking, elasticity should be viewed as a negative number. We'll follow a common convention of dropping the minus sign. That way, when we compare elasticities and say that one is larger, we'll be comparing absolute values.

In our example, elasticity has the value 3.0. But what, exactly, does that number mean? Here is a straightforward way to interpret the number:

The price elasticity of demand (E_D) tells us the percentage change in quantity demanded for each 1 percent change in price.

In our example, with $E_D = 3.0$, each 1 percent drop in price causes quantity demanded to rise by 3 percent. Given this interpretation, it's clear that an elasticity value of 3.0 implies greater price sensitivity than an elasticity value of 2.0, or one of 0.7.

More generally, *The greater the elasticity value, the more sensitive quantity demanded is to price.*

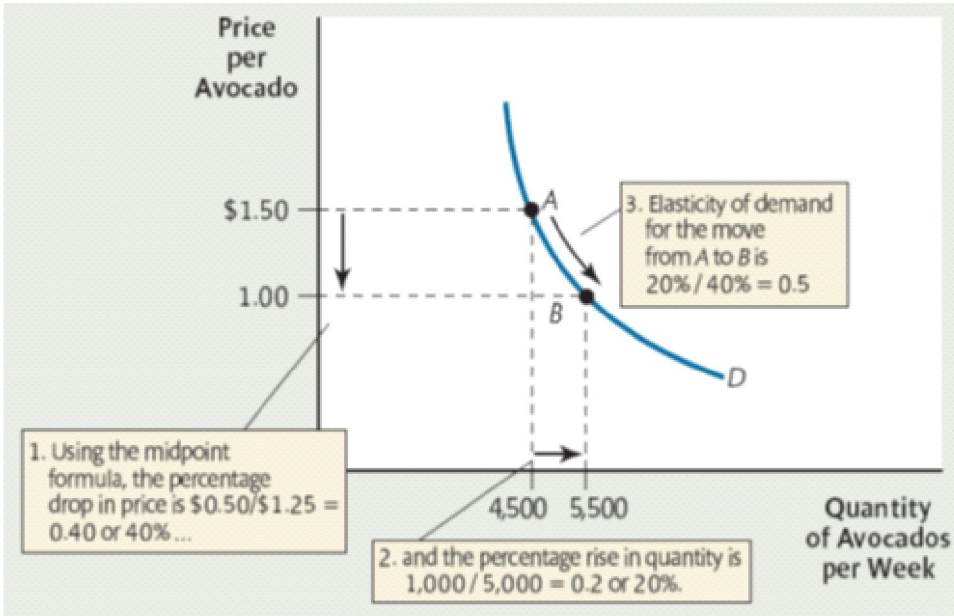
Calculating Price Elasticity of Demand

When we calculate price elasticity of demand, we imagine that *only price* is changing, while we hold constant all other influences on quantity demanded, such as buyers' incomes, the prices of other goods, and so on. Thus, we measure elasticity for a movement *along* an unchanging demand curve.

Figure 2, for example, shows a hypothetical demand curve in the market for avocados in a city. Suppose we want to measure elasticity along this demand curve between points *A* and *B*. As our first step, we'll calculate the percentage change in price.

Price elasticity of demand The sensitivity of quantity demanded to price; the percentage change in quantity demanded caused by a 1 percent change in price.

Figure 2: Using the Midpoint Formula for Elasticity



Let's suppose we move from point *A* to point *B*. Price falls by \$0.50. Since our starting price at point *A* was \$1.50, this would be a 33 percent drop in price. But wait . . . suppose we go in the reverse direction, from point *B* to *A*. Now our starting price would be \$1.00, so the \$0.50 price hike would be a 50 percent rise.

The percentage change in price (33 or 50 percent) depends on the direction we are moving. And the same will be true of quantity. Therefore, our elasticity value will also depend on which direction we move.

This presents us with a problem. Ideally, we'd like our measure of price sensitivity to be the same whether we go from *A* to *B* or from *B* to *A*, since each is simply the mirror image of the other. To accomplish this goal, elasticity calculations often use a special convention to get percentage changes: Instead of dividing the change in a variable by its *starting* value, we divide the change by the *average* of its starting and ending values. This is often called the "midpoint formula," because we are dividing the change by the midpoint between the old and new values.

When determining elasticities, we calculate the percentage change in a variable using the midpoint formula: the change in the variable divided by the average of the old and new values.

For example, in Figure 2, between points *A* and *B* the average of the old and new price is ($1.50 + $1.00) / 2 = $1.25. Using this average price as our base, the percentage change in price is $0.50 / $1.25 = 0.40 or 40 percent. With the midpoint formula, the percentage change in price is the same whether we move from *A* to *B,* or from *B* to *A*. More generally, when price changes from any value $P0$ to any other value $P1$, we define the percentage change in price as

% Change in Price = $(P1 - P0) \div (P1 - P0) / 2$

The term in the numerator is the change in price; the term in the denominator is the average of the two prices.

The percentage change in quantity demanded is calculated in a similar way.

When quantity demanded changes from $Q0$ to $Q1$, the percentage change is calculated as

% Change in Quantity Demanded = $(Q1 - Q0) \div (Q1 \_ Q0) / 2$

Once again, we are using the average of the initial and the new quantity demanded as our base quantity.

The midpoint formula is an approximation to the actual percentage change in a variable, but it has the advantage of giving us consistent elasticity values when we reverse directions. We will use the midpoint formula only when *calculating elasticity values from data on prices and quantities*. For all other purposes, we calculate percentage changes in the normal way, using the starting value as the base.

An Example

Let's calculate the price elasticity of demand for avocados along a part of the demand curve in Figure 2. As price falls from $1.50 to $1.00, quantity demanded rises from 4,500 to 5,500. Using the midpoint formula (and dropping negative signs):

% Change in Quantity Demanded = (5500 – 4500) ÷ (5500 + 4500 / 2)

= 1000 ÷ 5000

= 0.20 or 20%.

% Change in Price = [$1 - $1.5] ÷ [($1 + $1.5) / 2]

= $0.50- $1.25

= 0.40 or 40%.

Finally, we use these numbers to calculate the price elasticity of demand:

E_D = % Change in Quantity Demanded ÷ % Change in Price

= 20% ÷ 40%

= 0.5.

Or, in simple English, a 1 percent change in price causes a ½ percent change in quantity demanded.

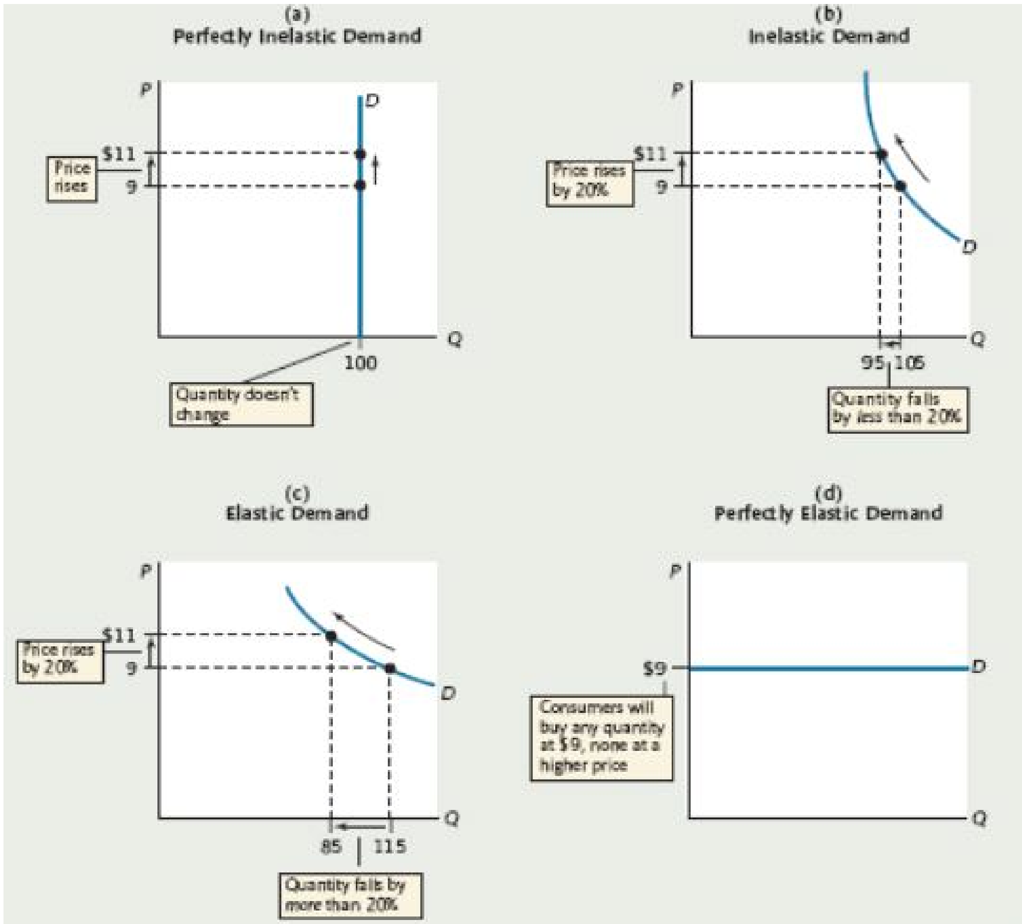
Categorizing Demand

Economists have found it useful to divide demand curves (or parts of demand curves) into categories, based on their elasticity values. These categories are illustrated in Figure 3.

Panel (a) shows an extreme theoretical case, called **perfectly inelastic demand,** where the elasticity has a value of zero. A perfectly inelastic demand curve is vertical, so a change in price causes *no* change in quantity demanded. In the figure, when price rises from $9 to $11 (20 percent using the midpoint formula), our formula for price elasticity of demand (E_D) gives us E_D = 0%/20% = 0.

Perfectly inelastic demand A price elasticity of demand equal to 0.

Figure 3: Categories of Demand Curves



Panel (b) shows a case where quantity demanded has *some* sensitivity to price, but not much. Here, the same 20 percent price increase causes quantity demand to fall from 105 to 95 (a 10 percent decrease using the midpoint formula). In this case, $E_D = 10\%/20\% = 0.5$. This is an example **inelastic demand,** which occurs whenever $E_D < 1$ (quantity changes by a smaller percentage than price).

Panel (c) shows a demand curve with more price sensitivity: the 20 percent rise in price causes quantity demanded to drop by 30 percent. Our elasticity calculation is $E_D = 30\%/20\% = 1.5$. This is an example of **elastic demand,** which occurs whenever $E_D > 1$ (quantity changes by a larger percentage than price changes).

Inelastic demand A price elasticity of demand between 0 and 1.

Elastic demand A price elasticity of demand greater than 1.

Finally, panel (d) shows another extreme case, called **perfectly elastic demand,** where the demand curve is horizontal. As long as the price stays at one particular value (where the demand curve touches the vertical axis), *any* quantity might be demanded. But even the tiniest price rise would cause quantity demanded to fall to zero. In this case, E_D = " (elasticity is infinite) because no matter how small we make the percentage change in price (in the denominator), the percentage change in quantity (in the numerator) will always be infinitely larger.

What about the special case when elasticity of demand is *exactly* equal to 1.0?

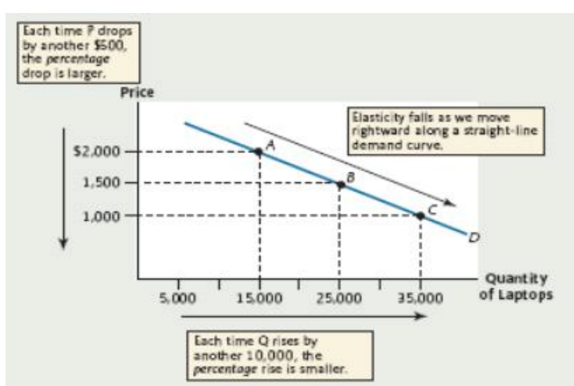
Then demand is neither elastic nor inelastic, but lies between these categories. We call this case **unit elastic.** Take a moment to draw a demand curve that is unit elastic for a price change from $9 to $11, choosing your numbers for quantity carefully.

Elasticity and Straight Line Demand Curve

Figure 4 shows a linear (straight-line) demand curve for laptop computers. Each time price drops by $500, the quantity of laptops demanded rises by 10,000. Because this behaviour remains constant all along the curve, is the price elasticity of demand also constant?

Actually, no. Elasticity is the ratio of *percentage* changes; what remains constant along a linear demand curve is the ratio of *absolute* or *unit* changes.

Figure 4: How Elasticity Changes along a Straight-Line Demand Curve



In fact, we can show that as we move rightward along a linear demand curve, the elasticity always decreases. For example, let's calculate the elasticity between points *A* and *B*. Price falls from \$2,000 to \$1,500, a 28.6 percent drop using the midpoint formula. Quantity rises from 15,000 to 25,000, which is a 50 percent rise using the midpoint formula. Taking the ratio of these changes, we find that the, elasticity for a move from point *A* to *B* is 50%/ 28.6% = 1.75.

Now let's calculate the elasticity between points *B* and *C*, where price falls from \$1,500 to \$1,000, and quantity rises from 25,000 to 35,000. For this change (as you can verify), price falls by 40 percent while quantity demanded rises by 33.3 percent (using the midpoint formula). So the elasticity for a move from point *B* to *C* is 33.3%/40% = 0.83.

Notice what's happened: as we've moved downward and rightward along this straight-line demand curve, elasticity has fallen from 1.75 to 0.83. Demand has become *less elastic*.

There is a good reason for this. As we travel down the demand curve, the average quantity we use as the base for figuring percentage changes keeps increasing. So a constant 10,000 increase in quantity becomes a smaller and smaller *percentage* increase. The opposite also happens with price: It keeps getting smaller, so the same \$500 decrease in price becomes a growing *percentage* decrease. Thus, as we travel down a linear demand curve, with $\%\Delta Q_D$ shrinking and $\%\Delta P$ growing, the ratio $\%\Delta Q_D / \%\Delta P$ decreases.

Elasticity of demand varies along a straight-line demand curve. More specifically, demand becomes less elastic (ED gets smaller) as we move downward and rightward.

This is a special conclusion about *linear* demand curves only. For *non*linear demand curves, moving down the curve can cause elasticity to rise, fall, or remain constant, depending on the shape of the curve.

Perfectly (infinitely) elastic demand A price elasticity of demand approaching infinity.

Unit elastic demand A price elasticity of demand equal to 1.

Questions

1. Explain Elasticity of Demand.
2. What is meant by Price Elasticity of Demand?

3. Derive the Slope of Demand Curve.
4. Explain in detail the Elasticity Approach.
5. How to Calculate the Price Elasticity of Demand?
6. What are different categories of demand?
7. Estimate Elasticity on Straight-Line Demand Curve.

This lesson will focus on the following:

1. Elasticity and Total Revenue
2. Determinants of Elasticity
3. Time Horizons and Demand Curves
4. Income Elasticity of Demand
5. Cross-Price Elasticity of Demand
6. Price Elasticity of Supply

Elasticity and Total Revenue

When the price of a good increases, the law of demand tells us that people will buy less of it. But this does not necessarily mean they will *spend* less on it. After the price rises, fewer units will be purchased in the market, but each unit will cost more. What happens to *total spending* on the good? Or, recognizing that total spending by all buyers equals the total revenue of all sellers, we can ask the same question this way:

When price rises, what happens to the *total combined revenue* of all firms that sell in the market?

Let's see. On the one hand, each unit sold can be sold for more, tending to increase revenue. On the other hand, fewer units will be sold, which works to decrease revenue. Which one will dominate?

The answer depends on the price elasticity of demand for the good. To see why, note that the total revenue of sellers in a market (TR) is the price per unit (P) times the quantity that people buy (Q):

$$TR = P - Q.$$

When we raise price, P goes up, but Q goes down. What happens to the product depends on which one changes by a larger percentage. Suppose that demand is *inelastic* ($E_D < 1$). Then a 1 percent rise in price will cause quantity demanded to fall by *less* than 1 percent. So the greater amount sellers get on each unit outweighs the impact of the drop in quantity, and total revenue will *rise*.

The behaviour of total revenue can be seen very clearly on a graph, once you learn how to interpret it. Look at the left panel of Figure 1, which duplicates the inelastic demand curve introduced earlier. On this demand curve, let's start at a price of \$9, and look at the rectangle with a corner at point A.

The height of the rectangle is the price of \$9, and the width is the quantity of 105, so its *area* (height x width = P x Q = \$9 x 105 = \$945) is the total revenue of sellers when price is \$9.

More generally, *At any point on a demand curve, sellers' total revenue is the area of a rectangle with height equal to price and width equal to quantity demanded.*

Now let's raise the price to \$11. The total revenue rectangle becomes the larger one, with a corner at point B. The area of this rectangle is TR = \$11 - 95 = \$1,045.

The rise in price has *increased* total revenue. Now suppose that demand is *elastic* ($E_D >$ 1). Once again, a 1 percent rise in price causes quantity demanded to fall, but this time it falls by *more* than 1 percent.

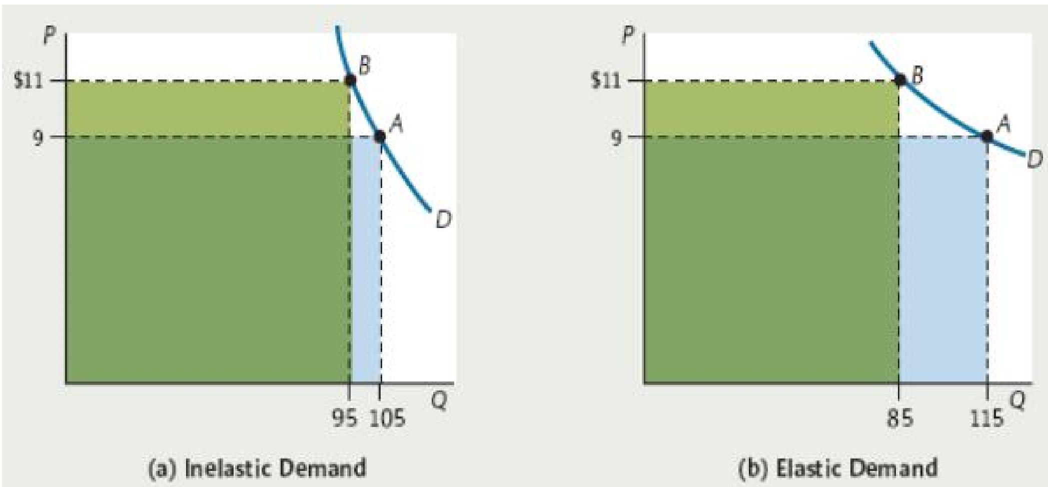
So the fact that sellers get more on each unit is outweighed by the drop in the quantity they sell, and total revenue *falls*.

This is shown in the right panel of Figure 1, using the example of elastic demand from a few pages earlier. When price is \$9, TR is the area of the rectangle with a corner at point A, equal to \$9 x 115 = \$1,035. When price rises to \$11, TR becomes the area of the taller rectangle with corner at point B. This area is \$11 x 85 = \$935. Because demand is elastic, the rise in price *decreases* total revenue.

We can conclude that:

An increase in price raises total revenue when demand is inelastic, and shrinks total revenue when demand is elastic.

Figure 1: Elasticity and Total Revenue



In panel (a), demand is inelastic, so a rise in price causes total revenue to increase. Specifically, at a price of $9 (point A*), total revenue is $9 x 105 = $945. When price rises to $11 (point* B*), total revenue increases to $11 x 95 = $1,045. In panel (b), demand is elastic, so a rise in price causes total revenue to decrease. Specifically, at a price of $9 (point* A*), total revenue is $9 x 115 = $1,035. When price rises to $11 (point* B*), total revenue falls to $11 x 85 = $935.*

Table 1: Effects of Price Changes on Revenue

Where Demand Is:	A Price Increase Will:	A Price Decrease Will:
inelastic ($E_D < 1$)	increase revenue	decrease revenue
unit elastic ($E_D = 1$)	cause no change in revenue	cause no change in revenue
elastic ($E_D > 1$)	decrease revenue	increase revenue

What if price fell instead of rose?

Then, in Figure 1 we'd be making the reverse move: from point *B* to point *A* on each curve. And logic tells us that if demand is inelastic, total revenue must fall. If demand is elastic, the drop in price will cause total revenue to rise.

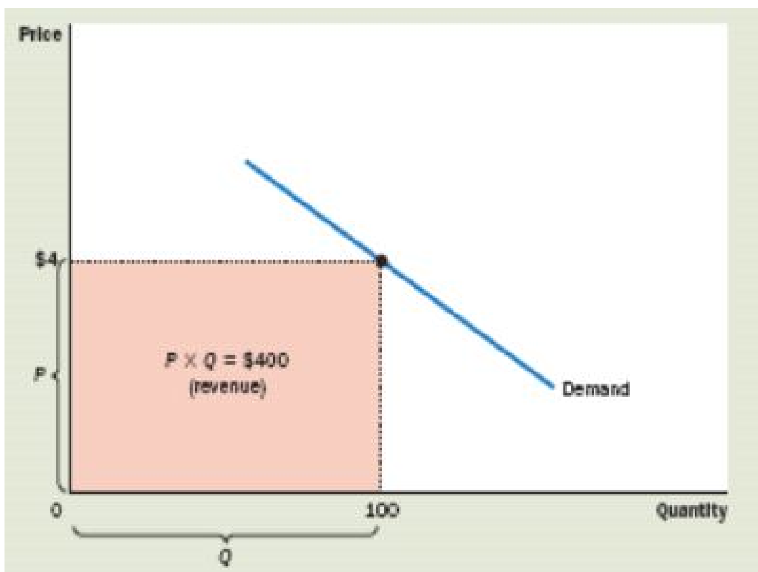
A decrease in price shrinks total revenue when demand is inelastic, and raises total revenue when demand is elastic.

What happens if demand is unit elastic?

You can probably guess. This would mean that a 1 percent change in price causes a 1 percent change in quantity, but in the opposite direction. The two effects on total revenue would cancel each other out, so total revenue would remain unchanged.

Table 1 summarizes these results about elasticity and total revenue. Don't try to memorize the table, but *do* use it to test yourself: Try to explain the logic for each entry.

Figure 2: Total Revenue

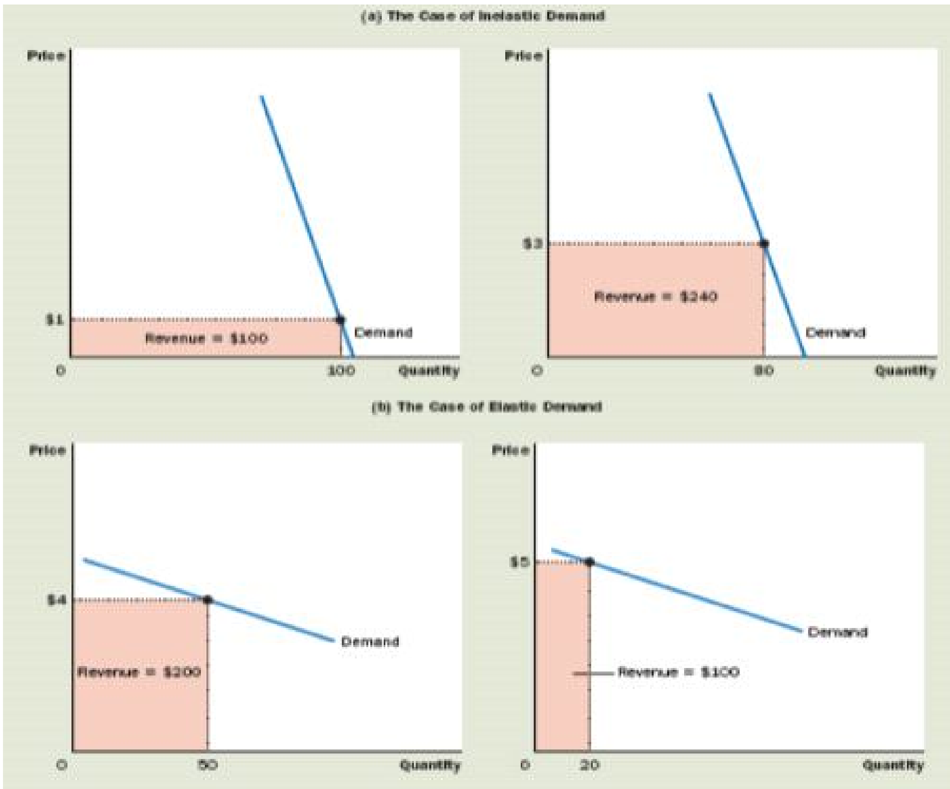


The total amount paid by buyers, and received as revenue by sellers, equals the area of the box under the demand curve, $P \times Q$. Here, at a price of $4, the quantity demanded is 100, and total revenue is $400.

Elasticity and Total Revenue along a Linear Demand Curve

Let's examine how elasticity varies along a linear demand curve, as shown in Figure 3. We know that a straight line has a constant slope. Slope is defined as "rise over run," which here is the ratio of the change in price ("rise") to the change in quantity ("run"). This particular demand curve's slope is constant because each $1 increase in price causes the same two-unit decrease in the quantity demanded.

Figure 3: How Total Revenue Changes When Price Changes



The impact of a price change on total revenue (the product of price and quantity) depends on the elasticity of demand. In panel (a), the demand curve is inelastic. In this case, an increase in the price leads to a decrease in quantity demanded that is proportionately smaller, so total revenue increases. Here an increase in the price from $1 to $3 causes the quantity demanded to fall from 100 to 80. Total revenue rises from $100 to $240. In panel (b), the demand curve is elastic. In this case, an increase in the price leads to a decrease in quantity demanded that is proportionately larger, so total revenue decreases. Here an increase in the price from $4 to $5 causes the quantity demanded to fall from 50 to 20. Total revenue falls from $200 to $100.

Even though the slope of a linear demand curve is constant, the elasticity is not. This is true because the slope is the ratio of *changes* in the two variables, whereas the elasticity is the ratio of *percentage changes* in the two variables.

You can see this by looking at the table 2, which shows the demand schedule for the linear demand curve in the graph. The table uses the midpoint method to calculate the price elasticity of demand. At points with a low price and high quantity, the demand curve is inelastic. At points with a high price and low quantity, the demand curve is elastic.

The table also presents total revenue at each point on the demand curve. These numbers illustrate the relationship between total revenue and elasticity. When the price is $1, for instance, demand is inelastic, and a price increase to $2 raises total revenue. When the price is $5, demand is elastic, and a price increase to $6 reduces total revenue. Between $3 and $4, demand is exactly unit elastic, and total revenue is the same at these two prices.

The linear demand curve illustrates that the price elasticity of demand need not be the same at all points on a demand curve. A constant elasticity is possible, but it is not always the case.

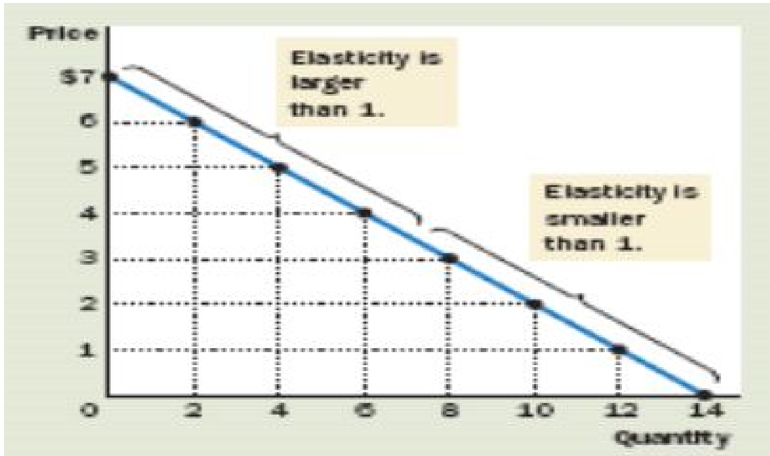
Elasticity of a Linear Demand Curve

The slope of a linear demand curve is constant, but its elasticity is not. The demand schedule in the table was used to calculate the price elasticity of demand by the midpoint method. At points with a low price and high quantity, the demand curve is inelastic. At points with a high price and low quantity, the demand curve is elastic.

Table 2: Elasticity of a Linear Demand Curve

Price	Quantity	Total Revenue (Price × Quantity)	Percentage Change in Price	Percentage Change in Quantity	Elasticity	Description
$7	0	$ 0				
			15	200	13.0	Elastic
6	2	12				
			18	67	3.7	Elastic
5	4	20				
			22	40	1.8	Elastic
4	6	24				
			29	29	1.0	Unit elastic
3	8	24				
			40	22	0.6	Inelastic
2	10	20				
			67	18	0.3	Inelastic
1	12	12				
			200	15	0.1	Inelastic
0	14	0				

Figure 4: Elasticity of a Linear Demand Curve



The impact of a price change on total revenue (the product of price and quantity) depends on the elasticity of demand. In panel (a), the demand curve is inelastic. In this case, an increase in the price leads to a decrease in quantity demanded that is proportionately smaller, so total revenue increases. Here an increase in the price from $1 to $3 causes the quantity demanded to fall from 100 to 80. Total revenue rises from $100 to $240. In panel (b), the demand curve is elastic. In this case, an increase in the price leads to a decrease in quantity demanded that is proportionately larger, so total revenue decreases. Here an increase in the price from $4 to $5 causes the quantity demanded to fall from 50 to 20. Total revenue falls from $200 to $100.

Determinants of Elasticity

Availability of Substitutes

When close substitutes are available for a good, demand for it will be more elastic. If the price of meat rises, with all other prices held constant, consumers can easily switch to chicken. But when the price of gasoline rises, the substitutes that are available (using mass transit, carpooling, biking, or even not going places) are not as close. Thus, it is not surprising that the demand for meat is more elastic than the demand for gasoline.

When close substitutes are available for a product, demand tends to be more elastic.

One factor that determines the closeness of substitutes is how narrowly or broadly we define the market we are analyzing. Demand in the market for beverages as a whole will be less elastic than demand in the market for soft drinks. And demand for soft drinks will be less elastic than the demand for Pepsi.

This is because when we determine the elasticity of demand in a market, we hold constant all prices outside of the market. So in determining the elasticity for Pepsi, we ask what happens when the price of Pepsi rises but the price of Coke remains constant. Since it is so easy to switch to Coke, demand is highly elastic.

But in determining the elasticity for soft drinks, we ask what happens when the price of *all* soft drinks rise together, holding constant only the prices of things that are *not* soft drinks. Demand is therefore less elastic.

Necessities versus Luxuries

Goods that we think of as necessary for our survival or general well-being, and for which there are no close substitutes, are often referred to as "necessities." Most people would include the broad categories "food," "housing," and "medical care" in this category. When we regard something as a necessity, demand for it will tend to be less elastic. This is another reason why the elasticity of demand for gasoline is so small: Many people regard gasoline as a necessity.

By contrast, goods that we can more easily do without—such as entertainment or vacation travel—are often referred to as "luxuries." Demand for these goods will tend to be more elastic, since people will cut back their purchases more when price rises.

Goods we regard as necessities tend to have less elastic demand than goods we regard as luxuries.

Importance in Buyers' Budgets

When a good takes up a large part of your budget initially, a rise in price has a large impact on how much you will have left to spend on other things. All else equal, this will tend to make demand more elastic. For example, a vacation trip to Paris would take a big bite out of most peoples' budgets. If the price of the vacation rises by, say, 20 percent, many people will start to consider other alternatives, since not doing so would mean a considerable sacrifice of other purchases.

Now consider the other extreme: ordinary table salt. A family with an income of $50,000 per year would spend less than 0.005 percent of its income on this good, so the price of salt could double or triple and have no significant impact on the ability to buy other goods. We would therefore expect the demand for table salt to be inelastic.

Demand for food is more elastic than the demand for eggs. Based on the narrowness of definition, we would expect the reverse. But eggs make up a rather small fraction of the typical family's budget, and certainly smaller than food as a whole. This tends to reduce the elasticity of demand for eggs.

When spending on a good makes up a larger proportion of families' budgets, demand tends to be more elastic.

Time Horizon

How much time we wait after a price change can have an important impact on the elasticity of demand. **Short-run elasticities**: the quantity response is measured for just a short time (usually a year or less) after the price change. A **long-run elasticity** measures the quantity response after more time has elapsed—typically a few years or more.

Demand is almost always more elastic in the long run than in the short run.

Why? Because the longer we wait after a sustained price change, the more time consumers have to make adjustments in their lives that affect their quantity demanded.

In general,

Short-run elasticity An elasticity measured just a short time after a price change.

Long-run elasticity An elasticity measured a year or more after a price change.

The longer we wait after a price change to measure the quantity response, the more elastic is demand. Therefore, long-run elasticities tend to be larger than short-run elasticities.

Table 2: Adjustments after a Rise in the Price of Gasoline

Short Run (less than a year)	Long Run (a year or more)
Use public transit more often	Buy a more fuel-efficient car
Arrange a car pool	Move closer to your job
Check tire pressure more often	Switch to a job closer to home
Drive more slowly on the highway	Move to a city where less driving is required
Eliminate unnecessary trips (use mail order instead of driving to stores; locate goods by phone or Internet instead of driving around; shop for food less often and buy more each time)	
If there are two cars, use the more fuel-efficient one	

It lists some of the ways people can adjust to a significant rise in the price of gasoline over the short run and the long run. Remember that the adjustments in the long-run column are *additional* adjustments people can make if given enough time.

Time Horizons and Demand Curves

Isn't price elasticity of demand measured *along* a demand curve? Indeed it is. But then how can we get two different elasticity measures from the same market?

The answer is: There can be more than one demand curve associated with a market. Whenever we draw a demand curve, we draw it for a *specific* time horizon.

Short-run demand curves show quantity demanded at different prices when people only have a short period of time (a few weeks or a few months) to adjust.

A *long-run* demand curve shows quantity demanded after buyers have had much longer—say, a year or more—to adjust to a price change.

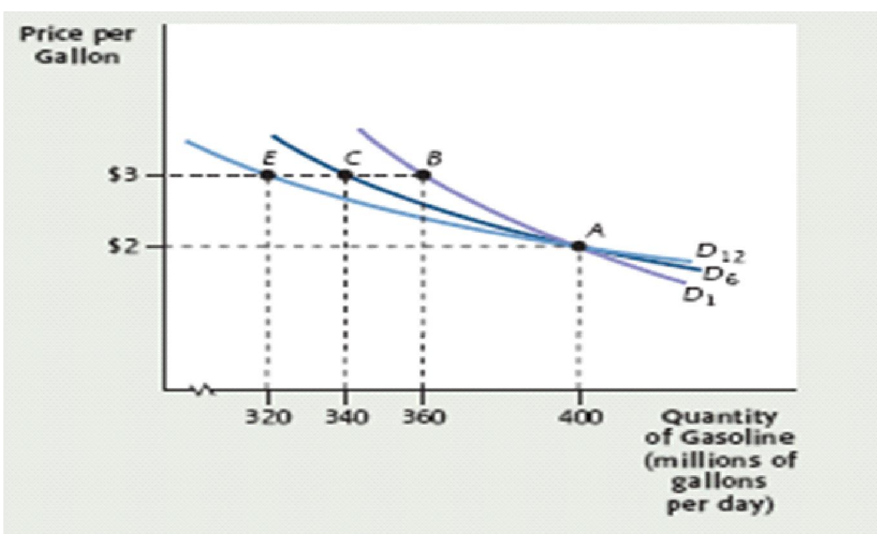
The three demand curves in Figure 5 illustrate how this works in the market for gasoline. We assume that initially we are at point *A,* with a price of $2 per gallon and quantity demanded of 400 million gallons per day. As long as the price—and every other influence on demand—remains the same, we would stay at point *A*.

But now, suppose the price rises from \$2 to \$3 and stays there. To find the daily quantity demanded *one month later,* we would move along the demand curve labeled *D*1. This demand curve has a very low elasticity of demand because there is not much time for gasoline buyers to adjust. For example, they might cut out some unnecessary trips, but they are unlikely to purchase a more fuel efficient car in this time frame. Using demand curve *D*1, we end up at point *B*, with gasoline demand equal to 360 million gallons per day.

To find the quantity demanded *six months later,* we would move along the demand curve labeled *D*6. Demand curve *D*6 is more elastic than *D*1, because we are allowing buyers more time to make adjustments that will reduce their purchases (e.g., for some gasoline buyers, 6 months is long enough to acquire a more fuel efficient car). So, if we wait 6 months, we'll find that we've moved from point A to point *C*, with consumers buying 340 million gallons of gas per day.

Finally, to find daily quantity demanded *12 months later,* we would move along the demand curve labeled *D*12. This demand curve is more elastic than the other two, because we've allowed buyers even more time to adjust to the higher price. So if we measure the quantity response after waiting a full year, we'll find that we've moved from point *A* to point *E*, where quantity demanded has fallen to 320 million gallons per day.

Figure 5: Different Demand Curves for Different Time Horizons



When the price of gasoline rises by $1, the decrease in quantity demanded (and the price elasticity of demand) depends on how long we wait before measuring buyers' response. If we wait just one month after the price change, we'd move along demand curve D1*, from point* A *to point* B*. If we wait six months, we'd move along demand curve* D6*, from point* A *to point* C*. The same rise in price causes a greater decrease in quantity demanded after six months, because buyers can make further adjustments. If we wait 12 months, we'd move from point* A *to point* E *along demand curve* D12*, with quantity demanded falling even more.*

Any demand curve is drawn for a particular time horizon (a waiting period before we observe the new quantity demanded after a price change). In general, the longer the time horizon, the more elastic the demand.

As a rule of thumb, demand curves drawn for time horizons less than one year are called short-run demand curves, while those drawn for time horizons of one year or longer are called long-run demand curves.

Price Elasticity of Demand and its Determinants

The law of demand states that a fall in the price of a good raises the quantity demanded. The **price elasticity of demand** measures how much the quantity demanded responds to a change in price. Demand for a good is said to be *elastic* if the quantity demanded responds substantially to changes in the price. Demand is said to be *inelastic* if the quantity demanded responds only slightly to changes in the price. The price elasticity of demand for any good measures how willing consumers are to buy less of the good as its price rises. Thus, the elasticity reflects the many economic, social, and psychological forces that shape consumer preferences. Based on experience, however, we can state some general rules about what determines the price elasticity of demand.

Availability of Close Substitutes

Goods with close substitutes tend to have more elastic demand because it is easier for consumers to switch from that good to others. For example, butter and margarine are easily substitutable. A small increase in the price of butter, assuming the price of margarine is held fixed, causes the quantity of butter sold to fall by a large amount. By contrast, because eggs are a food without a close substitute, the demand for eggs is less elastic than the demand for butter.

Necessities versus Luxuries

Necessities tend to have inelastic demands, whereas luxuries have elastic demands. When the price of a doctor's visit rises, people will not dramatically reduce the number of times they go to the doctor, although they might go somewhat less often. By contrast, when the price of sailboats rises, the quantity of sailboats demanded falls substantially. The reason is that most people view doctor visits as a necessity and sailboats as a luxury. Of course, whether a good is a necessity or a luxury depends not on the intrinsic properties of the good but on the preferences of the buyer. For avid sailors with little concern over their health, sailboats might be a necessity with inelastic demand and doctor visits a luxury with elastic demand.

Definition of the Market

The elasticity of demand in any market depends on how we draw the boundaries of the market. Narrowly defined markets tend to have more elastic demand than broadly defined markets because it is easier to find close substitutes for narrowly defined goods. For example, food, a broad category, has a fairly inelastic demand because there are no good substitutes for food. Ice cream, a narrower category, has a more elastic demand because it is easy to substitute other desserts for ice cream. Vanilla ice cream, a very narrow category, has a very elastic demand because other flavors of ice cream are almost perfect substitutes for vanilla.

Time Horizon

Goods tend to have more elastic demand over longer time horizons. When the price of gasoline rises, the quantity of gasoline demanded falls only slightly in the first few months. Over time, however, people buy more fuel efficient cars, switch to public transportation, and move closer to where they work. Within several years, the quantity of gasoline demanded falls more substantially.

Computing Price Elasticity of Demand

Now that we have discussed the price elasticity of demand in general terms, let's be more precise about how it is measured. Economists compute the price elasticity of demand as the percentage change in the quantity demanded divided by the percentage change in the price. That is,

Price elasticity of demand = Percentage change in quantity demanded ÷ Percentage change in price

For example, suppose that a 10 percent increase in the price of an ice-cream cone causes the amount of ice cream you buy to fall by 20 percent. We calculate elasticity of demand as

Price elasticity of demand = 20 percent

In this example, the elasticity is 2, reflecting that the change in the quantity demanded is proportionately twice as large as the change in the price. Because the quantity demanded of a good is negatively related to its price, the percentage change in quantity will always have the opposite sign as the percentage change in price.

In this example, the percentage change in price is a *positive* 10 percent (reflecting an increase), and the percentage change in quantity demanded is a *negative* 20 percent (reflecting a decrease). For this reason, price elasticities of demand are sometimes reported as negative numbers. We follow the common practice of dropping the minus sign and reporting all price elasticities of demand as positive numbers. (Mathematicians call this the *absolute value*).

Midpoint Method

If you try calculating the price elasticity of demand between two points on a demand curve, you will quickly notice an annoying problem: The elasticity from point A to point B seems different from the elasticity from point B to point A. For example, consider these numbers:

Point A: Price = $4 Quantity = 120

Point B: Price = $6 Quantity = 80

Going from point A to point B, the price rises by 50 percent, and the quantity falls by 33 percent, indicating that the price elasticity of demand is 33/50, or 0.66. By contrast, going from point B to point A, the price falls by 33 percent, and the quantity rises by 50 percent, indicating that the price elasticity of demand is 50/33, or 1.5. This difference arises because the percentage changes are calculated from a different base.

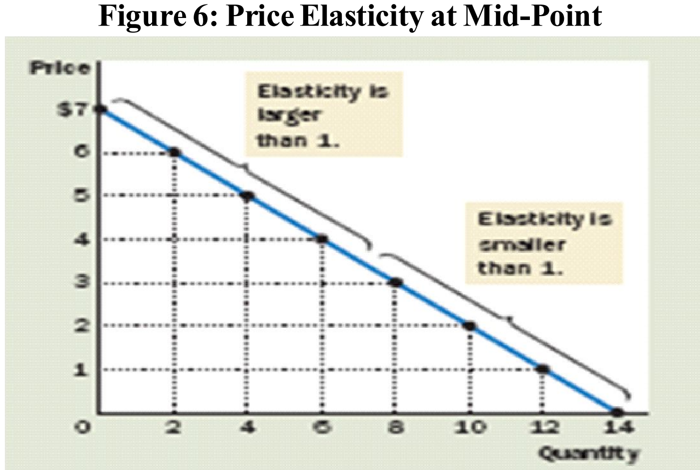
One way to avoid this problem is to use the *midpoint method* for calculating elasticities. The standard procedure for computing a percentage change is to divide the change by the initial level. By contrast, the midpoint method computes a percentage change by dividing the change by the midpoint (or average) of the initial and final levels. For instance, \$5 is the midpoint between \$4 and \$6. Therefore, according to the midpoint method, a change from \$4 to \$6 is considered a 40 percent rise because $(6 - 4) / 5 \times 100 = 40$. Similarly, a change from \$6 to \$4 is considered a 40 percent fall.

Because the midpoint method gives the same answer regardless of the direction of change, it is often used when calculating the price elasticity of demand between two points. In our example, the midpoint between point A and point B is:

Midpoint: Price = \$5 Quantity = 100

According to the midpoint method, when going from point A to point B, the price rises by 40 percent, and the quantity falls by 40 percent. Similarly, when going from point B to point A, the price falls by 40 percent, and the quantity rises by 40 percent. In both directions, the price elasticity of demand equals 1.

Figure 6: Price Elasticity at Mid-Point



The following formula expresses the midpoint method for calculating the price elasticity of demand between two points, denoted $(Q1, P1)$ and $(Q2, P2)$:

$$\text{Price elasticity of demand} = \frac{(Q2 - Q1) / [(Q2 + Q1) / 2]}{(P2 - P1) / [(P2 + P1) / 2]}$$

The numerator is the percentage change in quantity computed using the midpoint method, and the denominator is the percentage change in price computed using the midpoint method. If you ever need to calculate elasticities, you should use this formula.

Other Elasticities

The concept of *elasticity* is a very general one. It can be used to measure the sensitivity of virtually *any* variable to any other variable. All types of elasticity measures, however, share one thing in common: They tell us the percentage change in one variable caused by a 1 percent change in the other. Let's look briefly at three additional elasticity measures, and what each of them tells us.

Income Elasticity of Demand

The *income elasticity of demand* tells us how *sensitive* quantity demanded is to changes in buyers' incomes. The **income elasticity of demand** E_Y is the percentage change in quantity demanded divided by the percentage change in income, with all other influences on demand—including the price of the good— remaining constant:

Income elasticity of demand

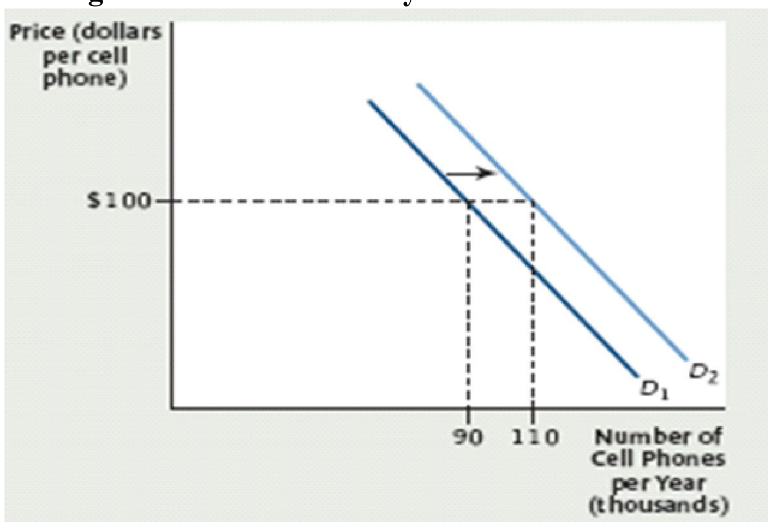
The percentage change in quantity demanded caused by a 1 percent change in income.

Income Elasticity = % Change in Quantity Demanded ÷ % Change in Income

Keep in mind that while price elasticity measures the sensitivity of demand to price as we *move along the demand curve* from one point to another, an income elasticity tells us the relative *shift* in the demand curve—the percentage increase in quantity demanded *at a given price*. Figure 7 illustrates how we might calculate an income elasticity in the market for cell phones. In the figure, we assume that the average income of buyers in the market rises by 10%. (Note that income itself is not shown in the graph.) As a result, *at a given price* ($100 in the figure), the quantity of cell phones demanded rises from 90 thousand to 110 thousand. Using our midpoint rule, we find that the percentage change in quantity is

20/100 _ 20%. So, using the formula, the income elasticity of demand would be 20%/10% = 2.0.

Figure 7: Income Elasticity and Demand Curves



Income elasticity is the percentage increase in demand (at a given price) divided by the percentage change in income. In the figure, we assume that income rises by 10%, causing quantity demanded at a price of $100 to rise from 90 to 110, or—using the midpoint rule—by 20/100 = 20%. Thus, the income elasticity of demand is 20%/10% = 2.0.

Note that with income elasticities (unlike price elasticities), the sign of the elasticity value matters. Income elasticity will be positive when people want more of a good as their income rises. Such goods are called *normal* goods. But income elasticity can also be negative, when a rise in income *decreases* demand for a good (*inferior* goods.) *Income elasticity is positive for normal goods, but negative for inferior goods.*

Inter-city bus travel is, in many markets, an inferior good. As household income rises, travelers are likely to shift away from cheaper bus travel to more expensive car, train, or airline travel. Similarly, as income rises, many households will shift from cheaper sources of calories (e.g., rice and beans) to more expensive items (steak, fresh fruit, and sushi).

An accurate knowledge of income elasticity can be crucial in predicting the growth in demand for a good as income grows over time. For example, economists know that different types of countries have different income elasticities of demand for oil. (In less-developed countries undergoing rapid industrialization, the income elasticity of demand for oil is typically twice as large as in developed countries.)

Cross-Elasticity of Demand

Cross-price elasticity relates the percentage change in quantity demanded for one good to the percentage change in the price of another good. More formally, we define the **cross-price elasticity of demand** between good X and good Z as:

% Change in Quantity Demanded of $X \div$ % Change in Price of Z

In words, a cross-price elasticity of demand tells us the percentage change in quantity demanded of a good for each 1 percent increase in the price of some other good, while all other influences on demand remain unchanged.

With cross-price elasticity (as with an income elasticity), the sign matters. A *positive* cross price elasticity means that the two goods are *substitutes:* A rise in the price of one good increases demand for the other good. For example, Coke and Pepsi are clearly substitutes, and the cross-price elasticity of Pepsi with Coke has, in one study, been estimated at 0.8.4 This means that a 1 percent rise in the price of Coke, holding constant the price of Pepsi, causes a 0.8 percent *rise* in the quantity of Pepsi demanded.

Similarly, gasoline and mass transit are substitutes. Cross-price elasticity of mass transit with the price of gasoline (and other trip-related automobile costs) was equal to 2.69. In simple English: A 10- percent rise in the cost of using an automobile for a trip would cause a 27 percent rise in mass transit use.

Negative cross-price elasticity means that the goods are *complements:* A rise in the price of one good *decreases* the demand for the other. Thus we'd expect higher gasoline prices to decrease the demand for large, fuel-inefficient cars. Indeed, when gasoline prices spiked in 2017 and 2018, the demand for SUVs and other gasguzzling vehicles plunged.

Price Elasticity of Supply

The law of supply states that higher prices raise the quantity supplied. The **price elasticity of supply** measures how much the quantity supplied responds to changes in the price. Supply of a good is said to be *elastic* if the quantity supplied responds substantially to changes in the price. Supply is said to be *inelastic* if the quantity supplied responds only slightly to changes in the price. The price elasticity of supply depends on the flexibility of sellers to change the amount of the good they produce. For example, beachfront land has an inelastic supply because it is almost impossible to produce more of it. By contrast, manufactured goods, such as books, cars, and televisions, have elastic supplies because

The **price elasticity of supply** is the percentage change in the quantity of a good supplied that is caused by a 1 percent change in the price of the good, with all other influences on supply held constant.

Price Elasticity of Supply = % Change in Quantity Supplied ÷ % Change in Price

The price elasticity of supply measures the sensitivity of quantity supplied to price changes as we move *along* the supply curve. A large value for the price elasticity of supply means that quantity supplied is very sensitive to price changes. For example, an elasticity value of 5 would imply that if price increased by 1 percent, quantity supplied would rise by 5 percent.

In most markets, a key determinant of the price elasticity of supply is the time period being considered. Supply is usually more elastic in the long run than in the short run. Over short periods of time, firms cannot easily change the size of their factories to make more or less of a good. Thus, in the short run, the quantity supplied is not very responsive to the price. By contrast, over longer periods, firms can build new factories or close old ones. In addition, new firms can enter a market, and old firms can shut down. Thus, in the long run, the quantity supplied can respond substantially to price changes.

Computing Price Elasticity of Supply

Economists compute the price elasticity of supply as the percentage change in the quantity supplied divided by the percentage change in the price. That is,

Price elasticity of supply = Percentage change in quantity supplied ÷ Percentage change in price

For example, suppose that an increase in the price of milk from $2.85 to $3.15 a gallon raises the amount that dairy farmers produce from 9,000 to 11,000 gallons per month. Using the midpoint method, we calculate the percentage change in price as

Percentage change in price = (3.15 – 2.85) / 3.00 × 100 = 10 percent.

Similarly, we calculate the percentage change in quantity supplied as percentage change in quantity supplied = (11,000 – 9,000) / 10,000 × 100 = 20 percent.

In this case, the price elasticity of supply is

Price elasticity of supply = 20 percent ÷ 10 percent

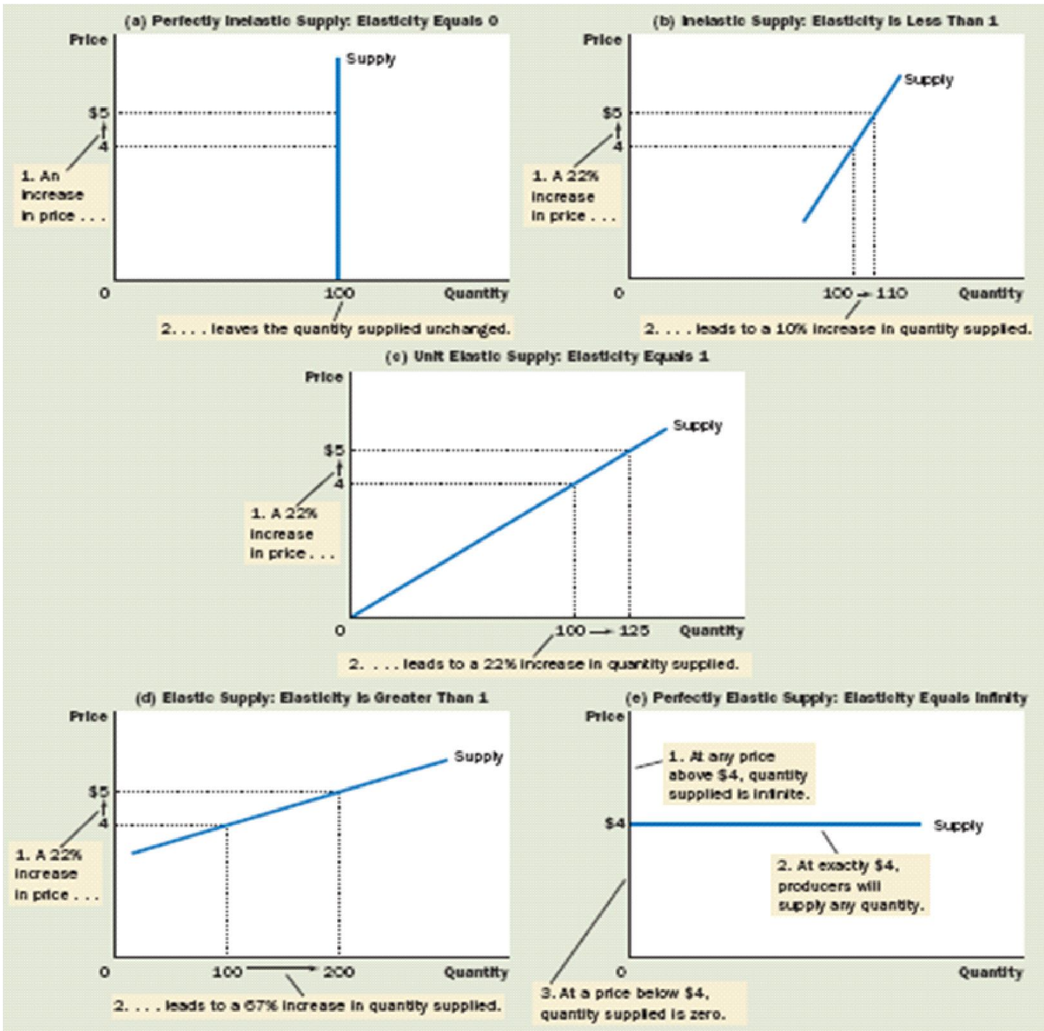
= 2.0.

In this example, the elasticity of 2 indicates that the quantity supplied changes proportionately twice as much as the price.

Variety of Supply Curves

Because the price elasticity of supply measures the responsiveness of quantity supplied to the price, it is reflected in the appearance of the supply curve. Figure 8 shows five cases. In the extreme case of zero elasticity, as shown in panel (a), supply is *perfectly inelastic*, and the supply curve is vertical. In this case, the quantity supplied is the same regardless of the price. As the elasticity rises, the supply curve gets flatter, which shows that the quantity supplied responds more to changes in the price. At the opposite extreme, shown in panel (e), supply is *perfectly elastic*. This occurs as the price elasticity of supply approaches infinity and the supply curve becomes horizontal, meaning that very small changes in the price lead to very large changes in the quantity supplied.

Figure 8: Price Elasticity of Supply



The price elasticity of supply determines whether the supply curve is steep or flat. Note that all percentage changes are calculated using the midpoint method. In some markets, the elasticity of supply is not constant but varies over the supply curve.

Figure 9 shows a typical case for an industry in which firms have factories with a limited capacity for production. For low levels of quantity supplied, the elasticity of supply is high, indicating that firms respond substantially to changes in the price. In this region, firms have capacity for production that is not being used, such as plants and equipment idle for all or part of the day. Small increases in price make it profitable for firms to begin using this idle capacity. As the quantity supplied rises, firms begin to reach capacity. Once capacity is fully used, increasing production further requires the construction of new plants. To induce firms to incur this extra expense, the price must rise substantially, so supply becomes less elastic.

Figure 9 presents a numerical example of this phenomenon. When the price rises from $3 to $4 (a 29 percent increase, according to the midpoint method), the quantity supplied rises from 100 to 200 (a 67 percent increase). Because quantity supplied changes proportionately more than the price, the supply curve has elasticity greater than 1. By contrast, when the price rises from $12 to $15 (a 22 percent increase), the quantity supplied rises from 500 to 525 (a 5 percent increase). In this case, quantity supplied moves proportionately less than the price, so the elasticity is less than 1.

Determinants of Price Elasticity of Supply

A major determinant of supply elasticity is the ease with which suppliers can find profitable activities that are *alternatives* to producing the good in question. In general, supply will tend to be more elastic when suppliers can switch to producing alternate goods more easily.

When can we expect suppliers to have easy alternatives? First, the nature of the good itself plays a role. All else equal, the supply of envelopes should be more elastic than the supply of microprocessor chips. This is because envelope producers can more easily modify their production lines to produce alternative paper products. Microprocessor suppliers, however, would be hard-pressed to produce anything other than computer chips.

Finally, the *time horizon* is important. The longer we wait after a price change, the greater the supply response to a price change. As we will see when we discuss the theory of the firm, there usually is *some* response to a price change right away. Existing firms simply speed up or slow down production with their current facilities. But further responses come about as firms have time to change their plant and equipment, and new firms have time to enter or leave an industry.

Conclusion

A useful tool for analyzing markets is *elasticity*: a measure of the sensitivity of one economic variable to another. The *price elasticity of demand* is defined as the percentage change in quantity demanded divided by the percentage change in price that caused it, without the negative sign. In general, price elasticity of demand varies along a demand curve. In the special case of a straight-line demand curve, demand becomes more and more elastic as we move upward and leftward along the curve. Along an elastic portion of any demand curve, a rise in price causes sellers' revenues (and consumers' expenditures) to fall.

Along an *inelastic* portion of any demand curve, a rise in price causes sellers' revenues and consumers' expenditures to increase. Generally speaking, demand for a good tends to be more elastic the less we regard the good as a "necessity," the easier it is to find substitutes for the good, the greater the share of households' budgets that is spent on the good, and the more time we allow for quantity demanded to respond to the price change.

In addition to price elasticity of demand, there are three other commonly used elasticities. The *income elasticity of demand* is the percentage change in quantity demanded divided by the percentage change in income that causes it.

The *cross-price elasticity of demand* is the percentage change in the quantity demanded of one good divided by the percentage change in the price of some other good. For income and price elasticities, the sign can be either positive or negative.

Finally, the *price elasticity of supply* is the percentage change in quantity supplied divided by the percentage change in price.

Conclusion

The price elasticity of demand measures how much the quantity demanded responds to changes in the price. Demand tends to be more elastic if close substitutes are available, if the good is a luxury rather than a necessity, if the market is narrowly defined, or if buyers have substantial time to react to a price change.

The price elasticity of demand is calculated as the percentage change in quantity demanded divided by the percentage change in price. If quantity demanded moves proportionately less than the price, then the elasticity is less than 1, and demand is said to be inelastic. If quantity demanded moves proportionately more than the price, then the elasticity is greater than 1, and demand is said to be elastic.

• Total revenue, the total amount paid for a good, equals the price of the good times the quantity sold. For inelastic demand curves, total revenue rises as price rises. For elastic demand curves, total revenue falls as price rises.

• The income elasticity of demand measures how much the quantity demanded responds to changes in consumers' income. The cross-price elasticity of demand measures how much the quantity demanded of one good responds to changes in the price of another good.

• The price elasticity of supply measures how much the quantity supplied responds to changes in the price. This elasticity often depends on the time horizon under consideration. In most markets, supply is more elastic in the long run than in the short run.

• The price elasticity of supply is calculated as the percentage change in quantity supplied divided by the percentage change in price. If quantity supplied moves proportionately less than the price, then the elasticity is less than 1, and supply is said to be inelastic. If quantity supplied moves proportionately more than the price, then the elasticity is greater than 1, and supply is said to be elastic.

• The tools of supply and demand can be applied in many different kinds of markets. This chapter uses them to analyze the market for wheat, the market for oil, and the market for illegal drugs.

Questions

1. Define the price elasticity of demand and the income elasticity of demand.

2. List and explain the four determinants of the price elasticity of demand.

3. What is the main advantage of using the midpoint method for calculating elasticity?

4. If the elasticity is greater than 1, is demand elastic or inelastic? If the elasticity equals 0, is demand perfectly elastic or perfectly inelastic?

5. On a supply-and-demand diagramme, show equilibrium price, equilibrium quantity, and the total revenue received by producers.

6. If demand is elastic, how will an increase in price change total revenue? Explain.
7. What do we call a good whose income elasticity is less than 0?
8. How is the price elasticity of supply calculated? Explain what it measures.
9. What is the price elasticity of supply of Picasso paintings?
10. Is the price elasticity of supply usually larger in the short run or in the long run? Why?
11. How did elasticity help explain why drug interdiction could reduce the supply of drugs, yet possibly increase drug-related crime?

This lesson will focus on the following:

1. Consumer Choice: Budget Constraint
2. Changes in Budget Line
3. Consumer Preferences
4. Consumer Decisions: Marginal Utility Approach
5. Combining Budget Constraint and Preferences

Consumer Choice: Budget Constraint

Making economic decisions involve *spending.* Economic choices require to allocate a scarce resource, for example, *time,* among different alternatives.

To understand the economic choices that individuals make, we must know what they are trying to achieve (their goals) and the limitations they face in achieving them (their constraints). Of course, we are all different from one another . . . when it comes to *specific* goals and *specific* constraints.

But at the highest level of generality, we are all very much alike. All of us, for example, would like to maximize our overall level of *satisfaction.* And all of us, as we attempt to satisfy our desires, come up against the same types of constraints: too little income or wealth to buy everything we might enjoy, and too little time to enjoy it all.

Choices about *spending:* how people decide what to buy. This is why the theory of individual decision making is often called "consumer theory."

The Budget Constraint

We all must face two facts of economic life:

(1) We have to pay for the goods and services we buy, and

(2) We have limited funds to spend.

These two facts are summarized by *budget constraint:*

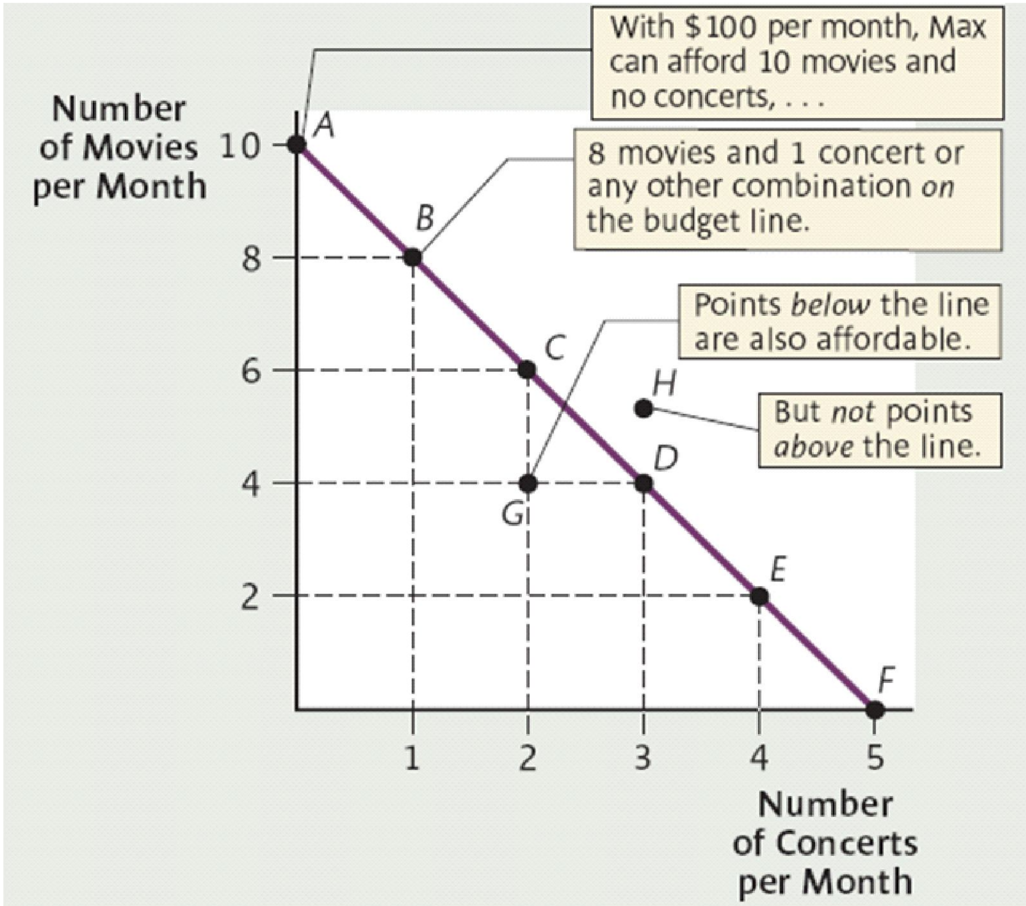
A consumer's ***budget constraint*** *identifies which combinations of goods and services the consumer can afford with a limited budget, at given prices.* Consider Max, a devoted fan of both movies and the local music scene, who has a total entertainment budget of $100 each month. The price of a movie is $10, while hearing a rock concert at his favorite local club costs him $20. If Max were to spend his $100 budget on concerts at $20 each, he could see at most five each month. If he were to spend it all on movies at $10 each, he could see 10 of them.

Table 1: Budget Line

Max's Consumption Possibilities with Income of $100

	Concerts at $20 each		Movies at $10 each	
	Quantity	Total Expenditure on Concerts	Quantity	Total Expenditure on Movies
A	0	$ 0	10	$100
B	1	$ 20	8	$ 80
C	2	$ 40	6	$ 60
D	3	$ 60	4	$ 40
E	4	$ 80	2	$ 20
F	5	$100	0	$ 0

Figure 1: Budget Constraint



But Max could also choose to spend *part* of his budget on concerts and *part* on movies. In this case, for each number of concerts, there is some *maximum* number of movies that he could see. For example, if he goes to one concert per month, it will cost him $20 of his $100 budget, leaving $80 available for movies. Thus, if Max were to choose one concert, the *maximum* number of films he could choose would be $80/$10 = 8.

Figure 1 lists, for each number of concerts, the maximum number of movies that Max could see. Each choice in the table is affordable for Max, since each will cost him exactly $100. Combination *A,* at one extreme, represents no concerts and 10 movies. Combination *F,* the other extreme, represents 5 concerts and no movies. In each of the combinations between *A* and *F,* Max attends both concerts and movies.

The Figure 1 plots the number of movies along the vertical axis and the number of concerts along the horizontal. Each of the points *A* through *F* corresponds to one of the combinations in the table. If we connect all of these points with a straight line, we have a graphical representation of Max's budget constraint, which we call Max's **budget line.**

Note that any point below or to the left of the budget line is affordable. For example, two concerts and four movies—indicated by point *G*—would cost only \$40 + \$40 = \$80. Max could certainly afford this combination. On the other hand, he *cannot* afford any combination *above* and to the right of this line. Point *H,* representing 3 concerts and 5 movies, would cost \$60 + \$50 = \$110, which is beyond Max's budget. The budget line therefore serves as a *border* between those combinations that are affordable and those that are not.

Let's look at Max's budget line more closely. The *vertical intercept* is 10, the number of movies Max could see if he attended zero concerts. Starting at the vertical intercept (point *A*), notice that each time Max increases one unit along the horizontal axis (attends one more concert), he must decrease 2 units along the vertical (see three fewer movies). Thus, the slope of the budget line is equal to - 2. The slope tells us Max's *opportunity cost* of one more concert. That is, the opportunity cost of one more concert is 2 movies foregone.

There is an important relationship between the *prices* of two goods and the opportunity cost of having more of one or the other. If we divide one money price by another money price, we get what is called a **relative price**, the price of one good *relative* to the other. Let's use the symbol P_c for the price of a concert and P_m for the price of a movie. Since P_c = \$20 and P_m = \$10, the *relative price of a concert* is the ratio P_c / P_m = \$20/\$10 = 2.

Notice that this same number, 2, is the opportunity cost of another concert in terms of movies; and, except for the minus sign, it is also the slope of the budget line. That is, *the relative price of a concert, the opportunity cost of another concert, and the slope of the budget line* have the same absolute value.

This is one example of a general relationship:

The slope of the budget line indicates the spending tradeoff between one good and another—the amount of one good that must be sacrificed in order to buy more of *another good. If P_y is the price of the good on the vertical axis and P_x is the price of the good on the horizontal axis, then the slope of the budget line is $-P_x / P_y$.*

Changes in Budget Line

To draw the budget line in Figure 1, we have assumed given prices for movies and concerts, and a given income that Max can spend on them. These "givens"—the prices of the goods and the consumer's income—are always *assumed constant* as we move along a budget line; if any one of them changes, the budget line will change as well. Let's see how.

Changes in Income

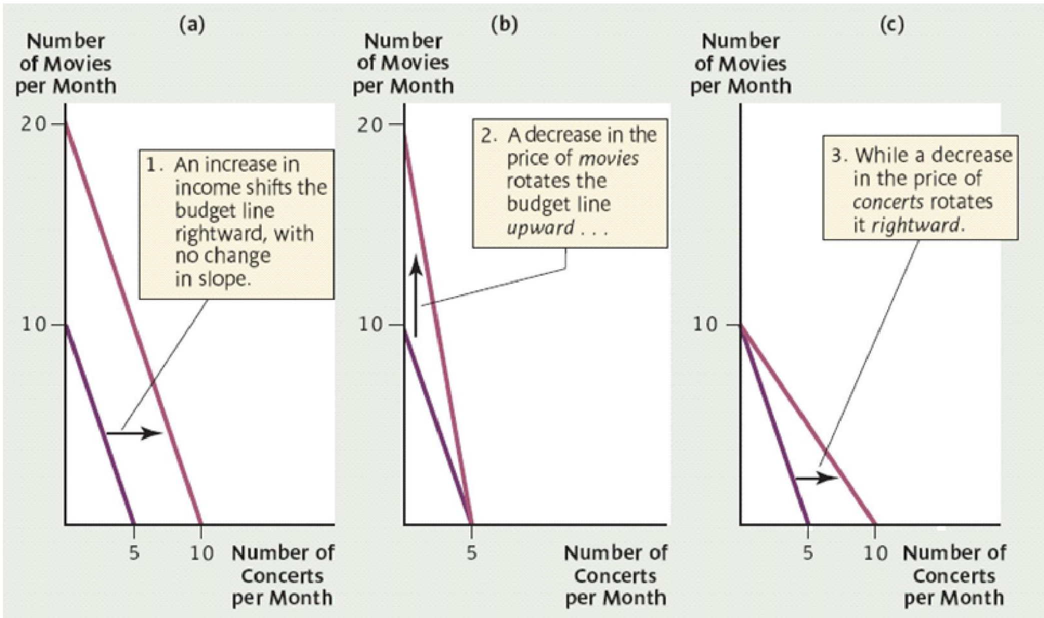
If Max's available income increases from $100 to $200 per month, then he can afford to see more movies, more concerts, or more of both, as shown by the change in his budget line in Figure 2(a). If Max were to devote *all* of his income to movies, he could now see 20 of them each month, instead of the 10 he was able to see before.

Devoting his entire income to concerts would enable him to attend 10, rather than 5. Moreover, for any number of concerts, he will be able to see more movies than before. For example, choosing 2 concerts would allow Max to see only 6 movies. Now, with a larger budget of $200, he can have 2 concerts and *16* movies.

Budget line The graphical representation of a budget constraint, showing the maximum affordable quantity of one good for given amounts of another good.

Relative price The price of one good relative to the price of another.

Figure 2: Change in Budget Line



Notice that the old and new budget lines in Figure 2(a) are parallel; that is, they have the same slope of - 2. We have changed Max's income but *not* the prices. Since the ratio P_c / P_m has not changed, the spending tradeoff between movies and concerts remains the same.

Thus, *an increase in income will shift the budget line upward (and rightward). A decrease in income will shift the budget line downward (and leftward). These shifts are parallel: Changes in income do not affect the budget line's slope.*

Changes in Price

Now let's go back to Max's original budget of $100 and explore what happens to the budget line when a price changes. Suppose the price of a movie falls from $10 to $5. The Figure 2(b) shows Max's old and new budget lines. When the price of a movie falls, the budget line rotates outward; that is, the vertical intercept moves higher. The reason is this: When a movie costs $10, Max could spend his entire $100 on them and see 10; now that they cost $5, he can see a maximum of 20. The horizontal intercept—representing how many concerts Max could see with his entire income—doesn't change at all, since there has been no change in the price of a concert. Notice that the new budget line is also *steeper* than the original one, with slope equal to $- P_c / P_m = -$ \$20/\$5 = - 4. Now, with movies costing \$5, the trade-off between movies and concerts is 4 to 1, instead of 2 to 1.

Panel (c) of Figure 2 illustrates another price change. This time, it's a fall in the price of a *concert* from \$20 to \$10. Once again, the budget line rotates, but now it is the horizontal intercept (concerts) that changes and the vertical intercept (movies) that remains fixed.

We could draw similar diagrams illustrating a *rise* in the price of a movie or a concert, but you should try to do this on your own. In each case, one of the budget line's intercepts will change, as well as its slope:

When the price of a good changes, the budget line rotates: Both its slope and one of its intercepts will change.

The budget constraint, as illustrated by the budget line, is one side of the story of consumer choice. It indicates the tradeoff consumers *are able to* make between one good and another. But just as important is the tradeoff that consumers *want to* make between one good and another, and this depends on consumers' *preferences,* the subject of the next section.

Preferences

How can we possibly speak systematically about people's preferences?

After all, people are different. They like different things. In spite of such wide differences in preferences, we can find some important common denominators—things that seem to be true for a wide variety of people. In our theory of consumer choice, we will focus on these common denominators.

Rationality

One common denominator—and a critical assumption behind consumer theory—is that people *have* preferences. More specifically, we assume that you can look at two alternatives and state either that you prefer one to the other or that you are entirely indifferent between the two—you value them equally.

Another common denominator is that preferences are *logically consistent,* or *transitive*. If, for example, you prefer a sports car to a jeep, and a jeep to a motorcycle, then we assume that you will also prefer a sports car to a motorcycle. When a consumer can make choices, and is logically consistent, we say that she has **rational preferences.**

Notice that rationality is a matter of how you make your choices, and not what choices you make. You can be rational and like apples better than oranges, or oranges better than apples. You can be rational even if you like chocolate-covered anchovies! What matters is that you make logically consistent choices, and most of us usually do.

More is Better

Another feature of preferences that virtually all of us share is this: We generally feel that *more is better*. Specifically, if we get more of some good or service, and nothing else is taken away from us, we will generally feel better off.

This condition seems to be satisfied for the vast majority of goods we all consume. Of course, there are exceptions. If you hate eggplant, then the more of it you have, the worse off you are. Similarly, a dieter who says, "Don't bring any ice cream into the house. I don't want to be tempted," also violates the assumption. The model of consumer choice in this chapter is designed for preferences that satisfy the "more is better" condition, and it would have to be modified to take account of exceptions like these.

So far, our characterization of consumer preferences has been rather minimal. We've assumed only that consumers are rational and that they prefer more rather than less of every good we're considering. But even this limited information allows us to say the following:

The consumer will always choose a point on the budget line, rather than a point below it.

To see why this is so, look again at Figure 1. Max would never choose point *G*, representing 2 concerts and 6 movies, since there are affordable points—on the budget line—that we know make him better off. For example, point *C* has the same number of concerts, but more movies, while point *D* has the same number of movies, but more concerts. "More is better" tells us that Max will prefer *C* or *D* to *G*, so we know *G* won't be chosen. Indeed, if we look at any point below the budget line, we can always find at least one point on the budget line that is preferred, as long as more is better.

Knowing that Max will always choose a point *on* his budget line is a start. But how does he find the *best* point on the line—the one that gives him the highest level of satisfaction?

This is where your *instructor's* preferences come in. There are two theories of consumer decision making, and they share much in common.

First, both assume that preferences are rational.

Second, both assume that the consumer would be better off with more of any good we're considering. This means the consumer will always choose a combination of goods *on*, rather than below, his budget line.

Finally, both theories come to the same general conclusions about consumer behavior. However, to *arrive* at those conclusions, each theory takes a different road.

Rational preferences

Preferences that satisfy two conditions:

(1) Any two alternatives can be compared, and one is preferred or else the two are valued equally, and

(2) The comparisons are logically consistent or transitive.

Consumer Decisions: Marginal Utility Approach

Economists assume that *any* decision maker—a consumer, the manager of a business firm, or officials in a government agency—tries to make the *best* out of any situation. Marginal utility theory treats consumers as striving to maximize their **utility**— an actual *quantitative* measure of well-being or satisfaction. Anything that makes the consumer better off is assumed to raise his utility. Anything that makes the consumer worse off will decrease his utility.

Utility and Marginal Utility

Figure 3 provides a graphical view of utility—in this case, the utility of a consumer named Lisa who likes ice cream cones. Look first at panel (a). On the horizontal axis, we'll measure the number of ice cream cones Lisa consumes each week. On the vertical axis, we'll measure the utility she derives from consuming each of them.

If Lisa values ice cream cones, her utility will increase as she acquires more of them, as it does in the figure. There we see that when she has one cone, she enjoys total utility of 30 "utils," and when she has two cones, her total utility grows to 50 utils, and so on. Throughout the figure, the total utility Lisa derives from consuming ice cream cones keeps rising as she gets to consume more and more of them.

But notice something interesting, and important: Although Lisa's utility increases every time she consumes more ice cream, the *additional* utility she derives from each *successive* cone gets smaller and smaller as she gets more cones. We call the *change in utility* derived from consuming an *additional unit* of a good the *marginal utility* of that additional unit.

***Marginal utility** is the change in utility an individual enjoys from consuming an additional unit of a good.*

Utility A quantitative measure of pleasure or satisfaction obtained from consuming goods and services.

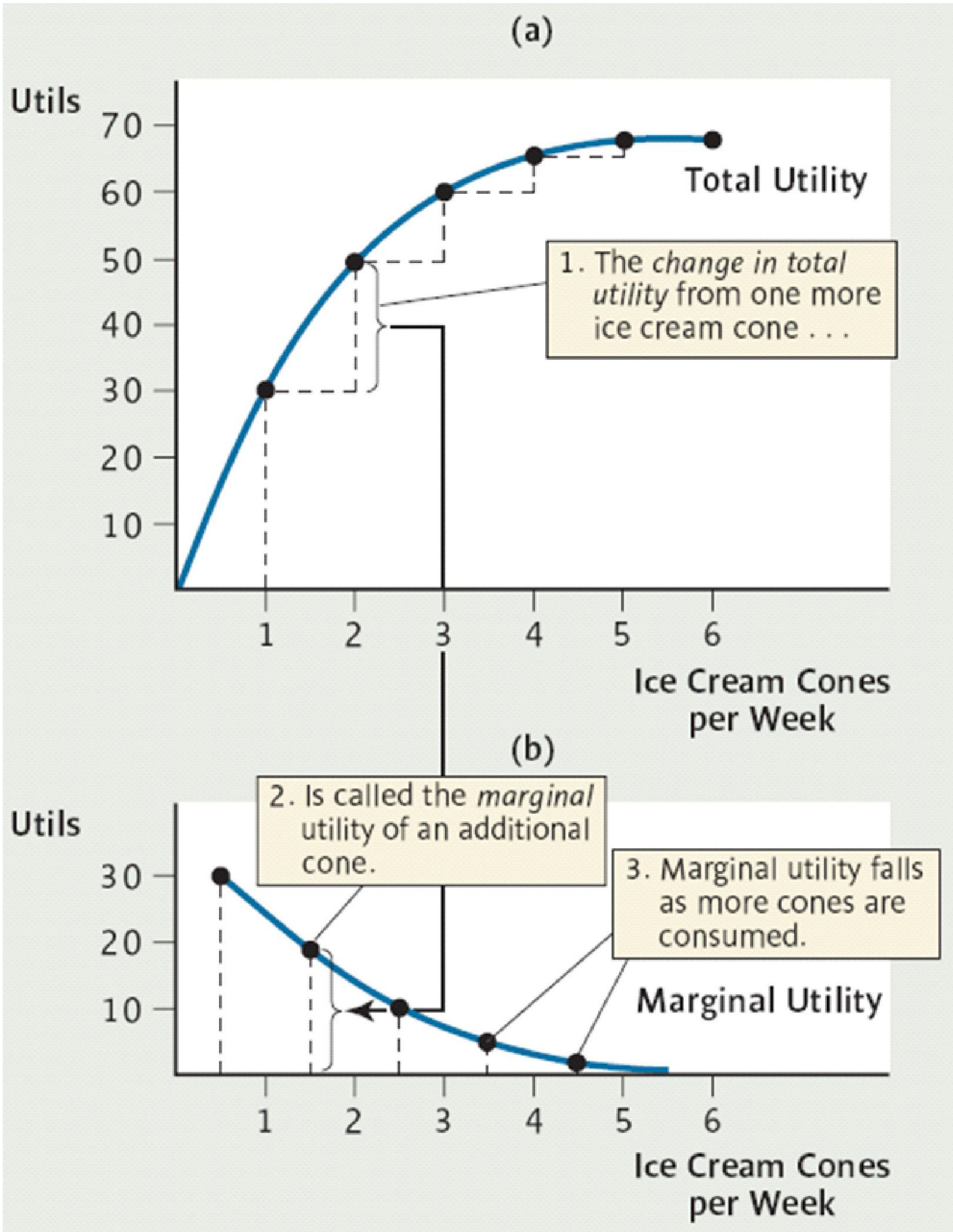
Marginal utility The change in total utility an individual obtains from consuming an additional unit of a good or service.

Table 3: Total and Marginal Utility

Lisa's Total and Marginal Utility from Consuming Ice Cream Cones

Number of Cones	Total Utility	Marginal Utility
0	0 utils	
		30 utils
1	30 utils	
		20 utils
2	50 utils	
		10 utils
3	60 utils	
		5 utils
4	65 utils	
		3 utils
5	68 utils	
		0 utils
6	68 utils	

Figure 3: Total and Marginal Utility



What we've observed about Lisa's utility can be restated this way: As she eats more and more ice cream cones in a given week, her *marginal utility* from another cone declines. We call this the **law of diminishing marginal utility**, which the great economist Alfred Marshall (1842–1924) defined this way:

The marginal utility of a thing to anyone diminishes with every increase in the amount of it he already has.

According to the law of diminishing marginal utility, when you consume your first unit of some good, like an ice cream cone, you derive some amount of utility. When you get your second cone that week, you enjoy greater satisfaction than when you only had one, but the extra satisfaction you derive from the second is likely to be smaller than the satisfaction you derived from the first. Adding the third cone to your weekly consumption will no doubt increase your utility further, but again the marginal utility you derive from that third cone is likely to be less than the marginal utility you derived from the second.

Law of diminishing marginal utility As consumption of a good or service increases, marginal utility decreases.

Figure 3 will again help us see what's going on. The table summarizes the information in the total utility graph. The first two columns show, respectively, the quantity of cones Lisa consumes each week and the total utility she receives each week from consuming them. The third column shows the *marginal* utility she receives from each successive cone she consumes per week. As you can see in the table, Lisa's total utility keeps increasing (marginal utility is always positive) as she consumes more cones (up to five per week), but the rate at which total utility increases gets smaller and smaller (her marginal utility diminishes) as her consumption increases.

Marginal utility is shown in panel (b) of Figure 3. Because marginal utility is the change in utility caused by a change in consumption from one level to another, we plot each marginal utility entry between the old and new consumption levels.

Notice the close relationship between the graph of total utility in panel (a) and the corresponding graph of marginal utility in panel (b). For every one-unit increment in Lisa's ice cream consumption her marginal utility is equal to the change in her total utility. Diminishing marginal utility is seen in both panels of the figure: in panel (b), by the downward sloping marginal utility curve, and in panel (a), by the positive but decreasing slope (flattening out) of the total utility curve.

One last thing about Figure 3: Because marginal utility diminishes for Lisa, by the time she has consumed a total of five cones per week, the marginal utility she derives from an additional cone has fallen all the way to zero. At this point, she is fully satiated with ice cream and gets no extra satisfaction or utility from eating any more of it in a typical week. Once this satiation point is reached, even if ice cream were free, Lisa would turn it down.

But remember from our earlier discussion that one of the assumptions we always make about preferences is that people prefer *more* rather than less of any good we're considering. So when we use marginal utility theory, we assume that marginal utility for every good is positive. For Lisa, it would mean she hasn't yet reached five ice cream cones per week.

Combing Budget Constraint and Preferences

The marginal utility someone gets from consuming more of a good tells us about his *preferences*. His budget constraint, by contrast, tells us only which combinations of goods he can *afford*. If we combine information about preferences (marginal utility values) with information about what is affordable (the budget constraint), we can develop a useful rule to guide us to an individual's utility maximizing choice.

To develop this rule, let's go back to Max and his choice between movies and concerts. Table 1 shows some utility numbers for movies and concerts. Notice that, as was the case with Lisa and her ice cream cones, Max's *total* utility rises with each concert or movie he consumes. But his *marginal* utility falls as more of either good is added. For example, a second concert per month adds 800 utils to Max's total utility. But the third concert adds only 600.

Remember that we want to find the best among the affordable combinations of these two goods. That means we'll need to consider the combinations of the two goods that are on Max's budget line, as shown in Figure 4 (reproduced from Figure 1). Figure 4 also shows some of the utility information in Table 1, but rearranged to more easily see Max's available choices. There's a lot going on in that table, so let's step through it carefully.

In column (1), the rows labeled A, B, C, etc. correspond to possible combinations on Max's budget line. For example, the row labeled C corresponds to point *C* on the budget line: 2 concerts and 6 movies per month. (The unlabeled rows in between would require Max to see and pay for a half of a concert, which we assume he can't do.)

Next, look at columns (2) and (5). Notice that the number of concerts runs from 0 to 5 as we travel down the rows. But the number of movies runs in the other direction, from 10 to 0. That's because—as we move along the budget line—attending more concerts means seeing fewer movies.

Now look at columns (3) and (6), which show the marginal utility from the "last" concert or movie. For example, in the row labeled C, the "last" concert Max sees is the second one, with marginal utility of 800 utils. In that same row, Max sees 6 movies, and the marginal utility from the last (sixth) movie is 200. These marginal utility numbers come from Table 3.

Table 3: Total and Marginal Utility of Concerts and Movies

Number of Concerts per Month	Total Utility	Marginal Utility	Number of Movies per Month	Total Utility	Marginal Utility
0	0		0	0	
1	1,000	1,000	1	600	600
2	1,800	800	2	1,050	450
3	2,400	600	3	1,400	350
4	2,800	400	4	1,700	300
5	3,050	250	5	1,950	250
6	3,250	200	6	2,150	200
			7	2,300	150
			8	2,400	100
			9	2,475	75
			10	2,535	60

Figure 4: Consumer's Decision Making

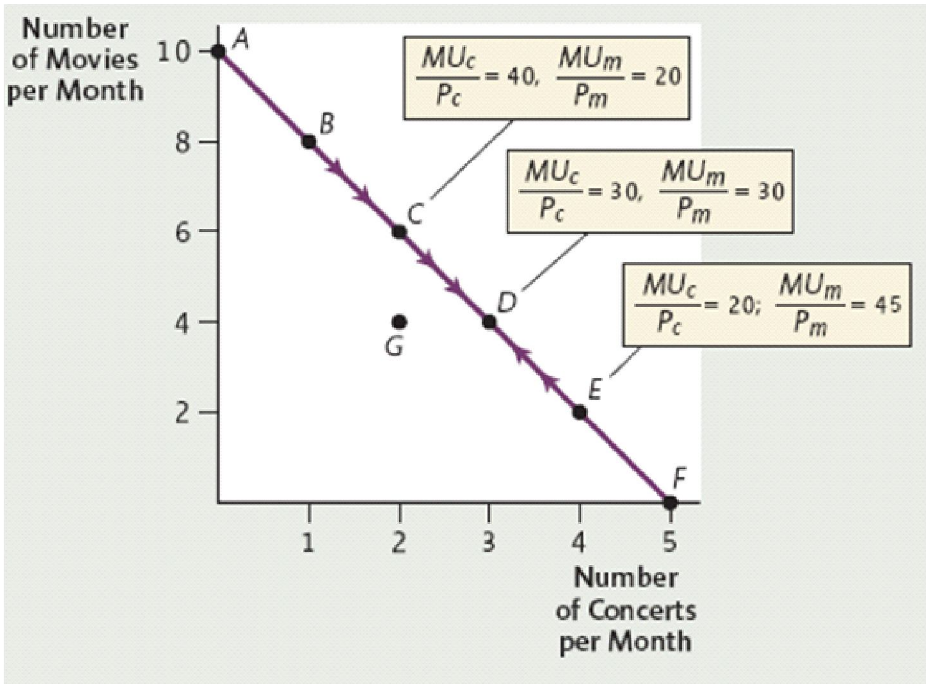


Table 4: Consumer's Decision Making

Budget = $100 per month						
	CONCERTS at $20 each			MOVIES at $10 each		
(1) Point on Budget Line	(2) Number of Concerts per Month	(3) Marginal Utility from Last Concert (MU_c)	(4) Marginal Utility per Dollar Spent on Last Concert $\left(\frac{MU_c}{P_c}\right)$	(5) Number of Movies per Month	(6) Marginal Utility from Last Movie (MU_m)	(7) Marginal Utility per Dollar Spent on Last Movie $\left(\frac{MU_m}{P_m}\right)$
A	0			10	60	6
				9	75	7.5
B	1	1000	50	8	100	10
				7	150	15
C	2	800	40	6	200	20
				5	250	25
D	**3**	**600**	**30**	**4**	**300**	**30**
				3	350	35
E	4	400	20	2	450	45
				1	600	60
F	5	250	12.5	0		

The budget line shows the maximum number of movies Max could attend for each number of concerts he attends. He would never choose an interior point like G *because there are affordable points—on the line—that make him better off. Max will choose the point on the budget line at which the marginal utilities per dollar spent on movies and concerts are equal. From the table, this occurs at point* D.

Now, look at column (4), which shows something new: the marginal utility *per dollar* spent on concerts. To get these numbers, we divide the marginal utility of the last concert (*MUc*) by the price of a concert, giving us *MUc*/*Pc*. This tells us the gain in utility Max gets *for each dollar he spends* on the last concert. For example, at point *C,* Max gains 800 utils from his second concert during the month, so his marginal utility *per dollar* spent on that concert is 800 utils/$20 = 40 utils per dollar.

Marginal utility per dollar, like marginal utility itself, declines as Max attends more concerts. After all, marginal utility itself decreases, and the price of a concert isn't changing. The last column gives us similar information for movies: the marginal utility per dollar spent on the last movie (*MUm*/*Pm*). As we travel *up* this column, Max attends more movies, and both marginal utility and marginal utility per dollar decline— once again, consistent with the law of diminishing marginal utility.

Now, Max's goal is to find the affordable combination of movies and concerts— the point on his budget line—that gives him the highest possible utility. As you are about to see, this will be the point at which *the marginal utility per dollar is the same for both goods*.

To see why, imagine that Max is searching along his budget line for the utility maximizing point, and he's currently considering point *B,* which represents 1 concert and 8 movies. Is he maximizing his utility? Let's see. Comparing the fourth and seventh entries in row *B* of the table, we see that Max's marginal utility per dollar spent on concerts is 50 utils, while his marginal utility per dollar spent on movies is only 10 utils. Since he gains more additional utility from each dollar spent on concerts than from each dollar spent on movies, he will have a net gain in utility if he shifts some of his dollars from movies to concerts. To do this, he must travel farther down his budget line.

Next suppose that, after shifting his spending from movies to concerts, Max arrives at point *C* on his budget line. What should he do then? At point *C,* Max's *MU* per dollar spent on concerts is 40 utils, while his *MU* per dollar spent on movies is 20 utils. Once again, he would gain utility by shifting from movies to concerts, traveling down his budget line once again.

Now suppose that Max arrives at point *D.* At this point, the *MU* per dollar spent on both movies and concerts is the same: 30 utils. There is no further gain from shifting spending from movies to concerts. At point *D,* Max has exploited all opportunities to make himself better off by moving down the budget line. He has maximized his utility.

What if Max had started at a point on his budget line *below* point *D?* Would he still end up at the same place? Yes, he would. Suppose Max finds himself at point *E,* with 4 concerts and 3 movies. Here, marginal utilities per dollar are 20 utils for concerts and 45 utils for movies. Now, Max could make himself better off by shifting spending away from concerts and toward movies. He will travel *up* the budget line, once again arriving at point *D,* where no further move will improve his well-being.

As you can see, it doesn't matter whether Max begins at a point on his budget line that's above point *D* or below it. Either way, if he keeps shifting spending toward the good with greater marginal utility per dollar, he will always end up at point *D.* And because marginal utility per dollar is the same for both goods at point *D,* there is nothing to gain by shifting spending in either direction.

What is true for Max and his choice between movies and concerts is true for *any* consumer and *any* two goods. We can generalize our result this way: For any two goods x and y, with prices Px and Py, whenever $MUx/Px = MUy/Py$, a consumer is made better off shifting spending away from y and toward x. When $MUy/Py = MUx/Px$, a consumer is made better off by shifting spending away from x and toward y. This leads us to an important conclusion: *A utility-maximizing consumer will choose the point on the budget line where marginal utility per dollar is the same for both goods ($MUx/Px = MUy/Py$). At that point, there is no further gain from reallocating expenditures in either direction.*

Conclusion

We can generalize this result to the more realistic situation of choosing combinations of more than two goods: different types of food, clothing, entertainment, transportation, and so on. No matter how many goods there are to choose from, when the consumer is doing as well as possible, it must be true that $MUx/Px = MUy/Py$ for *any* pair of goods x and *y*. If this condition is *not* satisfied, the consumer will be better off consuming more of one and less of the other good in the pair.

Questions

1. Explain Consumer Choice using the Budget Constraint.
2. Explain using suitable diagramme the Changes in Budget Line.
3. What do you understand by Consumer Preferences?
4. Examine the Consumer Decisions under the Marginal Utility Approach.
5. How Budget Constraint and Preferences are Combined to explain consumer decisions?

This lesson will focus on the following:

1. Consumer Decisions with Change in Income and Price
2. Consumer's Demand Curve
3. Income and Substitution Effects
4. Consumers in Markets

Consumer Decisions with Change in Income and Price

If every one of our decisions had to be made only once, life would be much easier. But that's not how life is. Just when you think you've figured out what to do, things change. In a market economy, as you've learned, prices can change for any number of reasons.

A consumer's income can change as well. These changes cause us to rethink our spending decisions: What maximized utility before the change is unlikely to maximize it afterward.

Changes in Income

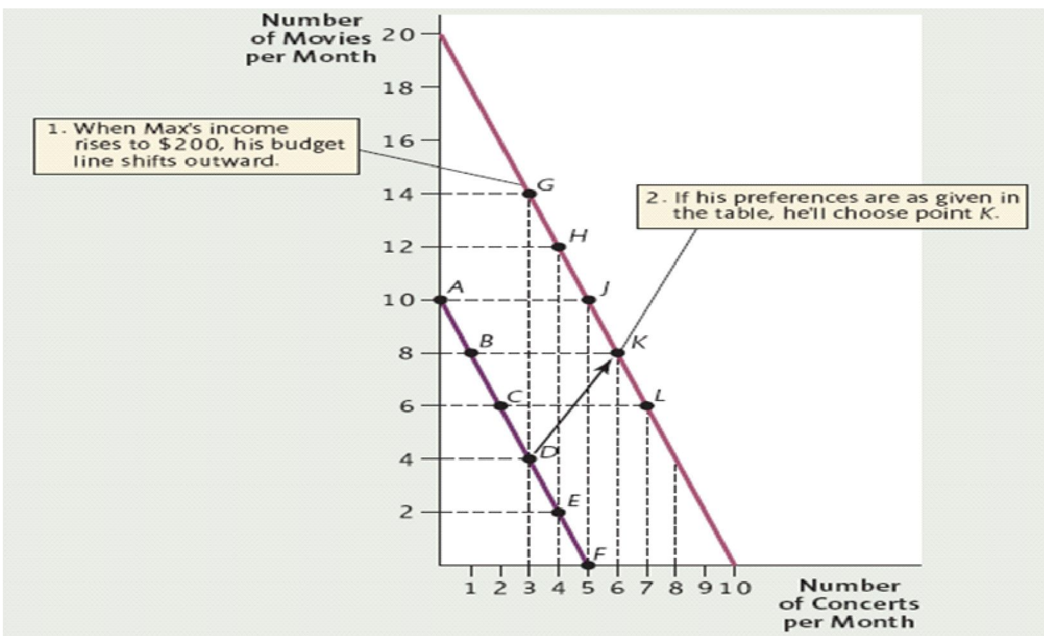
Figure 1 illustrates how an increase in income might affect Max's choice between movies and concerts. As before, we assume that movies cost $10 each, that concerts cost $20 each, and that these prices are not changing. Initially, Max has $100 in income to spend on the two goods, so his budget line is the line from point *A* to point *F*. As we've already seen, under these conditions,

Max would choose point *D* (three concerts and four movies) to maximize utility.

Table 1: Effect of Change in Income

Budget = $200 per month						
	CONCERTS at $20 each			MOVIES at $10 each		
(1) Point on New Budget Line	(2) Number of Concerts per Month	(3) Marginal Utility from Last Concert (MU_c)	(4) Marginal Utility per Dollar Spent on Last Concert $\left(\frac{MU_c}{P_c}\right)$	(5) Number of Movies per Month	(6) Marginal Utility from Last Movie (MU_m)	(7) Marginal Utility per Dollar Spent on Last Movie $\left(\frac{MU_m}{P_m}\right)$
G	3	600	30	14	38	3.8
				13	40	4
H	4	400	20	12	45	4.5
				11	50	5
J	5	250	12.5	10	60	6
				9	75	7.5
K	**6**	**200**	**10**	**8**	**100**	**10**
				7	150	15
L	7	180	9	6	200	20

Figure 1: Effects of an Increase in Income



Now suppose Max's monthly income (or at least the part he budgets for entertainment) increases to $200. Then his budget line will shift upward and outward in the figure. How will he respond? As always, he will search along his budget line until he finds the point where the marginal utility per dollar spent on both goods is the same. To help us find this point, Figure 5 includes some additional marginal utility values for combinations that weren't affordable before, but are now. For example, the *sixth* concert—which Max couldn't afford in Figure 1—is now assumed to have a marginal utility of 200 utils.

With the preferences described by these marginal utility numbers, Max will search along his budget line for the best choice. This will lead him directly to point *K,* enjoying 6 concerts and 8 movies per month. For this choice, *MU/P* is 10 utils per dollar for both goods, so total utility can't be increased any further by shifting dollars from one good to the other.

Now let's take a step back from these calculations and look at the figure itself. We see that an increase in income has changed Max's best choice from point *D* on the old budget constraint to point *K* on the new one. In moving from *D* to *K,* Max chooses to buy more concerts (6 rather than 3) and more movies (8 rather than 6). As discussed in Chapter 3, if an increase in income (with prices held constant) increases the quantity of a good demanded, the good is *normal*. For Max, with the marginal utility values we've assumed in Figure 5, both concerts and movies would be normal goods.

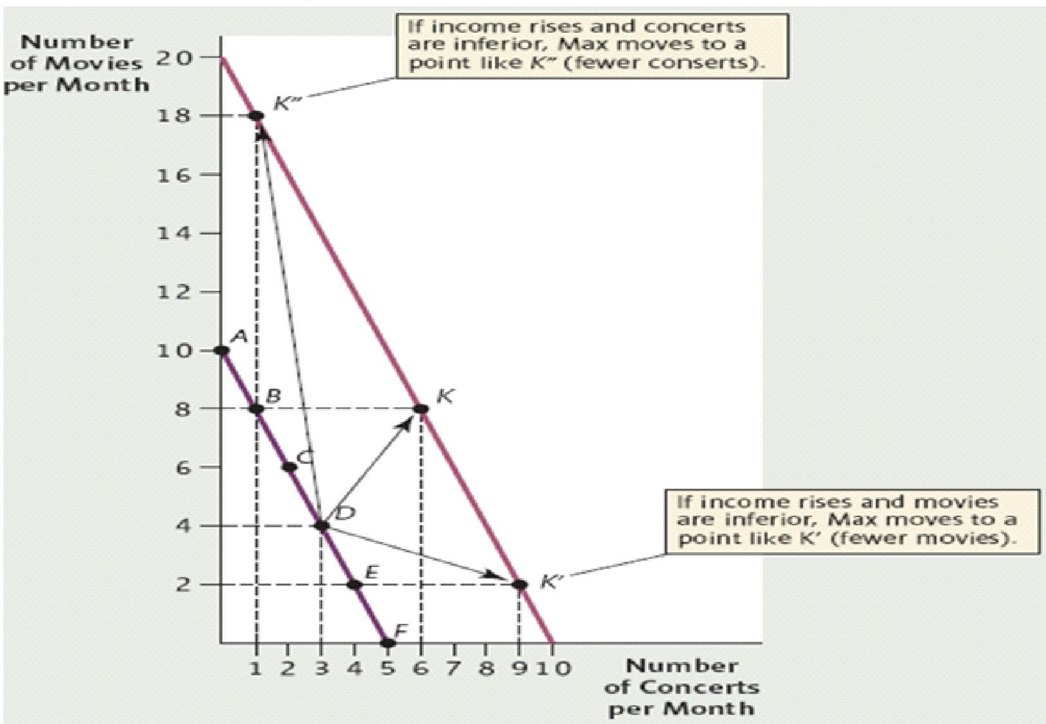
But this is not the only possible result. If Max's preferences (and marginal utility values) had been different than those in Figure 1, he might have chosen more of one good and *less* of the other.

These possibilities are illustrated in Figure 2. Our previous result—based on the marginal utility numbers in Figure 1—had Max moving from point *D* on his initial budget line to point *K* on the new, higher one when his income rose. But with different preferences (different marginal utility values), his marginal utilities per dollar might have been equal at a point like *K* ", with nine concerts and two movies. In this case, the increase in income would cause Max's consumption of concerts to increase (from 3 to 9), but his consumption of movies to *fall* (from 4 to 2). If so, movies would be an *inferior good* for Max—one for which demand decreases when income increases—while concerts would be a *normal* good.

Finally, let's consider another possible outcome for Max: point *K* " ". At this point, he attends more movies and fewer concerts compared to point *D.* If point *K* " "v is where Max's marginal utilities per dollar are equal after the increase in income, then *concerts* would be the inferior good, and movies would be normal.

A rise in income—with no change in prices—leads to a new quantity demanded for each good. Whether a particular good is normal (quantity demanded increases) or inferior (quantity demanded decreases) depends on the individual's preferences, as represented by the marginal utilities for each good, at each point along his budget line.

Figure 2: Normal and Inferior Goods



Changes in Price

Let's explore what happens to Max when the price of a concert decreases from $20 to $10, while his income and the price of a movie remain unchanged. The drop in the price of concerts rotates Max's budget line rightward, pivoting around its vertical intercept, as illustrated in the upper panel of Figure 3. What will Max do after his budget line rotates in this way? Again, he will select the combination of movies and concerts on his budget line that makes him as well off as possible. This will be the combination at which the marginal utility per dollar spent on both goods is the same.

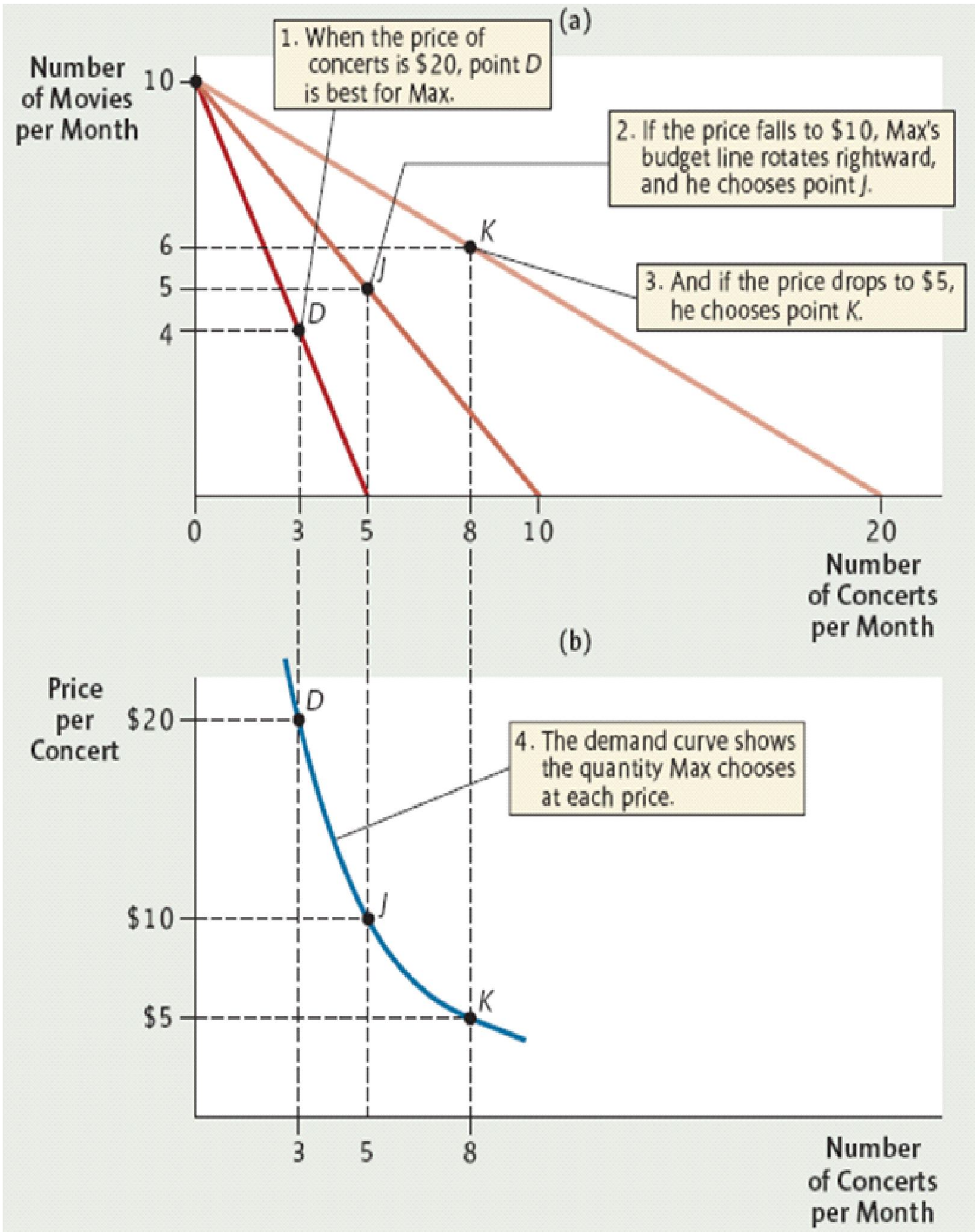
Once again, we've taken some of Max's marginal utility values from Figure 4 and added some additional numbers to construct the table in Figure 3. This table extends what we already knew about Max's preferences to cover the new, expanded possibilities.

With the preferences represented by these marginal utility numbers, Max will search along his budget line for the best choice. This will lead him directly to point *J,* where his quantities demanded are 5 concerts and 5 movies. Note that with each concert costing only $10 now, Max can afford this combination. Moreover, it satisfies our utility-maximizing rule: Marginal utility per dollar is 25 for both goods.

Table 2: Deriving Demand Curve

Max Choices with Budget of $100 per month						
	CONCERTS at $10 each			MOVIES at $10 each		
(1) Point on *New* Budget Line	(2) Number of Concerts per Month	(3) Marginal Utility from Last Concert (MU_c)	(4) Marginal Utility per Dollar Spent on Last Concert $\left(\frac{MU_c}{P_c}\right)$	(5) Number of Movies per Month	(6) Marginal Utility from Last Movie (MU_m)	(7) Marginal Utility per Dollar Spent on Last Movie $\left(\frac{MU_m}{P_m}\right)$
	3	600	60	7	150	15
	4	400	40	6	200	20
J	**5**	**250**	**25**	**5**	**250**	**25**
	6	180	18	4	300	30
	7	100	10	3	350	35

Figure 3: Deriving the Demand Curve



What if we dropped the price of concerts again—this time—to \$5? Then Max's budget line rotates further rightward, and he will once again find the utility maximizing point. In the figure, Max is shown choosing point *K,* attending 8 concerts and 6 movies.

Consumer's Demand Curve

Each time the price of concerts changes, so does the quantity of concerts Max will want to see. The lower panel of Figure 3 highlights this relationship by plotting the quantity of concerts demanded on the horizontal axis and the price of concerts on the vertical axis. For example, in both the upper and lower panels, point *D* tells us that when the price of concerts is \$20, Max will see three of them. When we connect points like *D, J,* and *K* in the lower panel, we get Max's demand curve, which shows *the quantity of a good he demands at each different price.* Notice that Max's demand curve for concerts slopes downward—a fall in the price of concerts increases the quantity demanded—showing that Max's responses to price changes obey the law of demand.

But if Max's preferences—and his marginal utility values—had been different, could his response to a price change have *violated* the law of demand? The answer is yes . . . and no. Yes, it is theoretically possible. (As a challenge, try identifying points on the three budget lines that would give Max an *upward-sloping* demand curve.) But no, it does not seem to happen in practice. To understand why and to gain other insights, the next section takes a deeper look into the effects of a price change on quantity demanded.

Income and Substitution Effects

Demand curve actually summarizes the impact of *two* separate effects of a price change on quantity demanded. These two effects sometimes work together, and sometimes oppose each other.

Substitution Effect

Suppose the price of a good falls. Then it becomes less expensive *relative to* other goods whose prices have not fallen. For example, when the price of concerts falls, so does its relative price (relative to movies). Max can now get more entertainment from his budget by substituting concerts in place of movies, so he will demand more concerts.

This impact of a price decrease is called a substitution effect: the consumer substitutes *toward* the good whose price has decreased, and away from other goods whose prices have remained unchanged.

*The **substitution effect** of a price change arises from a change in the relative price of a good, and it always moves quantity demanded in the opposite direction to the price change. When price decreases, the substitution effect works to increase quantity demanded; when price increases, the substitution effect works to decrease quantity demanded.*

Substitution effect As the price of a good falls, the consumer substitutes that good in place of other goods whose prices have not changed.

The substitution effect is a powerful force in the marketplace. For example, while the price of cell phone calls has fallen in recent years, the price of pay phone calls has remained more or less the same. This fall in the relative price of cell phone calls has caused consumers to substitute toward them and away from using regular pay phones. As a result, pay phones have almost completely disappeared in many areas of the country.

The substitution effect is also important from a theoretical perspective: It is the main factor responsible for the law of demand. Indeed, if the substitution effect were the *only* effect of a price change, the law of demand would be more than a law; it would be a logical necessity. But as we are about to see, a price change has another effect as well.

Income Effect

In Figure 3, when the price of concerts drops from \$20 to \$10, Max's budget line rotates rightward. Max now has a wider range of options than before: He can consume more concerts, more movies, or *more of both.* The price decline of *one* good increases his total purchasing power over *both* goods.

A price cut gives the consumer a gift, which is rather like an increase in *income*. Indeed, in an important sense, it *is* an increase in *available* income: Point *D* (3 concert and 4 movies) originally cost Max \$100, but after the decrease in the price of concerts, the same combination would cost him just (3 x \$10) + (4 x \$10) = \$70, leaving him with \$30 in *available income* to spend on more movies or concerts or both. This leads to the second effect of a change in price:

The ***income effect*** *of a price change arises from a change in purchasing power over both goods. A drop in price increases purchasing power, while a rise in price decreases purchasing power.*

How will a change in purchasing power influence the quantity of a good demanded? That depends. Recall that an increase in income will increase the demand for normal goods and decrease the demand for inferior goods. The same is true for the *income effect* of a price cut: It can work to either *increase* or *decrease* the quantity of a good demanded, depending on whether the good is normal or inferior. For example, if concerts are a normal good for Max, then the income effect of a price cut will lead him to consume more of them; if concerts are inferior, the income effect will lead him to consume fewer.

Combing Substitution and Income Effects

Now let's look again at the impact of a price change, considering the substitution and income effects together. A change in the price of a good changes both the relative price of the good (the substitution effect) and the overall purchasing power of the consumer (the income effect). The ultimate impact of the price change on quantity demanded will depend on *both* of these effects. For normal goods, these two effects work together to push quantity demanded in the same direction. But for inferior goods, the two effects oppose each other.

Normal Goods

When the price of a normal good falls, the substitution effect *increases* quantity demanded. The price drop will also increase the consumer's purchasing power and—for a normal good—*increase* quantity demanded even further. The opposite occurs when price increases: The substitution effect decreases quantity demanded, and the decline in purchasing power further decreases it.

Figure 4 summarizes how the substitution and income effects combine to make the price and quantity of a normal good move in opposite directions:

For normal goods, the substitution and income effects work together, causing quantity demanded to move in the opposite direction of the price. Normal goods, therefore, must always obey the law of demand.

Inferior Goods

Now let's see how a price change affects the demand for *inferior* goods. As an example, consider intercity bus service. For many consumers, this is an inferior good: with a higher income, these consumers would choose quicker and more comfortable alternatives (such as air or train travel), and therefore demand *less* bus service. Now, if the price of bus service falls, the substitution effect would work, as always, to *increase* quantity demanded. The price cut will also, as always, increase the consumer's purchasing power. But if bus service is inferior, the rise in purchasing power will *decrease* quantity demanded. Thus, we have two opposing effects: the substitution effect, increasing quantity demanded, and the income effect, decreasing quantity demanded. In theory, either of these effects could dominate the other, so the quantity demanded could move in either direction.

In practice, however, the substitution effect virtually always dominates for inferior goods.

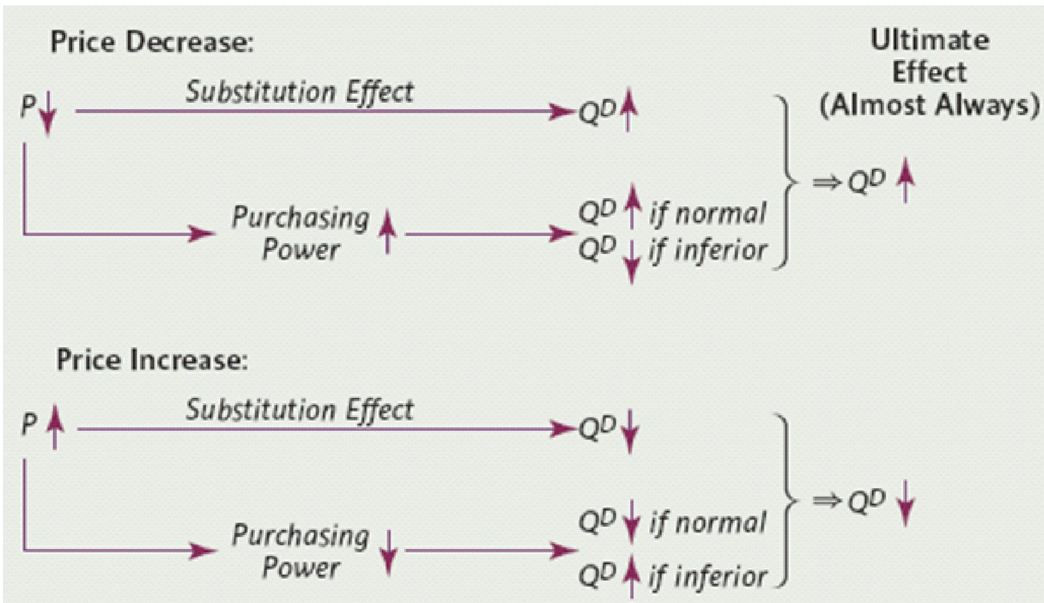
Why? Largely, because we consume such a wide variety of goods and services that a price cut in any one of them changes our purchasing power by only a small amount. For example, suppose you have an income of $20,000 per year, and you spend $500 per year on bus tickets. If the price of bus travel falls by, say, 20 percent, this would save you $100—like a gift of $100 in income. But $100 is only ½ percent of your income. Thus, a 20 percent fall in the price of bus travel would cause only a ½ percent rise in your purchasing power. Even if bus travel is, for you, an inferior good, we would expect only a tiny decrease in your quantity demanded when your purchasing power changes by such a small amount. Thus, the income effect should be very small. On the other hand, the *substitution* effect should be rather large: With bus travel now 20 percent cheaper, you will likely substitute away from other purchases and buy more bus travel.

For inferior goods, the substitution and income effects of a price change work against each other. The substitution effect moves quantity demanded in the opposite direction of the price, while the income effect moves it in the same direction as the price. But since the substitution effect virtually always dominates, consumption of inferior goods—like normal goods—will virtually always obey the law of demand.

Income effect

As the price of a good decreases, the consumer's purchasing power increases, causing a change in quantity demanded for the good.

Figure 4: Income and Substitution Effects



Consumers in Markets

Since the market demand curve tells us the quantity of a good demanded by *all* consumers in a market, it makes sense that we can derive it by adding up the individual demand curves of every consumer in that market.

Figure 5 illustrates how this can be done in a small local market for bottled water, where, for simplicity, we assume that there are only three consumers—Jerry, George, and Elaine. The first three diagrams show their individual demand curves.

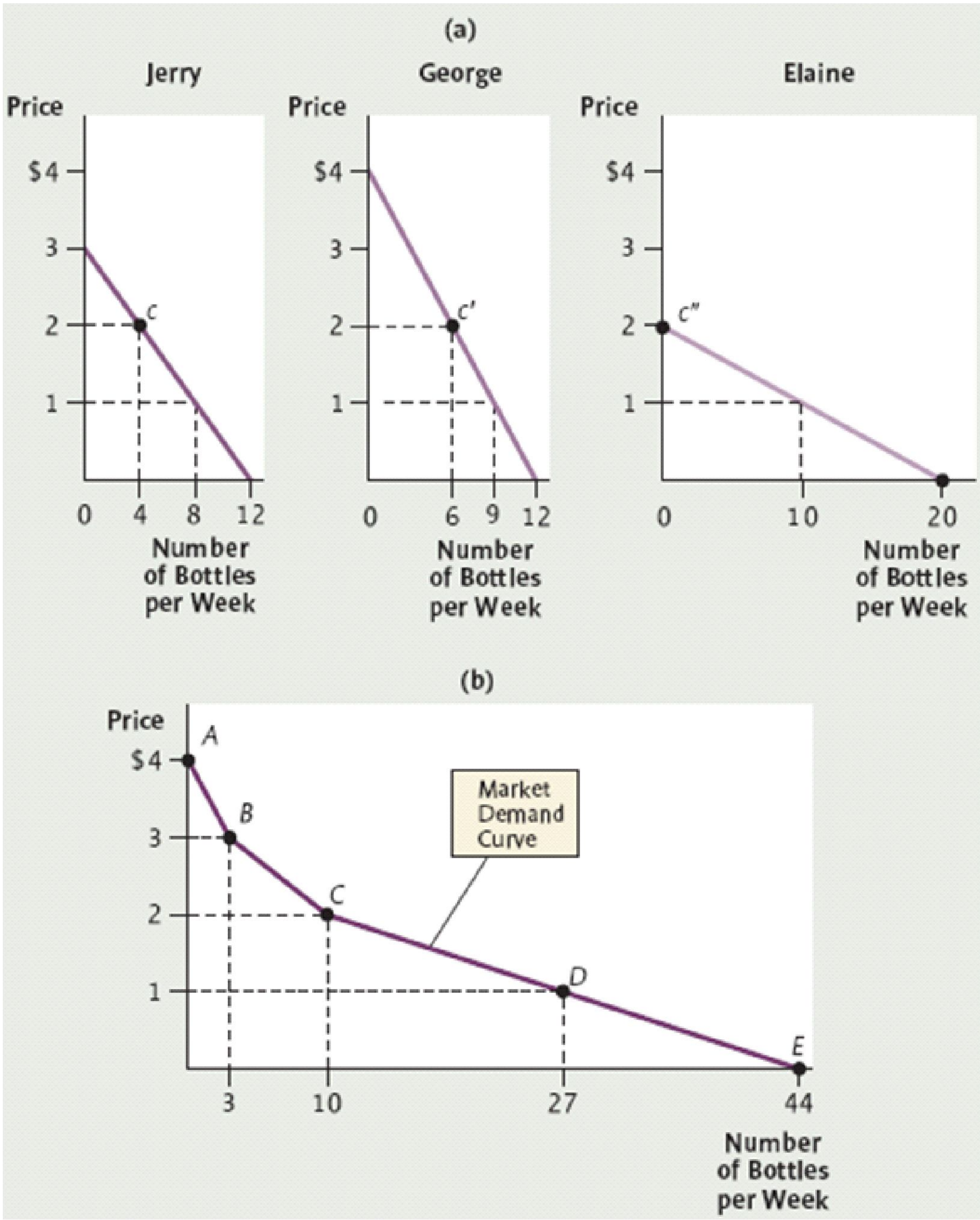
If the market price were, say, $2 per bottle, Jerry would buy 4 bottles each week (point *c*), George would buy 6 (point $c^{/}$), and Elaine would buy zero (point $c^{//}$. Thus, the market quantity demanded at a price of $2 would be 4 + 6 + 0 = 10, which is point *C* on the market demand curve. To obtain the entire market demand curve, we repeat this procedure at each different price, adding up the quantities demanded by each individual to obtain the total quantity demanded in the market. (Verify on your own that points *A, B, D,* and *E* have been obtained in the same way.) In effect, we obtain the market demand curve by summing horizontally across each of the individual demand curves:

The market demand curve is found by horizontally summing the individual demand curves of every consumer in the market.

Notice that as long as each individual's demand curve is downward sloping (and this will virtually always be the case), then the market demand curve will also be downward sloping. More directly, if a rise in price makes each consumer buy fewer units, then it will reduce the quantity bought by *all* consumers as well.

Indeed, the market demand curve can still obey the law of demand even when *some* individuals violate it. Thus, although we are already quite confident about the law of demand at the individual level, we can be even *more* confident at the market level. This is why we always draw market demand curves with a downward slope.

Figure 5: From Individual to Market Demand



The individual demand curves show how much bottled water will be demanded by Jerry, George, and Elaine at different prices. As the price falls, each demands more. The market demand curve in panel (b) is obtained by adding up the total quantity demanded by all market participants at different prices.

Conclusion

Graphically, the budget constraint is represented by the *budget line.* Only combinations on or below the budget line are affordable. An increase in income shifts the budget line outward. A change in the price of a good causes the budget line to rotate. Whenever the budget line shifts or rotates, the consumer moves to a point on the *new* budget line. The consumer will always choose the point that provides the greatest level of satisfaction or *utility,* and this will depend on the consumer's unique preferences. There are two alternative ways to represent consumer preferences, which lead to two different approaches to consumer decision making. The *marginal utility approach* is presented in the body of the chapter. In this approach, a utility-maximizing consumer chooses the combination of goods along her budget line at which the marginal utility per dollar spent is the same for all goods. When income or price changes, the consumer once again equates the marginal utility per dollar of both goods, resulting in a choice along the *new* budget line.

Questions

1. What would happen to the market demand curve for polyester suits, an inferior good, if consumers' incomes rose?

2. "If a good is inferior, a rise in its price will cause people to buy more of it, thus violating the law of demand." True or false? Explain.

3. Which of the following descriptions of consumer behaviour violates the assumption of *rational preferences*? Explain briefly.

   a. Dinesh is confused: He doesn't know whether he'd prefer to take a job now or go to college full-time.

   b. Aditi likes mustard on her pasta, in spite of the fact that pasta is not meant to be eaten with mustard.

c. Kavinder says, “I’d rather see an action movie than a romantic comedy, and I’d rather see a romantic comedy than a foreign film. But given the choice, I think I’d rather see a foreign film than an action movie.”

This lesson will focus on the following:

1. Indifference Curve Approach: Indifference Curve and Marginal Rate of Substitution
2. Properties of Indifference Curve
3. Indifference Map
4. Indifference Curve and Consumer Decision Making

Indifference Curve Approach

According to the ordinal utility theory, the benefit or satisfaction gained by consumers cannot be measured in quantitative form, but in terms of comparison to the consumption of other goods. Consumer behaviour in maximising satisfaction is depicted by indifference curve. This approach also stresses on comparison with consumption of other goods to determine the level of satisfaction.

Choice and Priority

Do you know the difference between the meaning of choice and priority? If you don't, let us differentiate choice and priority. Choice does not depend on price of goods or income. Choice might change but it is not based on the ability to pay. Even though you still cannot afford to own a big house or a luxury car, it does not mean that you cannot like both. Choice also shows our unlimited wants and needs because rational consumers will always choose something that is more compared to the least.

Priority is contrary to choice. Consumers may have their own choices or preferences, but out of the many choices that they have, consumers will have to choose only one which becomes their priority. Consumer will use the concept of priority when facing various choices and ability to pay. The simplest example will be the fees that you need to pay to take up this course.

Indifference Curve

We assume that an individual

(1) can compare any two options and decide which is best, or that both are equally attractive,

(2) makes choices that are logically consistent, and

(3) prefers more of every good to less.

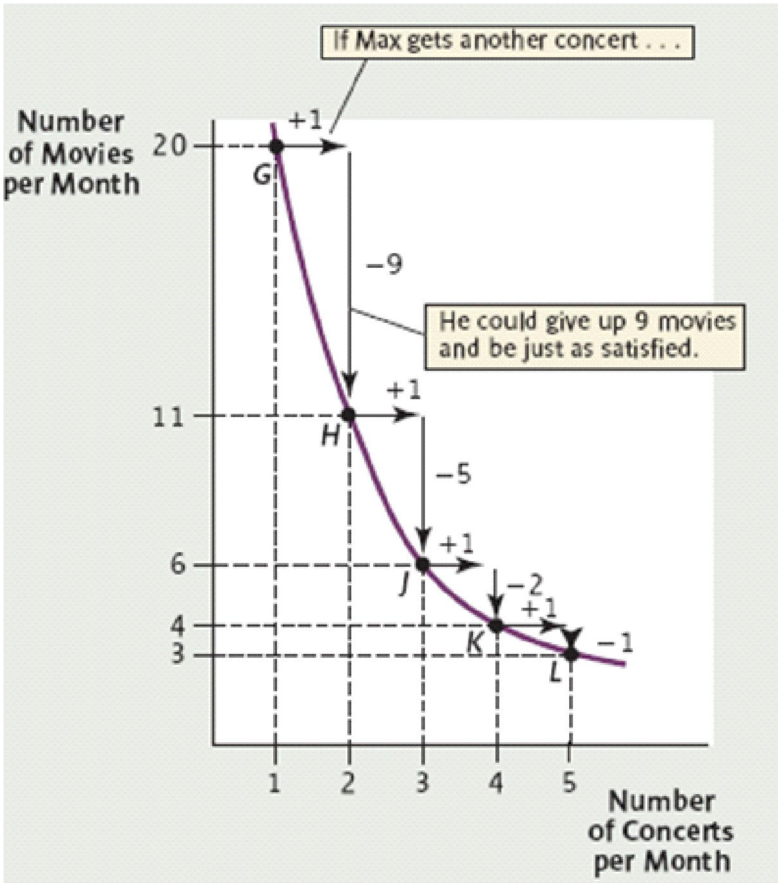
The first two assumptions are summarized as *rational preferences*; the third tells us that a consumer will always choose to be *on* her budget line, rather than below it. But now, we'll go a bit further.

An Indifference Curve

In Figure 1, look at point *G*, which represents 20 movies and 1 concert per month. Suppose we get Max to look at this figure with us, and ask him to imagine how satisfied he would be to have the combination at point *G*. Max thinks about it for a minute, then says, "Okay, I know how satisfied I would be."

Next, we say to Max, "Suppose you are at point *G* and we give you *another* concert each month, for a total of 2. That would make you even *more* satisfied, right?" Since Max likes concerts, he nods his head. But then we ask, "After giving you this additional concert, how many movies could we *take away* from you and leave you no better or worse off than you were originally, at point *G?*"

Figure 1: Indifference Curve



Starting at point *G,* adding 1 more concert and taking away 9 movies puts us at point *H.* But let's keep going. Now we get Max to imagine that he's at point *H,* and we ask him the same question, and this time he answers, "I could trade 5 movies for 1 more concert and be equally well off." Now Max is telling us that he is indifferent between point *H* and *J,* since *J* gives him 1 more concert and 5 fewer movies than point *H.*

So far, we know Max is indifferent between point *G* and point *H,* and between point *H* and point *J.* So long as he is rational, he must be entirely indifferent among all three points—*G, H,* and *J*—since all three give him the same level of satisfaction. By continuing in this way, we can trace out a set of points that—as far as Max is concerned—are equally satisfying. When we connect these points with a curved line, we get one of Max's *indifference curves*.

*An **indifference curve** represents all combinations of two goods that make the consumer equally well off.*

Notice that the indifference curve in Figure 1 slopes downward. This follows from our assumption about preferences that "more is better." Every time we give Max another concert, we make him better off. In order to find another point on his indifference curve, we must make him worse off by the same amount, *taking away* some movies.

Marginal Rate of Substitution (MRS)

When we move along an indifference curve, from one point to another, we discover the maximum number of movies that Max would *willingly trade* for one more concert. For example, going from point *G* to point *H,* Max gives up 9 movies for 1 concert and remains indifferent. Therefore, from point *G,* if he gave up *10* movies for 1 concert, he'd be *worse off,* and he would not willingly make that trade. Thus, at point *G,* the *greatest* number of movies he'd willingly sacrifice for another concert would be 9.

This notion of "willingness to trade," as you'll soon see, has an important role to play in our model of consumer decision making. And there's a technical term for it: the *marginal rate of substitution of movies for concerts.*

More generally, when the quantity of good *y* is measured on the vertical axis, and the quantity of good *x* is measured on the horizontal axis, *the **marginal rate of substitution of good y for good*** **x (MRSy,x)** *along any segment of an indifference curve is the maximum rate at which a consumer would willingly trade units of* y *for units of* x.

For example, say we move along the indifference curve from point *G* (20 movies and 1 concert) all the way to point *L* (3 movies and 5 concerts). Since Max ends up on the same indifference curve, he'd be willing to make that move. That is, he would willingly give up 17 movies to get 4 more concerts, so his MRS would be 17 ÷ 4 = 4¼. We could say that Max is willing to give up "4¼ movies per concert." If we were to draw a straight line (not shown) between points *G* and *L*, the slope of that line would be - 4¼, giving us a graphical representation of the MRS for that segment of the indifference curve.

However, the value of the MRS will depend on the *size* of the move we make. Suppose we start at the same point, *G*, but make a smaller movement this time—to point *J*. For this smaller move, Max is willing to give up 14 movies to get 2 more concerts, so his MRS would be $14 \div 2 = 7$ (i.e., he is willing to give up 7 movies per concert). The MRS is now the slope (without the minus sign) of the straight line drawn between point *G* and point *J*.

MRS at a Point

In consumer theory, we are often interested in very small changes: the rate at which the consumer is willing to trade one good for a "tiny bit more" of another good.

We imagine that the consumer makes a series of very tiny movements that, in total, account for the larger change we ultimately observe. Many goods (gasoline, electricity, or ground beef from a butcher) can, in fact, be consumed and traded off in arbitrarily small increments. As these increments shrink, the segment of the indifference curve we are considering shrinks as well. Eventually, the segment becomes so small that— for all intents and purposes—we are looking at a single point on the curve.

Until now, we've defined the MRS using a *segment* of the indifference curve. And you've seen that the MRS depends on the size of segment we are considering. Can we use the MRS as a measure of willingness to trade when the segment shrinks to a *point*, such as point *H* in Figure 1? Indeed we can, using the *slope* of the indifference curve itself *at point H*. To obtain that slope, we'd draw a straight line *tangent* to the indifference curve at point *H*, and use the slope of the tangent line.

The MRS at any point on the indifference curve is equal to the (absolute value of) the slope of the curve at that point. When measured at a point, the MRSy,x tells us the maximum rate at which a consumer would willingly trade good Y for a tiny bit more of good X. In our example, the "MRS at point *H*" in Figure 1 is the rate at which Max would trade movies for concerts when we offer him just *a tiny bit more concert* (say, one-tenth or one-hundredth of a concert) beyond the one he already has. Of course, it is not possible to trade fractions of movies for fractions of concerts.

The smallest trade Max could *actually* make would involve a whole concert. So for Max, as we shrink the segment we are considering, we could not realistically shrink it all the way to a single point. In that case, the slope at point *H* gives us only an *approximation* of Max's willingness to trade in the smallest increments possible: whole units. In the rest of our analysis, we'll use the slope at a point to approximate the MRS for the smallest movements that Max can realistically make.

How MRS Changes along Indifference Curve?

In Figure 1, notice that as we move downward and rightward along the indifference curve, it gets flatter—the absolute value of its slope decreases. Another way of saying this is: As we move down an indifference curve, the *MRS* (the number of movies Max would be willing trade for another concert) gets smaller and smaller.

To see why the *MRS* behaves this way, consider point *G,* high on Max's indifference curve. At this point, Max is seeing a lot of movies and relatively few concerts compared to points lower down, such as *J, K,* or *L.* With so few concerts, he'd value another one very highly. And with so many movies, each one he gives up doesn't harm him much. So, at a point like *G,* he'd be willing to trade a large number of movies for even one more concert. Using "*m*" for movies and "*c*" for concerts, his $MRS_{m,c}$ is relatively large. Since the *MRS* is the absolute value of the indifference curve's slope, the curve is relatively steep at point *G.*

But as we continue traveling down his indifference curve, from *G* to *H* to *J* and so on, movies become scarcer for Max, so each one given up hurts him a bit more. At the same time, he's attending more and more concerts, so adding another one doesn't benefit him as much as before. At a point like *K,* then, Max is more reluctant to trade movies for concerts. To get another concert, he'd willingly trade fewer movies at point *K* than at point *G.* So at point *K,* the *MRS* is relatively small and the curve is relatively flat.

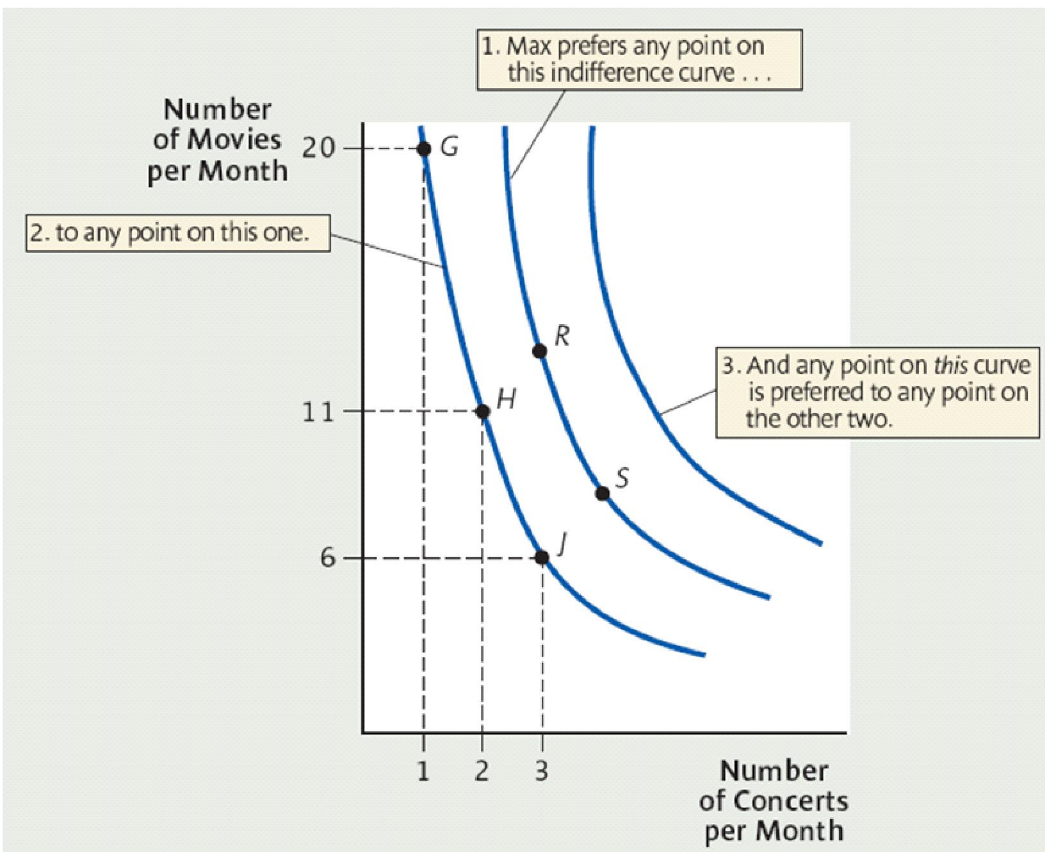
Indifference Map

To trace out the indifference curve in Figure 1, we began at a specific point—point *G.* Figure 2 reproduces that same indifference curve through *G, H,* and *J.* But now consider the new point *R,* which involves more movies *and* more concerts than at point *H.* We know that point *R* is preferred to point *H* ("more is better"), so it is not on the indifference curve that goes through *H.*

However, we can use the same procedure we used earlier to find a *new* indifference curve, connecting all points indifferent to point *R.* Indeed, we can repeat this procedure for any initial starting point we might choose, tracing out dozens or even hundreds of Max's indifference curves, as many as we'd like.

The result would be an **indifference map**, a set of indifference curves that describe Max's preferences, like the three curves in Figure 2. We know that Max would always prefer any point on a higher indifference curve to any point on a lower one. For example, consider the points *H* and *S*. *S* represents more concerts but fewer movies than *H*. But Max's indifference map tells us that he *must* prefer *S* to *H*. Why? We know that he prefers *R* to *H*, since *R* has more of both goods. We also know that Max is indifferent between *R* and *S*, since they are on the same indifference curve. Since he is indifferent between *S* and *R*, but prefers *R* to *H*, then he must also prefer *S* to *H*.

Figure 2: Indifference Map



The same technique could be used to show that *any point on a higher indifference curve is preferred to any point on a lower one.* Thus, Max's indifference map tells us how he ranks all possible alternatives. An indifference map gives us a complete characterization of someone's preferences: It allows us to look at any two points and—just by seeing which indifference curves they are on—immediately know which, if either, is preferred.

Consumer Decision Making

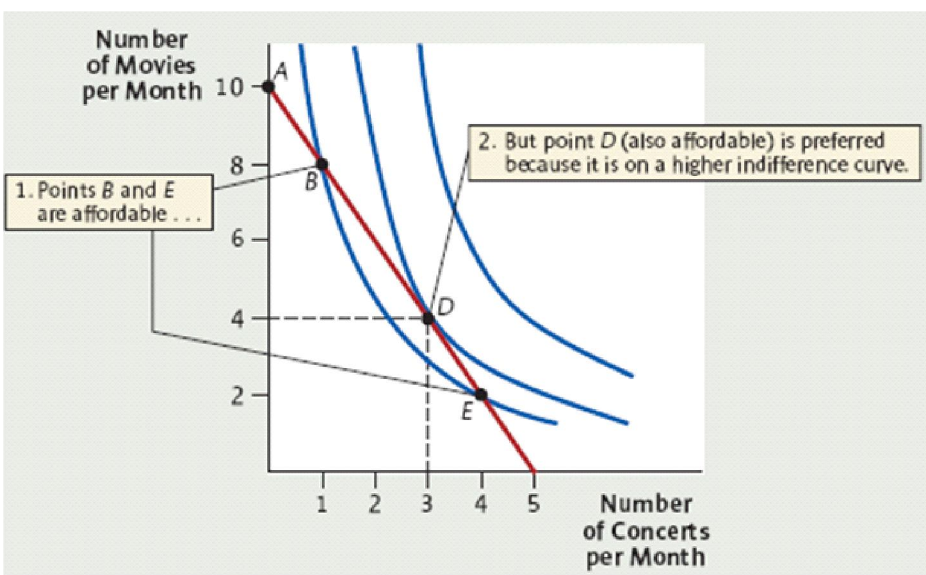
Now we can combine everything you've learned about budget lines in the chapter, and what you've learned about indifference curves in this appendix, to determine the combination of movies and concerts that Max should choose. Figure A.3 adds Max's budget line to his indifference map. In drawing the budget line, we assume that Max has a monthly entertainment budget of $100, and that a concert costs $20 and a movie costs $10.

We assume that Max—like any consumer— wants to make himself as well off as possible. His optimal combination of movies and concerts will satisfy two criteria:

(1) it will be a point *on* his budget line; and

(2) it will lie on the highest indifference curve possible.

Figure 3: Consumer Decision Making with Indifference Curves



Max can find this point by traveling down his budget line from *A*. As he does so, he will pass through a variety of indifference curves. (To see this clearly, you can pencil in additional indifference curves *between* the ones drawn in the figure.)

At first, each indifference curve is higher than the one before, until he reaches the highest curve possible. This occurs at point *D,* where Max sees three concerts and four movies each month. Any further moves down the budget line will put him on lower indifference curves, so these moves would make him worse off. Point *D* is Max's optimal choice.

Notice something interesting about point *D.* First, it occurs where the indifference curve and the budget line are tangent—where they touch but don't cross. As you can see in the diagram, when an indifference curve actually crosses the budget line, we can always find some other point on the budget line that lies on a higher indifference curve.

Second, at point *D,* the slope of the indifference curve is the same as the slope of the budget line. Does this make sense? Yes, when you think about it this way: The absolute value of the indifference curve's slope—the *MRS*—tells us the rate at which Max would *willingly* trade movies for concerts. The slope of the budget line, by contrast, tells us the rate at which Max is *actually able* to trade movies for concerts.

If there's any difference between the rate at which Max is *willing* to trade one good for another and the rate at which he is *able* to trade, he can always make himself better off by moving to another point on the budget line.

For example, suppose Max were at point *B* in Figure 3. The indifference there is steeper than his budget line. In fact, the indifference curve appears to have a slope of about - 4, so Max's MRS there is about 4; he'd *willingly* give up 4 movies for 1 more concert. But his budget line—as you learned earlier in the chapter—has a slope of - 2. So according to his budget line, he is *able* to trade just 2 movies for each concert. If trading 4 movies for a concert would leave him indifferent, then trading just 2 movies for a concert must make him better off. We conclude that when Max's indifference curve is steeper than his budget line, he should spend more on concerts and less on movies.

Using similar reasoning, convince yourself that Max should make the opposite move—spending less on concerts and more on movies—if his indifference curve is *flatter* than his budget line, as it is at point *E.* Only when the indifference curve and the budget line have the same slope—when they touch but do not cross—is Max as well off as possible. This is the point where the indifference curve is *tangent* to the budget line. When Max, or any other consumer, strives to be as well off as possible, he will follow this rule:

The optimal combination of goods for a consumer is the point on the budget line where an indifference curve is tangent to the budget line.

We can also express this decision-making rule in terms of the *MRS* and the prices of two goods. Recall that the slope of the budget line is = - Px/Py, so the absolute value of the budget line's slope is Px/Py. As you've just learned, the absolute value of the slope of an indifference curve is $MRSy,x$. This allows us to state the decision-making rule as follows:

The optimal combination of two goods x and y is that combination on the budget line for which $MRSy,x = Px/Py$.

If this condition is not met, there will be a difference between the rate at which a consumer is *willing* to trade good *y* for good *x,* and the rate at which he is *able* to trade them. This would leave the consumer with an opportunity to make himself better off.

Deriving Demand Curves

We use consumer theory to show by how much the quantity demanded of a good falls as its price rises. An individual chooses an optimal bundle of goods by picking the point on the highest indifference curve that touches the budget line.

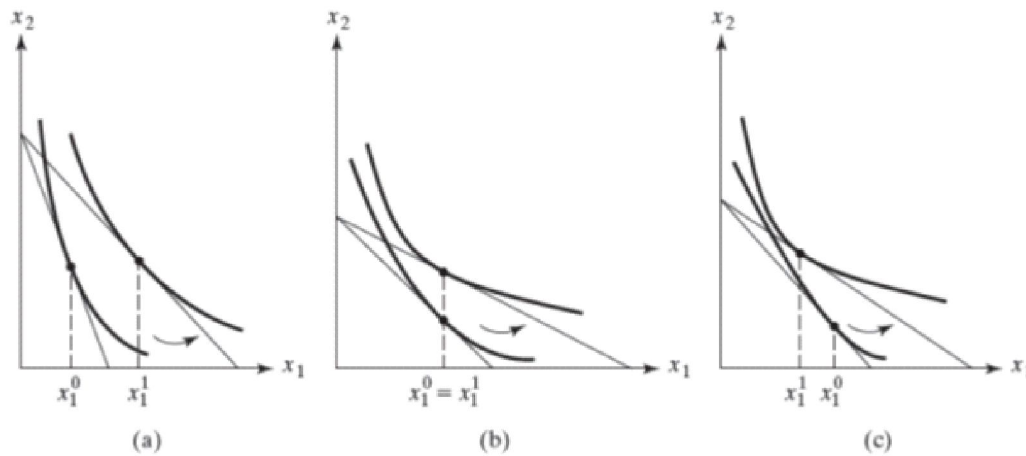
When a price changes, the budget constraint the consumer faces shifts, so the consumer chooses a new optimal bundle. By varying one price and holding other prices and income constant, we determine how the quantity demanded changes as the price changes, which is the information we need to draw the demand curve. After deriving an individual's demand curve, we show the relationship between consumer tastes and the shape of the demand curve, which is summarized by the elasticity of demand.

Income and Substitution Effects

An important question in our model of consumer behaviour concerns the response we should expect in quantity demanded when price changes. Ordinarily, we tend to think a consumer will buy more of a good when its price declines and less when its price increases, other things being equal. That this need not always be the case is illustrated in Figure 4.

In each panel, a utility-maximising consumer with strictly monotonic, convex preferences faces market-determined prices. In Figure 4(a), a decrease in the price of good 1 causes the quantity of good 1 bought to increase, as we would usually expect. By contrast, in Figure 4(b), a decrease in price causes no change in the amount of good 1 bought, whereas in Figure 4(c), a decrease in price causes an absolute *decrease* in the amount of good 1 bought. Each of these cases is fully consistent with our model. What, then – if anything – does the theory predict about how someone's demand behaviour responds to changes in (relative) prices?

Figure 4: Response of Quantity Demanded to a Change in Price



Let us approach it intuitively first. When the price of a good declines, there are at least two conceptually separate reasons why we expect some change in the quantity demanded. First, that good becomes relatively cheaper compared to other goods. Because all goods are desirable, even if the consumer's total command over goods were unchanged, we would expect him to substitute the relatively cheaper good for the now relatively more expensive ones. This is the **substitution effect** (*SE*). At the same time, however, whenever a price changes, the consumer's command over goods in general is *not* unchanged. When the price of any one good declines, the consumer's total command over all goods is effectively increased, allowing him to change his purchases of *all* goods in any way he sees fit. The effect on quantity demanded of this generalised increase in purchasing power is called the **income effect** (*IE*).

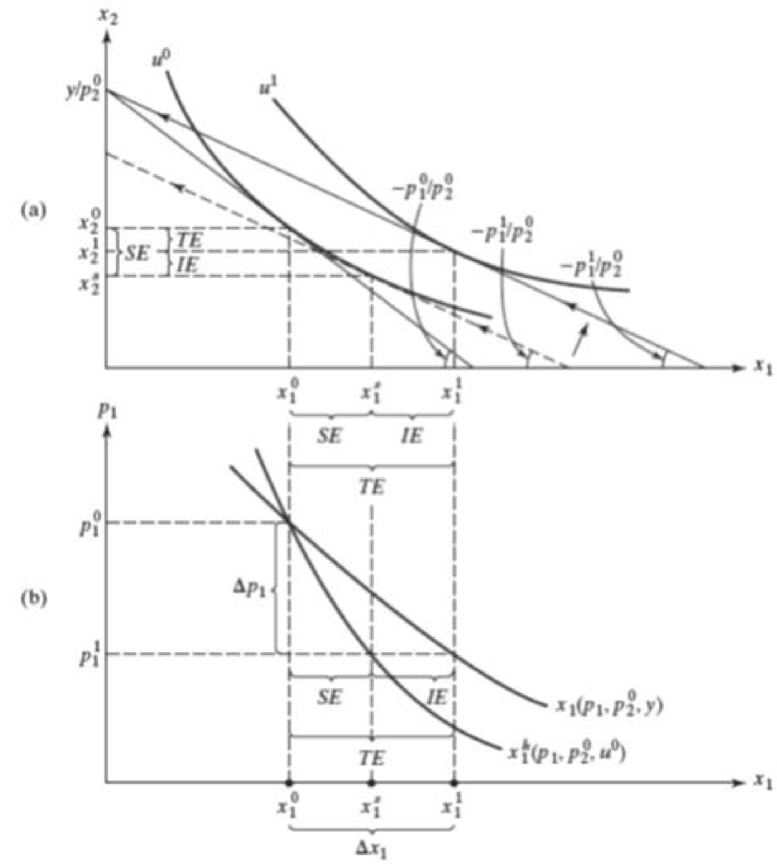
Although intuition tells us we can in some sense decompose the total effect (*TE*) of a price change into these two separate conceptual categories, we will have to be a great deal more precise if these ideas are to be of any analytical use. Different ways to formalise the intuition of the income and substitution effects have been proposed. We shall follow that proposed by Hicks (1939).

The Hicksian decomposition of the total effect of a price change starts with the observation that the consumer achieves some level of utility at the original prices before any change has occurred. The formalisation given to the intuitive notion of the substitution effect is the following: the substitution effect is that (hypothetical) change in consumption that *would* occur if relative prices were to change to their new levels but the maximum utility the consumer can achieve were kept the same as before the price change. The income effect is then defined as whatever is left of the total effect after the substitution effect.

Notice that because the income effect is defined as a residual, the total effect is always completely explained by the sum of the substitution and the income effect. At first, this might seem a strange way to do things, but a glance at Figure 5 should convince you of at least two things: its reasonable correspondence to the intuitive concepts of the income and substitution effects, and its analytical ingenuity.

Look first at Figure 5(a), and suppose the consumer originally faces prices p^0_1 and p^0_2 and has income y. He originally buys quantities x^0_1 and x^0_2 and achieves utility level u^0.

Figure 5: Hicksian Decomposition of a Price Change



Suppose the price of good 1 falls to $p^1_1 < p^0_1$ and that the total effect of this price change on good 1 consumption is an increase to x^0_1, and the total effect on good 2 is a decrease to x^1_2. To apply the Hicksian decomposition, we first perform the hypothetical experiment of allowing the price of good 1 to fall to the new level p^1_1 *while holding the consumer to the original u^0 level indifference curve*. It is as if we allowed the consumer to face the new relative prices but reduced his income so that he faced the dashed hypothetical budget constraint and asked him to maximise against it. Under these circumstances, the consumer would *increase* his consumption of good 1 – the now relatively cheaper good – from x^0_1 to x^s_1, and would *decrease* his consumption of good 2 – the now relatively more expensive good – from x^0_2 to x^s_2. These hypothetical changes in consumption are the Hicksian substitution effects on good 1 and good 2, and we regard them as due 'purely' to the change in relative prices with no change whatsoever in the well-being of the consumer.

Look now at what is left of the total effect to explain. After hypothetical changes from x^0_2 and x^0_2 to x^s_1 and x^s_2, the changes from x^s_1 and x^s_2 to x^1_1 and x^1_2 remain to be explained. Notice, however, that these are precisely the consumption changes that would occur if, at the new prices and the original level of utility $u0$, the consumer were given an *increase in real income* shifting his budget constraint from the hypothetical dashed one out to the final, post price-change line tangent to u^1. It is in this sense that the Hicksian income effect captures the change in consumption due 'purely' to the income-like change that accompanies a price change.

Look now at Figure 5(b), which ignores what is happening with good 2 and focuses exclusively on good 1. Clearly, (p^0_1, x^0_1) and (p^1_1, x^1_1) are points on the Marshallian demand curve for good 1. Similarly, (p^0_1, x^0_1) and (p^1_1, x^s_1) are points on the Hicksian demand curve for good 1, relative to the original utility level u^0. We can see that the Hicksian demand curve picks up *precisely* the pure *Hicksian* substitution effect of an own-price change.

Conclusion

The Marshallian demand curve picks up the total effect of an own-price change. The two diverge from one another precisely because of, and in an amount equal to, the Hicksian income effect of an own-price change. The Hicksian decomposition gives us a neat analytical way to isolate the two distinct forces working to change demand behaviour following a price change. We can take these same ideas and express them much more precisely, much more generally, and in a form that will prove more analytically useful. The relationships between total effect, substitution effect, and income effect are summarised in the *Slutsky equation*. The Slutsky equation is sometimes called the 'Fundamental Equation of Demand Theory', so what follows merits thinking about rather carefully.

Questions

1. Define Indifference Curve and Marginal Rate of Substitution.
2. Explain the Properties of Indifference Curve and draw Indifference Map.
3. Using Indifference Curve, explain Consumer Decision Making.

This lesson will focus on the following:

1. Change in Income and Consumer Decisions
2. Deriving Demand Curve with Indifference Curves
3. Consumer Decision and Revealed Preference Approach
4. Consumer Surplus

What Happens When Things Change?

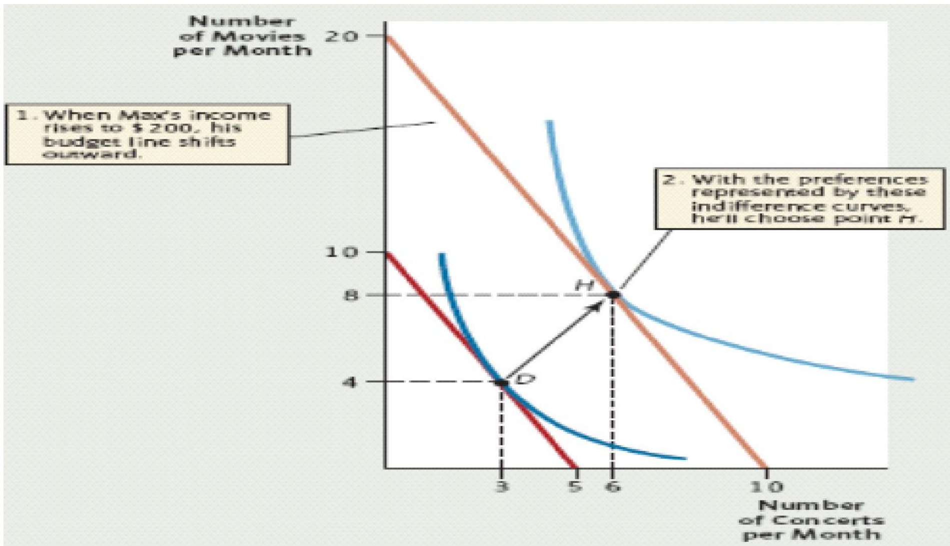
So far, as we've examined Max's search for the best combination of movies and concerts, we've assumed that Max's income, and the prices of each good, have remained unchanged. But in the real world, an individual's income, and the prices of the things they buy, *can* change. How would these changes affect his choice?

Change in Income and Consumer Decisions

Initially, Max has \$100 to spend on the two goods, so his budget line is the lower line through point *D.* As we've already seen, under these conditions, the optimal combination for Max is point *D* (3 concerts and 4 movies).

Now suppose Max's income increases to \$200. Then his budget line will shift upward and rightward in the figure. How will he respond? As always, he will search along his budget line until he arrives at the highest possible indifference curve, which will be tangent to the budget line at that point.

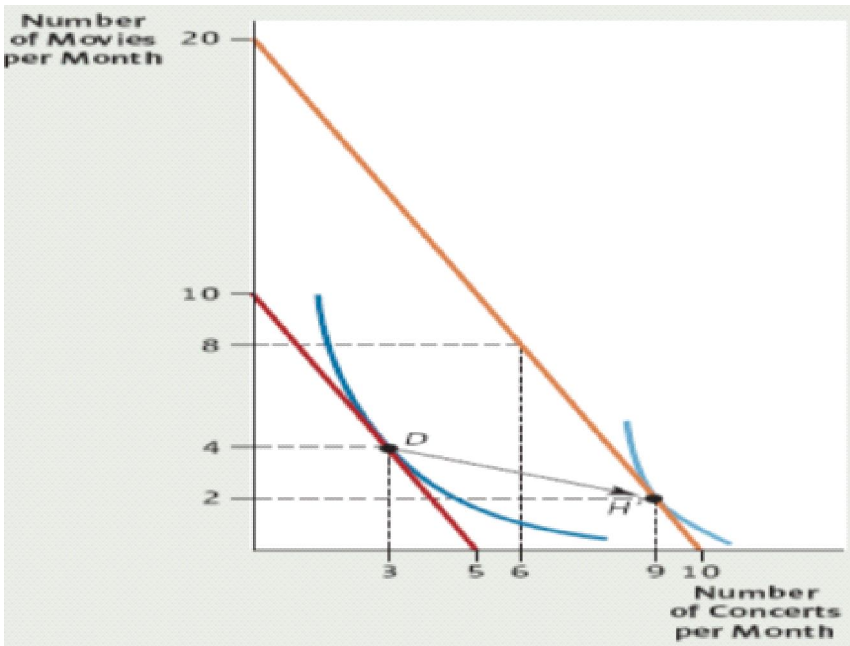
Figure 1: An Increase in Income



But where will this point be? There are several possibilities, and they depend on Max's preferences, as reflected in his indifference map. In the figure, we've used an indifference map for Max that leads him to point *H,* enjoying 6 concerts and 8 movies per month. As you can see in the figure, at this point, he has reached the highest possible indifference curve that his budget allows. It's also the point at which $MRSm,c = Pc / Pm = 2$.

Notice that, in moving from *D* to *H,* Max chooses to buy more concerts (6 rather than 3) and more movies (8 rather than 6). As discussed in Chapter 3, if an increase in income (with prices held constant) increases the quantity of a good demanded, the good is *normal*. For Max, with the indifference map we've assumed in Figure 1, both concerts and movies would be normal goods.

Figure 2: Income Rises and Movies are Inferior



When income rises, whether demand for a good rises or falls depends on preferences, as represented by the consumer's indifference map. In this figure, Max's preferences make movies an inferior good. So when income rises from \$100 to \$200, he moves from point D *to point* H. *Concerts increase from 3 to 9, but movies decrease from 8 to 2.*

But what if Max's preferences, and his indifference map, were as shown in Figure 5?

Here, after income rises, the tangency between his budget line and the highest indifference curve he could reach occurs at point like *H'*, with 9 concerts and 2 movies. In this case, the increase in income would cause Max's consumption of concerts to increase (from 3 to 9), but his consumption of movies to *fall* (from 6 to 2). If so, movies would be an *inferior good* for Max, one for which demand decreases when income increases, while concerts would be a normal good.

It's also possible for Max to have preferences that lead him to a different point—with more movies and *fewer* concerts than at point *D*. In this case, *concerts* would be the inferior good and movies would be normal.

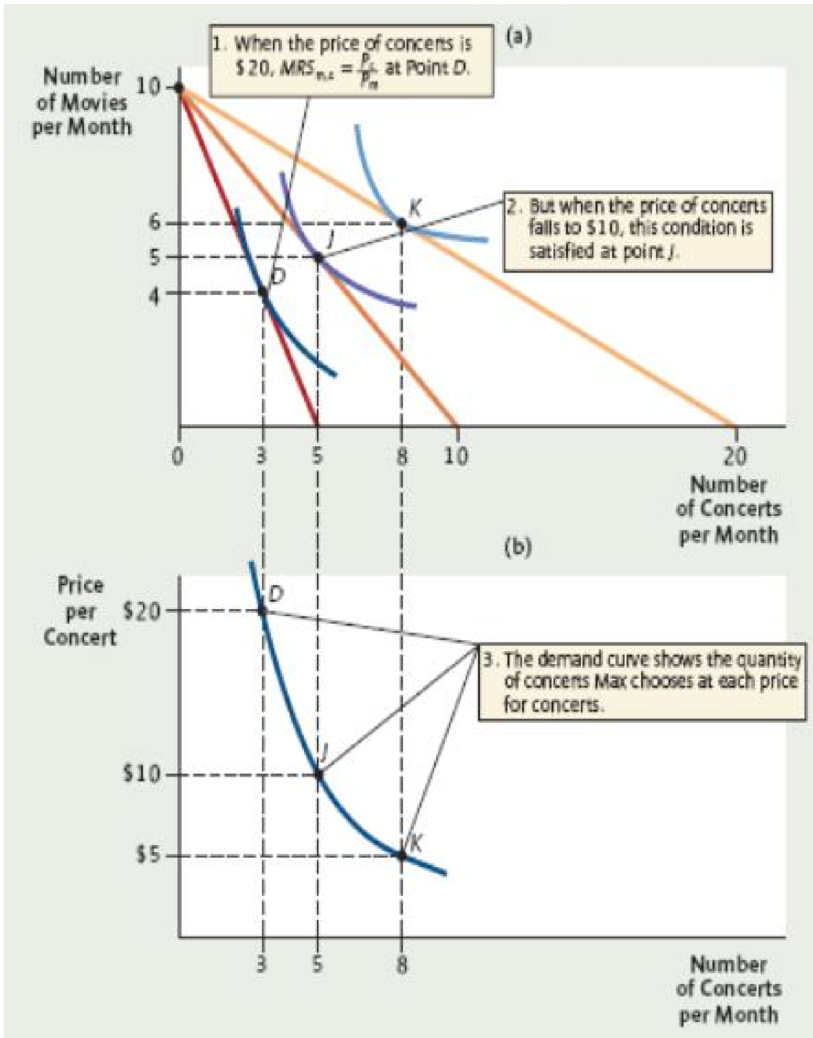
A rise in income, with no change in prices, leads to a new quantity demanded for each good. Whether a particular good is normal (quantity demanded increases) or inferior (quantity demanded decreases) depends on the individual's preferences, as represented by his indifference map.

Change in Price and Consumer Decisions & Deriving Demand Curve with Indifference Curve

Let's explore what happens to Max when the price of a concert decreases from \$20 to \$10, while his income and the price of a movie remain unchanged. The drop in the price of concerts rotates Max's budget line rightward, pivoting around its vertical intercept, as illustrated in the upper panel of Figure 3. What will Max do after his budget line rotates in this way? Based on his indifference curves—as they appear in the figure—he'd choose point *J.* This is the new combination of movies and concerts on his budget line that makes him as well off as possible (puts him on the highest possible indifference curve that he can afford). It's also the point at which $MRS_{m,c} = P_c/P_m = 1$, since movies and concerts now have the same price.

What if we dropped the price of concerts again, this time, to \$5? Then Max's budget line rotates further rightward, and he will once again find the best possible point. In the figure, Max is shown choosing point *K,* attending 8 concerts and 6 movies.

Figure 3: Deriving the Demand Curve with Indifference Curves



The lower panel of Figure 3 illustrates this relationship by plotting the quantity of concerts demanded on the horizontal axis and the *price* of concerts on the vertical axis. For example, in both the upper and lower panels, point *D* tells us that when the price of concerts is \$20, Max will see three of them. When we connect points like *D, J,* and *K* in the lower panel, we get Max's demand curve, which shows the quantity of a good he demands at each different price. Notice that Max's demand curve for concerts slopes downward—a fall in the price of concerts increases the quantity demanded—showing that for Max, concerts obey the law of demand.

But if Max's preferences—and his indifference map— had been different, could his response to a price change have *violated* the law of demand? The answer is yes . . . and no. Yes, it is theoretically possible. But no, it does not seem to happen in practice. To find out why, it's time to go back to the body of the chapter, to the section titled, "Income and Substitution Effects."

Indifference curve A curve representing all combinations of two goods that make the consumer equally well off.

Indifference map A set of indifference curves that represent an individual's preferences.

Marginal rate of substitution (*MRSy,x*) The maximum amount of good *y* a consumer would willingly trade for one more unit of good *x*. Also, the slope of a segment of an indifference curve.

Consumer Decision and Revealed Preference Theory (RPT)

Concept of Revealed Preference

Prof. Samuelson has invented an alternative approach to the theory of consumer behaviour which, in principle, does not require the consumer to supply any information about himself. If his tastes do not change, this theory, known as the Revealed Preference Theory (RPT), permits us to find out all we need to know just by observing his market behaviour, by seeing what he buys at different prices, assuming that his acquisitions and buying experiences do not change his preference patterns or his purchase desires. Given enough such information, it is even theo-retically possible to reconstruct the consumer's indifference map.

Assumptions

With the help of the simple principle of RP, we may build up a powerful theory of consumer demand. The assumptions that we shall make here are:

(i) The consumer buys and uses only two goods (X and Y). The quantities x and y of these goods are continuous variables.

(ii) Both these goods are of MIB (more-is-better) type. This assumption is also known as the assumption of monotonicity. This assumption implies that the ICs of the consumer are nega-tively sloped.

(iii) The consumer's preferences are strictly convex. This assumption implies that the ICs of the consumer would be convex to the origin, which again implies that there would be obtained only one point (the point of tangency) on the budget line of the consumer that would be chosen by him over all other affordable combinations.

This assumption is very important. On the basis of this assumption, we shall obtain a one-to-one relation between the consumer's price-income situation or budget line and his equilib-rium choice—for any particular budget line of the consumer, there would be obtained one and only one equilibrium combination of goods and for any combination to be an equilibrium one, there would be obtained one and only one budget line.

(iv) The fourth assumption of the RP theory is known as the weak axiom of RP (WARP). Here we assume that if the consumer chooses the combination $E_1(x_1, y_1)$ over another afford-able combination $E_2(x_2, y_2)$ in a particular price-income situation, then under no circumstances would he choose E_2 over E_1 if E_1 is affordable. In other words, if a combination E_1 is revealed preferred to E_2, then, under no circumstances, E_2 can be revealed preferred to E_1.

(v) The fifth assumption of the RP theory is known as the strong axiom of RP (SARP). According to this assumption, if the consumer, under different price-income situations, reveals the combination E_1 as preferred to E_2, E_2 to E_3,…, E_{k-1} to E_k, then E_1 would be revealed preferred to E_k and E_k would never (under no price-income situation) be revealed preferred to E_1.

Revealed Preference—Direct and Indirect

If RP is confined to only two combinations of goods, E_1 and E_2, and if, in a particular price- income situation, E_1 (x_1,y_1) is revealed preferred to combination E_2 (x_2, y_2), then it is said that E_1 is directly revealed preferred to E_2.

But if preferences are considered for more than two combinations and if preferences are established by way of transitivity of RP, then it is a case of indirectly revealed preference. For example, if E_1 is revealed preferred to E_2,…, E_{k-1} to E_k, then by SARP, we say E_1 is indirectly revealed preferred to E_k.

Assumptions

1. Rationality

The consumer is assumed to behave rationally, in that he prefers bundles of goods that include more quantities of the commodities.

2. Consistency

The consumer behaves consistently, that is, if he chooses bundle A in a situation in which bundle B was also available to him he will not choose B in any other situation in which A is also available. Symbolically if A > B, then B !> A

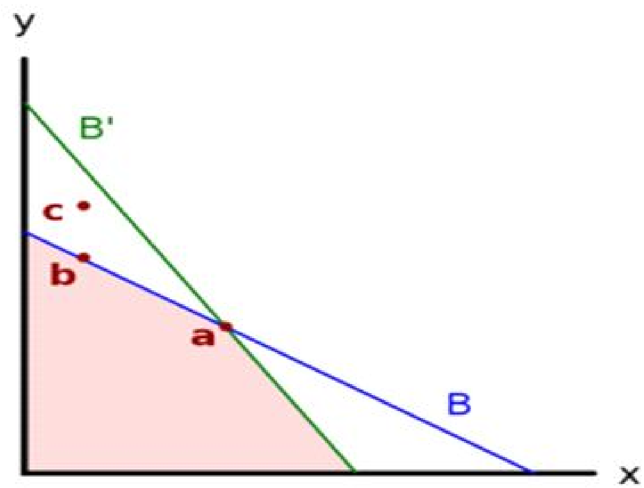
3. Transitivity

If in any particular situation A > B and B > C, then A > C.

4. The revealed preference axiom

The consumer, by choosing a collection of goods in any one budget situation, reveals his preference for that particular collection. The chosen bundle is revealed to be preferred among all other alternative bundles available under the budget constraint. The chosen 'basket of goods' maximizes the utility of the con-sumer. The revealed preference for a particular collection of goods implies (axiomatically) the maximization of the utility of the consumer.

If bundle b is revealed preferred over bundle a in budget set B, then bundle a can not be strictly revealed preferred over bundle b in any budget set B'. This would be equally true had a been located anywhere else in the pink area. The bundle c will not violate even if it is chosen in budget set B', because it is not in the pink (shaded) area.



Critique of the revealed preference hypothesis

We have already said that Samuelson's revealed preference theory is a major advance-ment to the theory of demand. It provides a direct way to the derivation of the demand curve, which does not require the use of the concept of utility. The theory can prove the existence and convexity of the indifference curves under weaker assumptions than the earlier theories. It has also provided the basis for the construction of index numbers of the cost of living and their use for judging changes in consumer welfare in situations where prices change.

Consumer Surplus

In order to explain the concept of consumer surplus, we have to know what is willingness to pay?

Willingness to pay

Each consumer's maximum price for a commodity is called his **willingness to pay**, and it measures how much that buyer values the good. Each buyer would be eager to buy at a price less than his willingness to pay, and he would refuse to buy the commodity at a price greater than his willingness to pay. At a price equal to his willingness to pay, the buyer would be indifferent about buying the good. If the price is exactly the same as the value he places on the commodity, he would be equally happy buying it or keeping his money. C**onsumer surplus** is the amount a buyer is willing to pay for a good minus the amount the buyer actually pays for it

How a Lower Price Raises Consumer Surplus

Because buyers always want to pay less for the goods they buy, a lower price makes buyers of a good better off. But how much does buyers' well-being rise in response to a lower price? We can use the concept of consumer surplus to answer this question precisely. In a market with many buyers, the resulting steps from each buyer dropping out are so small that they form, in essence, a smooth curve. Although this curve has a different shape, the ideas we have just developed still apply:

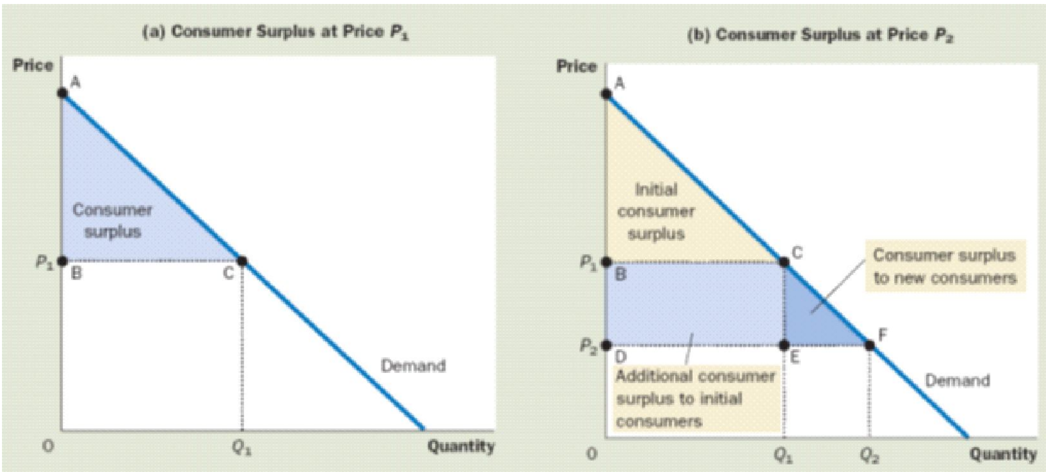
Consumer surplus is the area above the price and below the demand curve. In panel (a), consumer surplus at a price of $P1$ is the area of triangle ABC. Now suppose that the price falls from $P1$ to $P2$, as shown in panel (b). The consumer surplus now equals area ADF. The increase in consumer surplus attributable to the lower price is the area BCFD. This increase in consumer surplus is composed of two parts. First, those buyers who were already buying $Q1$ of the good at the higher price $P1$ are better off because they now pay less.

The increase in consumer surplus of existing buyers is the reduction in the amount they pay; it equals the area of the rectangle BCED. Second, some new buyers enter the market because they are willing to buy the good at the lower price. As a result, the quantity demanded in the market increases from $Q1$ to $Q2$. The consumer surplus these newcomers receive is the area of the triangle CEF.

What does Consumer Surplus Measure?

Our goal in developing the concept of consumer surplus is to make judgments about the desirability of market outcomes. Now that you have seen what consumer surplus is, let's consider whether it is a good measure of economic well-being.

Figure 1: How the Price Affects Consumer Surplus



Imagine that you are a policymaker trying to design a good economic system. Would you care about the amount of consumer surplus?

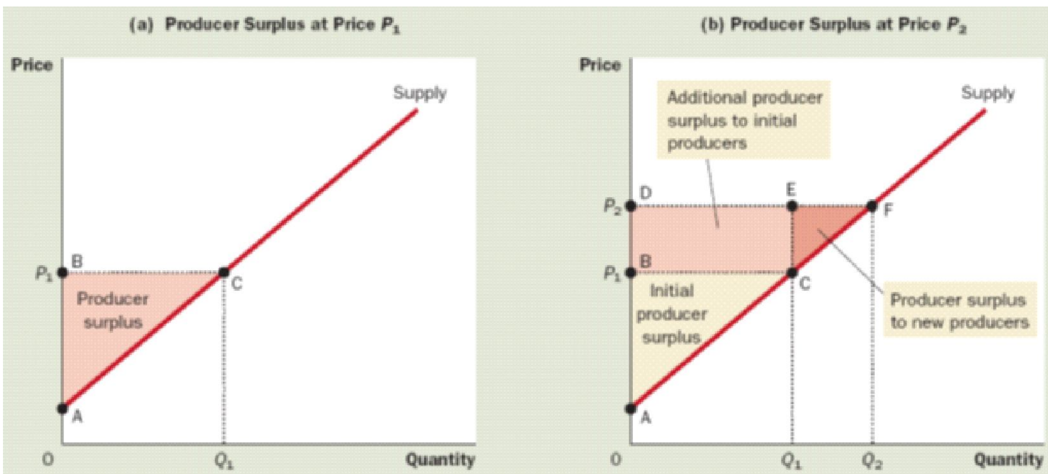
Consumer surplus, the amount that buyers are willing to pay for a good minus the amount they actually pay for it, measures the benefit that buyers receive from a good *as the buyers themselves perceive it*. Thus, consumer surplus is a good measure of economic wellbeing if policymakers want to respect the preferences of buyers.

In some circumstances, policymakers might choose not to care about consumer surplus because they do not respect the preferences that drive buyer behavior. For example, drug addicts are willing to pay a high price for heroin. Yet we would not say that addicts get a large benefit from being able to buy heroin at a low price (even though addicts might say they do). From the standpoint of society, willingness to pay in this instance is not a good measure of the buyers' benefit, and consumer surplus is not a good measure of economic well-being, because addicts are not looking after their own best interests.

In most markets, however, consumer surplus does reflect economic well-being. Economists normally assume that buyers are rational when they make decisions. Rational people do the best they can to achieve their objectives, given their opportunities. Economists also normally assume that people's preferences should be respected. In this case, consumers are the best judges of how much benefit they receive from the goods they buy.

In panel (a), the price is $P1$, the quantity demanded is $Q1$, and consumer surplus equals the area of the triangle ABC. When the price falls from $P1$ to $P2$, as in panel (b), the quantity demanded rises from $Q1$ to $Q2$, and the consumer surplus rises to the area of the triangle ADF. The increase in consumer surplus (area BCFD) occurs in part because existing consumers now pay less (area BCED) and in part because new consumers enter the market at the lower price (area CEF).

Figure 2: How the Price Affects Producer Surplus



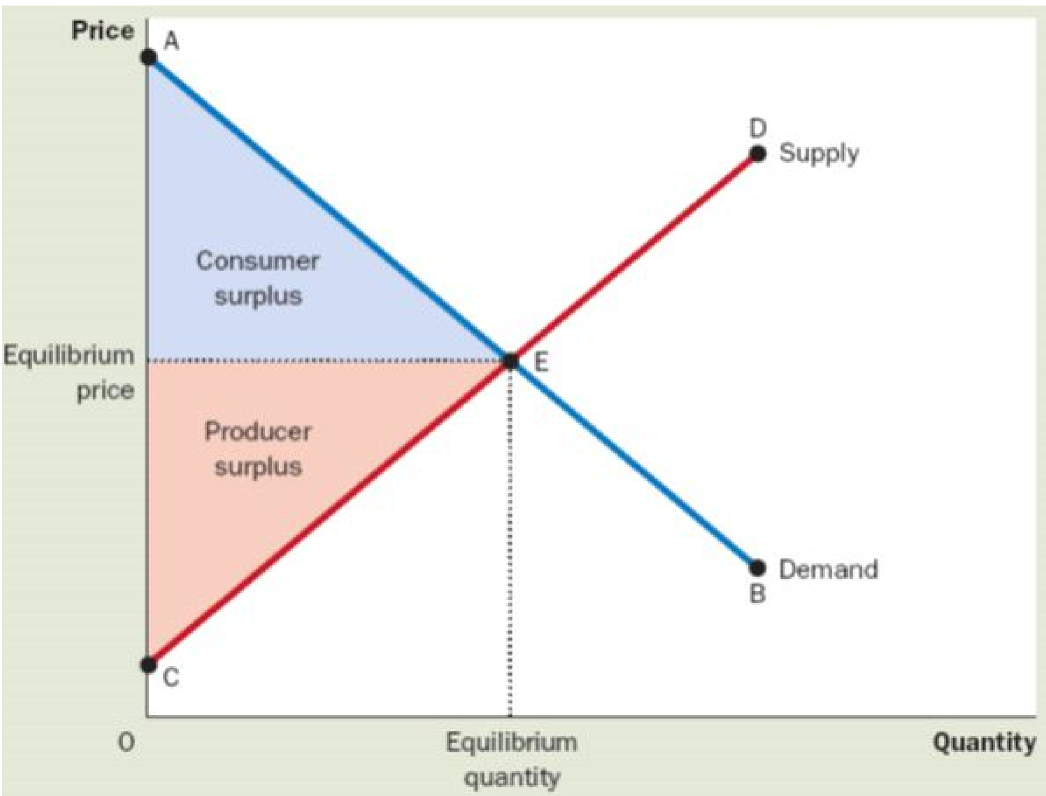
In panel (a), the price is $P1$, the quantity supplied is $Q1$, and producer surplus equals the area of the triangle ABC. When the price rises from $P1$ to $P2$, as in panel (b), the quantity supplied rises from $Q1$ to $Q2$, and the producer surplus rises to the area of the triangle ADF. The increase in producer surplus (area BCFD) occurs in part because existing producers now receive more (area BCED) and in part because new producers enter the market at the higher price (area CEF).

Market Equilibrium

Figure 3 shows consumer and producer surplus when a market reaches the equilibrium of supply and demand. Recall that consumer surplus equals the area above the price and under the demand curve and producer surplus equals the area below the price and above the supply curve. Thus, the total area between the supply and demand curves up to the point of equilibrium represents the total surplus in this market.

Is this equilibrium allocation of resources efficient? That is, does it maximize total surplus? To answer this question, recall that when a market is in equilibrium, the price determines which buyers and sellers participate in the market. Those buyers who value the good more than the price (represented by the segment AE on the demand curve) choose to buy the good; buyers who value it less than the price (represented by the segment EB) do not. Similarly, those sellers whose

Figure 3: Consumer and Producer Surplus in the Market Equilibrium



Total surplus—the sum of consumer and producer surplus—is the area between the supply and demand curves up to the equilibrium quantity. costs are less than the price (represented by the segment CE on the supply curve) choose to produce and sell the good; sellers whose costs are greater than the price (represented by the segment ED) do not.

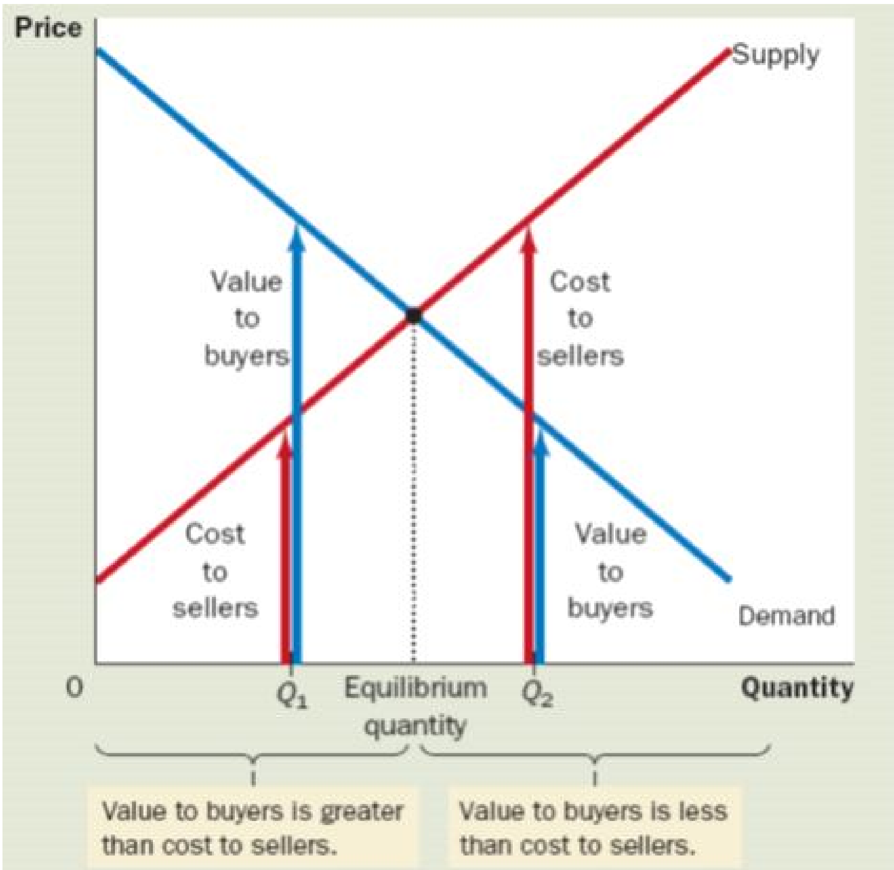
These observations lead to two insights about market outcomes:

1. Free markets allocate the supply of goods to the buyers who value them most highly, as measured by their willingness to pay.

2. Free markets allocate the demand for goods to the sellers who can produce them at the least cost. Thus, given the quantity produced and sold in a market equilibrium, the social planner cannot increase economic well-being by changing the allocation of consumption among buyers or the allocation of production among sellers. But can the social planner raise total economic well-being by increasing or decreasing the quantity of the good? The answer is no, as stated in this third insight about market outcomes:

3. Free markets produce the quantity of goods that maximizes the sum of consumer and producer surplus. To interpret this figure, keep in mind that the demand curve reflects the value to buyers and the supply curve reflects the cost to sellers. At any quantity below the equilibrium level, such as $Q1$, the value to the marginal buyer exceeds the cost to the marginal seller. As a result, increasing the quantity produced and consumed raises total surplus. This continues to be true until the quantity reaches the equilibrium level. Similarly, at any quantity beyond the equilibrium level, such as $Q2$, the value to the marginal buyer is less than the cost to the marginal seller. In this case, decreasing the quantity raises total surplus, and this continues to be true until quantity falls to the equilibrium level. To maximize total surplus, the social planner would choose the quantity where the supply and demand curves intersect.

Figure 4: The Efficiency of the Equilibrium Quantity



At quantities less than the equilibrium quantity, such as *Q*1, the value to buyers exceeds the cost to sellers. At quantities greater than the equilibrium quantity, such as *Q*2, the cost to sellers exceeds the value to buyers. Therefore, the market equilibrium maximizes the sum of producer and consumer surplus.

Conclusion

Together, these three insights tell us that the market outcome makes the sum of consumer and producer surplus as large as it can be. In other words, the equilibrium outcome is an efficient allocation of resources. The benevolent social planner can, therefore, leave the market outcome just as he finds it. This policy of leaving well enough alone goes by the French expression *laissez faire,* which literally translates to "allow them to do."

Society is lucky that the planner doesn't need to intervene. Although it has been a useful exercise imagining what an all-knowing, all-powerful, well- intentioned dictator would do, let's face it: Such characters are hard to come by. Dictators are rarely benevolent, and even if we found someone so virtuous, he would lack crucial information.

Suppose our social planner tried to choose an efficient allocation of resources on his own, instead of relying on market forces. To do so, he would need to know the value of a particular good to every potential consumer in the market and the cost of every potential producer. And he would need this information not only for this market but for every one of the many thousands of markets in the economy. The task is practically impossible, which explains why centrally planned economies never work very well.

The planner's job becomes easy, however, once he takes on a partner: Adam Smith's invisible hand of the marketplace. The invisible hand takes all the information about buyers and sellers into account and guides everyone in the market to the best outcome as judged by the standard of economic efficiency. It is, truly, a remarkable feat. That is why economists so often advocate free markets as the best way to organize economic activity.

Questions

1. Explain how buyers' willingness to pay, consumer surplus, and the demand curve are related.
2. Explain how sellers' costs, producer surplus, and the supply curve are related.
3. In a supply-and-demand diagram, show producer and consumer surplus in the market equilibrium.
4. What is efficiency? Is it the only goal of economic policymakers?
5. What does the invisible hand do?

This lesson will focus on the following:

1. Production Function
2. Short-Run versus Long-Run Production Decisions
3. Production in Short Run
4. Marginal Returns to Labour

Production Function

The word *production* encompasses more than just manufacturing. *Production is the process of combining* inputs *to make* goods and services. Notice that the definition refers to goods *and* services. Some production does indeed create physical *goods,* like automobiles, aircraft, or ice cream. But production also creates *services*. Indeed, many of India's largest corporations are service providers.

Profit Total revenue minus total cost.

Business firm An organization, owned and operated by private individuals, that specializes in production.

Production and Cost

What about the inputs that are used to produce things? These include the four resources (land, labour, capital, and entrepreneurship), as well as other things. A book publishing business firm uses several resources, including *labour* (of the authors, editors, art designers, and others), *capital* (buildings, office furniture, computers), and *land* (under the buildings). But the company also used *other* inputs that were produced by other firms, including raw materials like paper and ink, transportation and telecommunications services, and more.

Technology and Production

When you hear the word "technology," you are likely to think of the latest electronic gadget that someone told you about. But in economics technology has aspecific meaning:

A firm's ***technology*** *refers to the methods it can use to turn inputs into outputs (produced goods or services).*

If the firm's technology allows it to use *different* methods for producing the same level of output, we assume the firm will choose the cheapest method it can find. You'll learn more about how a firm makes this choice later.

To keep things simple, we'll spell out the production technology for a mythical firm that uses only two inputs: *capital* and *labour*. Our firm is Spotless Car Wash, whose output is a service: washing cars. The firm's capital consists of automated car-washing lines. It's labour is full-time workers who drive the cars on to the line, drive them out, towel them down at the end, and deal with customers.

Short-Run versus Long-Run Decisions

When a firm changes its level of production, it will want to adjust the amounts of inputs it uses. But these adjustments depend on the *time horizon* the firm's managers are thinking about. Some inputs can be adjusted relatively quickly. Most firms, for example, can hire more labour and purchase more raw materials within a few weeks or less. But at many firms, some inputs take longer to adjust. It may take a year or longer before an automobile firm can purchase and install new assembly lines, or acquire additional factory space. And legal obligations, like leases or rental agreements, can delay efforts to downsize operations, because the firm will have to continue paying for equipment and buildings for some time, whether it uses them or not. Thus, when we ask, "What quantities of what inputs will the firm use to produce a given level of output?"

Technology The methods available for combining inputs to produce agood or service.

If it's next month, a firm may be stuck with the factory and equipment it currently has, so it can only adjust its labour and raw materials. If we're asking about next year, there is more flexibility—enough time to adjust capital equipment as well.

These considerations make it useful to divide the different time horizons firms can use into two broad categories: the *long run* and the *short run*.

The ***long run*** *is a time period long enough for a firm to change the quantity of* all *of its inputs.*

Another way to say this is that, over the long run, all the inputs the firm uses a reviewed as

Variable inputs—inputs that can be adjusted up or down as the quantity of output changes.

At Spotless Car Wash, we'll imagine it takes a year to acquire and install a new automated line, or to sell the lines it already has. Thus, for Spotless the long run isa time horizon of one year or longer. When its managers make long-run decisions, they regard all inputs (labour and capital in this case) as variable inputs.

What about shorter time periods? The company may be stuck with the current quantities of some inputs. We call these **fixed inputs**—inputs that, over the time period we're considering, cannot be adjusted as output changes. Using this terminology, we can define the short run as follows:

The ***short run*** *is a time period during which at least one of the firm's inputs is fixed*

For Spotless Car Wash, the short run is any time period less than a year, because that is how long it is stuck with its current quantity of automated lines. Over the short-run, Spotless's labour is a variable input, but its capital is a fixed input.

Production in the Short Run

When firms make decisions over the short run, there is nothing they can do about their fixed inputs: They are stuck with whatever quantity they have. They can, however, make choices about their variable inputs. Indeed, we see examples of such short-run decisions all the time. If Levi Strauss decides to increase production of blue jeans over the next quarter, it may use additional workers, cotton cloth, and sewing machines. But it continues to make do with the same factories because there isn't time to expand them or acquire new ones. Here, workers, cloth, and sewing machines are all variable over the quarter, while the factory buildings are fixed.

At Spotless Car Wash, over the short run, labour is the only variable input, and capital is the only fixed input. The three columns in Table 1 describe Spotless's production choices in the short run. Column 1 shows the quantity of the fixed input, capital (K); column 2 the quantity of the variable input, labour (L). Note that in the short run, Spotless is stuck with one unit of capital—one automated line—but it can take on as many or as few workers as it wishes. Column 3 shows the firm's *total product* (Q).

***Total product** is the maximum quantity of output that can be produced from a given combination of inputs.*

Long run A time horizon long enough for a firm to vary all of its inputs.

Variable input An input whose usage can change as the level of output changes.

Fixed input An input whose quantity must remain constant, regardless of how much output is produced.

Short run A time horizon during which at least one of the firm's inputs cannot be varied.

Table 1: Short-Run Production at Spotless Car Wash

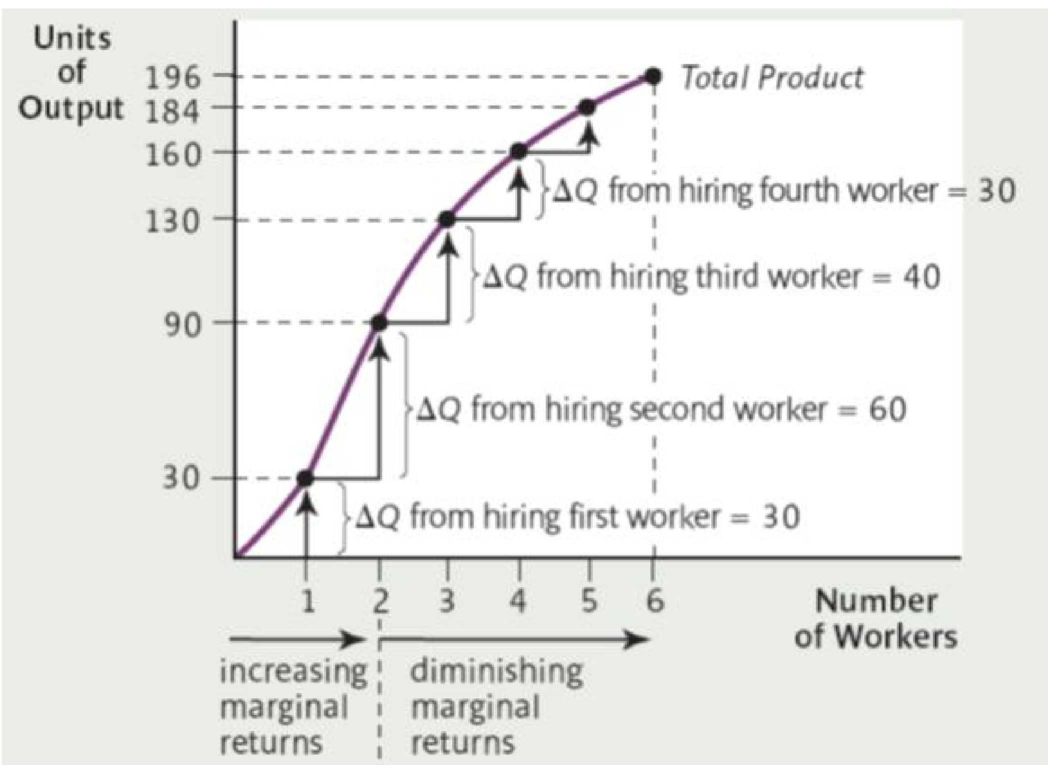
Quantity of Capital	Quantity of Labor	Total Product (Cars Washed per Day)
1	0	0
1	1	30
1	2	90
1	3	130
1	4	160
1	5	184
1	6	196

For example, the table shows us that with one automated line but no labour, total product is zero. With one line and six workers, total product is 196 cars washed per day. The total product numbers in the table tell us the *maximum* output for each number of workers. We can also reverse this logic, and say that for each value of the total product, the labour column shows us the *lowest* number of workers that can produce it. Since labour is the only variable input, this lowest number of workers will also be the *least-cost* method of producing any level of output.

Figure 1 shows Spotless's *total product curve*. The horizontal axis represents the number of workers, while the vertical axis measures total product. (The amount of capital—which is held fixed at one automated line—is not shown on the graph.)

Notice that each time the firm hires another worker, output increases, so the total product curve slopes upward. The vertical arrows in the figure show precisely *how much* output increases with each one-unit rise in employment. We call this rise in output the *marginal product of labour*.

Figure 1: Total and Marginal Product



The total product curve shows the total amount of output that can be produced using various numbers of workers. The marginal product of labour (MPL) is the change in total product when another worker is hired. The MPL for each change in employment is indicated by the length of the vertical arrows.

The ***marginal product of labor*** (MPL) *is the change in total product* (ΔQ) *divided by the change in the number of workers employed* (ΔL):

$$\text{MPL} = \frac{\Delta Q}{\Delta L}$$

The MPL *tells us the rise in output produced when one more worker is hired.*

Using the Greek letter Δ ("delta") to stand for "change in," we can define marginal product this way:

The ***marginal product of labour*** **(MPL)** *is the change in total product* (ΔQ) *divided by the change in the number of workers employed* (ΔL)*:*

$$\text{MPL} = \Delta Q \div \Delta L$$

The MPL *tells us the rise in output produced when one more worker is hired.*

For example, if employment rises from 2 to 3 workers, total product rises from 90 to 130, so the marginal product of labour for *that* change in employment is calculated as (130 - 90)/1 = 40 units of output.

Marginal Returns to Labour

Look at the vertical arrows in Figure 1, which measure the marginal product of labour, and you may notice something interesting. As more and more workers are hired, the *MPL* first increases (the vertical arrows get longer) and then decreases (the arrows get shorter). This pattern is believed to be typical at many types of firms, so it's worth exploring.

Increasing Marginal Returns to Labour

When the marginal product of labour rises as more workers are hired, there are **increasing marginal returns to labour.** Each time a worker is hired, total output rises by more than it did when the previous worker was hired. Why might this happen? Additional workers may allow production to become more specialized.

For example, Figure 1 tells us that Spotless Car Wash experiences increasing returns to labour up to the hiring of the second worker. While one worker *could* operate the car wash alone, he or she would have to do everything: drive the cars on and off the line, towel them down, and deal with customers. Much time would be wasted switching from one task to another. Table 1 tells us that one worker can wash only 30 cars each day. Add a second worker, though, and now specialization is possible. One worker can collect money and drive the cars onto the line, and the other can drive them off and towel them down. Thus, with two workers, output rises all the way to 90 car washes per day; the second worker adds more to production (60 car washes) than the first (30 car washes) by making *both* workers more productive.

Diminishing Marginal Returns to Labour

When the marginal product of labour is decreasing, we say there are **diminishing marginal returns to labour.** Output still rises when another worker is added, so marginal product is positive. But the rise in output is smaller and smaller with each successive worker.

Why does this happen? For one thing, as we keep adding workers, additional gains from specialization will be harder and harder to achieve. Moreover, each worker will have less and less of the fixed inputs with which to work.

This last point is worth stressing. It applies not just to labour but to any variable input. In all kinds of production, if we keep increasing the quantity of any one input, while holding the others fixed, diminishing marginal returns will eventually set in.

For example, if a farmer keeps adding additional pounds of fertilizer to fixed amounts of land and labour, the yield may continue to increase, but eventually the *size* of the increase—the marginal product of fertilizer—will begin to come down. This tendency is so pervasive and widespread that it has been deemed a law:

Conclusion

The ***law of diminishing (marginal) returns*** *states that as we continue to add more of any one input (holding the other inputs constant), its marginal product will* *eventually decline.* The law of diminishing returns is a physical law, not an economic one. It is based on the nature of production—on the physical relationship between inputs and outputs with a given technology. Figure 1 tells us that at Spotless diminishing returns set in after two workers have been hired. Beyond this point, the firm is crowding more and more workers into a car wash with just one automated line. Output continues to increase, since there is usually *something* an additional worker can do to move the cars through the line more quickly, but the increase is less dramatic each time.

Questions

1. Define Production Function.
2. Explain Short-Run versus Long-Run Production Decisions.
3. Explain Production in Short Run.
4. Discuss Marginal Returns to Labour.

This lesson will focus on the following:

1. Concept of Cost
2. Sunk Costs
3. Explicit versus Implicit Costs
4. Cost in Short Run: Measuring Short-Run Costs
5. Shape of Marginal Cost Curve
6. Relationship between Average and Marginal Costs

Concept of Costs

Production is the *physical* relationship between inputs and outputs. But a more critical concern for a firm is: What will it *cost* to produce any level of output? Everything you've learned about production will help you understand the behaviour of costs.

Opportunity cost—what we must give up in order to do something. This concept applies to the firm as well:

A firm's total cost of producing a given level of output is the opportunity cost of the owners—everything they must give up in order to produce that amount of output.

Using the concept of opportunity cost can help us understand which costs matter— and which don't—when making business decisions.

Sunk Costs

Suppose that last year, Acme Pharmaceutical Company spent $10 million developing a new drug to treat acne that, if successful, would have generated millions of dollars in annual sales revenue. At first, the drug seemed to work as intended. But then, just before launching production, management discovered that the new drug didn't cure acne at all—

but was remarkably effective in treating a rare underarm fungus. In this smaller, less lucrative market, annual sales revenue would be just $30,000. Now management must decide: Should they sell the drug as an anti fungus remedy?

When confronted with a problem like this, some people will say: "Acme should *not* sell the drug. You don't sell something for $30,000 a year when it cost you $10 million to make it." Others will respond this way: "Of course Acme should sell the drug. If they don't, they'd be wasting that huge investment of $10 million." But to an economist, neither approach to answering this question is correct, because both use the $10 million development cost to reach a conclusion—and that cost is completely *irrelevant* to the decision.

Law of diminishing (marginal) returns As more and more of any input is added to a fixed amount of other inputs, its marginal product will eventually decline.

The $10 million already spent on developing the drug is an example of a *sunk cost*. More generally, *a **sunk cost** is one that already has been paid, or must be paid,* regardless *of any future action being considered.*

In the case of Acme, the development cost has been paid already. The firm will not get this money back, whether it chooses to sell the drug in this new smaller market or not. Because the $10 million is not part of the opportunity cost of either choice—something that would have to be sacrificed *for* that choice—it should have no bearing on the decision. For Acme, as for any business, *Sunk costs should not be considered when making decisions.*

What *should* be considered are the costs that *do* depend on the decision about producing the drug, namely, the cost of actually manufacturing it and marketing it for the smaller market. If these costs are less than the $30,000 Acme could earn in annual revenue, Acme should produce the drug. Otherwise, it should not.

Look again at the definition of sunk cost and you'll see that even a *future* payment can be sunk, if an *unavoidable commitment to pay it has already been made*. Suppose, for example, Acme Pharmaceuticals has signed an employment contract with a research scientist, legally binding the firm to pay her annual salary for three years even if she is laid off. Although some or all of the payments haven't yet been made, all three years of salary are sunk costs for Acme because they *must* be made no matter what Acme does. As sunk costs, they are irrelevant to Acme's decisions.

Explicit versus Implicit Costs

There are two types of costs: *explicit* (involving actual payments) and *implicit* (no money changes hands). The same distinction applies to costs for a business firm.

Suppose you've opened a restaurant in a building that you already owned. You don't pay any rent, so there's no explicit rental cost. Does this mean that using the building is free?

Sunk cost A cost that has been paid or must be paid, regardless of any future action being considered.

Table 1 summarizes our discussion by listing some common categories of costs that business firms face, both explicit (on the left) and implicit (on the right).

Table 1: A Firm's Cost

Explicit Costs	Implicit Costs
Rent paid out	Opportunity cost of:
Interest on loans	Owner's land and buildings (rent foregone)
Managers' salaries	Owner's money (investment income foregone)
Hourly workers' wages	
Cost of raw materials	Owner's time (labor income foregone)

To an accountant—who focuses on actual money payments—the answer is yes. But to an economist—who thinks of opportunity cost—the answer is *absolutely not.* By using your own building for your restaurant, you are sacrificing the opportunity to rent it to someone else. This *foregone rent* is an *implicit cost,* and it is as much a cost of production as the rent you would pay if you didn't own a building yourself.

In both cases, something is given up to produce your output. Now suppose that instead of borrowing money to set up your restaurant, you used $100,000 of your own money. Therefore, you aren't paying any interest. But there is an opportunity cost: your $100,000 *could* have been put in the bank or lent to someone else, where it would be earning interest for you. If the going interest rate is 5 percent, then each year that you run your restaurant, you are giving up $5,000 in interest you could have instead. This *foregone interest* is another implicit cost of your business.

Finally, suppose you decided to manage your restaurant yourself. Have you escaped the costs of hiring a manager? Not really, because you are still bearing an opportunity cost: You *could* do something else with your time. We measure the value of your time as the income you *could* earn by devoting your labour to your next-best income-earning activity. This *foregone labour income*—the wage or salary you could be earning elsewhere—is an implicit cost of your business, and therefore part of its opportunity cost.

Cost in the Short Run

Managers must answer questions about costs over different time horizons. One question might be, "How much will it cost to produce a given level of output *this year?*" Another might be, "How much will it cost us to produce a given level of output *three years from now and beyond?*"

In this section, we'll explore managers' view of costs over a short-run time horizon—a time period during which *at least one* of the firm's inputs is fixed. That is, we'll be looking at costs with a *short-run* planning horizon.

Remember that no matter how much output is produced, the quantity of a fixed input *must* remain the same. Other inputs, by contrast, can be varied as output changes. Because the firm has these two different types of inputs in the short run, it will also face two different types of costs.

The costs of a firm's fixed inputs are called, not surprisingly, **fixed costs**. Like the fixed inputs themselves, fixed costs must remain the same no matter what the level of output. Typically, we treat rent and interest—whether explicit or implicit—as fixed costs, since producing more or less output in the short run will not cause these costs to change. Managers typically refer to fixed costs as their *overhead costs,* or simply, *overhead.*

The costs of obtaining the firm's variable inputs are its **variable costs.** These costs, like the usage of variable inputs themselves, will rise as output increases. Most businesses treat the wages of hourly employees and the costs of raw materials as variable costs, because quantities of labour and raw materials can usually be adjusted rather rapidly.

Measuring Short-Run Costs

In Table 2, we return to our example—Spotless Car Wash—and ask: What happens to costs as output changes in the short run? The first three columns of the table tell us the inputs Spotless will use for each output level, just as in Table 1. But there is one slight difference: In Table 3, we've reversed the order of the columns, putting total output first. We are changing our perspective slightly: Now we want to observe how a change in the quantity of *output* causes the firm's *inputs*—and therefore its *costs*—to change.

We also need to know one more thing before we can analyze Spotless's costs: what it must *pay* for its inputs. In Table 2, the price of labour is set at $120 per worker per day, and the price of each automated car-washing line at $150 per day.

How do Spotless's short-run costs change as its output changes? Get ready, because there are a surprising number of different ways to answer that question, as illustrated in the remaining columns of Table 2.

Total Costs

Columns 4, 5, and 6 in the table 2 show three different types of total costs. In column 4, we have Spotless's **total fixed cost (*TFC*),** the cost of all inputs that are fixed in the short run.

We'll assume that the cost of purchasing and installing an automated line is $912,500, and that the annual interest rate is 6%. So for one automated line, Spotless's owners sacrifice interest of .06 x $912,500 = $54,750 per year, or $150 per day. That is Spotless's total fixed cost per day. Running down the column, you can see that this cost—because it is fixed—remains the same no matter how many cars are washed each day.

Column 5 shows **total variable cost (*TVC*),** the cost of all variable inputs. For Spotless, labour is the only variable input. As output increases, more labour will be needed, so *TVC* will rise. For example, to wash 90 cars each day requires 2 workers, and each worker must be paid $120 per day, so *TVC* will be 2 x $120 = $240. But to wash 130 cars requires 3 workers, so *TVC* will rise to 3 x $120 = $360.

Table 2: Short-Run Costs for Spotless Car Wash

Labor cost = $120 per day					Capital cost = $150 per day				
(1) Output (per Day)	(2) Capital	(3) Labor	(4) TFC	(5) TVC	(6) TC	(7) MC	(8) AFC	(9) AVC	(10) ATC
0	1	0	$150	$ 0	$150		—	—	—
						$ 4.00			
30	1	1	$150	$120	$270		$5.00	$4.00	$9.00
						$ 2.00			
90	1	2	$150	$240	$390		$1.67	$2.67	$4.33
						$ 3.00			
130	1	3	$150	$360	$510		$1.15	$2.77	$3.92
						$ 4.00			
160	1	4	$150	$480	$630		$0.94	$3.00	$3.94
						$ 5.00			
184	1	5	$150	$600	$750		$0.82	$3.26	$4.08
						$10.00			
196	1	6	$150	$720	$870		$0.77	$3.67	$4.44

Finally, column 6 shows us that ***total cost*** **(TC)** *is the sum of all fixed and variable costs:*

$$TC = TFC + TVC$$

For example, at 90 units of output, *TFC* = $150 and *TVC* = $240, so *TC* = $150 + $240 = $390. Because total variable cost rises with output, total cost rises as well. Now look at Figure 1, where we've graphed all three total cost curves for Spotless Car Wash. Both the *TC* and *TVC* curves slope upward, since these costs increase along with output. *TFC* is represented in two ways in the graph. One is the *TFC* curve, which is a horizontal line, since *TFC* has the same value at any level of output. The other is the *vertical distance* between the rising *TVC* and *TC* curves, since *TFC* is always the *difference* between *TVC* and *TC*. In the graph, this vertical distance must remain the same, at $150, no matter what the level of output.

Average Costs

While total costs are important, sometimes it is more useful to track a firm's costs *per unit* of output, which we call its *average cost.* There are three different types of average cost, each obtained from one of the total cost concepts just discussed.

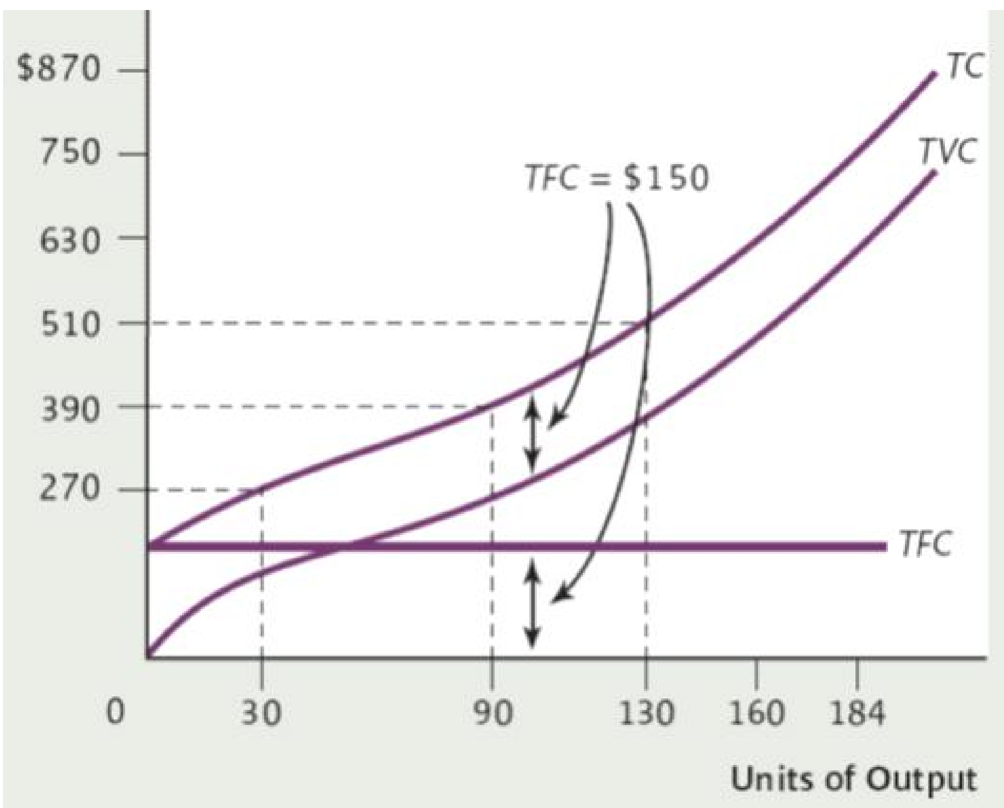
*The firm's **average fixed cost** (**AFC**) is its total fixed cost divided by the quantity* (Q) *of output:*

$$AFC = TFC \div Q$$

.**Total cost** The costs of all inputs— fixed and variable.

Average fixed cost Total fixed cost divided by the quantity of output produced.

Figure 1: A Firm's Total Cost Curve



At any level of output, total cost (TC) *is the sum of total fixed cost* (TFC) *and total variable cost* (TVC).

No matter what kind of production or what kind of firm, *AFC* will always fall as output rises. Why? Because *TFC* remains constant, so a rise in *Q must* cause the ratio *TFC/Q* to fall. Business managers often refer to this decline in *AFC* as "spreading their overhead" over more output. For example, a restaurant has overhead costs for its buildings, furniture, and cooking equipment. The more meals it serves, the lower will be its overhead cost per meal.

For Spotless Car Wash, look at column 8 of the table. When output is 30 units, *AFC* is \$150/30 = \$5.00. But at 90 units of output, *AFC* drops to \$150/90 = \$1.67. And *AFC* keeps declining as we continue down the column. The more output produced, the lower is fixed cost per unit of output.

Next is average *variable* cost.

Average variable cost **(AVC)** *is the cost of the variable inputs per unit of output:*

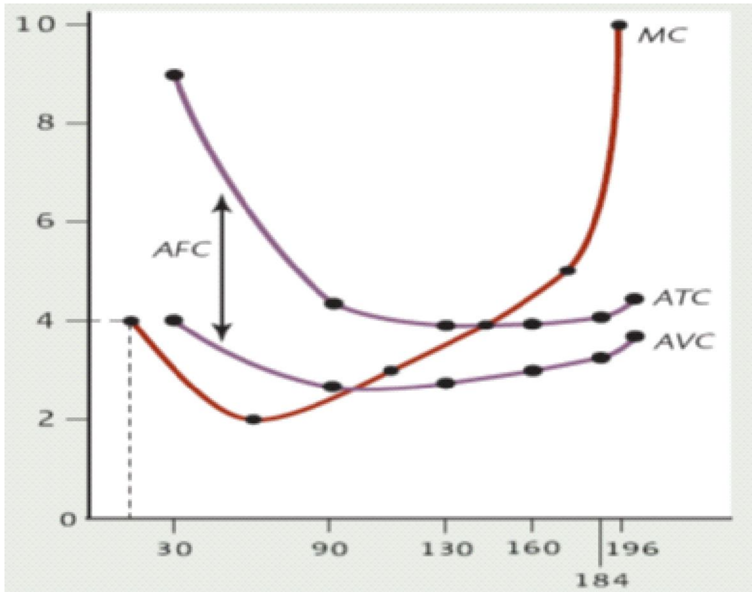
$$AVC = TVC \div Q$$

.*AVC* is shown in column 9 of the table. For example, at 30 units of output, *TVC* = \$120, so *AVC* = *TVC/Q* = \$120/30 = \$4.00.

What happens to *AVC* as output rises? If you run your finger down the *AVC* column in Table 3, you'll see a pattern: The *AVC* numbers first decrease and then increase. Economists believe that this pattern of decreasing and then increasing average variable cost is typical at many firms. When plotted in Figure 2, this pattern causes the *AVC* curve to have a U shape. We'll discuss the reason for this characteristic U shape a bit later.

Average variable cost Total variable cost divided by the quantity of output produced.

Figure 2: Average and Marginal Costs



Average variable cost (AVC) *and average total cost* (ATC) *are U-shaped, first decreasing and then increasing. Average fixed cost* (AFC), *the vertical distance between* ATC *and* AVC*, becomes smaller as output increases. The marginal cost* (MC) *curve is also U-shaped, reflecting first increasing and then diminishing marginal returns to labour.* MC *passes through the minimum points of both the* AVC *and* ATC *curves.*

The last average cost measure is average *total* cost.

Average total cost **(ATC)** *is the total cost per unit of output:* ATC = TC ÷ Q

.Values for *ATC* are listed in column 10 of Table 2. For example, at 90 units of output, *TC* = \$390, so *ATC* = *TC/Q* = \$390/90 = \$4.33. And a quick look at column 10 shows that as output rises, *ATC* first falls and then rises. So the *ATC* curve—like the *AVC* curve—is U-shaped. However—as you can see in Figure 2 - it is not identical to the *AVC* curve. At each level of output, the vertical distance between the two curves is equal to average *fixed* cost (*AFC*). Since *AFC* declines as output increases, the *ATC* curve and the *AVC* curve must get closer and closer together as we move rightward.

Marginal Cost

The total and average costs we've considered so far tell us about the firm's cost at a particular *level* of output. For many purposes, however, we are more interested in how cost *changes* when output *changes*. This information is provided by another cost concept:

Average total cost Total cost divided by the quantity of output produced.

***Marginal cost* (MC)** *is the change in total cost* (ΔTC) *divided by the change in output* (ΔQ):

$$\mathbf{MC = \Delta TC \div \Delta Q}$$

.It tells us how much cost rises per unit increase in output.

For Spotless Car Wash, marginal cost is entered in column 7 of Table 2 and graphed in Figure 2. Since marginal cost tells us what happens to total cost when output *changes*, the entries in the table are placed *between* one output level and another. For example, when output rises from 0 to 30, total cost rises from \$150 to \$270. For this change in output, we have ΔTC = \$270 - \$150 = \$120, while ΔQ = 30, so MC = \$120/30 = \$4.00. This entry is listed *between* the output levels 0 and 30 in the table 2.

Shape of Marginal Cost Curve

Look at the graph of marginal cost in Figure 2. As in the table, each value of marginal cost is plotted *between* output levels. For example, the marginal cost of increasing output from 0 to 30 is \$4, and this is plotted at output level 15—midway between 0 and 30. Similarly, when going from 30 to 90 units of output, the *MC* is plotted midway between 30 and 90.

If you look carefully at the *MC* curve in Figure 2, you'll see that *MC* first declines and then rises. Why is this? Here, we can use what we learned earlier about marginal returns to labour. At low levels of employment and output, there are increasing marginal returns to labour: $MPL = \Delta Q/\Delta L$ is rising. That is, each worker hired adds more to production than the worker before. But that means *fewer additional workers are needed to produce an additional unit of output*, so the *cost* of an additional unit of output (*MC*) must be falling. Thus, as long as *MPL* is rising, *MC* must be falling.

For Spotless, since *MPL* rises when employment increases from zero to one and again from one to two workers, *MC* must fall as the firm's output rises from zero to 30 units (produced by one worker) and then from 30 to 90 units (produced by two workers).

At higher levels of output, we have the opposite situation: Diminishing marginal returns set in and the marginal product of labour ($\Delta Q/\Delta L$) falls. Therefore, additional units of output require *more* and *more* additional labour. As a result, each additional unit of output costs more and more to produce. Thus, as long as *MPL* is falling, *MC* must be rising.

For Spotless, diminishing marginal returns to labour occur for all workers beyond the second, so *MC* rises for all increases in output beyond 90.

To sum up:

When the marginal product of labour (MPL) *rises, marginal cost* (MC) *falls. When* MPL *falls,* MC *rises. Since* MPL *ordinarily rises and then falls,* MC *will do the opposite: It will fall and then rise. Thus, the* MC *curve is U-shaped.*

Marginal cost The increase in total cost from producing one more unit of output.

Relationship between Average and Marginal Costs

Although marginal cost and average cost are not the same, there is an important relationship between them. Look again at Figure 2 and notice that all three curves— *MC*, *AVC*, and *ATC*—first fall and then rise, but not all at the same time. The *MC* curve bottoms out before either the *AVC* or *ATC* curve. Further, the *MC* curve intersects each of the average curves *at their lowest points*. These graphical features of Figure 2 are no accident; indeed, they follow from the laws of mathematics. To understand this, let's consider a related example with which you are probably more familiar.

An Example: Average and Marginal Test Scores

Suppose you take five tests in your economics course during the term, with the results listed in Table 3. To your immense pleasure, you score 100 on your first test. Your total score—the total number of points you have received thus far during the term—is 100. Your marginal score—the *change* in your total caused by the most recent test—will also be 100, since your total rose from 0 to 100. Your average score so far is 100 as well.

Now suppose that, for the second test, you forget to study actively. Instead, you just read the text while simultaneously watching music videos and eaves dropping on your roommate's phone conversations. As a result, you get a 50, which is your marginal score. Since this score is lower than your previous average of 100, the second test will *pull your average down.* Indeed, whenever your score is lower than your previous average, it will pull down your average. In the table, we see that your average after the second test falls to 75.

Now you start to worry, so you turn off the TV while studying, and your performance improves a bit: You get a 60. Does the improvement in your score—from 50 to 60—increase your *average* score? No . . . your average will decrease once again, because your *marginal* score of 60 is *still* lower than your previous average of 75. As we know, when you score lower than your average, it pulls the average down, even if you're improving. In the table, we see that your average now falls to 70. For your fourth exam, you study a bit harder and score a 70. This time, since your score is precisely *equal* to your previous average, the average remains unchanged at 70.

Table 3: Average and Marginal Test Scores

Number of Tests Taken	Total Score	Marginal Score	Average Score
0	0		—
		100	
1	100		100
		50	
2	150		75
		60	
3	210		70
		70	
4	280		70
		80	
5	360		72

Finally, on your fifth and last test, your score improves once again, this time to 80. This time, you've scored *higher* than your previous average, pulling your average up from 70 to 72.

This example may be easy to understand because you are used to figuring out your average score in a course as you take additional exams. But the relationship between marginal and average spelled out here is universal: It is the same for grade point averages, batting averages—*and* costs.

Average and Marginal Cost

Now let's apply these insights to a firm's cost curves. We'll start with the relationship between the *MC* and *AVC* curves, because both curves reflect changes in the costs of variable inputs only. We already know that marginal cost first decreases and then increases. At low levels of output, as marginal cost decreases, it is *lower* than average variable cost, so it will pull the average down: *AVC* decreases. But then marginal cost rises (due to diminishing returns to labour). Eventually it rises above *AVC*, pulling the average up: *AVC* rises. Because *AVC* first decreases and then rises, the *AVC* curve is U-shaped.

The U-shape of the AVC *curve results from the U-shape of the* MC *curve, which in turn is based on increasing and then diminishing marginal returns to labour.*

There is a similar relationship between *MC* and *ATC*, except for one additional complication. *ATC* is the sum of *AVC* and *AFC*. *AFC always* falls as output rises. So at low levels of output, when both *AVC* and *AFC* are falling, *ATC* decreases— even more rapidly than AVC does. When *AVC* starts to rise, the rising *AVC* and falling *AFC* compete with each other. But eventually, the rise in *AVC* wins out, and *ATC* begins to rise as well. This explains why the *ATC* curve is U-shaped.

The U shape of the ATC *curve results from the behaviour of both* AVC *and* AFC*. At low levels of output,* AVC *and* AFC *are both falling, so the* ATC *curve slopes downward. At higher levels of output, rising* AVC *overcomes falling* AFC*, and the* ATC *curve slopes upward.*

Conclusion

The relationships tell us something important about the crossing point between the *MC* curve and the average curves in Figure 2. Whenever the *MC* curve is below one of the average curves, the average curve slopes downward. Whenever the *MC* curve is above the average curve, the average curve slopes upward. Therefore, when *MC* goes from below the average to above the average—that is, where the *MC* curve *crosses* the average curve—the average must be at its very *minimum* (where it changes from a downward slope to an upward slope).

The MC *curve crosses both the* AVC *curve and the* ATC *curve at their respective minimum points.*

If you look at Table 3 you'll see that when Spotless's output rises from 30 to 90, *MC* is below *AVC*, and *AVC* falls. When output rises from 90 to 130, *MC* is above *AVC*, and *AVC* rises. As a result, in Figure 2, the *MC* curve crosses the *AVC* curve where *AVC* bottoms out. The same relationship holds for the *MC* and *ATC* curves. But because of the competing affects of *AFC* and *AVC* on *ATC*, it takes longer for the *ATC* curve to hit bottom than the *AVC* curve. That's why minimum *ATC* occurs at a higher output than does minimum *AVC*.

Questions

1. Define Cost.
2. What is meant by Sunk Costs.
3. Distinguish between Explicit and Implicit Costs.
4. What are Short Run costs?
5. Draw Marginal Cost Curve.
6. Explain the Relationship between Average and Marginal Costs.

This lesson will focus on the following:

1. Production and Cost in Long Run
2. Relationship between Long-Run and Short-Run Costs
3. Explaining the Shape of *LRATC* Curve

Production and Cost in the Long Run

Most of the business firms you have contact with—such as your supermarket, the stores where you buy clothes, your telephone company, and your Internet service provider—plan to be around for quite some time. They have a long-term planning horizon, as well as a short-term one. But so far, we've considered the behaviour of costs only in the short run.

In the long run, costs behave differently, because the firm can adjust *all* of its inputs in any way it wants: *In the long run, there are no fixed inputs or fixed costs; all inputs and all costs are variable.*

How will the firm choose the inputs to use for any given output level? It will follow the *least cost rule: To produce any given level of output, the firm will choose the input mix with the lowest cost.*

Let's apply the least cost rule to Spotless Car Wash. Suppose we want to know the cost of washing 196 cars per day. In the short run, of course, Spotless does not have to worry about what input mix to use: It is stuck with one automated line, and if it wants to wash 196 cars, it must hire six workers (see Table 2 previous lesson 11). Total cost in the short run will be 6 x \$120 + \$150 = \$870.

In the long run, however, Spotless can vary the number of automated lines as well as the number of workers. Suppose, based on its production technology, Spotless can use four different input combinations to wash 196 cars per day. These are listed in Table 4. Combination *A* uses the least capital and the most labour—no automated lines at all and nine workers washing the cars by hand. Combination *D* uses the most capital and the least labour—three automated lines with only three workers.

Table 1: Four Ways to Wash 196 Cars per Day

Method	Quantity of Capital	Quantity of Labor	Cost
A	0	9	$1,080
B	1	6	$ 870
C	2	4	$ 780
D	3	3	$ 810

Since each automated line costs $150 per day and each worker costs $120 per day, it is easy to calculate the cost of each production method. Spotless will choose the one with the lowest cost: combination *C,* with two automated lines and four workers, for a total cost of $780 per day.

Retracing our steps, we have found that if Spotless wants to wash 196 cars per day, it will examine the different methods of doing so and select the one with the least cost. Once it has determined the cheapest production method, the other, more expensive methods can be ignored.2

Table 5 shows the results of going through this procedure for several different levels of output. The second column, **long-run total cost (LRTC)**, tells us the cost of producing each quantity of output *when the least-cost input mix is chosen.* For each output level, different production methods are examined, the cheapest one is chosen, and the others are ignored.

Notice that the *LRTC* of zero units of output is $0. This will always be true for any firm. In the long run, all inputs can be adjusted as the firm wishes, and the cheapest way to produce zero output is to use *no* inputs at all. (For comparison, what is the *short*-run total cost of producing zero units? Why can it never be $0?)

The third column in Table 2 gives the **long-run average total cost (LRATC),** the cost per unit of output in the long run:

$$LRATC = LRTC \div Q$$

Long-run average total cost is similar to average total cost, which was defined earlier. Both are obtained by dividing total cost by the level of output. There is one important difference, however: To calculate *ATC,* we used total cost (*TC*), which pertains to the short run, in the numerator. In calculating *LRATC,* we use *long-run* total cost (*LRTC*). Thus, *LRATC* tells us the cost per unit when the firm can vary *all* of its inputs and always chooses the cheapest input mix possible. *ATC,* however, tells us the cost per unit when the firm is stuck with some collection of fixed inputs and is able only to vary its remaining inputs, such as labour.

Long-run total cost The cost of producing each quantity of output when all inputs are variable and the least-cost input mix is chosen.

Long-run average total cost The cost per unit of producing each quantity of output in the long run, when all inputs are variable.

Table 2: Long-Run Costs for Spotless Car Wash

Output	*LRTC*	*LRATC*
0	$ 0	—
30	$ 200	$6.67
90	$ 390	$4.33
130	$ 510	$3.92
160	$ 608	$3.80
184	$ 720	$3.91
196	$ 780	$3.98
250	$1,300	$5.20
300	$2,400	$8.00

Relationship between Long-Run and Short-Run Costs

If you compare Table 2 (long run) with Table 2 previous lesson 11 (short run), you will see something important: For some output levels, *LRTC* is smaller than *TC.* For example, Spotless can wash 196 cars for an *LRTC* of \$780. But earlier, we saw that in the short run, the *TC* of washing these same 196 cars was \$870. There is a reason for this difference.

Look back at Table 1, which lists the four different ways of washing 196 cars per day. In the short run, the firm is stuck with just one automated line, so its only option is method *B.* In the long run, however, the firm can choose any of the four methods of production, including method *C,* which is cheapest. The freedom to choose among different production methods usually enables the firm to select a cheaper input mix in the long run than it can in the short run. Thus, in the long run, the firm may be able to save money.

But not always. At some output levels, the freedom to adjust all inputs doesn't save the firm a dime. In our example, the long-run cost of washing 130 cars is \$510—the same as the short-run cost (compare Table 2 of lesson 12 and Table 2 previous lesson 11). For this output level, it just so happens that the least-cost output mix uses one automated line, which is what Spotless is stuck with in the short run. So if Spotless wants to wash 130 cars, it cannot do so any more cheaply in the long run than in the short run.

More generally, *the long-run total cost of producing a given level of output can be less thanor equal to, but not greater than, the short-run total cost (*LRTC TC).

We can also state this relationship in terms of *average* costs. That is, we can divide both sides of the inequality by *Q* and obtain *LRTC / Q TC / Q*. Using our definitions, this translates to *LRATC ATC.*

The long-run average cost of producing a given level of output can be less than or equal to, but not greater than, the short-run average total cost (LRATC ATC).

Average Cost and Plant Size

Often, economists refer to the collection of inputs that are fixed in the short run as the firm's plant. For example, the plant of a computer manufacturer such as Dell might include its factory buildings and the assembly lines inside them. The plant of the Hertz car-rental company would include all of its automobiles and rental offices.

For Spotless Car Wash, we've assumed that the plant is simply the company's capital equipment the automated lines for washing cars. If Spotless were to add to its capital, then each time it acquired another automated line, it would have a different, and larger, plant. Viewed in this way, we can distinguish between the long run and the short run as follows: *In the long run, the firm can change the size of its plant; in the short run, it is stuck with its current plant.*

Now think about the *ATC* curve, which tells us the firm's average total cost in the short run. This curve is always drawn for a specific plant. That is, the *ATC* curve tells us how average cost behaves in the short run, *when the firm uses a plant of a given size*. If the firm had a different-size plant, it would be moving along a different *ATC* curve.

In fact, there is a different *ATC* curve for each different plant the firm could have. In the long run, then, the firm can choose to operate on *any* of these *ATC* curves. To produce any level of output, it will always choose that *ATC* curve— among all of the *ATC* curves available—with the lowest possible average total cost. This insight tells us something about the relationship between the firm's *ATC* curves and its *LRATC* curve.

Graphing the *LRATC* Curve

Look at Figure 1, which shows several different *ATC* curves for Spotless Car Wash. There is a lot going on in this figure, so let's take it one step at a time. First, find the curve labeled *ATC*1. This is our familiar *ATC* curve—the same one shown in

Figure 1—which we used to find Spotless's average total cost in the short run, when it was stuck with one automated line.

The other *ATC* curves refer to *different* plants that Spotless *might* have had instead. For example, the curve labeled *ATC*0 shows how average total cost would behave if Spotless had a plant with *zero* automated lines washing all cars manually.

*ATC*2 shows average total cost with *two* automated lines, and so on. Since, in the long run, the firm can choose which size plant to operate, it can also choose on which of these *ATC* curves it wants to operate. And, as we know, in the long run, it will always choose the plant with the lowest possible average total cost for any output level it produces.

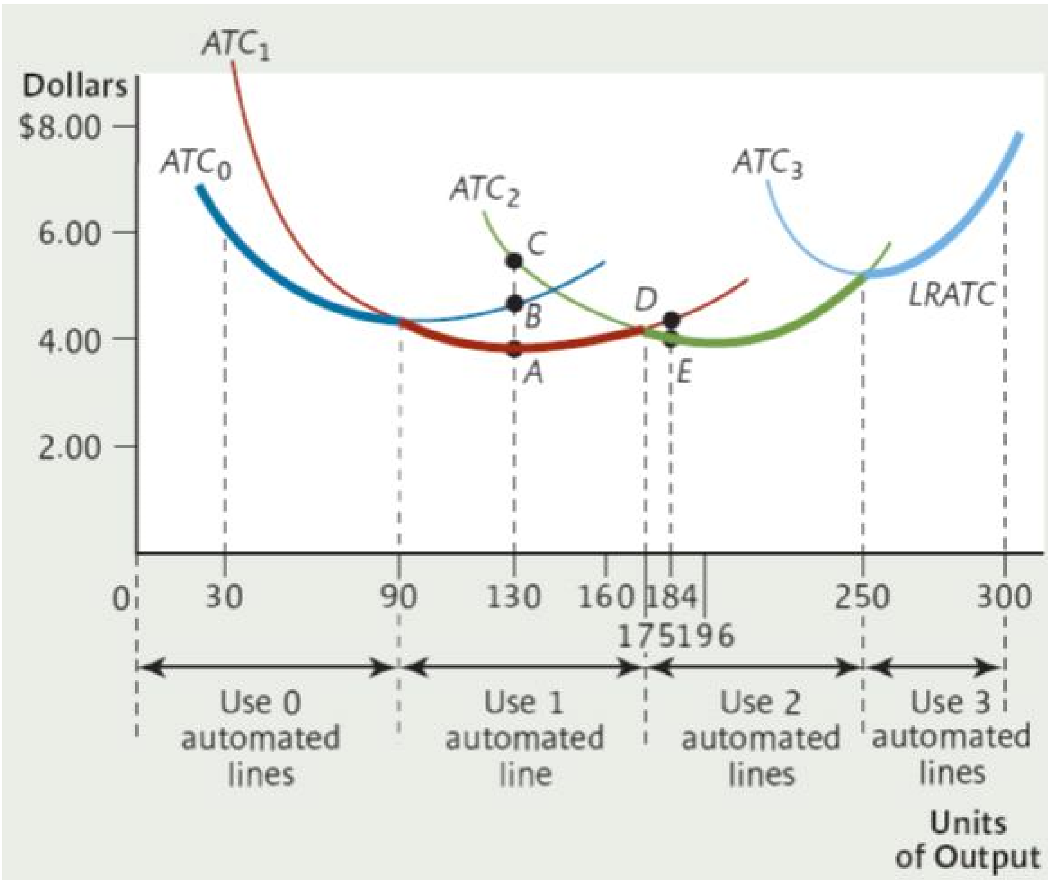
Let's take a specific example. Suppose that Spotless is planning to wash 130 cars per day. In the long run, what size plant should it choose? Scanning the different *ATC* curves in Figure 1, we see that the lowest possible per-unit cost—\$3.92 per car—is at point *A* along *ATC*1. The best plant for washing 130 cars per day, therefore, will have just one automated line.

For this output level, Spotless would never choose a plant with zero lines, because it would then have to operate on *ATC*0 at point *B*. Since point *B* is higher than point *A*, we know that point *B* represents a larger per-unit cost. Nor would the firm choose a plant with two lines—operating on *ATC*2 at point *C*—for this would mean a still larger per-unit cost. Of all the possibilities for producing 130 units in the long run, Spotless would choose to operate at point *A* on *ATC*1. So point *A* represents the *LRATC* of 130 units.

Now, suppose instead that Spotless wanted to produce 184 units of output in the long run. A plant with one automated line is no longer the best choice. Instead, the firm would choose a plant with *two* automated lines. How do we know? For an output of 184, the firm could choose point *D* on *ATC*1, or point *E* on *ATC*2. Since point *E* is lower, it is the better choice. At this point, average total cost would be \$1.96, so this would be the *LRATC* of 184 units.

Plant The collection of fixed inputs at a firm's disposal.

Figure 1: Long-Run Average Total Cost



Average-total cost curves ATC0, ATC1, ATC2, *and* ATC3 *show average costs when the firm has zero, one, two, and three automated lines, respectively. The* LRATC *curve combines portions of all the firm's* ATC *curves. In the long run, the firm will choose the lowest cost* ATC *curve for each level of output.*

Continuing in this way, we could find the *LRATC* for *every* output level Spotless might produce. To produce any given level of output, the firm will always operate on the *lowest ATC* curve available. As output increases, it will move along an *ATC* curve until another, lower *ATC* curve becomes available—one with lower costs. At that point, the firm will increase its plant size, so it can move to the lower *ATC* curve. In the graph, as Spotless increases its output level from 90 to 175 units of output, it will continue to use a plant with one automated line and move along *ATC*1. But if it wants to produce *more* than 175 units in the long run, it will increase its plant to *two* automated lines and begin moving along *ATC*2.

Thus, we can trace out Spotless's *LRATC* curve by combining just the lowest portions of all the *ATC* curves from which the firm can choose. In Figure 4, this is the thick, scallop-shaped curve.

A firm's LRATC *curve combines portions of each* ATC *curve available to the firm in the long run. For each output level, the firm will always choose to operate on the* ATC *curve with the lowest possible cost.*

Figure 1 also gives us a view of the different options facing the firm in the short run and the long run. Once Spotless builds a plant with one automated line, its options in the short run are limited: It can only move along *ATC*1. If it wants to increase its output from 130 to 184 units, it must move from point *A* to point *D*. But in the long run, it can move along its *LRATC* curve—from point *A* to point *E*—by changing the size of its plant.

More generally, *in the short run, a firm can only move along its current* ATC *curve. In the long run, however, it can move from one* ATC *curve to another by varying the size of its plant. As it does so, it will also be moving along its* LRATC *curve.*

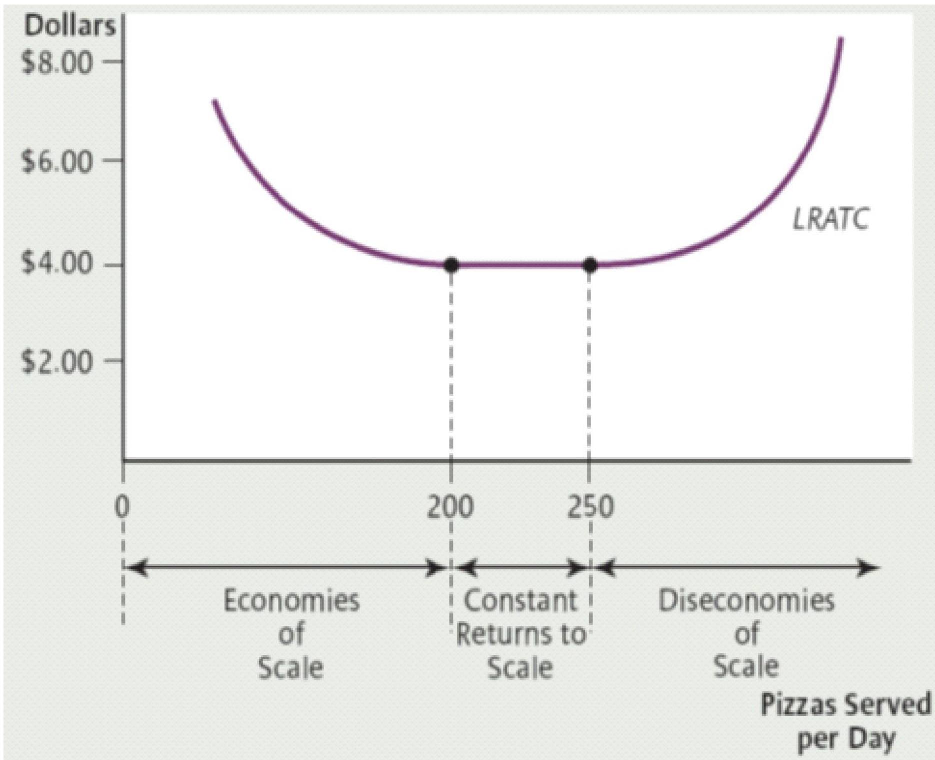
Explaining Shape of *LRATC* Curve

In Figure 1, the *LRATC* curve has a scalloped look because the firm can only choose among four different plants. But many firms can adjust their plant size in smaller increments. Each different plant size would be represented by a different *ATC* curve, so there would be hundreds of *ATC* curves crowded into the figure. As a result, the scallops would disappear, and the *LRATC* curve would appear as a smooth curve.

Figure 2 shows what the *LRATC* curve might look like for Mike's Pizza Restaurant. The horizontal axis measures the number of pizzas served per day. The vertical axis measures cost per pizza. Note that as we move along this curve, we are looking at *long-run* average total cost. In the long run, as output rises, not only can Mike's use more cooks, ingredients, and wait-staff, it can also adjust the size of its "plant"—its restaurant facility.

The *LRATC* curve for Mike's Pizza is U-shaped—much like the *AVC* and *ATC* curves you learned about earlier. That is, as output increases, long-run average costs first decline, then remain constant, and finally rise. Although there is no law or rule of logic that requires an *LRATC* curve to have all three of these phases, in many industries this seems to be the case. Let's see why, by considering each of the three phases in turn.

Figure 2: Shape of LRATC



If long-run total cost rises proportionately less than output, production reflects economies of scale, and LRATC *slopes downward. If cost rises proportionately more than output, there are diseconomies of scale, and* LRATC *slopes upward. Between those regions, cost and output rise proportionately, yielding constant returns to scale.*

Economies of Scale

When an increase in output causes *LRATC* to decrease, we say that the firm is enjoying **economies of scale:** The more output produced, the lower the cost per unit.

Mike's Pizza has economies of scale for all output levels up to 200. On a purely mathematical level, economies of scale means that long-run total cost is rising by a smaller proportion than output. For example, if a doubling of output (Q) can be accomplished with less than a doubling of costs, then the ratio $LRTC / Q = LRATC$ will decline. Is this economies of scale?

When long-run total cost rises proportionately less than output, production is characterized by economies of scale, and the LRATC *curve slopes downward.* But in the real world, *why* should total costs ever increase by a smaller proportion than output? Why should a firm experience economies of scale?

Gains from Specialization

One reason for economies of scale is gains from specialization. At very low levels of output, workers may have to perform a greater variety of tasks, slowing them down and making them less productive. But as output increases and workers are added, more possibilities for specialization are created.

For example, at low levels of output, Mike's Pizza might have a very small facility with just one employee. This one worker would do everything himself: cook the pizzas, take orders, clean the tables, accept payments, order ingredients, and so on.

But as output expands, Mike can run a larger operation with more workers, each specializing in one of these tasks. Since each worker is more productive, output will increase by a greater proportion than costs.

You've learned that increased specialization also plays a role in costs in the short run: it is one of the reasons why marginal cost (and therefore average costs) can decrease as output expands from low levels. But the ability of specialization to reduce costs is even greater in the long run. Remember that, in the short run, output expands by adding more and more variable inputs to unchanging amounts of fixed inputs. At some point, the fixed inputs cause diminishing returns to set in, overwhelming any further gains from specialization. In the long run, however, *all* inputs can be increased as output expands—including factory size, capital equipment, managers, and more. This opens up many more ways to re-arrange production to take full advantage of specialization.

The greatest opportunities for increased specialization occur when a firm is starting at a relatively low level of output, with a relatively small plant and small workforce. Thus, economies of scale are more likely to occur at lower levels of output.

Spreading Costs of Lumpy Inputs

Another explanation for economies of scale involves the "lumpy" nature of many types of plant and equipment. **Lumpy inputs** are inputs that cannot be increased in tiny increments, but rather must be increased in large jumps. In some cases, a minimal amount of the inputs is needed to produce any output at all.

A medical practice, for example, needs the use of at least one X-ray machine in order to serve patients. And it must buy a *whole* machine, not a half or a fifth of an X-ray machine. The more patients the practice serves, the lower will be the cost of the machine per patient.

We see this phenomenon in many types of businesses: Plant and equipment must be purchased in large lumps, and a low cost per unit is achieved only at high levels of output. Other inputs besides equipment can also be lumpy in this way. A theater must have at least one ticket taker and one projectionist, regardless of how many people come to see the show. A restaurant must pay a single license fee to the city each year, no matter how many meals it serves. In all of these cases, an increase in output allows the firm to spread the cost of lumpy inputs over greater amounts of output, lowering the cost *per unit of output*.

Economies of scale Long-run average total cost decreases as output increases.

Lumpy input An input whose quantity cannot be increased gradually as output increases, but must instead be adjusted in large jumps.

Spreading the costs of lumpy inputs has more impact on *LRATC* at low levels of output when these costs make up a greater proportion of the firm's total costs. At higher levels of output, the impact is smaller. For example, suppose Mike's restaurant must pay a yearly license fee of $3,650, which amounts to $10 per day. If output doubles from 10 to 20 pizzas per day, license costs per meal served will fall from $1 to $0.50. But if output doubles from 200 to 400, license costs per meal drop from $0.05 to $0.025—a hardly noticeable difference. Thus, spreading lumpy inputs across more output—like the gains from specialization—is more likely to create economies of scale at relatively low levels of output. This is another reason why the typical *LRATC* curve—as illustrated in Figure 4—slopes downward at relatively low levels of output.

Diseconomies of Scale

As output continues to increase, most firms will reach a point where bigness begins to cause problems. Large firms may require more layers of management, so communication and decision making become more time consuming and costly. Huge corporations like Ford, Microsoft, and Verizon each have several hundred high level managers, and thousands more at lower levels.

Large firms may also have a harder time screening out misfits among new hires and monitoring those already working at the firm. This leads to more mistakes, shirking of responsibilities, and even theft from the firm. If Mike expands his facility so he can serve hundreds of pizzas per day, with dozens of employees, some of them might start sneaking pizzas home at the end of the day, others might take extra long breaks without anyone noticing, and so on. As output continues to rise and the firm has exhausted the cost-saving opportunities from increasing its scale of operations, these sorts of problems will eventually dominate, causing *LRATC* to rise.

When LRATC rises with an increase in output, we have **diseconomies of scale.** Mathematically, *when long-run total cost rises more than in proportion to output, there are diseconomies of scale, and the* LRATC *curve slopes upward.*

While economies of scale are more likely at low levels of output, *dis*economies of scale are more likely at higher output levels. In Figure 6, Mike's Pizza does not experience diseconomies of scale until it is serving more than 250 pizzas per day.

Constant Returns to Scale

In Figure 2, for output levels between 200 and 250, the smoothed-out *LRATC* curve is roughly flat. Over this range of output, *LRATC* remains approximately constant as output increases. Here, output and *LRTC* rise by roughly the same proportion:

When both output and long-run total cost rise by the same proportion, production is characterized by ***constant returns to scale,*** *and the* LRATC *curve is flat.*

Why would a firm experience constant returns to scale? We have seen that as output increases, cost savings from specialization and spreading the costs of lumpy inputs will eventually be exhausted. But production may still have room to expand before the costly problems of "bigness" kick in. The firm will then have a range of output over which average cost neither rises nor falls as production increases— constant returns to scale. Notice that constant returns to scale, if present at all, are most likely to occur at some *intermediate* range of output.

In sum, when we look at the behaviour of *LRATC,* we often expect a pattern like the following: economies of scale (decreasing *LRATC*) at relatively low levels of output, constant returns to scale (constant *LRATC*) at some intermediate levels of output, and diseconomies of scale (increasing *LRATC*) at relatively high levels of output. This is why *LRATC* curves are typically U-shaped. Of course, even U-shaped *LRATC* curves will have different appearances for firms in different industries. And as you're about to see, these differences in *LRATC* curves have much to tell us about the economy.

Diseconomies of scale Long-run average total cost increases as output increases.

Constant returns to scale Long run average total cost is unchanged as output increases.

Cost: A Summary

This lesson has presented a number of new terms and concepts. As you first learn them, it's easy to get them confused. Table 3 provides a useful summary, which you can use both as a reference and a self-test.

Table 3: Types of Costs

Term	Symbol and/or Formula	Definition
Explicit cost		A cost where an actual payment is made
Implicit cost		An opportunity cost, but no actual payment is made
Sunk cost		An irrelevant cost because it cannot be affected by any current or future decision
Lumpy input cost		The cost of an input that can only be adjusted in large, indivisible amounts
Short-run costs		
Total fixed cost	*TFC*	The cost of all inputs that are fixed (cannot be adjusted) in the short run
Total variable cost	*TVC*	The cost of all inputs that are variable (can be adjusted) in the short run
Total cost	$TC = TFC + TVC$	The cost of all inputs in the short run
Average fixed cost	$AFC = TFC / Q$	The cost of all fixed inputs per unit of output
Average variable cost	$AVC = TVC/Q$	The cost of all variable inputs per unit of output
Average total cost	$ATC = TC/Q$	The cost of all inputs per unit of output
Marginal cost	$MC = \Delta TC/\Delta Q$	The change in total cost for each one-unit rise in output
Long-run costs		
Long-run total cost	*LRTC*	The cost of all inputs in the long run, using the least-cost method of producing any given output level
Long-run average	$LRATC = LRTC/Q$	Cost per unit in the long run, using the least-cost method of producing any given output level

Conclusion

Business firms combine inputs to produce outputs. A firm's production *technology* determines the maximum output it can produce using different quantities of inputs. In the *short run,* at least one of the firm's inputs is fixed. In the *long run,* all inputs can be varied.

A firm's *cost of production* is the opportunity cost of its owners—everything they must give up in order to produce output. In the short run, some costs are *fixed* and independent of the level of production. Other costs— *variable costs*—can change as production changes.

Marginal cost is the change in total cost from producing one more unit of output. The *marginal cost curve* has a U shape, reflecting the underlying marginal product of labour. A variety of average cost curves can be defined. The *average variable cost curve* and the *average total cost curve* are each U-shaped, reflecting the relationship between average and marginal cost the marginal cost curve must cross each of the average curves at their minimums.

In the long run, all costs are variable. The firm's *long run total cost curve* indicates the cost of producing each quantity of output with the least-cost input mix. The related *long-run average total cost (LRATC) curve* is formed by combining portions of different *ATC* curves, each portion representing a different plant size. The *LRATC* curve slopes downward when there are economies of scale, slopes upward when there are diseconomies of scale, and is flat when there are constant returns to scale. Economies of scale can play a role in explaining mergers and acquisitions, especially when there are too many firms for each to operate at its *minimum efficient scale*.

Questions

1. Explain Production and Cost in Long Run.
2. Explain Relationship between Long-Run and Short-Run Costs.
3. Explain the Shape of *LRATC* Curve.

This lesson will focus on the following:

1. Concept of Isoquant and its Properties
2. Marginal Rate of Technical Substitution
3. Isocost Lines and Properties of Isocost Lines
4. Least-Cost Input Combination
5. Firms Decisions: Goal of Profit Maximization
6. Firm's Constraints: Demand Curve Facing Firm, Cost Constraint

Concept of Isoquant and its Properties

Imagine that you own an artichoke farm, and you are free to vary two inputs: labour and land. Your output is measured in "boxes of artichokes per month." Your farm's production technology determines the maximum possible number of boxes you could produce in a given month using different combinations of labour and land.

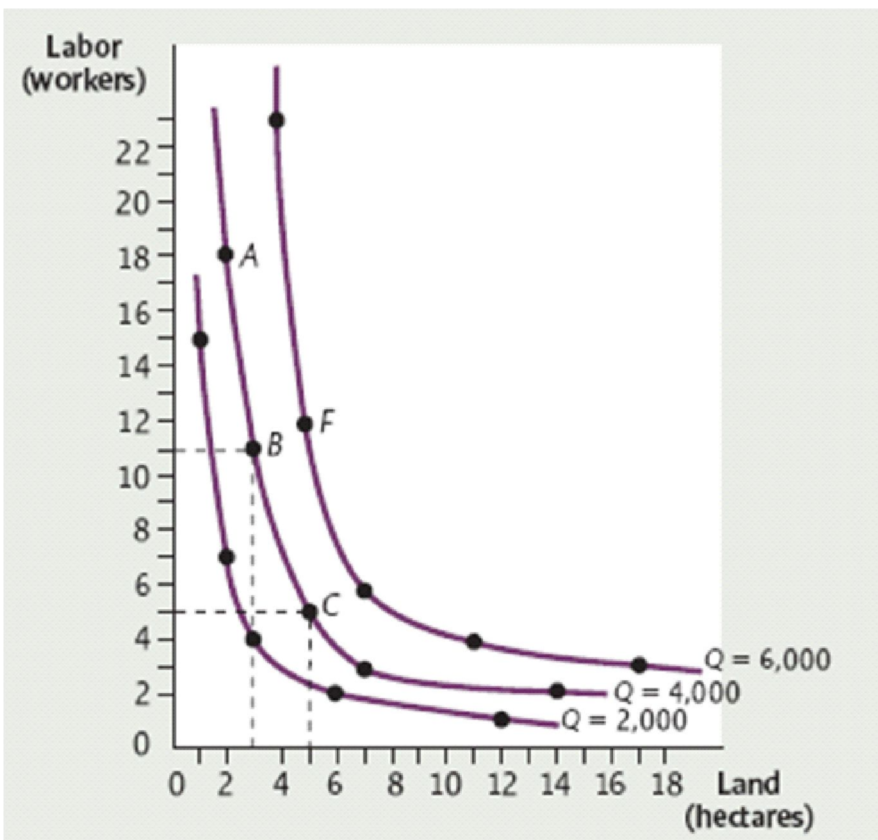
Alternatively, it tells us all the different input mixes that could be used to produce any given quantity of output. Table 1 lists some of the information we could obtain based on the technology of production on your farm. Notice that, to produce each of the three output levels included in the table, there are many different combinations of inputs you could use. For example, the table tells us that your farm could produce 4,000 boxes of artichokes using 2 hectares of land and 18 workers, or 3 hectares and 11 workers, or 5 hectares and 5 workers, and so on.

The information in the table can also be illustrated with a graph. In Figure 1, the quantity of land is plotted along the horizontal axis, and the number of workers on the vertical axis. Each combination of the two inputs is represented by a point. For example, the combination *3 hectares, 11 workers* is represented by the point labeled *B*, while the combination *5 hectares, 12 workers* is represented by point *F*.

Table 1: Production Technology of a Firm

2,000 Boxes of Artichokes per Month		4,000 Boxes of Artichokes per Month		6,000 Boxes of Artichokes per Month	
Hectares of Land	*Number of Workers*	*Hectares of Land*	*Number of Workers*	*Hectares of Land*	*Number of Workers*
1	15	2	18	4	23
2	7	3	11	5	12
3	4	5	5	7	6
6	2	7	3	11	4
12	1	14	2	17	3

Figure 1: An Isoquant Map



Each of the curves in the figure is an isoquant, showing all combinations of labour and land that can produce a given output level. The middle curve, for example, shows that 4,000 units of output can be produced with 11 workers and 3 hectares of land (point B*), with 5 workers and 5 hectares of land (point* C*), as well as other combinations of labour and land. Each isoquant is drawn for a different level of output. The higher the isoquant line, the greater the level of output.*

Now let's focus on a single output level: 4,000 boxes per month. The middle columns of Table 1 shows 5 of the different input combinations that can produce this output level, each represented by a point in Figure 1.

When we connect all 5 points with a smooth line we get the curve labeled "Q _ 4,000" in Figure 1. This curve is called an **isoquant** ("iso" means "same," and "quant" stands for "quantity of output").

Every point on an isoquant represents an input mix that produces the same quantity of output.

Figure 1 also shows two additional isoquants. The higher one is drawn for the output level $Q = 6{,}000$, and the lower one for the output level $Q = 2{,}000$. When these curves are shown together on a graph, we have an **isoquant map** for the firm.

Knowing Isoquants

As we move along any isoquant, the quantity of output remains the same, but the combination of inputs changes. More specifically, as we move along an isoquant, we are *substituting one input for another*. For example, as we from point B to point C along the isoquant labeled $Q = 4{,}000$, the quantity of land rises from 3 to5 hectares, while the number of workers falls from 11 to 5.

You are substituting land for labour, while maintaining the same level of output. Since each of the two inputs contributes to production, every time you increase one input, you must decrease the other in order to maintain the same level of output.

An increase in one input requires a decrease in the other input to keep total production unchanged. This is why isoquants always slope downward.

What happens as we move from isoquant to isoquant?

Whenever we move to a higher *isoquant* (moving northeasterly in Figure 1), the quantity of output increases. Moving directly northward means you are using more labour with the same amount of land, and moving directly eastward means you are using more land with the same amount of labour. When you move both north and east simultaneously (as in the move from point *B* to point *F*), you are using more of *both* inputs. For all of these movements, output increases. For the same reason, if we move southwestward, output decreases.

Marginal Rate of Technical Substitution

The (absolute value of the) slope of an isoquant is called the **marginal rate of technical substitution (MRTS).** As the name suggests, it measures the rate at which a firm can substitute one input for another while keeping output constant. In our example, if we use "L" for labour and "N" for land, the $\text{MRTS}_{L,N}$ tells us how many *fewer* workers you can employ each time you use *one more hectare of land*, and still maintain the same level of output.

For example, if you move from point *A* to point *B* along isoquant $Q = 4{,}000$, you use 1 more hectare and 7 fewer workers, so the $\text{MRTS}_{L,N} = 7/1 = 7$ for that move. Going from point *B* to point *C,* you use 2 more hectares of land, and 6 fewer workers, so the $\text{MRTS}_{L,N} = 6/2 = 3$.

Using this new term, the changing slope of an isoquant can be expressed this way:

As we move rightward along any given isoquant, the marginal rate of technical substitution decreases.

But why does the $\text{MRTS}_{L,N}$ decrease? To answer this question, it helps to understand the relationship between the *MRTS* and the *marginal products* of land and labour. You've already learned that the marginal product of labour (MP_L) is a firm's additional output when one more worker is hired and all other inputs remain constant.

The marginal product of land (MP_N, using "N" for land) is defined in a similar way: It's the additional output a firm can produce with one additional unit of land (one more hectare, in our example), holding all other inputs constant.

Suppose that, starting from a given input mix, you discover that your MP_N is 21 boxes of artichokes, and your MP_L is 7 boxes. Then conduct the following mental experiment: Add one hectare of land, with no change in labour, and your output *increases* by 21 boxes. Then, give up 3 workers, with no change in land, and your output *decreases* by 3 x 7 = 21 boxes. In this case, adding 1 hectare of land, and hiring 3 fewer workers leaves your output unchanged. The slope of the isoquant for a move like this would be $\Delta L/\Delta N = -3/1 = -3$.

More generally, each time we change the amount of labour (L), the firm's output will change by $\Delta L \times MPL$. Each time we change the firm's land (N), the change in output will be $\Delta N \times MPN$. If we want the net result to be zero change in output, we must have

$$\Delta L \times MPL + \Delta N \times MPN = 0$$

$$\text{or } \Delta L \times MPL = -\Delta N \times MPN.$$

Rearranging this equation gives us:

$$\Delta L/\Delta N = -MPN/MPL.$$

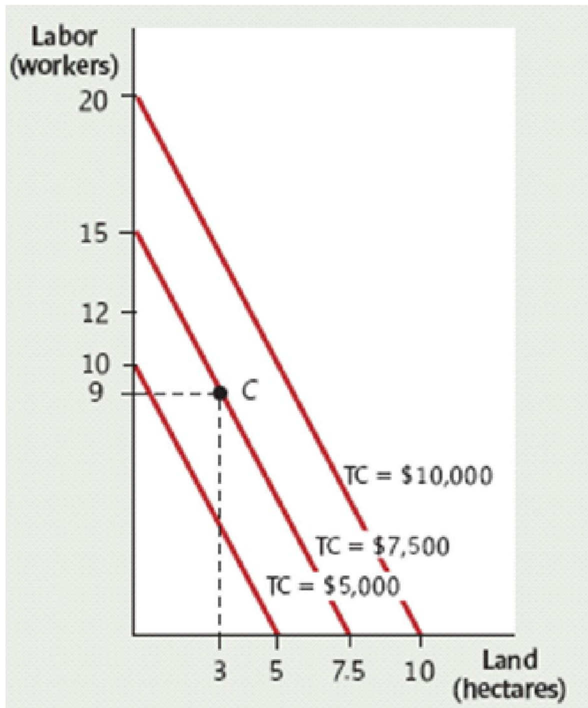
The left-hand side is the ratio of the change in labour to the change in land needed to keep output unchanged, that is, *the slope of the isoquant*. The right-hand side tells us that this slope is equal to the ratio of the marginal products of land and labour, except for the sign, which is negative.

That is,

At each point along an isoquant with land measured horizontally, and labour measured vertically, the (absolute value of the) slope of the isoquant, which we call the MRTSL,N, *is the ratio of the marginal products,* MPN/MPL.

Now, what does this have to do with the shape of the isoquant? As we move rightward and downward along an isoquant, the firm is acquiring more and more land, and using less and less labour. The marginal product of land will decrease—since land is becoming more plentiful—and the marginal product of labour will increase—since labour is becoming more and more scarce. Taken together, these changes tell us that the ratio *MPN/ MPL* must fall and so must the slope of the isoquant.

Figure 2: Isocost Lines



Each of the lines in the figure is an isocost line, showing all combinations of labour and land that have the same total cost. The middle line, for example, shows that total cost will be $7,500 if 9 workers and 3 hectares of land are used (point C*). All other combinations of land and labour on the middle line have the same total cost of $7,500. Each isocost line is drawn for a different value of total cost. The higher the isocost line, the greater is total cost.*

An isoquant becomes flatter as we move rightward because the MPN *decreases, while the* MPL *increases, so the ratio*—MPN/MPL—*decreases.*

Isocost Lines

An isoquant map shows us the different input mixes capable of producing different amounts of output. But how should the firm *choose* among all of these input mixes? In order to answer that question, we must know something about input *prices*. After all, if you own an artichoke farm, you must *pay* for your land and labour. To keep the math simple, let's use round numbers. We'll suppose that the price of labour—the wage—is$500 per month

(PL = \$500), and the price of land— what you must pay in rent to its owner, or your implicit cost if you own the land yourself—is \$1,000 per hectare per month (PN = \$1,000).

An **isocost line** ("same cost" line) tells us all combinations of the two inputs that would require the same total outlay for the firm. It is very much like the *budget line* you learned about in Chapter 5, which showed all combinations of two *goods* that resulted in the same cost for the consumer.

The difference is that an isocost line represents total cost to a *firm* rather than a consumer, and is based on paying for *inputs* rather than goods.

Figure 2 shows three isocost lines for your artichoke farm. The middle line (labeled TC = \$7,500) tells us all combinations of land and labour that would cost \$7,500 per month. For example, point G represents the combination *3 hectares, 9 workers,* for a total cost of 3 x \$1,000 = 9 x \$500 = \$7,500.

Knowing Isocost Lines

Notice that all three isocost lines in Figure 2 *slope downward.* Why is this? As you move rightward in the figure, you are using more land. If you continued to use an unchanged amount of labour, your cost would therefore increase. But an isocost line shows us input combinations with the *same* cost. Thus, to keep your cost unchanged as you use more land (move rightward), you must also employ *fewer* workers (move downward).

If you use more of one input, you must use less of the other input in order to keep your total cost unchanged. This is why isocost lines always slope downward.

Notice, though, that the *slope* of the isocost line remains *constant* as we move along it. That is, isocost lines are *straight lines.* Why? Let's find an expression for the slope of the isocost line. Each time you change the number of workers by ΔL, your total cost will change by $P_L \times \Delta L$. Each time you change the amount of land you use by ΔN, your total cost will change by $P_N \times \Delta N$. In order for your total cost to remain the same as you change the amounts of both land and labour, the changes must satisfy the equation:

$$P_L \times \Delta L + P_N \times \Delta N = 0,$$

or

$$P_L \times \Delta L = - P_N \times \Delta N$$

which can be rearranged to

$\Delta L/\Delta N = - P_N / P_L$

The term on the left is the change in labour divided by the change in land that leaves total cost unchanged—the slope of the isocost line. The term on the right is the (negative of the) ratio of the inputs' prices. In our example, with P_N = $1,000 and P_L = $500, the slope of the isocost line is - $1,000/$500 = - 2.

Now you can see why the isocost line is a straight line: As long as the firm can continue to buy its inputs at unchanged prices, the ratio $- P_N / P_L$ will remain constant. Therefore, the slope of the isocost line will remain constant as well.

The slope of an isocost line with land (N) *on the horizontal axis and labour* (L) *on the vertical axis is* - P_N / P_L. *This slope remains constant as we move along the line.*

Finally, there is one more thing to note about isocostlines. As you move in a northeasterly direction in Figure 2, to higher isocost lines, you are paying for greater amounts of land and labour, so your total cost must rise. For the same reason, as you move in a southwesterly direction, you are paying for smaller amounts of land and labour, so your total costs fall.

Higher isocost lines represent greater total costs for the firm than lower isocost lines.

In Figure 2, the highest line represents all inputs combinations with a total cost of $10,000, and the lowest line represents all combinations with a total cost of$5,000.

Least-Cost Input Combination

Now we are ready to combine what we know about a firm's production—represented by its isoquants—with our knowledge of the firm's costs—represented by its isocost lines. Together, these will allow us to find the least-cost input combination for producing any level of output a firm might choose to produce.

Suppose you want to know what is the best way to produce 4,000 boxes of artichokes per month. Figure 3 reproduces the isoquant labeled $Q = 4{,}000$ from Figure 1, along with the three isocost lines from Figure 2. You would like to find the input combination that is *capable* of producing 4,000 boxes (an input combination *on the isoquant $Q = 4{,}000$*), with the lowest possible cost (an input combination on the lowest possible isocost line).

As you can see in the diagramme, there is only one input combination that satisfies both requirements: point *C*. At this point, the firm uses 5 hectares of land, and 5 workers, for a total cost of 5 x \$1,000 + 5 x \$500 = \$7,500. As you can see, while there are other input combinations that can also produce 4,000 boxes, such as point *J* or point *K*, each of these lie on a higher isocost line (*TC* = \$10,000) and will require a greater total outlay than the least-cost combination at point *C*.

The least-cost combination will always be found where the isocost line is *tangent* to the isoquant. This is where the two lines touch each other at a single point, and both lines *have the same slope*.

The least-cost input combination for producing any level of output is found at the point where an isocost line is tangent to the isoquant for that output level.

This result will prove very useful. We already know that the slope of the isoquant at any point is equal to -*MPN*/ *MPL*. And we know that the slope of the isocost line is equal to $-P_N/P_L$. Putting the two together, we know that when you have found the least-cost input combination for any output level,

$$-MPN/MPL = -P_N/P_L$$

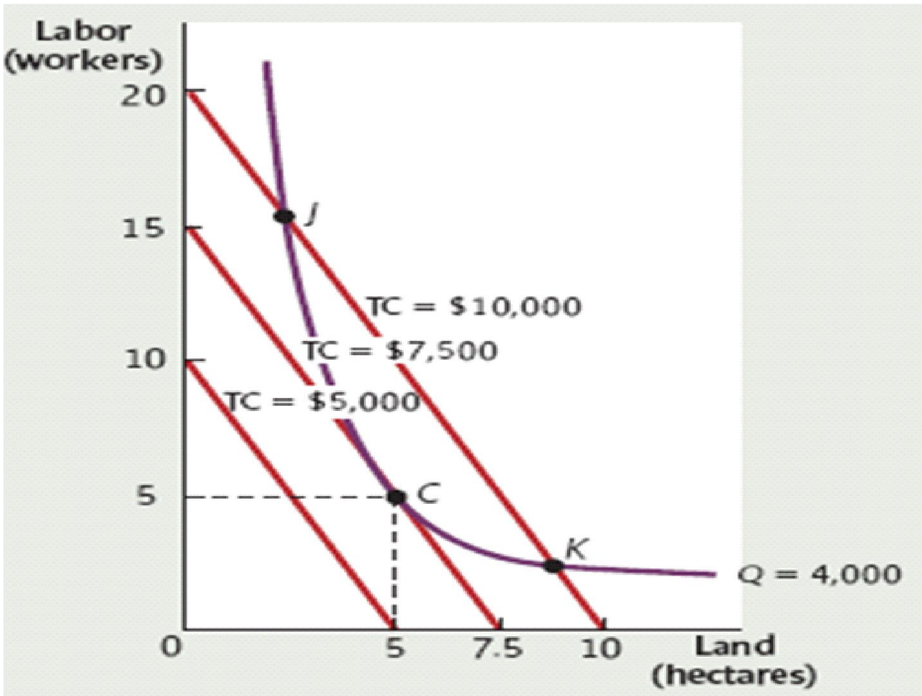
or

$$MPN/MPL = P_N/P_L.$$

The term on the left-hand side is just the $\text{MRTS}_{L,N}$ the marginal rate of technical substitution between labour and land. We conclude that:

When a firm is using the least-cost combination of two inputs (*L and N*) *for a particular output level, the firm's* MRTS *between the two inputs* (MPN/MPL) *will equal the ratio of input prices* (P_N/P_L).

Figure 3: The Least-Cost Input Combination for a Given Output Level



To produce any given level of output at the least possible cost, the firm should use the input combination where the isoquant for that output level is tangent to an isocost line. In the figure, the input combinations at points J, C, *and* K *can all be used to produce 4,000 units of output. But the combination at point* C *(5 workers and 5 hectares of land), where the isoquant is tangent to the isocost line, is the least expensive input combination for that output level.*

In our example, $P_N / P_L = \$1{,}000/\$500 = 2$. This tells us that, at point C, the ratio $MPN/MPL = 2$ as well. Finally, we can rearrange the equation $MPN/MPL = P_N / P_L$ to get:

$$MPN/P_N = MPL/P_L.$$

This form of the equation gives us another insight. It says that when you have found the least-cost input mix for any output level, the marginal product of land divided by the price of land will be equal to the marginal product of labour divided by the price of labour.

How can we interpret the marginal product of an input divided by its price? It gives us the additional output from spending one more *dollar* on the input. For example, if the (monthly) price of a hectare of land is \$1,000, and using one more hectare increases your output by 21 boxes (*MPN* = 21), then an additional *dollar* spent on land will give you 1/1,000 of a hectare, which, in turn, will increase your output by (1/1,000) x 21 = 21/1,000 or .021 boxes. So, MPN/P_N = 21/1,000 is the additional output from one more dollar spent on land. As a kind of shorthand, we'll call *MPN/PN* the "marginal product per dollar" of land.

Using this language, we can state our result this way:

When a firm is using the least-cost combination of land and labour for any output level, the marginal product per dollar of land (MPN/P_N) *must equal the marginal product per dollar of labour* (MPL/P_L).

Case of More than Two Inputs

When a firm can vary three or more inputs, we cannot illustrate isoquants and isocost lines on a two dimensional graph. Nevertheless, the conclusions we reached for the two-input case can be generalized to any number of inputs.

Suppose a firm has several variable inputs, which we can label *A*, *B*, *C*, . . . , with marginal products *MPA*, *MPB*, *MPC*, . . . and input prices P_A, P_B, P_C, ... Then for any level of output, the least-cost combination of all of these inputs will always satisfy:

$$MPA/P_A = MPB/P_B = MPC/P_C = \ldots$$

That is, *When a firm with many variable inputs has found its least-cost input mix, the marginal product per dollar of any input will be equal to the marginal product per dollar of any other input.*

How do we know this must always be true? First, remember that MPA/P_A tells us the additional output the firm will produce *per additional dollar spent on input A*. Next, suppose we have two inputs, *A* and *B*, for which MPA/P_A is *not* equal to MPB/P_B. Then we can show that the firm can always shift its spending from one input to another, lowering its cost while leaving its output unchanged.

Let's take a specific example. Suppose that $MPA/P_A = 2$, and $MPB/P_B = 3$. Then the firm can easily save money by shifting dollars away from input *A* toward input *B*. Each dollar shifted away from *A* causes output to decrease by 2 units, while each dollar shifted toward input B causes output to rise by 3 units. Thus, the firm could shift dollars away from input A, and use only *some* of those dollars to increase the amount of input B, and still keep its production unchanged.

The same holds for any other two inputs we might compare: Whenever the marginal product per dollar is different for any two inputs, the firm can always shift its spending from the input with the lower marginal product per dollar to the input with the higher marginal product per dollar, achieving lower total cost with no change in output.

Goal of Profit Maximization

To analyze decision making at the firm, let's start with a very basic question: What is the firm trying to maximize?

Economists have given this question a lot of thought. Some firms—especially large ones—are complex institutions in which many different groups of people work together. A firm's owners will usually want the firm to earn as much profit as possible.

But the workers and managers who actually run the firm may have other agendas. They may try to divert the firm away from profit maximization in order to benefit themselves. For now, let's assume that workers and managers are faithful servants of the firm's owners. That is,

We will view the firm as a single economic decision maker whose goal is to maximize its owners' profit.

Why do we make this assumption? Because it has proven so *useful* in understanding how firms behave. True, this assumption leaves out the details of these other agendas that often are present in real-world firms. But remember that every economic model *abstracts* from reality. To stay simple and comprehensible, it leaves out many real-world details and includes only what is relevant for the purpose at hand. If the purpose is to explain conflict within the firm or deviations from profit-maximizing behaviour, then the differing goals of managers and owners should be a central element of the model.

But when the purpose is to explain how firms decide what price to charge and how much to produce, or whether to temporarily shut down the firm or continue operating, or whether to enter a new market or permanently leave a current one, the assumption of profit maximization has proven to be very useful. It explains what firms actually do with reasonable—and sometimes remarkable—accuracy.

Why? Part of the reason is that managers who deviate *too* much from profit maximizing for *too* long are typically replaced. The managers may be sacked either by the current dissatisfied owners or by other firms that acquire the underperforming firm.

Another reason is that so many managers are well trained in the tools of profit maximization. This is in contrast to our model of consumer behaviour, in which we asserted that consumers act *as if* they are using the model's graphs and calculations— although we recognize that most consumers never actually do. The basic economic model of the firm's behaviour, however, is well understood *and used* by most managers, who have often taken several economics courses as part of their management education. In fact, economists' thinking about firm behaviour has so permeated the language and culture of modern business that it's sometimes hard to distinguish where theory ends and practice begins.

Understanding Profit

Profit is defined as the firm's *sales revenue* minus its *costs of production*. There is widespread agreement over how to measure the firm's revenue—the flow of money into the firm. But there are two different conceptions of the firm's costs, and eachof them leads to a different definition of profit.

Definitions of Profit

One conception of costs is the one used by accountants. With a few exceptions, accountants consider only *explicit* costs, where money is actually paid out. If we deduct only the costs recognized by accountants, we get one definition of profit:

Accounting profit = Total revenue - Accounting costs.

But economics, as you have learned, has a much broader view of cost— *opportunity cost.* For the firm's owners, opportunity cost is the total value of *everything* sacrificed to produce output. This includes not only the explicit costs recognized by accountants—such as wages and salaries and outlays on raw materials—but also *implicit costs,* when something is given up but no money changes hands. For example, if an owner contributes his own time or money to the firm, there will be foregone wages or foregone investment income—both implicit costs for the firm.

Accounting profit Total revenue minus accounting costs. This broader conception of costs leads to a second definition of profit:

Economic profit = Total revenue - *All* costs of production

= Total revenue - (Explicit costs + Implicit costs)

The difference between economic profit and accounting profit is an important one; when they are confused, some serious (and costly) mistakes can result. An example might help make the difference clear.

Suppose you own a firm that produces T-shirts and you want to calculate your profit over the year. Your bookkeeper provides you with the following information:

Table 2: Accounting Profit

Total Revenue from Selling T-shirts		$300,000
Cost of raw materials	$ 80,000	
Wages and salaries	150,000	
Electricity and phone	20,000	
Advertising cost	40,000	
Total Explicit Cost		290,000
Accounting Profit		$ 10,000

From the looks of things, your firm is earning a profit, so you might feel pretty good. Indeed, if you look only at money coming in and money going out, you have indeed earned a profit: $10,000 for the year . . . an accounting profit.

But suppose that in order to start your business you invested $100,000 of your own money—money that *could* be earning $6,000 in interest if you sold the business and got it back. Also, you are using two extra rooms in your own house as a factory—rooms that *could* be rented out for $4,000 per year. Finally, you are managing the business full-time, without receiving a separate salary, and you could instead be working at a job earning $40,000 per year. All of these costs—the interest, rent, and salary you *could* have earned—are implicit costs that have not been taken into account by your bookkeeper. They are part of the opportunity cost of your firm because they are sacrifices you made to operate your business.

Now let's look at this business from the economist's perspective and calculate your *economic* profit.

Table 3: Economic Profit

Total Revenue from Selling T-shirts		$300,000
Cost of raw materials	$ 80,000	
Wages and salaries	150,000	
Electricity and phone	20,000	
Advertising cost	40,000	
Total Explicit Costs	$290,000	
Investment income foregone	$ 6,000	
Rent foregone	4,000	
Salary foregone	40,000	
Total Implicit Costs	$ 50,000	
Total Costs		$340,000
Economic Profit		−$ 40,000

From an economic point of view, your business is not profitable at all, but is actually losing $40,000 per year! But wait—how can we say that your firm is suffering a loss when it takes in more money than it pays out? Because, as we've seen, your *opportunity cost*—the value of what you are giving up to produce your output—includes more than just money costs. When *all* costs are considered— implicit as well as explicit—your total revenue is not sufficient to cover what you have sacrificed to run your business. You would do better by shifting your time, your money, and your spare room to some alternative use.

Which of the two definitions of profit is the correct one? Either one of them, depending on the reason for measuring it. For tax purposes, the government is interested in profits as measured by accountants. The government cares only about the money you've earned, not what you *could* have earned had you done something else with your money or your time.

However, for our purposes—understanding the behaviour of firms—economic profit is clearly better. Should your T-shirt factory stay in business? Should it expand or contract in the long run? Will other firms be attracted to the T-shirt industry? It is economic profit that will help us answer these questions, because it is economic profit that you and other owners care about.

The proper measure of profit for understanding and predicting the behaviour of firms is economic profit*. Unlike accounting profit, economic profit recognizes all the opportunity costs of production—both explicit costs and implicit costs.*

Why are there Profits?

When you look at the income received by households in the economy, you see a variety of payments. Those who provide firms with land receive *rent*—the payment for land. Those who provide labour receive a wage or salary. And those who lend firms money so they can purchase capital equipment receive interest. The firm's profit goes to its owners. But what do the owners of the firm provide that earns them this payment?

Economists view profit as a payment for two contributions of entrepreneurs, which are just as necessary for production as are land, labour, or machinery. These two contributions are *risk taking* and *innovation*.

Consider a restaurant that happens to be earning profit for its owner. The land, labour, and capital the restaurant uses to produce its meals did not simply come together magically. Someone—the owner—had to be willing to take the initiative to set up the business, and this individual assumed the risk that the business might failand the initial investment be lost. Because the consequences of loss are so severe, the reward for success must be large in order to induce an entrepreneur to establish a business.

On a larger scale, when two Stanford students (Larry Page and Sergey Brin) started Google, they spent considerable time designing an effective search algorithm and planning their future company. As entrepreneurs, their contribution was *innovation.*

But *other* entrepreneurs—the individuals and venture capital partners that provided funds to launch the new company—played the role of *risk-takers.* Had Google not been successful, they would have lost part or all of their investments. In hindsight, the fact that Google was such a good investment seems inevitable, as successful investments often do. But in the early days, Google—like any startup— was a risky venture.

Innovation and risk taking can also be more subtle, and they are more common than you might think. When you pass by a successful laundromat, you may not give it a second thought. But someone, at some time, had to be the first one to realize, "I bet a laundromat in this neighborhood would do well"—an innovation. And someone had to risk the funds needed to get the business up and running.

Firm's Constraints

If the firm were free to earn whatever level of profit it wanted, it would earn virtually infinite profit. This would make the owners very happy. Unfortunately for owners, though, the firm is not free to do this; it faces *constraints* on both its revenue and its costs.

Demand Curve facing Firm

The constraint on the firm's revenue arises from a familiar concept: the demand curve. This curve always tells us the quantity of a good buyers wish to buy at different prices. But which buyers? And from which firms are they buying? Depending on how we answer these questions, we might be talking about different types of demand curves.

Market demand curves tell us the quantity demanded by *all* consumers from *all* firms in a market. In this chapter, we look at another kind of demand curve:

*The **demand curve facing the firm** tells us, for different prices, the quantity of output that customers will choose to purchase from that firm.*

Notice that this new demand curve—the demand curve facing the firm—refers to only *one* firm, and to *all buyers* who are potential customers of that firm.

Let's consider the demand curve faced by Ned, the owner and manager of Ned's Beds, a manufacturer of bed frames. Figure 1 lists the different prices that Ned could charge for each bed frame and the number of them (per day) he can sell at each price. The figure also shows a graph of the demand curve facing Ned's firm. For each price (on the vertical axis), it shows us the quantity of output the firm can sell (on the horizontal axis). Notice that, like the other types of demand curves we have studied, the demand curve facing the firm slopes downward. In order to sell more bed frames, Ned must lower his price.

The definition of the demand curve facing the firm suggests that once it selects a price, the firm has also determined how much output it will sell. But, as you saw a few chapters ago, we can also flip the demand relationship around: Once the firm has selected an output level, it has also determined the maximum price it can charge. This leads to an alternative definition:

The demand curve facing the firm shows us the maximum price the firm can charge to sell any given amount of output.

Looking at Figure 4 from this perspective, we see that the horizontal axis shows alternative levels of output and the vertical axis shows the price Ned should charge if he wishes to sell each quantity of output.

These two different ways of defining the firm's demand curve show us that it is, indeed, a constraint for the firm. The firm can freely determine *either* its price *or* its level of output. But once it makes the choice, the other variable is automatically determined by the firm's demand curve. Thus, the firm has only *one* choice to make.

Selecting a particular price *implies* a level of output, and selecting an output level *implies* a particular price. Economists typically focus on the choice of output level, with the price implied as a consequence.

Demand and Total Revenue

A firm's **total revenue** is the total inflow of receipts from selling output. Each time the firm chooses a level of output, it also determines its total revenue. Why? Because once we know the level of output, we also know the highest price the firm can charge. Total revenue, which is the number of units of output times the price per unit, follows automatically.

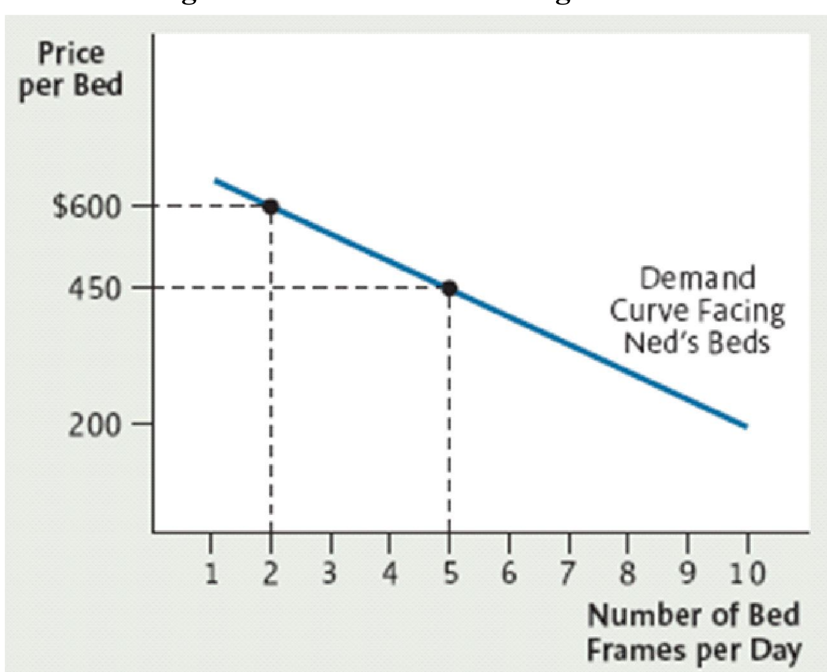
Demand curve facing the firm A curve that indicates, for different prices, the quantity of output that customers will purchase from a *products, or its level of output; once it chooses price, its level of output is determined, and vice versa.*

Table 4: Demand Curve facing a Firm

(1) Price	(2) Output	(3) Total Revenue	(4) Total Cost	(5) Profit
>$650	0	0	$ 300	–$ 300
$650	1	$ 650	$ 700	–$ 50
$600	2	$1,200	$ 900	$ 300
$550	3	$1,650	$1,000	$ 650
$500	4	$2,000	$1,150	$ 850
$450	5	$2,250	$1,350	$ 900
$400	6	$2,400	$1,600	$ 800
$350	7	$2,450	$1,900	$ 550
$300	8	$2,400	$2,250	$ 150
$250	9	$2,250	$2,650	–$ 400
$200	10	$2,000	$3,100	–$1,100

The table 4 presents information about Ned's Beds. Data from the first two columns are plotted in the figure to show the demand curve facing the firm. At any point along that demand curve, the product of price and quantity equals total revenue, which is given in the third column of the table.

Figure 4: Demand Curve facing a Firm



The third column in Table 4 lists the total revenue of Ned's Beds. Each entry is calculated by multiplying the quantity of output (column 2) by the price per unit (column 1).

For example, if Ned's firm produces 2 bed frames per day, he can charge $600 for each of them, so total revenue will be 2 x $600 = $1,200. If Ned increases output to 3 units, he must lower the price to $550, earning a total revenue of 3 x $550 = $1,650.

Because the firm's demand curve slopes downward, Ned must lower his price each time his output increases, or else he will not be able to sell all he produces.

With more units of output, but each one selling at a lower price, total revenue could rise or fall. Scanning the total revenue column, we see that for this firm, total revenue first rises and then begins to fall. This will be discussed in greater detail later on.

Cost Constraint

Every firm struggles to reduce costs, but there is a limit to how low costs can go. These limits impose a second constraint on the firm. Where do the limits come from?

First, the firm has a given production technology, which determines the different combinations of inputs the firm can use to produce its output.

Second, the firm must pay *prices* for each of the inputs that it uses, and we assume there is nothing the firm can do about those prices. Together, the firm's technology and the prices of the inputs determine the cheapest way to produce any given level of output. And once the firm finds this least-cost method, it has driven the cost of producing that output level as low as it can go.

The fourth column of Table 4 lists Ned's total cost—the lowest possible cost of producing each quantity of output. More output always means greater costs, so the numbers in this column are always increasing. For example, at an output of zero, total cost is $300. This tells us we are looking at costs in the short run, over which some of the firm's costs are *fixed.* (What would be the cost of producing 0 units if this were the long run?) If output increases from 0 to 1 bed frame, total cost rises from $300 to $700. This increase in total costs—$400—is caused by an increase in *variable* costs, such as labour and raw materials.

Questions

1. Define Isoquant. Explain its Properties.
2. What is meant by Marginal Rate of Technical Substitution?
3. Define Isocost Lines. Discuss Properties of Isocost Lines.
4. Explain Least-Cost Input Combination of a Firm.
5. Explain the Goal of Profit Maximization of a Firm.
6. What are Firm's Constraints?

This lesson will focus on the following:

1. Profit-Maximizing Output Level: Total Revenue (TR) and Total Cost (TC) Approach
2. Marginal Revenue (MR) and Marginal Cost (MC) Approach
3. Profit Maximization Using Graphs: TR-TC Approach
4. MR-MC Approach
5. Dealing with Losses: Short Run and Shutdown Rule
6. Long Run and Exit Decision

Profit-Maximizing Output Level: Total Revenue and Total Cost Approach

At any given output level, the data in Table 13 of lesson tell us

(1) how much revenue the firm will earn and

(2) its cost of production.

We can then easily see how much profit the firm earns at each output level, which is the difference between total revenue (*TR*) and total cost (*TC*).

In the total revenue and total cost approach, we see the firm's profit as the difference between TC *and* TR *at each output level. The firm chooses the output level where profit is greatest.*

Let's see how this works for Ned's Beds. Column 5 of Table 4 lists total profit at each output level. If the firm were to produce no bed frames at all, total revenue (*TR*) would be 0, while total cost (*TC*) would be \$300. Total profit would be *TR* - *TC* = 0 - \$300 = -\$300. We would say that the firm earns a profit of negative \$300 or a loss of \$300 per day. Producing one bed frame would raise total revenue to \$650 and total cost to \$700, for a loss of $50. Not until the firm produces 2 bed frames does total revenue rise above total cost and the firm begin to make a profit. At 2 bed frames per day, *TR* is $1,200 and *TC* is $900, so the firm earns a profit of $300. Remember that as long as we have been careful to include *all* costs in *TC*—implicit as well as explicit—the profits and losses we are calculating are *economic* profits and losses.

In the total revenue and total cost approach, locating the profit-maximizing output level is straightforward: We just scan the numbers in the profit column until we find the largest value, $900, and the output level at which it is achieved, 5 units per day. We conclude that the profit-maximizing output for Ned's Beds is 5 units per day.

Marginal Revenue and Marginal Cost Approach

There is another way to find the profit-maximizing level of output. This approach, which uses *marginal* concepts, gives us some powerful insights into the firm's decision-making process. It is also closer to the trial-and-error procedure at some firms, in which small experimental changes are made to determine the impact on profit. Recall that *marginal* cost is the *change* in total cost per unit increase in output.

Now, let's consider a similar concept for revenue.

Marginal revenue **(MR)** *is the change in the firm's total revenue* (TR) *divided by the change in its output* (ΔQ)*:*

$$MR = \Delta TR \div \Delta Q$$

MR *tells us how much revenue rises per unit increase in output.*

Table 1 reproduces the *TR* and *TC* columns from Table 4 of lesson 13, but adds columns for marginal revenue and marginal cost. (In the table, output is always changing by one unit, so we can use *TR* alone as our measure of marginal revenue.) For example, when output changes from 2 to 3 units, total revenue rises from $1,200 to $1,650. For this output change, *MR* = $450. As usual, marginal are placed *between* different output levels because they tell us what happens as output *changes* from one level to another.

There are two important things to notice about marginal revenue. First, when *MR* is *positive,* an increase in output causes total revenue to *rise.* In the table, *MR* is positive for all increases in output from 0 to 7 units. When *MR* is *negative,* an increase in output causes total revenue to *fall,* as occurs for all increases beyond 7 units.

The second thing to notice about *MR* is a bit more complicated: Each time output increases, *MR* is *smaller* than the price the firm charges at the new output level.

For example, when output increases from 2 to 3 units, the firm's total revenue rises by \$450—even though it sells the third unit for a price of \$550. This may seem strange to you. After all, if the firm increases output from 2 to 3 units, and it gets \$550 for the third unit of output, why doesn't its total revenue rise by \$550?

Loss The difference between total cost (*TC*) and total revenue (*TR*), when *TC* > *TR*.

Marginal revenue The change in total revenue from producing one more unit of output.

Table 1: MC and MR Approach

Output	Total Revenue	Marginal Revenue	Total Cost	Marginal Cost	Profit
0	0		\$ 300		−\$ 300
		\$650		\$400	
1	\$ 650		\$ 700		−\$ 50
		\$550		\$200	
2	\$1,200		\$ 900		\$ 300
		\$450		\$100	
3	\$1,650		\$1,000		\$ 650
		\$350		\$150	
4	\$2,000		\$1,150		\$ 850
		\$250		\$200	
5	\$2,250		\$1,350		\$ 900
		\$150		\$250	
6	\$2,400		\$1,600		\$ 800
		\$ 50		\$300	
7	\$2,450		\$1,900		\$ 550
		−\$ 50		\$350	
8	\$2,400		\$2,250		\$ 150
		−\$150		\$400	
9	\$2,250		\$2,650		−\$ 400
		−\$250		\$450	
10	\$2,000		\$3,100		−\$1,100

The answer is found in the firm's downward-sloping demand curve, which tells us that to sell more output, the firm must cut its price. When output increases from 2 to 3 units, the firm must lower its price from $600 to $550. Moreover, the new price of $550 will apply to *all three* units the firm sells.4 This means it *gains* some revenue—$550—by selling that third unit. But it also *loses* some revenue—$100—by having to lower the price by $50 on each of thetwo units of output it could have otherwise sold at $600. Marginal revenue willalways equal the *difference* between this gain and loss in revenue—in this case, $550 - $100 = $450.

When a firm faces a downward-sloping demand curve, each increase in output causes a revenue gain, from selling additional output at the new price, and a revenue loss, from having to lower the price on all previous units of output. Marginal revenue is therefore less than the price of the last unit of output.

Using *MR* and *MC* to Maximize Profits

Now we'll see how marginal revenue, together with marginal cost, can be used to find the profit-maximizing output level. The logic behind the *MC* and *MR* approach is this:

An increase in output will always raise profit as long as marginal revenue is greater than marginal cost (MR > MC).

Notice the word *always*. Let's see why this rather sweeping statement must be true. Table 1 tells us that when output rises from 2 to 3 units, *MR* is $450, while *MC* is $100. This change in output causes both total revenue and total cost to rise, but it causes revenue to rise by *more* than cost ($450 > $100). As a result, profit must increase. Indeed, looking at the profit column, we see that increasing output from 2 to 3 units *does* cause profit to increase, from $300 to $650.

The converse of this statement is also true:

An increase in output will always lower profit whenever marginal revenue is less than marginal cost (MR < MC).

For example, when output rises from 5 to 6 units, *MR* is $150, while *MC* is $250. For this change in output, both total revenue and total cost rise, but cost rises *more,* so profit must go down.

These insights about *MR* and *MC* lead us to the following simple guideline the firm should use to find its profit-maximizing level of output:

To find the profit-maximizing output level, the firm should increase output whenever MR > MC, *and decrease output when* MR < MC.

Let's apply this rule to Ned's Beds. In Table 1 we see that when moving from 0 to 1 unit of output, *MR* is \$650, while *MC* is only \$400. Since *MR* is larger than *MC*, making this move will increase profit. Thus, if the firm is producing 0 beds, it should always increase to 1 bed. Should it stop there? Let's see. If it moves from 1 to 2 beds, *MR* is \$550, while *MC* is only \$200. Once again, *MR* > *MC,* so the firm should increase to 2 beds. You can verify from the table that if the firm finds itself producing 0, 1, 2, 3, or 4 beds, *MR* > *MC* for an increase of 1 unit, so it will always make greater profit by increasing production.

Until, that is, output reaches 5 beds. At this point, the picture changes: From 5 to 6 beds, *MR* is \$150, while *MC* is \$250. For this move, *MR* > *MC*, so profits would decrease. Thus, if the firm is producing 5 beds, it should *not* increase to 6.

The same is true at every other output level beyond 5 units: The firm should *not* raise its output, since *MR* < *MC* for each increase. We conclude that Ned maximizeshis profit by producing 5 beds per day—the same answer we got using the *TR* and *TC* approach earlier.

Profit Maximization using Graphs

Both approaches to maximizing profit (using totals or using marginals) can be seen even more clearly when we use graphs. In Figure 2(a) and (b), the data from Table 1 have been plotted—the *TC* and *TR* curves in the upper panel, and the *MC* and *MR* curves in the lower one.

The marginal revenue curve has an important relationship to the total revenue curve. As you can see in Figure 2(a), total revenue (*TR*) is plotted on the vertical axis, and quantity (*Q*) on the horizontal axis, so the slope along any interval is just $\Delta TR/\Delta Q$. But this is exactly the definition of marginal revenue.

The marginal revenue for any change in output is equal to the slope *of the total revenue curve along that interval.*

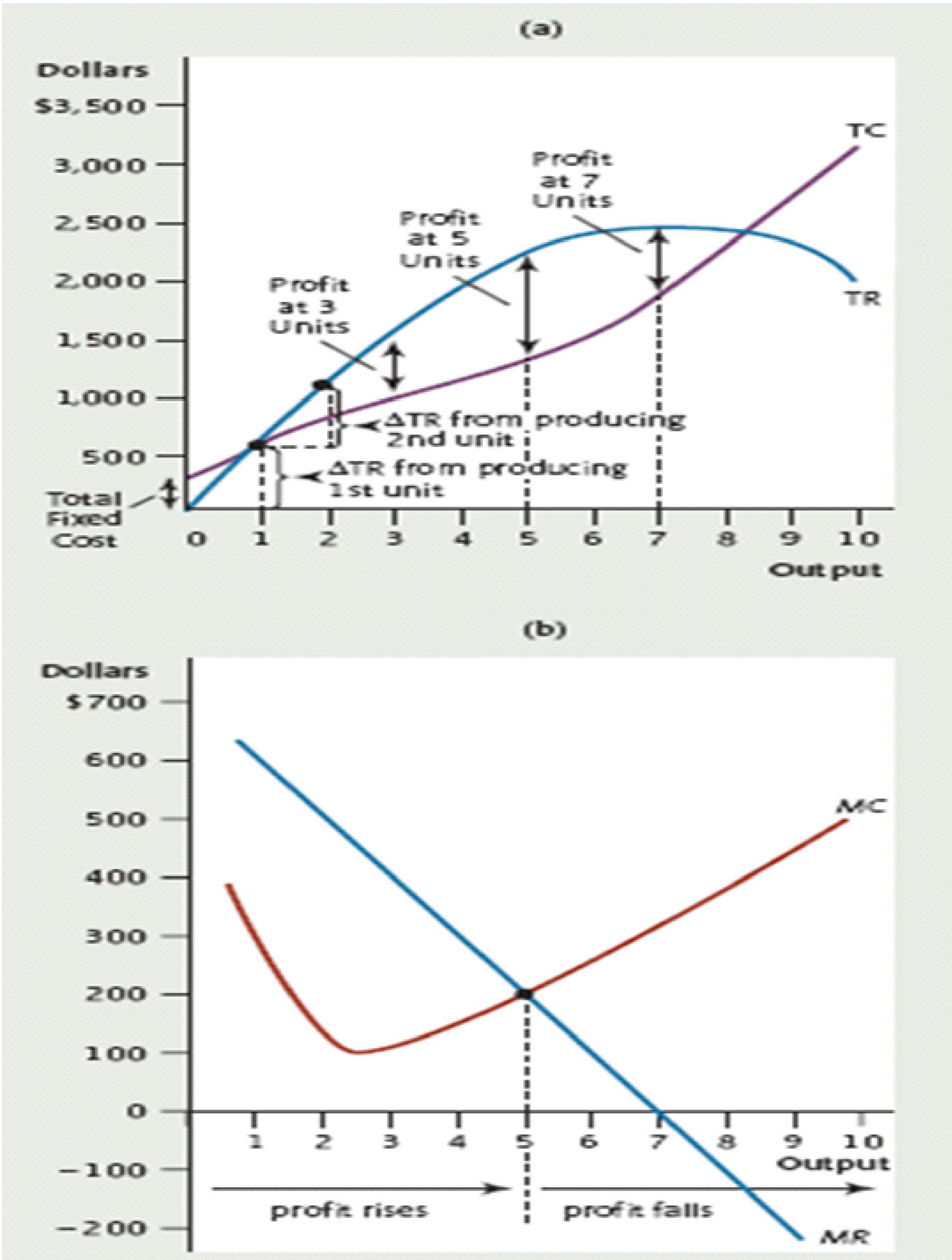
Thus, as long as the *MR* curve lies above the horizontal axis (*MR* > *0*), *TR* must be increasing and the *TR* curve must slope upward. In the figure, *MR* > 0, and the *TR*curve slopes upward from zero to 7 units. When the *MR* curve dips below the horizontal axis (*MR* < *0*), *TR* is decreasing, so the *TR* curve begins to slope downward.

In the figure, this occurs beyond 7 units of output. As output increases in Figure 2, *MR* is first positive and then turns negative, so the *TR* curve will first *rise*and then *fall*.

TR and TC Approach Using Graphs

Now let's see how we can use the *TC* and *TR* curves to guide the firm to its profit maximizing output level. We know that the firm earns a profit at any output level where *TR* > *TC*—where the *TR* curve lies *above* the *TC* curve. In Figure 1(a), you can see that all output levels from 2 through 8 units are profitable for the firm. The *amount* of profit is simply the *vertical distance* between the *TR* and *TC* curves, whenever the *TR* curve lies above the *TC* curve. Since the firm cannot sell part of a bed frame, it must choose whole numbers for its output, so the profit-maximizing output level is simply the whole-number quantity at which this vertical distance is greatest—5 units of output. Of course, the *TR* and *TC* curves in Figure 2 were plotted from the data in Table 1, so we should not be surprised to find the same profit maximizing output level—5 units—that we found before when using the table.

Figure 1: Profit Maximization



Panel (a) shows the firm's total revenue (TR) *and total cost* (TC) *curves. Profit is the vertical distance between the two curves at any level of output. Profit is maximized when that vertical distance is greatest—at 5 units of output. Panel (b) shows the firm's marginal revenue* (MR) *and marginal cost* (MC) *curves. (As long as* MR *lies above the horizontal axis, the* TR *curve slopes upward.) Profit is maximized at the level of output closest to where the two curves cross—at 5 units of output.*

We can sum up our graphical rule for using the *TR* and *TC* curves this way:

To maximize profit, the firm should produce the quantity of output where the vertical distance between the TR *and* TC *curve is greatest and the* TR *curve lies above the* TC *curve.*

MR and MC Approach Using Graphs

Figure 1 also illustrates the *MR* and *MC* approach to maximizing profits. As usual, the marginal data in panel (b) are plotted *between* output levels, since they tell us what happens as output changes from one level to another.

In the diagram, as long as output is less than 5 units, the *MR* curve lies above the *MC* curve (*MR* > *MC*), so the firm should produce more. For example, if we consider the move from 4 to 5 units, we compare the *MR* and *MC* curves at the midpoint between 4 and 5. Here, the *MR* curve lies above the *MC* curve, so increasing output from 4 to 5 will increase profit.

But now suppose the firm is producing 5 units and considering a move to 6. At the midpoint between 5 and 6 units, the *MR* curve has already crossed the *MC* curve, and now it lies *below* the *MC* curve. For this move, *MR* < *MC*, so raising output would *decrease* the firm's profit. The same is true for every increase in output beyond 5 units: The *MR* curve always lies below the *MC* curve, so the firm will decrease its profits by increasing output. Once again, we find that the profit maximizing output level for the firm is 5 units.

Notice that the profit-maximizing output level—5 units—is the level closest to where the *MC* and *MR* curves cross. This is no accident. For each change in output that *increases* profit, the *MR* curve will lie above the *MC* curve. The first time that an output change *decreases* profit, the *MR* curve will cross the *MC* curve and dip below it. Thus, the *MC* and *MR* curves will always cross closest to the profit maximizing output level.

With this graphical insight, we can summarize the *MC* and *MR* approach this way:

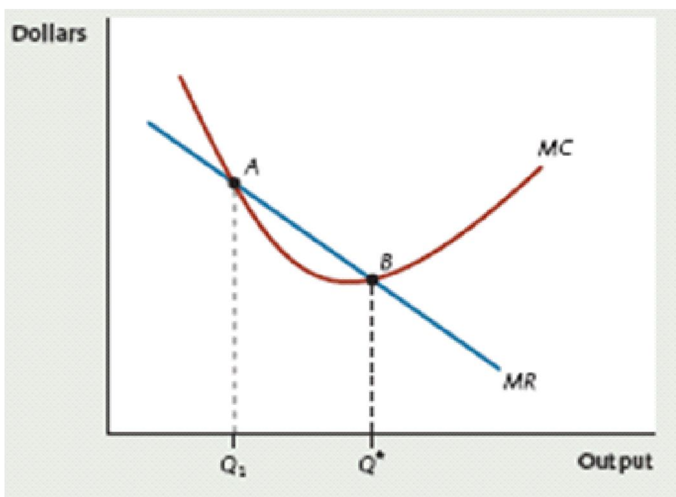
To maximize profit, the firm should produce the quantity of output closest to the point where MC = MR—*that is, the quantity of output at which the* MC *and* MR *curves intersect.*

This rule is very useful, since it allows us to look at a diagram of *MC* and *MR* curves and *immediately* identify the profit-maximizing output level.

"The profit-maximizing output level is where *MC* equals *MR*," translate to "The profit-maximizing output level is closest to the point where the *MC* curve crosses the *MR* curve."

A Proviso. There is, however, one important exception to this rule. Sometimes the *MC* and *MR* curves cross at two different points. In this case, the profit-maximizing output level is the one at which the *MC* curve crosses the *MR* curve *from below*.

Figure 2: Two Points of Interaction



Sometimes the MR *and* MC *curves intersect twice. Theprofit-maximizing level ofoutput is always found where* MC *crosses* MR *from below.*

Figure 2 shows why. At point *A,* the *MC* curve crosses the *MR* curve from *above*.Our rule tells us that the output level at this point, *Q*1, is *not* profit maximizing. Why not? Because at output levels lower than *Q*1, *MC* > *MR*, so profit *falls* as output increases toward *Q*1. Also, profit *rises* as output increases *beyond* *Q*1, since *MR* > *MC* for these moves. Since it never pays to increase *to* $Q1$, and profit rises when increasing *from* $Q1$, we know that $Q1$ cannot possibly maximize the firm's profit.

But now look at point *B,* where the *MC* curve crosses the *MR* curve from below. You can see that when we are at an output level lower than Q^*, it always pays to increase output, since *MR* _ *MC* for these moves. You can also see that, once we have arrived at Q^*, further increases will reduce profit, since $MC > MR$. Q^* is thus the profit-maximizing output level for this firm—the output level at which the *MC* curve crosses the *MR* curve *from below.*

Average Costs

We have not yet referred to *average* cost in this lesson. There is a good reason for this. We have been concerned about how much the firm should produce if it wishes to earn the greatest possible level of profit. To achieve this goal, the firm should produce more output whenever doing so *increases* profit, and it needs to know only *marginal* cost and *marginal* revenue for this purpose. The different types of average cost (*ATC, AVC,* and *AFC*) are simply irrelevant. Indeed, a common error—sometimes made even by business managers—is to use *average* cost in place of *marginal* cost in making decisions.

For example, suppose a yacht maker wants to know how much his total cost will rise in the short run if he produces another unit of output. It is tempting—*but wrong*—for the yacht maker to reason this way: "My cost per unit (*ATC*) is currently $50,000 per yacht. Therefore, if I increase production by 1 unit, my total cost will rise by $50,000; if I increase production by 2 units, my total cost will rise by $100,000, and so on."

There are two problems with this approach.

First, *ATC* includes many costs that are *fixed* in the short run—including the cost of all fixed inputs such as the factory and equipment and the design staff. These costs will *not* increase when additional yachts are produced, and they are therefore irrelevant to the firm's decision making in the short run.

Second, *ATC changes* as output increases. The cost per yacht may rise above $50,000 or fall below $50,000, depending on whether the *ATC* curve is upward or downward sloping at the current production level. Note that the first problem— fixed costs—could be solved by using *AVC* instead of *ATC*. The second problem— changes in average cost—remains even when *AVC* is used.

The correct approach, as we've seen in this lesson, is to use the *marginal cost* of a yacht and to consider increases in output one unit at a time. The firm shouldproduce the output level where its *MC* curve crosses its *MR* curve from below.

Average cost doesn't help at all; it only confuses the issue.

Does this mean that all of your efforts to master *ATC* and *AVC*—their definitions, their relationship to each other, and their relationship to *MC*—were a waste of time? Far from it. As you'll see, average cost will prove *very* useful in the chapters to come. You'll learn that whereas marginal values tell the firm *what* to do, averages can tell the firm *how well* it has done. But average cost should *not* be used in place of marginal cost as a basis for decisions.

Marginal Approach to Profit

The *MC* and *MR* approach for finding the profit-maximizing output level is actually a very specific application of a more general principle:

*The **marginal approach to profit** states that a firm should take any action that adds more to its revenue than to its costs.*

In this lesson, the action being considered is whether to increase output by 1 unit. We've learned that the firm should take this action whenever $MR > MC$.

But the same logic can be applied to *any other decision* facing the firm. Should a restaurant owner take out an ad in the local newspaper? Should a convenience store that currently closes at midnight stay open 24 hours instead? Should a private kindergarten hire another teacher? Should an inventor pay to produce an infomercial for her new gizmo? Should a bank install another ATM? The answer to all of these questions is yes—*if* the action would add more to revenue than to costs. In future chapters, we'll be using the marginal approach to profit to analyze some other types of firm decisions.

Dealing with Losses

So far, we have dealt only with the pleasant case of profitable firms and how they select their profit-maximizing output level. But what about a firm that cannot earn a positive profit at *any* output level? What should it do? The answer depends on what time horizon we are looking at.

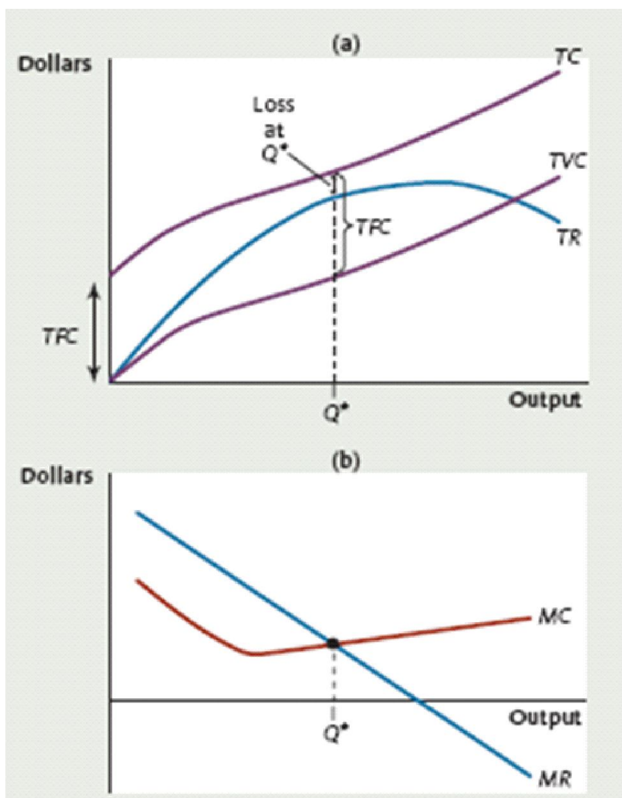
Marginal approach to profit A firm maximizes its profit by taking any action that adds more to its revenue than to its cost.

Short Run and Shutdown Rule

In the short run, the firm must pay for its fixed inputs, because there is not enough time to sell them or get out of lease and rental agreements. But the firm can *still*make decisions about production. And one of its options is to *shut down*—to stop producing output, at least temporarily.

At first glance, you might think that a loss-making firm should always shut down its operation in the short run. After all, why keep producing if you are not making any profit? In fact, it makes sense for some unprofitable firms to continue operating.

Figure 3: Loss Minimization

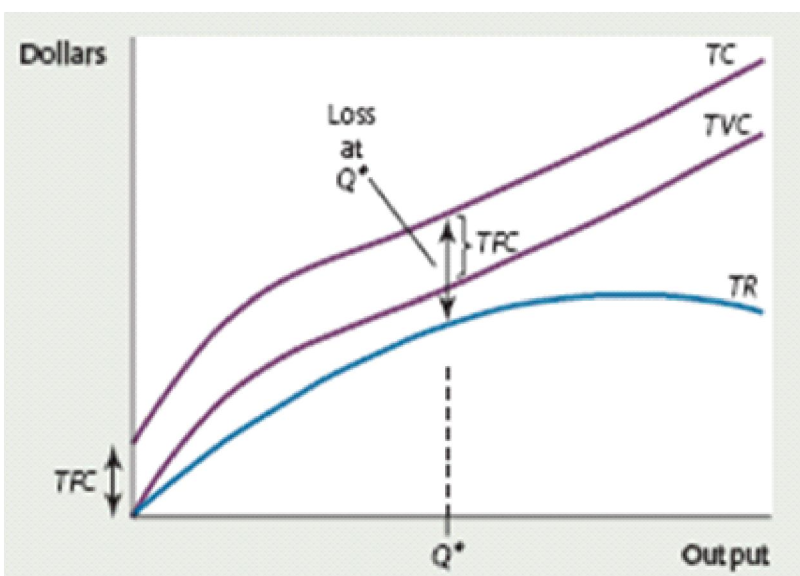


Imagine a firm with the *TC* and *TR* curves shown in the upper panel of Figure 3 (ignore the *TVC* curve for now). No matter what output level the firm produces, the *TC* curve lies above the *TR* curve, so it will suffer a loss—a negative profit. For this firm, the goal is still profit maximization. But now, the highest profit will be the one with the *least negative value.* In other words, profit maximization becomes *loss minimization*.

If the firm keeps producing, then the smallest possible loss is at an output level of *Q**, where the distance between the *TC* and *TR* curves is smallest. *Q** is also the output level we would find by using our marginal approach to profit (increasing output whenever that adds more to revenue than to costs). This is why, in the lower panel of Figure 4, the *MC* and *MR* curves must intersect at (or very close to) *Q**.

The question is: Should this firm produce at *Q** and suffer a loss? The answer is yes—*if* the firm would lose even *more* if it stopped producing and shut down its operation. Remember that, in the short run, a firm must continue to pay its total fixed cost (*TFC*) no matter what level of output it produces—even if it produces nothing at all. If the firm shuts down, it will therefore have a loss equal to its *TFC,* since it will not earn any revenue. But if, by producing some output, the firm can cut its loss to something *less* than *TFC,* then it should stay open and keep producing.

Figure 4: Shut Down



At Q**, this firm's total variable cost exceeds its total revenue. The best policy is to shut down, produce nothing, and suffer a loss equal to* TFC *in the short run.*

To understand the shutdown decision more clearly, let's think about the firm's total variable costs. Business managers often call *TVC* the firm's *operating cost,* since the firm only pays these variable costs when it continues to operate. If a firm, by staying open, can earn *more* than enough revenue to cover its operating costs, then it is making an *operating profit* (*TR* > *TVC*). It should not shut down because its operating profit can be used to help pay its fixed costs. But if the firmcannot even cover its operating cost when it stays open—that is, if it would suffer an *operating loss* (*TR* < *TVC*)—it should definitely shut down. Continuing to operate only *adds* to the firm's loss, increasing the total loss beyond fixed costs. This suggests the following guideline—called the **shutdown rule**—for a loss making firm:

Let Q* *be the output level at which* MR = MC. *Then, in the short run:*

> *If* TR > TVC *at* Q**, the firm should keep producing.*
>
> *If* TR <TVC *at* Q**, the firm should shut down.*
>
> *If* TR = TVC *at* Q**, the firm should be indifferent between shutting down and producing.*

Look back at Figure 3. At *Q*,* the firm is making an operating profit, since its *TR* curve is above its *TVC* curve. This firm, as we've seen, should continue to operate.

Figure 4 is drawn for a different firm, one that has different cost curves and a different *TR* curve than the firm in Figure 4. This firm *cannot* earn an operating profit, since its *TR* curve lies below its *TVC* curve everywhere—even at *Q**. This firm should shut down.

The shutdown rule is a powerful predictor of firms' decisions to stay open or cease production in the short run. It tells us, for example, why some seasonal businesses— such as ice cream shops in summer resort areas—shut down in the winter, when *TR* drops so low that it becomes smaller than *TVC*. And it tells us why producers of steel, automobiles, agricultural goods, and television sets will often keep producing output for some time even when they are losing money.

Long Run and Exit Decision

The shutdown rule applies only in the short run, a time horizon too short for the firm to escape its commitments to pay for fixed inputs such as plant and equipment. In fact, we only use the term *shut down* when referring to the short run. But a firm can also decide to stop producing in the long run. In that case, we say the firm has decided to exit the industry.

The long-run decision to exit is different than the short-run decision to shut down. That's because in the long run, there *are* no fixed costs, since all inputs can be varied. Therefore, a firm that exits, by reducing all of its inputs to zero, will have *zero* costs (an option not available in the short run). And since exit also means zero revenue, a firm that exits will earn zero profit. When would a firm decide to exit and earn zero profit? When its only other alternative is to earn *negative* profit.

A firm should exit the industry in the long run when—at its best possible output level—it has any loss at all.

Shutdown rule In the short run, the firm should continue to produce if total revenue exceeds total variable costs; otherwise, it should shut down.

Exit A permanent cessation of production when a firm leaves an industry.

Conclusion

In economics, we view the firm as a single economic decision maker with the goal of maximizing the owners' profit. Economic profit is total revenue minus *all* costs of production, explicit and implicit. In their pursuit of maximum profit, firms face two constraints. One is embodied in the demand curve the firm faces; it indicates the maximum price the firm can charge to sell any amount of output. This constraint determines the firm's revenue at each level of production. The other constraint is imposed by costs: More output always means greater costs. In choosing the profit-maximizing output, the firm must consider both revenues and costs.

One approach to choosing the optimal level of output is to measure profit as the difference between total revenue and total cost at each level of output, and then select the output level at which profit is greatest. An alternate approach uses *marginal revenue* (*MR*), the change in total revenue from producing one more unit of output, and *marginal cost* (*MC*), the change in total cost from producing one more unit. The firm should increase output whenever $MR < MC$, and lower output when $MR > MC$. The profit maximizing output level is the one closest to the point where $MR = MC$.

If profit is negative, but total revenue exceeds total variable cost, the firm should continue producing in the short run. Otherwise, it should shut down and suffer a loss equal to its fixed cost. A firm with negative profit in the long run should exit the market.

Questions

1. Explain Profit-Maximizing Output Level using Total Revenue (TR) and Total Cost (TC) Approach.
2. Explain Profit-Maximizing Output Level Marginal Revenue (MR) and Marginal Cost (MC) Approach.
3. How firms deal with Losses: Short Run and Shutdown Rule
4. Explain Long Run and Exit Decision using suitable diagramme.

This lesson will focus on the following:

1. Perfect Competition: Characteristics and Profit Maximization
2. Profit-Maximizing Output Level: TR-TC Approach and MC-MR Approach
3. Measuring Profit or Loss
4. Firm's Short-Run Supply Curve and Shut Down Price

Perfect Competition

In microeconomics, we can divide markets for goods and services into four basic kinds of market structure:

a. *perfect competition,*

b. monopoly,

c. monopolistic competition,

d. *oligopoly*

What is Perfect Competition?

The supply and demand model explains how prices are determined in *perfectly competitive markets*.

Now we're going to take a much deeper and more comprehensive look at perfectly competitive markets. By the end of this chapter, you will understand very clearly how perfect competition and the supply and demand model are related.

Meaning of competition

When you hear the word *competition*, you may think of an intense, personal rivalry, like that between two boxers competing in a ring or two students competing for the best grade in a small class. But there are other, less personal forms of competition.

Meaning of perfect competition

A market structure in which there are many buyers and sellers, the product is standardized, sellers can easily enter or exit the market, and buyers and sellers are well-informed.

Meaning of market structure

*By **market structure**, we mean all the characteristics of a market that influence the behavior of buyers and sellers when they come together to trade.*

Characteristics of Perfect Competition

***Perfect competition** is a market structure with four important characteristics:*

1. *There are large numbers of buyers and sellers, and each buys or sells only a tiny fraction of the total quantity in the market.*
2. *Sellers offer a standardized product.*
3. *Sellers can easily enter into or exit from the market.*
4. *Buyers and sellers are well-informed.*

Four conditions of perfect competition

A Large Number of Buyers and Sellers

In perfect competition, there must be many buyers and sellers. How many? It would be nice if we could specify a number. Unfortunately, we cannot. What constitutes a large number of buyers and sellers can be different under different conditions.

In a perfectly competitive market, the number of buyers and sellers is so large that no individual decision maker can significantly affect the price of the product by changing the quantity it buys or sells.

Example: World market for wheat

On the selling side, there are hundreds of thousands of individual wheat farmers. Each of these farmers produces only a tiny fraction of the total market quantity.

If any one of them were to double, triple, or even quadruple production. The impact on total market quantity and market price would be negligible. The same is true on the buying side. There are so many small buyers. No one of them can affect the market price by increasing or decreasing its quantity demanded. Most agricultural markets conform to the large-number-of-small-firms requirement.

Another example: Market for athletic shoes.

Four large firms - Nike, Adidas (including Reebok), New Balance, and Puma - accounted for more than 70 percent of worldwide sales in this market. If any one of these producers decided to change its output by even 10 percent, the impact on total quantity supplied and market price would be *very* noticeable. The market for athletic shoes fails the large number of-small-firms requirement. So it is not an example of perfect competition.

A Standardized Product Offered by Sellers

In a perfectly competitive market, buyers do not perceive differences between the products of one seller and another. For example, buyers of wheat will ordinarily have no preference for one farmer's wheat over another's, so wheat would surely pass the standardized product test.

The same is true of many other agricultural products. For example, corn syrup and soybeans. It is also true of commodities like crude oil and precious metals like gold or silver.

When differences among firms' products matter to buyers, the market is not perfectly competitive. For example, most consumers perceive differences among the various brands of coffee. They may have strong preferences for one particular brand. Coffee, therefore, fails the standardized product test of perfect competition. Other goods and services that would fail this test include automobiles.

Easy Entry into and Exit from the Market

Entry into a market is rarely free. A new seller must always incur *some* costs to set up shop, begin production, and establish contacts with customers. But a perfectly competitive market has no *significant* barriers or special costs to discourage new entrants.

Any firm wishing to enter can do business on the same terms as firms that are already there. For example, anyone who wants to start a wheat farm can do so, facing the same costs for land, farm equipment, seeds, fertilizer, and hired labour as existing farms.

The same is true of anyone wishing to open up a dry cleaning shop and restaurant. These examples would pass the easy entry test of perfect competition.

Perfect competition also requires easy *exit.* A firm suffering a long-run loss must be able to sell off its plant and equipment and leave the industry without obstacles. Some markets satisfy this requirement, and some do not.

Well-Informed Buyers and Sellers

In perfect competition, both buyers and sellers have all information relevant to their decision to buy or sell. For example, they know about the quality of the product, and the prices being charged by competitors.

Market example: Agricultural and commodities markets

In most other types of markets, buyers and sellers are *reasonably* well informed. But in some markets, this assumption may not be realistic.

Is Perfect Competition Realistic?

The four assumptions a market must satisfy to be perfectly competitive.

Do any markets satisfy all these requirements?

How broadly can we apply the model of perfect competition when we think about the real world?

In some cases, the model fits remarkably well. Example: Market for wheat, for example, passes all four tests for a competitive market: many buyers and sellers, standardized output, easy entry and exit, and well-informed buyers.

But in the vast majority of markets, one or more of the assumptions of perfect competition will, in a strict sense, be violated. This might suggest that the model can be applied only in a few limited cases. Economists use the perfect competition model more than any other market model.

Why is this?

First, the model of perfect competition is powerful. Using simple techniques, it leads to important predictions about a market's response to changes in consumer tastes, technology, and government policies. While other types of market structure models also yield valuable predictions, they are often more cumbersome and their predictions less definitive.

Second, many markets, while not strictly perfectly competitive, come *reasonably* close. The more closely a real-world market fits the model, the more accurate our predictions will be when we use it.

Perfect competition approximates conditions and yields accurate-enough predictions in a wide variety of markets. This is why you will often find economists using the model to analyze the markets for crude oil, consumer electronic goods, fast-food meals, and medical care, even though in each of these cases one or more of the requirements is not strictly satisfied.

Perfectly Competitive Firm

A market is a collection of individual decision makers. The decisions made by individuals, collectively, affect the market. And the market, in turn, influences the choices made by individuals. This is why, in learning about perfectly competitive markets, we'll be going back and forth between the competitive firm and the market in which it operates.

In Figure 1(a), we start with the market—specifically, the competitive market in which gold is produced and sold. The intersection of the market supply and demand curves determines the market price of gold which, in the figure, is $800 per troy ounce.

Competitive Firm's Demand Curve

Panel (b) of Figure 1 shows the demand curve facing Small Time Gold Mines. Notice the special shape of this curve: It is horizontal, or perfectly price elastic. This tells us that no matter how much gold Small Time produces, it will always sell it at the same price—$800 per troy ounce.

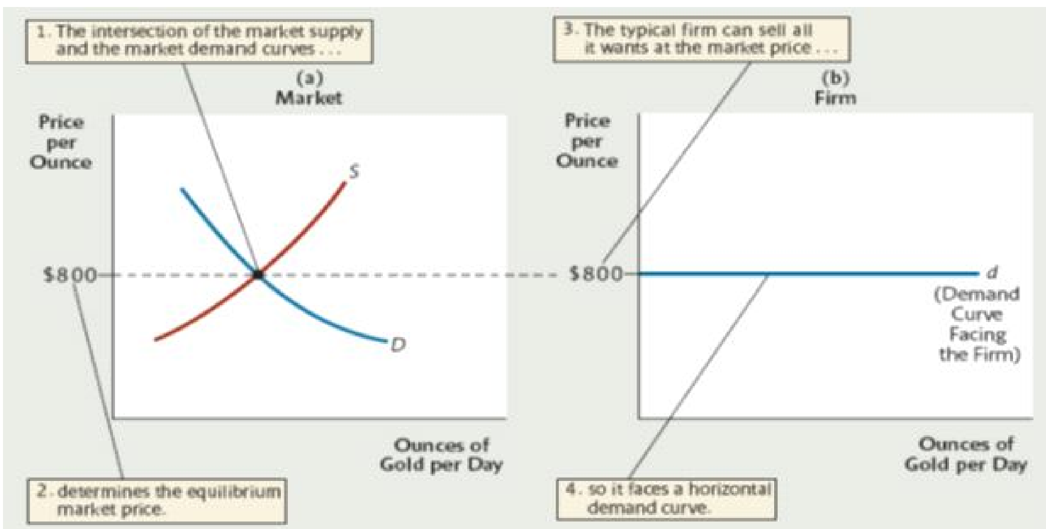
A perfectly competitive firm faces a demand curve that is horizontal (perfectly elastic) at the market price.

Why should this be?

First, in perfect competition, output is standardized—buyers do not distinguish the gold of one mine from that of another. If Small Time were to charge a price even a tiny bit higher than other producers, it would lose all of its customers. They would simply buy from Small Time's competitors. The horizontal demand curve captures this effect. It tells us that if Small Time raises its price above $800, it will not just sell *less* output, it will sell *no* output.

Second, Small Time is only a tiny producer relative to the entire gold market. No matter how much it produces and sells, it cannot make a noticeable difference in market quantity supplied. So it cannot affect the market price. Once again, the horizontal demand curve describes this effect very well. The firm can increase its production without having to lower its price.

Figure 1: The Competitive Industry and Firm



The horizontal demand curve facing the firm and the corresponding price-taking behaviour of firms are hallmarks of perfect competition. When a manager thinks, "If we produce more output, we will have to lower our price in order to sell it" then the firm faces a *downward sloping* demand curve and it is *not* a competitive firm.

The manager of a competitive firm will instead think, "We can sell all the output we want at the going price, so how much should we produce?"

Notice that, since a competitive firm takes the market price as given, its only decision is *how much output to produce and sell*. And that decision will determine the firm's cost of production, as well as its total revenue. Let's see how this works in practice with Small Time Gold Mines.

Cost and Revenue for a Competitive Firm

Table 1 shows cost and revenue data for Small Time. In the first two columns are different quantities of gold that Small Time could produce each day and the selling price per ounce. Because Small Time is a competitive firm (a price taker), the price remains constant at $800 per ounce, no matter *how* much gold it produces.

Price is always $800. Each time the firm produces another ounce of gold, total revenue rises by $800. **Marginal revenue** is the additional revenue from selling one more ounce of gold. This remains constant at $800.

Figure 2 plots Small Time's total revenue and marginal revenue.

Notice that the total revenue (*TR*) curve in the upper panel is a *straight line* that slopes upward. Each time output increases by one unit, *TR* rises by the same $800. The slope of the *TR* curve is equal to the price of output.

The marginal revenue (*MR*) curve in the lower panel is a *horizontal* line at the market price. In fact, the *MR* curve is the same horizontal line as the demand curve facing the firm. Why? **Marginal revenue** is the additional revenue the firm earns from selling an additional unit of output. For a price-taking competitive firm, that additional revenue will always be the unchanging price it gets for each unit. In this case, it is $800.

For a competitive firm, marginal revenue is the same as the market price. For this reason, the marginal revenue curve and the demand curve facing the firm are the same: a horizontal line at the market price.

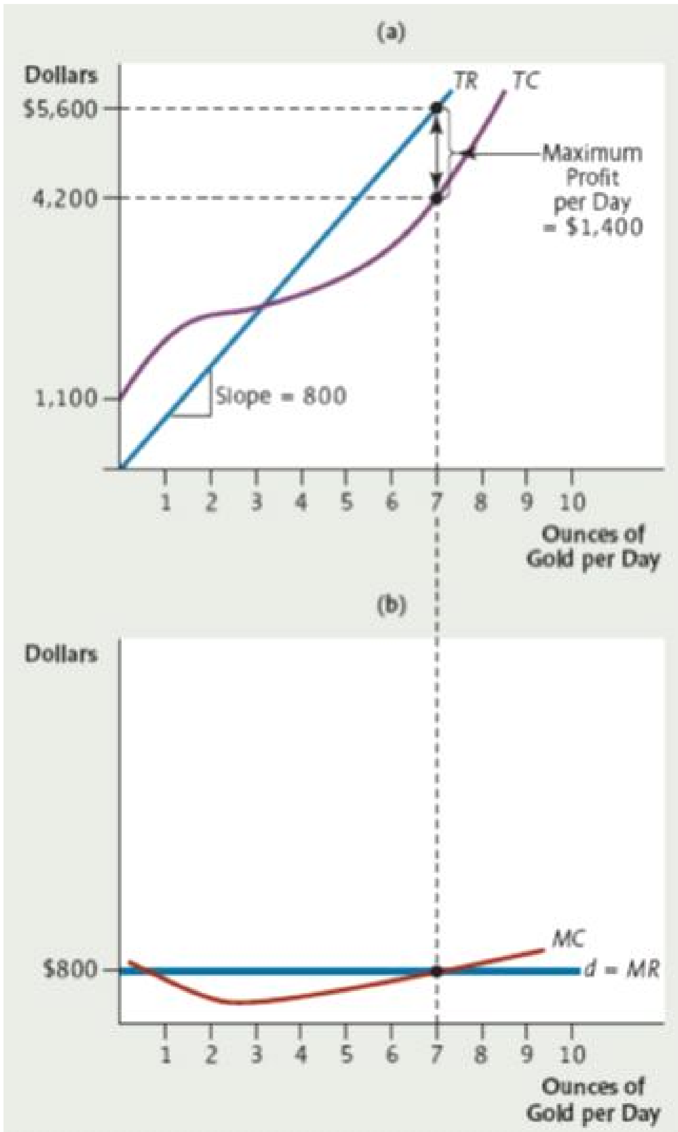
In panel (b), we have labeled the horizontal line "*d* = *MR*," since this line is both the firm's demand curve (*d*) *and* its marginal revenue curve (*MR*).

Columns 5 and 6 of Table 1 show total cost and marginal cost for Small Time. There is nothing special about cost data for a competitive firm. In Figure 2, marginal cost (*MC*) first falls and then rises. Total cost, therefore, rises first at a decreasing rate and then at an increasing rate.

Table 1: Cost and Revenue for Small Time Gold Mines

(1) Output (Troy Ounces of Gold per Day)	(2) Price (per Troy Ounce)	(3) Total Revenue	(4) Marginal Revenue	(5) Total Cost	(6) Marginal Cost	(7) Profit
0	$800	$ 0		$1,100		−$1,100
			$800		$ 900	
1	$800	$ 800		$2,000		−$1,200
			$800		$ 400	
2	$800	$1,600		$2,400		−$ 800
			$800		$ 100	
3	$800	$2,400		$2,500		−$ 100
			$800		$ 200	
4	$800	$3,200		$2,700		$ 500
			$800		$ 300	
5	$800	$4,000		$3,000		$1,000
			$800		$ 500	
6	$800	$4,800		$3,500		$1,300
			$800		$ 700	
7	**$800**	**$5,600**		**$4,200**		**$1,400**
			$800		$ 900	
8	$800	$6,400		$5,100		$1,300
			$800		$1,100	
9	$800	$7,200		$6,200		$1,000
			$800		$1,300	
10	$800	$8,000		$7,500		$ 500

Figure 2: Profit Maximization in Perfect Competition



Panel (a) shows a competitive firm's total revenue (TR) *and total cost* (TC) *curves.*

TR *is a straight line with slope equal to the market price.*

Profit is maximized at 7 ounces per day, where the vertical distance between TR *and TC is greatest.*

Panel (b) shows that profit is maximized where the marginal cost (MC) *curve intersects the marginal revenue* (MR) *curve, which is also the firm's demand curve.*

Profit-Maximizing Output Level

A competitive firm—like any other firm—wants to earn the highest possible profit. We can use either Table 1 or Figure 2 to find the profit-maximizing output level. And we can use the techniques you have already learned:

the total revenue and total cost approach, or

the marginal revenue and marginal cost approach

Total Revenue and Total Cost Approach

The *TR* and *TC* approach is the most direct way of viewing the firm's search for the profit-maximizing output level.

Quite simply, at each output level, subtract total cost from total revenue to get total profit:

Total Profit = *TR* - *TC*

Then we just scan the different output levels to see which one gives the highest number for profit. In Table 1, total profit is shown in the last column. A simple scan of that column tells us that $1,400 is the highest daily profit that Small Time Gold Mines can earn.

To earn this profit, the first column tells us that Small Time must produce 7 ounces per day, its profit-maximizing output level. The same approach to maximizing profit can be seen graphically, in the upper panel of Figure 2. There, total profit at any output level is the distance between the *TR* and *TC* curves. As you can see, this distance is greatest when the firm produces 7 units, verifying what we found in the table.

This approach is simple and straightforward, but it hides the way that *changes* in output cause total revenue and total cost to change. The other approach to finding the profit-maximizing output level focuses on these changes.

Marginal Revenue and Marginal Cost Approach

In the *MR* and *MC* approach, the firm should continue to increase output as long as marginal revenue is greater than marginal cost. You can verify, using Table 1, that if the firm is initially producing 1, 2, 3, 4, 5, or 6 units, it will find that *MR* < *MC* when it raises output by one unit, so producing more will raise profit. Once the firm is producing 7 units, however, *MR* < *MC,* so further increases in output will reduce profit.

Alternatively, using the graph in panel (b) of Figure 2, we look for the output level at which $MR = MC$. As the graph shows, there are two output levels at which the MR and MC curves intersect. However, we can rule out the first crossing point because there, the MC curve crosses the MR curve from above. Remember that the profit-maximizing output is found where the MC curve crosses the MR curve from *below*, at 7 units of output.

Both its demand curve and its marginal revenue curve sloped *downward*. Small Time, however, operates under perfect competition, so its demand and MR curves are the same horizontal line.

Measuring Total Profit

One way to measure a firm's total profit on a graph is the vertical distance between the TR and TC curves. There is another graphical way to measure profit.

To do this, we start with the firm's *profit per unit,* which is the revenue it gets on each unit minus the cost per unit. Revenue per unit is just the price (P) of the firm's output, and cost per unit is our familiar average total cost, so we can write:

Profit per unit = $P - ATC$.

In Figure 3(a), Small Time's ATC curve has been plotted (calculated from the data in Table 1). When the firm is producing at the profit-maximizing output level, 7 units, its ATC is TC/Q = \$4,200/7 = \$600.

Since the price of output is \$800,

Profit *per unit* = P - ATC = \$800 - \$600 = \$200

Graphically, this is the vertical distance between the firm's demand curve and its ATC curve at the profit-maximizing output level.

Once we know Small Time's profit per unit, it is easy to calculate its *total* profit: Just multiply profit per unit by the number of units sold.

Small Time is earning \$200 profit on each ounce of gold, and it sells 7 ounces in all, so total profit is \$200 x 7 = \$1,400.

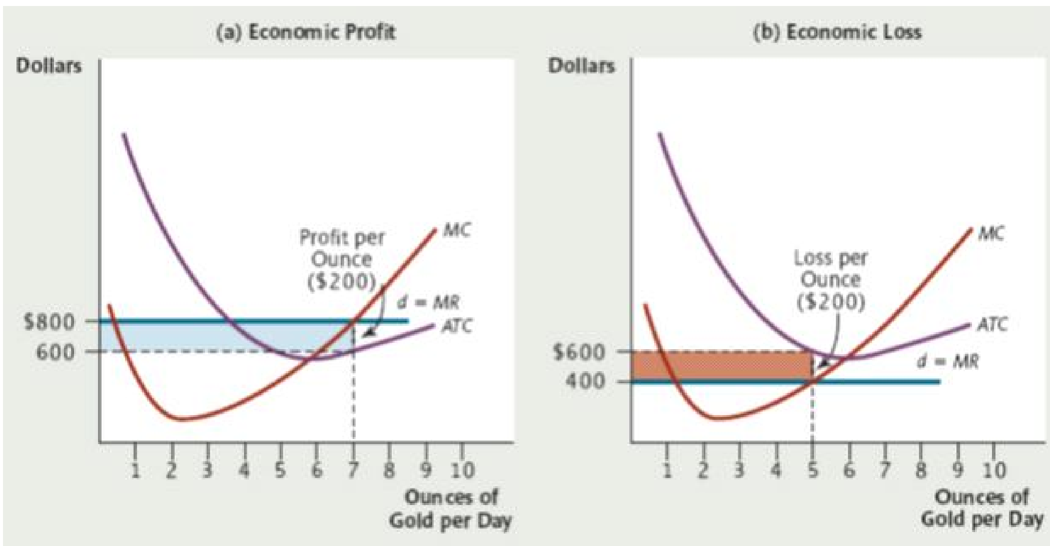
Now look at the shaded rectangle in Figure 3(a). The height of this rectangle is profit per unit, and the width is the number of units produced.

The *area* of the rectangle = height x width = equals Small Time's profit:

A firm earns a profit whenever P > ATC. *Its total profit at the best output level equals the area of a rectangle with height equal to the distance between* P *and* ATC, *and width equal to the quantity of output.*

In the figure, Small Time is fortunate: At a price of $800, there are several output levels at which it can earn a profit. Its problem is to select the one that makes its profit as large as possible. (We should all wish for such problems.)

Figure 3: Measuring Profit or Loss



The competitive firm in panel (a) produces where marginal cost equals marginal revenue, or 7 units of output per day. Profit per unit at that output level is equal to revenue per unit ($800) minus cost per unit ($600), or $100 per unit.

Total profit (indicated by the blue-shaded rectangle) is equal to profit per unit times the number of units sold, $200 x *7 = $1,400. In panel (b), we assume that the market price is lower, at $400 per ounce.*

The best the firm can do is to produce 5 ounces per day and suffer a loss shown by the red area. It loses $200 per ounce on each of those 5 ounces produced, so the total loss is $1,000—the area of the shaded rectangle (panel b).

But what if the price had been lower than $800—so low, in fact, that Small Time could not make a profit at *any* output level? Then the best it can do is to choose the smallest possible loss. Just as we did in the case of profit, we can measure the firm's total loss using the *ATC* curve.

Panel (b) of Figure 3 reproduces Small Time's *ATC* and *MC* curves from panel (a). This time, however, we have assumed a lower price for gold— $400—so the firm's *d* _ *MR* curve is the horizontal line at $400. Since this line lies everywhere below the *ATC* curve, profit per unit ($P = ATC$) is always negative: Small Time cannot make a positive profit at *any* level.

With a price of $400, the *MC* curve crosses the *MR* curve from below at 5 units of output. Unless Small Time decides to shut down (we'll discuss shutting down for competitive firms later), it should produce 5 units.

At that level of output, *ATC* is $600, and profit per unit is $P = ATC$ = $400 - $600 = - $200, a *loss* of $200 per unit. The total loss is loss per unit (negative profit per unit) times the number of units produced, or - $200 x 5 = - $1,000. This is the area of the red-shaded rectangle in Figure 3(b), with height of $200 and width of 5 units:

A firm suffers a loss whenever P < ATC *at the best level of output. Its total loss equals the area of a rectangle with height equal to the distance between* P *and* ATC*, and width equal to the quantity of output.*

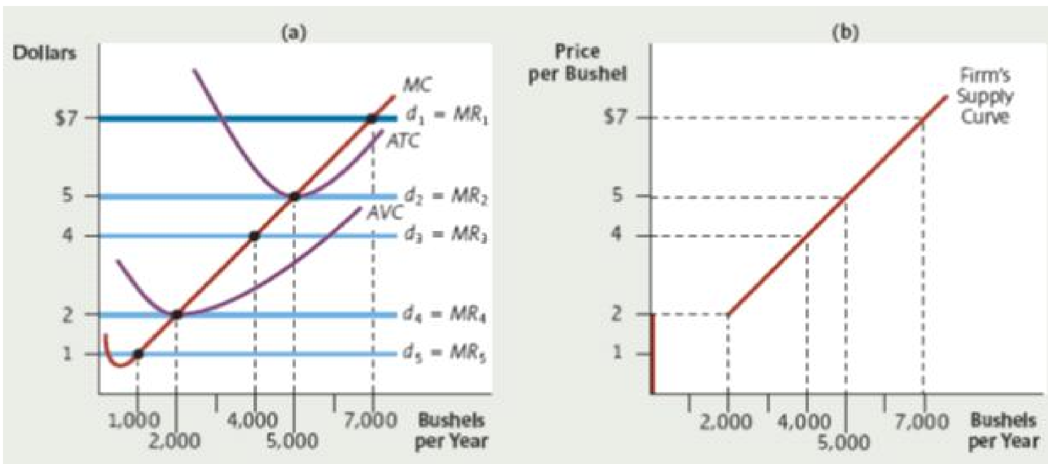
Firm's Short-Run Supply Curve

A competitive firm is a price taker: It takes the market price as given and then decides how much output it will produce at that price. If the market price changes for any reason, the price taken as given by the firm will change as well. The firm will then have to find a new profit-maximizing output level. Let's see how the firm's profit-maximizing output changes as the market price rises or falls.

Figure 4(a) shows *ATC, AVC,* and *MC* curves for a competitive producer of wheat. The figure also shows five hypothetical demand curves the firm might face, each corresponding to a different market price for wheat. If the market price were $7 per bushel, the firm would face demand curve *d*1, and its profit-maximizing output level—where *MC* and *MR* intersect—would be 7,000 bushels per year. If the price dropped to $5 per bushel, the firm would face demand curve *d*2, and its profit-maximizing output level would drop to 5,000 bushels.

Profit-maximizing output level is always found by traveling from the price, across to the firm's *MC* curve, and then down to the horizontal axis. In other words, *as the price of output changes, the firm will slide along its* MC *curve in deciding how much to produce.*

Figure 4: Short-Run Supply under Perfect Competition



Panel (a) shows a typical competitive firm facing various market prices.

For prices between $2 and $7 per bushel, the profit maximizing quantity is found by sliding along the MC *curve.*

Below $2 per bushel, the firm is better off shutting down, because P < AVC*, Panel (b) shows that the firm's supply curve consists of two segments.*

Above the shutdown price of $2 per bushel it follows the MC *curve; below that price, it is coincident with the vertical axis.*

But there is one problem with this: If the firm is suffering a loss—a loss large enough to justify shutting down—then it will *not* produce along its *MC* curve; it will produce zero units instead. Thus, in order to know for certain how much output the firm will produce, we must bring in the shutdown rule.

Shutdown Price

A firm should shut down in the short run if, at its best positive output level, it finds that $TR < TVC$. In words, if the firm cannot even cover its operating costs, it should not continue to operate.

If $TR > TVC$, the firm should continue to operate.

But when we use a graph such as Figure 4(a), which has different prices *per unit* on the vertical axis and has curves showing cost *per unit,* it will be helpful to express this shutdown rule in "per unit" terms.

Shut down if $TR < TVC$

Next, with lowercase q representing the individual firm's output level, we divide both sides of the inequality by q:

Shut down if $(TR/q) < (TVC/q)$

Finally, we recognize that TR/q is just revenue per unit, or the price (P), and TVC/q is the firm's average variable cost (AVC), giving us

Shut down if $P < AVC$

Now let's apply the shutdown rule to the firms in Figure 4(a).

Suppose the price drops down to $4 per bushel.

At this price, the best output level is 4,000 bushels, and the firm suffers a loss, since $P < ATC$. Should the firm shut down?

Let's see. At 4,000 bushels, it is also true that $P > AVC$, since the demand curve lies above the AVC curve at this output level.

Thus, at a price of $4, the firm will stay open and produce 4,000 units of output.

Now, suppose the price drops all the way down to $1 per bushel. At this price, $MR = MC$ at 1,000 bushels. But notice that here $P < AVC$. Therefore, at a price of $1, this firm will shut down and produce *zero* units of output.

Finally, let's consider a price of \$2. At this price, $MR = MC$ at 2,000 bushels, and here we have $P = AVC$. At \$2, therefore, the firm will be indifferent between staying open and shutting down. We call this price the firm's **shut-down price,** since it will shut down at any price lower and stay open at any price higher.

The shutdown price is found at the *minimum* of the AVC curve. Why?

As the price decreases, the best output level is found by sliding along the MC curve, until MC and AVC cross. At that point, the firm will shut down. MC will always cross AVC at its minimum point.

Now let's recapitulate what we've found about the firm's output decision. For all prices above the minimum point on the AVC curve, the firm will stay open and will produce the level of output at which $MR = MC$. For these prices, the firm slides

Shutdown price The price at which a firm is indifferent between producing and shutting down.

Questions

1. Assume that the market for cardboard is perfectly competitive (if not very exciting). In each of the following scenarios, should a typical firm continue to produce or should it shut down in the short run?

   Draw a diagram that illustrates the firm's situation in each case.

   a. Minimum $ATC = 2.00$

      Minimum $AVC = 1.50$

      Market price = 1.75

   b. $MR = 1.00$

      Minimum $AVC = 1.50$

      Minimum $ATC = 2.00$

2. "*A profit-maximizing* competitive firm will produce the quantity of output at which price *exceeds* cost per unit by the greatest possible amount." True or false? Explain briefly.

3. Elaborate the characteristics of perfect competition.

4. Explain profit-maximizing output level of perfectly competitive firm using:

    a. TR-TC Approach

    b. MC-MR Approach

5. Explain diagrammatically shut down price in perfect competition.

This lesson will focus on the following:

1. Competitive Markets: Short-Run and Long-Run Equilibrium
2. Perfect Competition and Plant Size
3. Competitive Firm and Changes in Demand

Competitive Markets in the Short Run

Short run is a time period too short for the firm to vary its fixed inputs. Short run is also insufficient time for a *new* firm to acquire those fixed inputs and *enter* the market. Similarly, it is too short a period for firms to reduce their fixed inputs to zero and *exit* the market. *In the short run, the number of firms in the industry is fixed.*

Market supply curve

To obtain the ***market supply curve,*** *we add up the quantities of output supplied by all firms in the market at each price.*

Suppose there are 100 identical wheat farms. Each one has the supply curve shown in Figure 5(a). If the price is \$7, each firm produces 7,000 bushels. With 100 such firms, the market quantity supplied is 7,000 x 100 = 700,000 bushels. If the price is \$5, each firm supplies 5,000 bushels, so market supply is 500,000.

Market supply curve is shown in panel (b) of Figure 5. Once the price drops below \$2—the shutdown price for each firm—the market supply curve jumps to zero.

The market supply curve in the figure is a *short-run* market supply curve, since it gives us the combined output level of just those firms *already* in the industry.

Two things are assumed constant:

(1) the fixed inputs of each firm and

(2) the number of firms in the market.

Short-run Equilibrium

How does a perfectly competitive market achieve equilibrium?

Firm's supply curve

A curve that shows the quantity of output a competitive firm will produce at different prices.

Market supply curve

A curve indicating the quantity of output that all sellers in a market will produce at different prices in the short run.

Figure 1 Deriving the Market Supply Curve

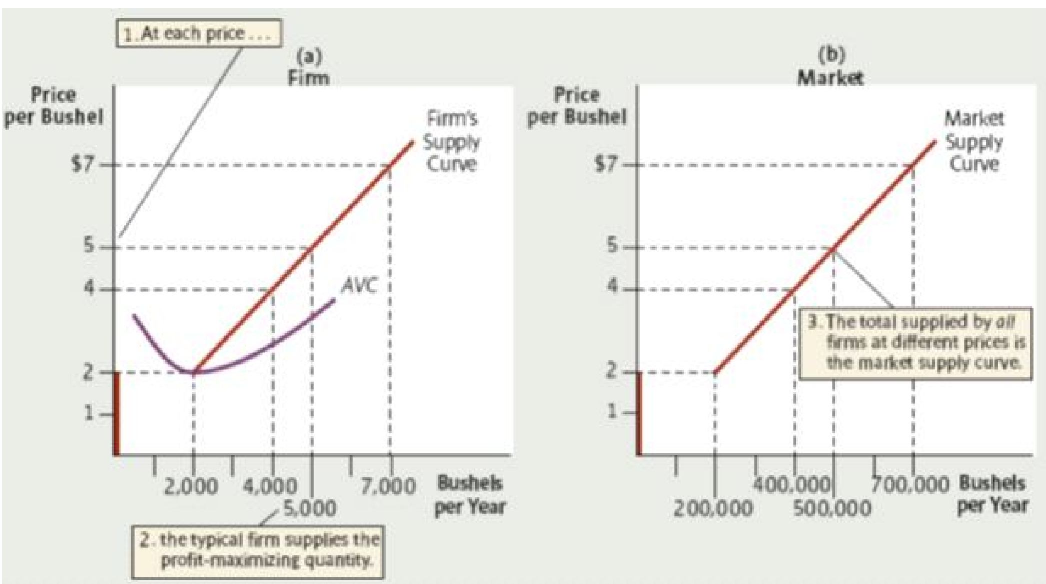


Figure 2 puts information on consumer choice. It paints a complete picture of how a competitive market arrives at its short-run equilibrium. On the right side, we add up the quantities supplied by all firms to obtain the market supply curve. On the left side, we add up the quantities demanded by all consumers to obtain the market demand curve.

At this stage, the market supply and demand curves show if/then relationships:

If the price were such and such, *then* firms would supply this much and consumers would buy that much.

But once we bring the two curves together and find their intersection point, we know the *equilibrium* price at which trading will actually take place.

Finally, when we confront each firm and each consumer with the equilibrium price, we find the actual quantity each consumer will buy and the actual quantity each firm will produce.

Figure 3 gets more specific, illustrating two possible short-run equilibriums in the wheat market, depending on the position of the market demand curve.

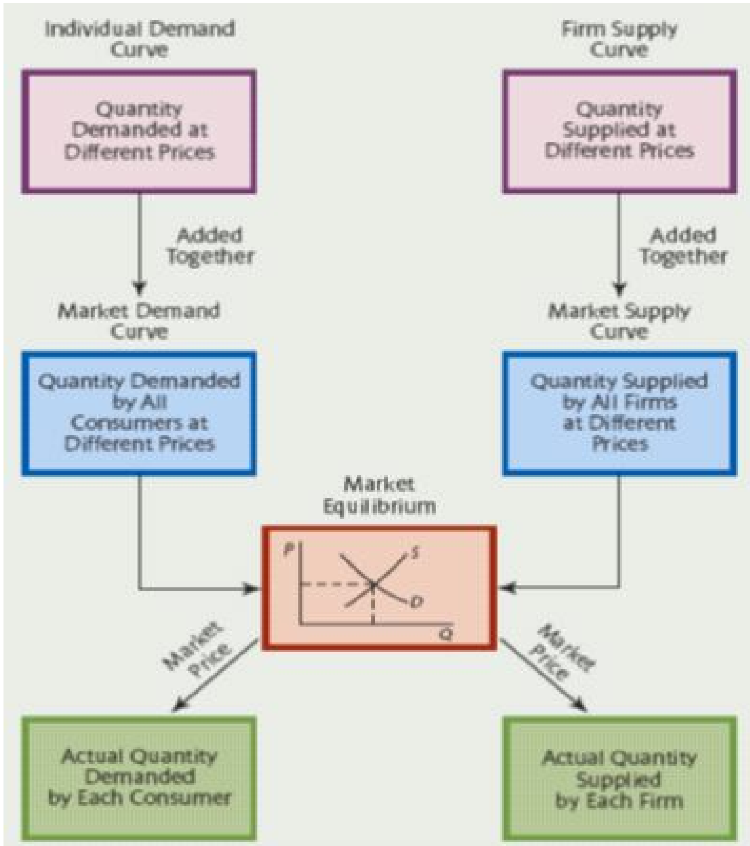
In panel (a), if the market demand curve were $D1$, the short-run equilibrium price would be $7.

Each firm would face the horizontal demand curve $d1$ [panel (b)] and decide to produce 7,000 bushels. With 100 such firms, the equilibrium market quantity would be 700,000 bushels. Notice that, at a price of $7, each firm is enjoying an economic profit, since $P = ATC$.

If the market demand curve were $D2$ instead, the equilibrium price would be $4. Each firm would face demand curve $d2$ and produce 4,000 bushels. With 100 firms, the equilibrium market quantity would be 400,000. Here, each firm is suffering an economic loss, since $P = ATC$.

These two examples show us that *in short-run equilibrium, competitive firms can earn an economic profit or suffer an economic loss.*

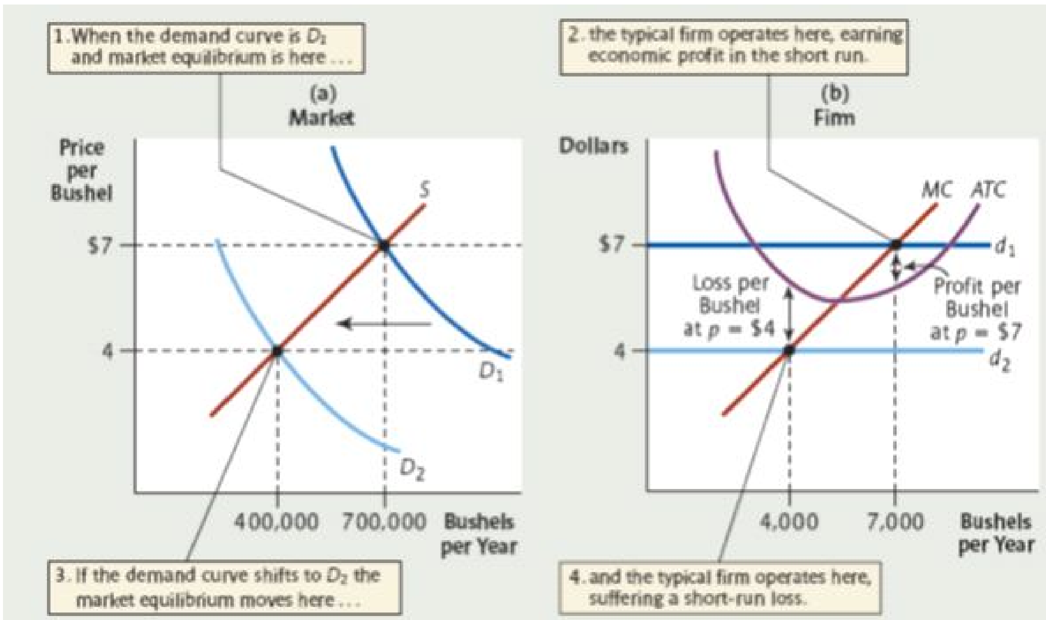
Figure 2 Perfect Competition



Equilibrium

We are about to leave the short run and turn our attention to what happens in a competitive market over the long run. But before we do, let's look once more at how a short-run equilibrium is established. One part of this process—combining supply and demand curves to find the market equilibrium—has been familiar to you all along. But now you can see how much information is contained within each of these curves. And you can appreciate what an impressive job the market does—coordinating millions of decisions made by people who may never even meet each other.

Figure 3: Short-Run Equilibrium in Perfect Competition



So many individual consumers and firms, each with its own agenda, trading in the market.

Not one of them has any power to decide or even influence the market price.

Rather, the price is determined by *all* of them, adjusting until *total* quantity supplied is equal to *total* quantity demanded.

Then, facing this equilibrium price, each consumer buys the quantity he or she wants, each firm produces the output level that it wants, and we can be confident that all of them will be able to realize their plans.

Each buyer can find willing sellers, and each seller can find willing buyers.

In perfect competition, the market sums up the buying and selling preferences of individual consumers and producers, and determines the market price. Each buyer and seller then takes the market price as given, and each is able to buy or sell the desired quantity.

Competitive Markets in the Long Run

The long run is a time horizon sufficiently long for firms to vary *all* of their inputs. This includes inputs that were treated as fixed in the short run, such as plant and equipment. Logically, then, the long run must be enough time for *new* firms to acquire those inputs and enter the market as *new* suppliers. And it is also long enough for existing firms to sell all such inputs and exit the market.

In the long run, new firms can enter a competitive market, and existing firms can exit the market.

But what makes firms want to enter or exit a market? The driving force behind entry is economic profit, and the force behind exit is economic loss.

Profit and Loss and Long Run

Economic profit is the amount by which total revenue exceeds *all* costs of doing business. The costs we deduct include implicit costs like foregone investment income or foregone wages for an owner who devotes money or time to the business. Thus, when a firm earns positive economic profit, we know the owners are earning *more* than they could by devoting their money and time to some other activity.

A temporary episode of positive economic profit will not have much impact on a competitive industry, other than the temporary pleasure it gives the owners of competitive firms. But when positive profit reflects basic conditions in the industry and is expected to continue, major changes are in the works. Outsiders, hungry for profit themselves, will want to enter the market and—since *there are no barriers to entry*—they can do so.

On the other hand, if firms already in the industry are suffering economic losses, they are not earning enough revenue to cover all their costs. There must be other opportunities that would more adequately compensate the owners for their money or time. If this situation is expected to continue over the firm's long-run planning horizon—a period long enough to vary *all* inputs—there is only one thing for the firm to do: exit the industry by selling off its plant and equipment, thereby reducing its loss to zero.

In a competitive market, economic profit and loss are the forces driving long-run change. The expectation of continued economic profit causes outsiders to enter the *market; the expectation of continued economic losses causes firms in the market to exit.*

Long run exit from a market can occur in different ways. An example is when a firm (such as this retailer) goes entirely out of business.

Long Run Equilibrium

Entry and exit—however they occur—are powerful forces in real-world competitive markets. They determine how these markets change over the long run, how much output will be available to consumers, and the prices they must pay. To explore these issues, let's see how entry and exit move a market to its long-run equilibrium from different starting points.

From Short-Run Profit to Long-Run Equilibrium

Suppose that the market for wheat is initially in a short-run equilibrium like point *A* in panel (a) of Figure 4, with market supply curve *S*1.

The initial equilibrium price is $9 per bushel.

In panel (b), we see that a typical competitive firm—producing 9,000 bushels—is earning economic profit, since $P = ATC$ at that output level. As long as we remain in the short run, with no new firms entering the market, this situation will not change.

But as we enter the long run, much will change. First, economic profit will attract new entrants, increasing the number of firms in the market. When we draw a market supply curve like *S*1, we draw it for some *given* number of firms, and we hold that number constant. But in the long run, as the number of firms increases, the market supply curve will *shift rightward;* a greater quantity will be supplied at any given price. As the market supply curve shifts rightward, several things happen:

1. The market price begins to fall—from $9 to $8 to $7 and so on.

2. As market price falls, the horizontal demand curve facing each firm shifts downward.

3. Each firm—striving as always to maximize profit—will slide down its marginal cost curve, decreasing output.

This process of adjustment, in the market and the firm, continues until the *reason* for entry—positive profit—no longer exists. That is, it will continue until the market supply curve shifts rightward enough, and the price falls enough, so that *each existing firm is earning zero economic profit*.

Panels (c) and (d) in Figure 4 show the final, long-run equilibrium.

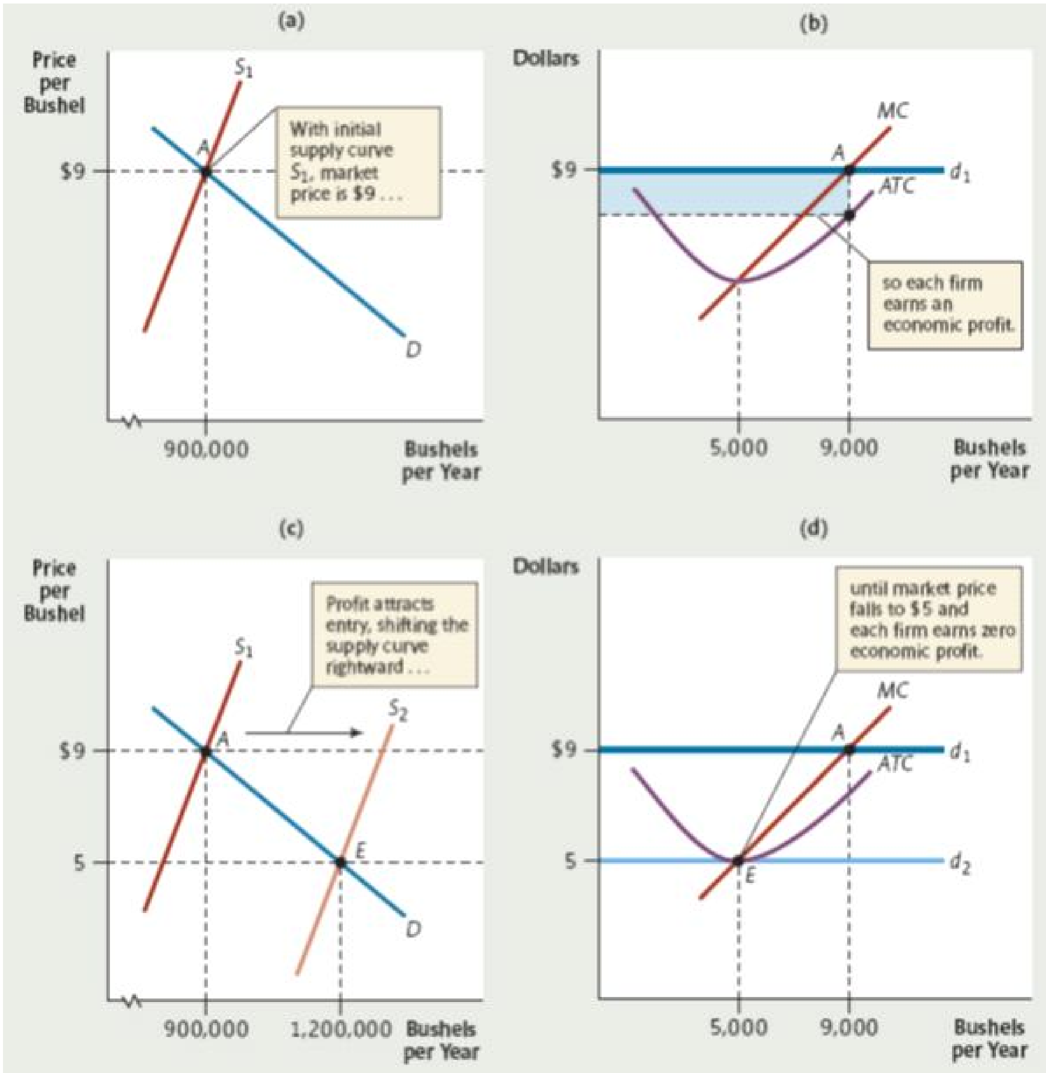
First, look at panel (c), which shows long-run market equilibrium at point *E*. The market supply curve has shifted to *S*2, and the price has fallen to $5 per bushel.

Next, look at panel (d), which tells us why the market supply curve stops shifting when it reaches *S*2. With that supply curve, each firm is producing at the lowest point of its *ATC* curve, with $P = ATC = \$5$, and each is earning zero economic profit.

With no economic profit, there is no further reason for entry, and no further shift in the market supply curve.

In a competitive market, positive economic profit continues to attract new entrants until economic profit is reduced to zero.

Figure 4: From Short-Run Profit to Long-Run Equilibrium



With no significant barriers to entry, we can be confident that economic profit at the typical firm will attract new firms to the industry, driving down the market price until the economic profit disappears.

If a permanent barrier— legal or otherwise—prevented new firms from coming into the market, this mechanism would not work, so long-run economic profit would be possible.

Before proceeding further, take a close look at Figure 4. As the market moves to its long-run equilibrium [point *E* in panels (c) and (d)], output at each firm *decreases* from 9,000 to 5,000 bushels. But in the market as a whole, output *increases* from 900,000 to 1,200,000 bushels. How can this be?

From Short-Run Loss to Long-Run Equilibrium

We have just seen how, beginning from a position of short-run profit at the typical firm, a competitive market will adjust until the profit is eliminated. But what if we begin from a position of loss? As you might guess, the same type of adjustments will occur, only in the opposite direction.

What happens in the market, and at each firm, as economic loss causes some firms to exit.

In a competitive market, economic losses continue to cause exit until the losses are reduced to zero.

When there are no significant barriers to exit, we can be confident that economic loss will eventually drive firms from the industry, raising the market price until the typical firm breaks even again. Significant barriers to exit which would prevent this mechanism from working, and economic losses could persist even in the long run.

Concept of Zero Profit in Perfect Competition

In the long run, firms can expect zero economic profit.

Zero *economic* profit is not the same as zero *accounting* profit. When a firm is making zero *economic* profit, it is still making some accounting profit. In fact, the accounting profit is just enough to cover all of the owner's implicit costs, including compensation for any foregone investment income or foregone salary.

Suppose, for example, that a farmer paid \$100,000 for land and works 40 hours per week. Suppose, too, that the \$100,000 *could* be invested in some other way and earn \$6,000 per year, and the farmer *could* work equally pleasantly elsewhere and earn \$50,000 per year. Then the farm's implicit costs will be \$56,000, and zero economic profit means that the farm is earning \$56,000 in *accounting profit* each year. This won't make a farmer ecstatic, but it will make it worthwhile to keep working the farm. After all, if the farmer quits and takes up the next best alternative, he or she will do no better.

To emphasize that zero economic profit is not an unpleasant outcome, economists often replace it with the term **normal profit,** which is a synonym for "zero economic profit," or "just enough accounting profit to cover implicit costs."

We can summarize long-run conditions at the typical firm this way:

In the long run, the competitive firm will earn normal profit—that is, zero economic profit.

Perfect Competition and Plant Size

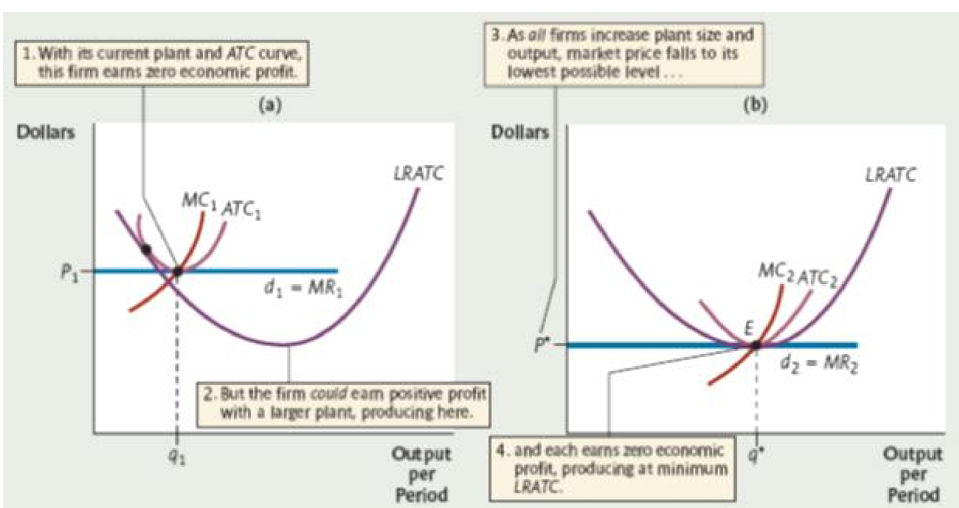
Plant size is characteristic of competitive markets in the long run. Entry and exit—that cause all firms to earn zero economic profit *also* determine the size of each firm's plant.

In long-run equilibrium, a competitive firm will operate with the plant and output level that bring it to the bottom of its LRATC *curve.*

To see why, let's consider what would happen if this condition were violated. Figure 5(a) illustrates a firm in a perfectly competitive market. The firm faces a market price of $P1$ and produces quantity $q1$, where $MC1 = MR1$. With its current plant, the firm has average costs given by $ATC1$. Note that the firm is earning zero profit, since average cost is equal to $P1$ at the best output level.

Normal profit Another name for zero economic profit.

Figure 5: Perfect Competition and Plant Size



But panel (a) does *not* show a true long-run equilibrium.

How do we know this?

First, in the long run, the typical firm will want to expand.

Why?

Because by increasing its plant size, it could slide down its *LRATC* curve and produce more output at a lower cost per unit. Since it is a perfectly competitive firm—a small participant in the market—it can expand in this way *without* worrying about affecting the market price.

As a result, the firm, after expanding, could operate on a new, lower *ATC* curve, so that *ATC* is less than *P*. That is, by expanding, the firm could potentially earn an economic profit.

Second, this same opportunity to earn positive economic profit will attract new entrants that will establish larger plants from the outset.

Expansion by existing firms and entry by new ones increase market output and bring down the market price. The process will stop—and a long-run equilibrium will be established—only when there is no potential to earn positive economic profit with *any* plant size.

As you can see in panel (b), this condition is satisfied only when each firm is operating at the minimum point on its *LRATC* curve, using the plant represented by $ATC2$, and producing output of q^*. Entry and expansion must continue in this market until the price falls to P^* because only then will each firm—doing the best that it can do—earn zero economic profit.

Summary of Competitive Firm in the Long Run

Panel (b) of Figure 5 summarizes about the competitive firm in long-run equilibrium.

The typical firm, taking the market price P^* as given, produces the profit-maximizing output level q^*, where $MR = MC$.

Since this is the long run, each firm will be earning zero economic profit, so we also know that $P^* = ATC$.

But since $P^* = MC$ and $P^* = ATC$, it must also be true that $MC = ATC$.

MC and ATC are equal only at the minimum point of the ATC curve. Thus, we know that each firm must be operating at the lowest possible point on the ATC curve for the plant it is operating.

Finally, each firm selects the plant that makes its *LRATC* as low as possible, so each operates at the minimum point on its *LRATC* curve.

In long-run equilibrium, the competitive firm operates where MC = *minimum* ATC = *minimum* LRATC = P.

In Figure 5(b), this equality is satisfied when the firm produces at point *E,* where its demand, marginal cost, *ATC,* and *LRATC* curves all intersect.

Figure 5(b) also explains one of the important ways in which perfect competition benefits consumers.

In the long run, each firm is driven to the plant size and output level at which its cost per unit is as low as possible. This lowest possible cost per unit is also the price per unit that consumers will pay. If price were any lower than P^*, it would not be worthwhile for firms to continue producing the good in the long run. Thus, given the *LRATC* curve faced by each firm in this industry—a curve that is determined by the technology of production and the prices of its inputs—P^* is the lowest possible price that will ensure the continued availability of the good. In perfect competition, consumers are getting the best deal they could possibly get.

What Happens When Things Change?

A Change in Demand

In Figure 6, panel (a) shows a competitive market that is initially in long-run equilibrium at point *A*, where the market demand curve $D1$ and supply curve $S1$ intersect. Panel (b) shows conditions at the firm, which faces demand curve $d1$ and produces the profit-maximizing quantity $q1$.

But now suppose that the market demand curve shifts rightward to $D2$ and remains there.

In the *short run*, the shift in demand moves the market equilibrium to point *B,* with market output *Q*SR and price *P*SR. At the same time, the demand curve facing each firm shifts upward, and each firm raises output to the new profit-maximizing level *q*SR. At this output level, $P = ATC$, so each firm is earning economic profit.

Thus, the short-run impact of an increase in demand is

(1) a rise in market price,

(2) a rise in market quantity, and

(3) economic profits

In the long run, economic profit will attract the entry of new firms. An increase in the number of firms shifts the market supply curve rightward, which drives down the price until the economic profit is eliminated.

But how far must the price fall in order to bring this about?

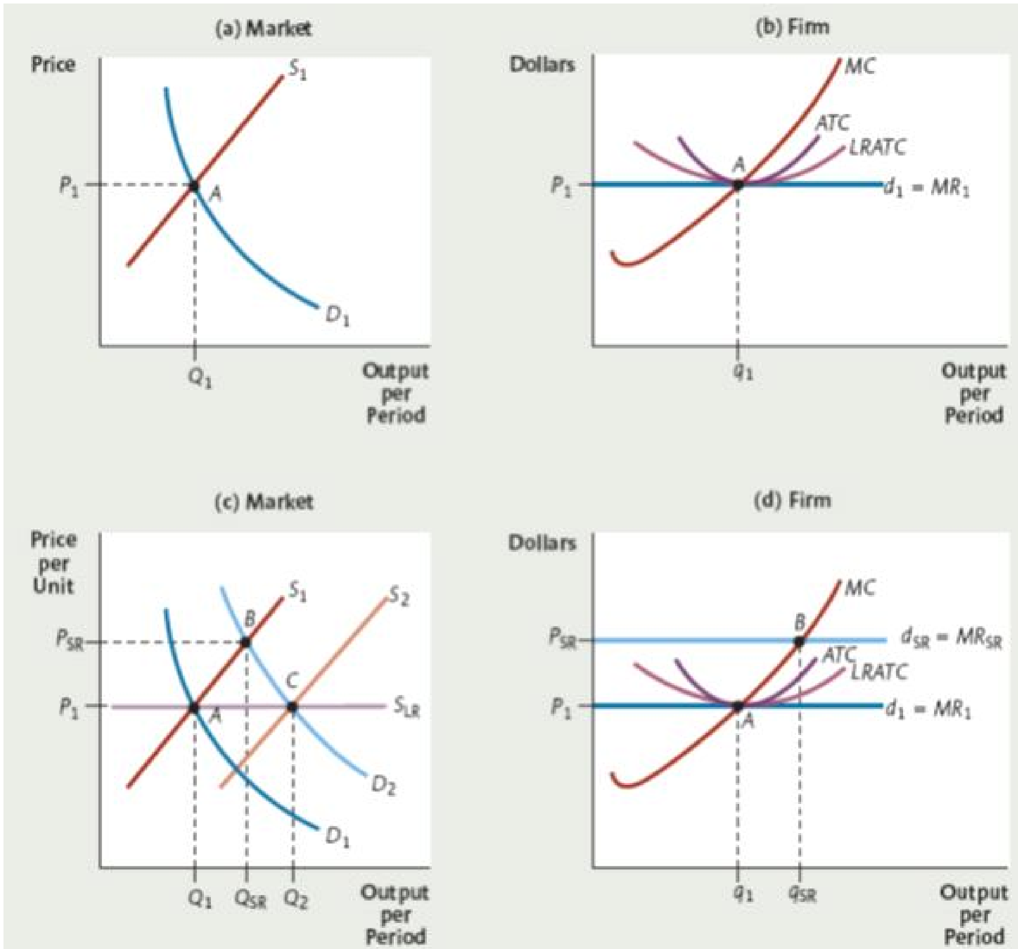
How far can we expect the market supply curve to shift?

That depends on whether or not the expansion of the industry causes each firm's cost curves to shift.

A Constant Cost Industry

Let's assume, for now, that a change in industry output (such as when new firms enter) has no impact on the cost curves of the individual firm. This is called a **constant cost industry**. Then in panel (c), entry will continue—and the supply curve will continue shifting rightward—until the price returns to *P*1, its original level. (At any higher price, each firm would still be earning economic profit, and new firms would still be entering.) Our new long-run equilibrium occurs at point *C*, with the supply curve *S*2, price *P*1, and market quantity *Q*2. Panel (d) shows what happens at the typical firm: The price moves back to *P*1, so the demand curve facing the firm shifts back to *d*1, and the typical firm returns to its original level of output *q*1.

Figure 6: A Constant Cost Industry



At point A *in panel (a), the market is in long-run equilibrium. The typical firm in panel (b) operates at the minimum of its* ATC *and* LRATC *curves, and earns zero economic profit.*

The lower panels show what happens if demand increases. In the short run, the market reaches a new equilibrium at point B *in panel (c), and the typical firm in panel (d) earns economic profit at the higher price* PSR*. In the long run, profit attracts entry, increasing market supply and lowering price.*

Entry continues until economic profit at the typical firm in panel (d) is reduced to zero, which requires the price to drop to P1*, its original level.*

In panel (d), the typical firm returns to point A*, and in panel (c), the new long-run market equilibrium is point* C*. The increase in demand raises output, but leaves price unchanged, as shown by the horizontal long-run supply curve connecting points* A *and* C*.*

What happens in the *long run* after the demand curve shifts rightward?

The answer is: The market equilibrium will move from point *A* to point *C*. A line drawn through these two points tells us, in the long run, the market price we can expect for any quantity the market provides.

In Figure 6, this is the thin line, which is called the *long-run supply curve* (*S*LR).

The ***long-run supply curve*** *shows the relationship between market price and market quantity produced after all long-run adjustments have taken place.*

In case of a constant cost industry, the long-run supply curve is horizontal.

In a constant cost industry, in which industry output has no effect on individual firms' cost curves, the long-run supply curve is horizontal. In the long-run, the industry will supply any amount of output demanded at an unchanged price.

An Increasing Cost Industry

In an **increasing cost industry**, the entry of new firms that use the same inputs as existing firms drives up input prices. This, in turn, causes each firm's *LRATC* curve to shift upward.

If the demand for wheat increased significantly, existing wheat farms would expand, and new farms would enter the industry. The price of farmland would rise.

Because *every* farm in the industry—the existing ones as well as new entrants— would have to pay more for farmland, their *LRATC* curves would shift upward (greater cost per unit at each output level).

Let's see how this changes the graphical analysis of an increase in demand.

Panel (a) in Figure 7 shows a competitive market in an initial long-run equilibrium at point *A*.

Panel (b) shows the situation of a single competitive firm in this market, facing demand curve $d1$ and producing output level $q1$. To keep the diagram simple, we've left out the *MC* and *ATC* curves for the firm and show the only cost curve that will matter to our analysis: the *LRATC* curve. Initially, the firm operates at the minimum point of *LRATC*1.

Now suppose the demand curve shifts rightward to *D*2 [panel (a)]. As a result, the short-run market equilibrium moves to point *B*, and price rises to *P*SR. Because the typical firm enjoys economic profit (not shown), entry will occur in the long run, and the market supply curve shifts rightward. As usual, the supply curve will continue shifting rightward until economic profit is eliminated.

But this time, the entry of new firms and the rise in industry output causes the typical firm's *LRATC* curve to shift *upward* to *LRATC*2. With higher long-run average cost, zero profit will occur at a price *higher* than the original price *P*1. In Figure 7, the supply curve stops shifting when the price reaches *P*2, with the new market equilibrium at point *C*. As panel (b) shows, once the price reaches *P*2, the typical firm—facing the horizontal demand curve *d*2—operates at the minimum point on *LRATC*2, earning zero economic profit.

Increasing cost industry An industry in which the long-run supply curve slopes upward because each firm's *LRATC* curve shifts upward as industry output increases.

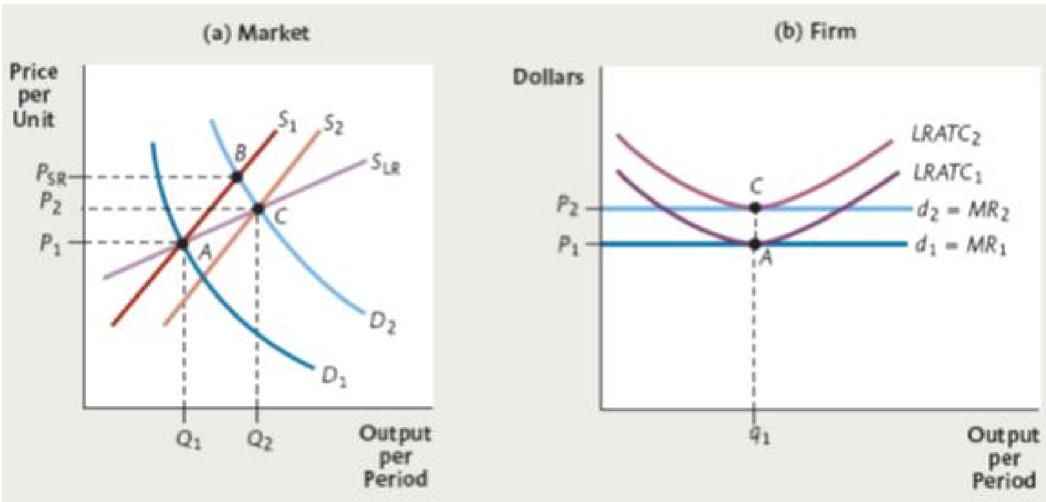
Long-run supply curve A curve indicating price and quantity combinations in an industry after all long run adjustments have taken place.

Let's now concentrate on just the long-run impact of the change in demand, which moves the equilibrium from point *A* to point *C*. Connecting these two equilibrium points gives us the long-run supply curve for this industry. As you can see, the curve slopes *upward,* telling us that the industry will supply greater output, but only with a higher price.

In an increasing cost industry, *a rise in industry output shifts up each firm's* LRATC *curve, so that zero economic profit occurs at a higher price. The long-run supply curve slopes upward.*

The long-run supply curve tells us that an increasing cost industry will deliver more output, but only at a higher price. It also tells us that, if industry output decreases, the price will drop. This is because a decrease in output would cause each firm's *LRATC* curve to shift *downward* so that zero profit would be established at a *lower* price than initially.

Figure 7: An Increasing Cost Industry



Point A *in both panels shows the initial long-run market equilibrium, with the typical firm earning zero economic profit. After demand increases, the market reaches a new short-run equilibrium at point* B *in panel (a).*

At the higher price, the typical firm earns economic profit (not shown). In the long run, profit attracts entry, supply increases and price begins to fall. But in an increasing cost industry, the rise in industry output also causes costs to rise, shifting up the LRATC *curve. In the final, long-run market equilibrium (point* C *in both panels), price at* P2 *is higher than originally, and the typical firm once again earns zero economic profit.*

The increase in demand raises both output and price, as shown [in panel (a)] by the upward-sloping long-run supply curve.

A Decreasing Cost Industry

In a **decreasing cost industry,** a rise in industry output causes input prices to *fall,* and the *LRATC* curve to shift downward at each firm. This might occur for a number of reasons.

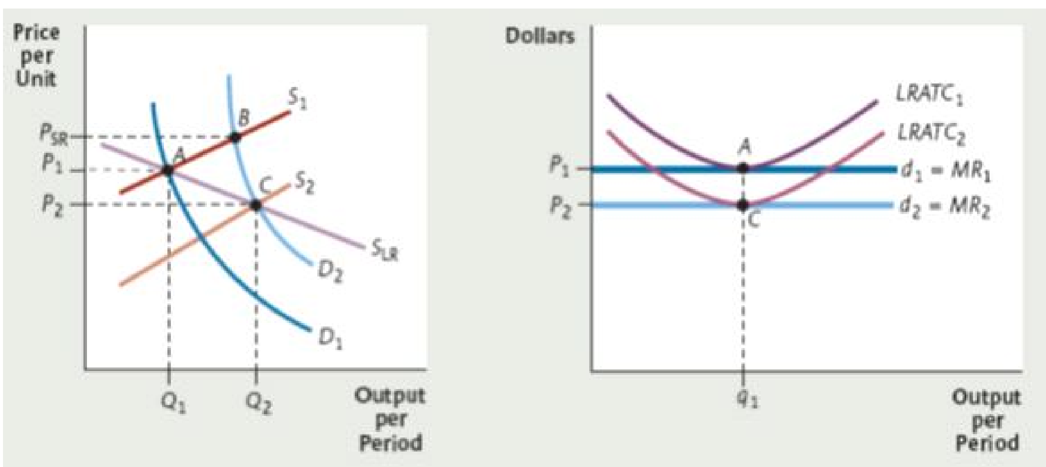
As an industry expands, there might be more workers in the area with the needed skills, making it easier and less expensive for each firm to find and recruit qualified employees.

Or transportation costs might decrease. For example, suppose that a modest size city has just a few restaurants. Periodically, a partially loaded truck makes a special trip from a distant larger city to deliver raw fish, and other special ingredients to these few restaurants. Transportation costs—part of the price of the ingredients—will be rather high.

Now suppose that demand increases. Profits at the existing restaurants attract entry. With more restaurants ordering ingredients, the same delivery truck makes the same trip, but now it is fully loaded and the transportation costs are shared among more restaurants.

As a result, transportation costs at *each* restaurant decrease—and each restaurant's *LRATC* curve shifts down. Competition among the restaurants then ensures that prices will drop to match the lower *LRATC*. As a result, the long-run effect of an increase in demand is a *lower* price for eating sushi at a restaurant—a downward sloping long-run supply curve.

Figure 8: A Decreasing Cost Industry



Point A *in both panels shows the initial long-run market equilibrium, with the typical firm earning zero economic profit.*

After demand increases, the market reaches a new short-run equilibrium at point B in panel (a). At the higher price, the typical firm earns economic profit (not shown).

In the long run, profit attracts entry, supply increases and price begins to fall. But in a decreasing cost industry, the rise in industry output causes costs to fall, shifting down the LRATC *curve.*

In the final, long-run market equilibrium (point C *in both panels), price at* P2 *is lower than originally, and the typical firm once again earns zero economic profit. The increase in demand raises output but lowers price, as shown [in panel (a)] by the downward-sloping long-run supply curve.*

Figure 8 illustrates how a decreasing cost industry behaves after an increase in demand. In panel (a), after the demand curve shifts rightward, the market equilibrium moves from *A* to *B* in the short run. The typical firm earns economic profit (not shown). In the long run, profit causes entry. But now, as the industry expands, the *LRATC* curve at each firm shifts *downward.* With lower cost per unit, zero economic profit occurs at a long-run equilibrium price *lower than the original price.* In the figure, the market reaches its new long-run equilibrium at point *C,* at the new, lower price *P*2.

When we draw a line through the initial equilibrium at point *A* and the new long-run equilibrium at point *C,* we get the long-run supply curve for this industry. As you can see, the curve slopes downward: In a decreasing cost industry, as industry output rises, the *price drops.*

Conclusion

In a decreasing cost industry, *a rise in industry output shifts down each firm's* LRATC *curve, so that zero economic profit occurs at a lower price. The long-run supply curve slopes downward.*

The long-run supply curve tells us that in a decreasing cost industry, the more output produced, the lower the price. On the other hand, if industry output were to fall, the price would rise. This is because a decrease in output would cause each firm's *LRATC* curve to shift *upward,* so that zero profit would be established at a *higher* price than initially.

Questions

1. Explain short-run and long-run equilibrium of competitive markets.
2. How perfect competition influences the plant size>
3. How competitive firm behave with changes in demand
4. Which of the following is not the short-run impact of an increase in demand

(a) a rise in market price,

(b) a rise in market quantity, and

(c) economic profits

(d) None of the above

5. In a perfectly competitive, increasing cost industry, is the long-run supply curve always flatter than the short-run market supply curve? Explain.

This lesson will focus on the following:

1. Market Signals and Economy
2. Change in Demand and Reallocation of Resources
3. Technological Change in Perfect Competition

Market Signals and the Economy

An *increase* in demand always leads to an *increase* in market output in the short run, as existing firms raise their output levels, and an even *greater* increase in output in the long run, as new firms enter the market.

What happens when demand *decreases*?

The leftward shift of the demand curve will cause a drop in output in the short run and an even greater drop in the long run. The effect on price will depend on the nature of the industry, (i.e., whether it is a constant, increasing, or decreasing cost industry).

In the real world, the demand curves for different goods and services are constantly shifting. For example, over the last couple of decades, people have developed an increased taste for bottled water.

As a consequence, the *production* of bottled water has increased dramatically. Consumers want more bottled water and the economy provides it.

As demand increases or decreases in a market, *prices change*. And price changes act as *signals* for firms to enter or exit an industry. How do these signals work?

When demand increases, the price tends to initially *overshoot* its long-run equilibrium value during the adjustment process, creating sizable temporary profits for existing firms. Similarly, when demand decreases, the price falls *below* its long-run equilibrium value, creating sizable losses for existing firms. These exaggerated, temporary movements in price, and the profits and losses they cause, are almost irresistible forces, pulling new firms into the market or driving existing firms out. In this way, the economy is driven to produce whatever collection of goods consumers prefer.

Figure 1 illustrates the process. In the upper panel, as people shifted their tastes toward bottled water, the market demand curve for this good shifted rightward from D1 to D2. Initially, the price rose *above* its new long-run equilibrium value, to P2, leading to high profits at existing bottled water firms.

High profits, in the long run, attracted entry—especially the entry of new brands from established firms not previously selling bottled water. Entry shifted the supply curve rightward, to *S*2, bringing the price back down to *P*1.

We are viewing bottled water as a constant cost industry.

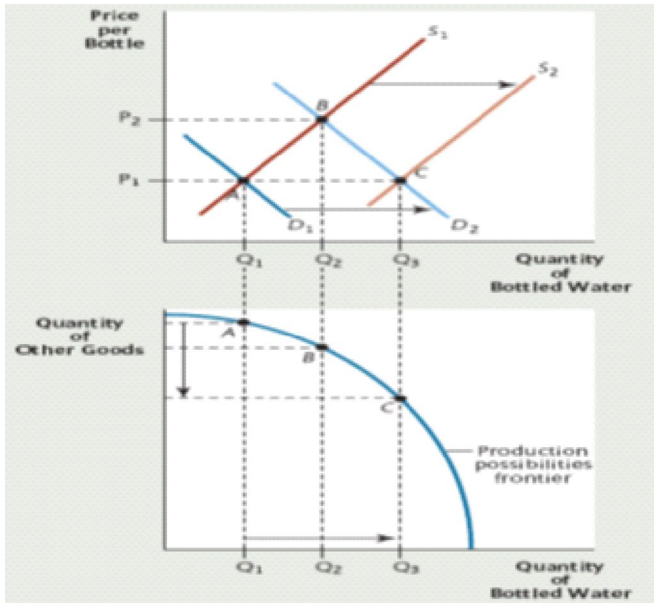
As a result, production expanded to match the increase in demand by consumers. More of our land, labor, capital, and entrepreneurial skills are now used to produce bottled water. Where did these resources come from? In large part, they were freed up from those industries that experienced a *decline* in demand. In these industries, lower prices have caused exit, freeing up land, labour, capital, and entrepreneurship to be used in other, expanding industries, such as the bottled water industry.

The lower panel of Figure 1 shows a production possibilities frontier (PPF) for bottled water and other goods. As production of bottled water increases from *Q*1 to *Q*2 to *Q*3, the production of other things decreases.

If we wanted to illustrate economic growth at the same time, the entire PPF would shift outward. In that case, increased production of bottled water would mean that production of other things would *increase* by less than otherwise. This leads us to an important observation:

*In a market economy, price changes act as **market signals,** ensuring that the pattern of production matches the pattern of consumer demands. When demand increases, a rise in price signals firms to enter the market, increasing industry output. When demand decreases, a fall in price signals firms to exit the market, decreasing industry output.*

Figure 1: How a Change in Demand Reallocates Resources



Importantly, in a market economy, no single person or government agency directs this process. There is no central command post where information about consumer demand is assembled, and no one tells firms how to respond. Instead, existing firms and new entrants, in their *own* search for higher profits, respond to market signals and help move the overall market in the direction it needs to go. This is what Adam Smith meant when he suggested that individual decision makers act— as if guided by an *invisible hand*—for the overall benefit of society, even though, as individuals, they are merely trying to satisfy their own desires.

In the upper panel, an increased desire for bottled water shifts the market demand curve rightward, from D1 *to* D2. *Price and quantity rise in the short run, and we move from* A *to* B *along short-run supply curve* S1. *The lower panel shows the corresponding short-run movement from* A *to* B *along the economy's PPF: Greater production of bottled water, less production of other things.*

In the long run, the higher price creates economic profit, attracting new firms, and shifting the supply curve rightward (upper panel). Price falls and quantity rises further. In the figure, we assume bottled water is a constant cost industry, so entry brings the price back to its initial value of P1 *at point* C. *In the lower panel, the* *further long-run increase in bottled water production moves us along the PPF, from* B *to* C.

A Change in Technology

How competitive markets ensure that the benefits of technological advances are enjoyed by consumers.

One industry that has experienced especially rapid technological changes is farming. By using genetically altered seeds, farmers are able to grow crops that are more resistant to insects and more tolerant of herbicides. This lowers the total—and average—cost of producing any given amount of the crop.

Figure 2 illustrates the market for corn, but it could just as well be the market for soybeans, cotton, or many other crops. In panel (a), the market begins at point *A,* where the price of corn is $6 per bushel. In panel (b), the typical farm produces 1,000 bushels per year and—with long-run average cost curve *LRATC*1—earns zero economic profit.

Now let's see what happens when new, higher-yield corn seeds are made available. Suppose first that only one farm uses the new technology. This farm will enjoy a downward shift in its *LRATC* curve from *LRATC*1 to *LRATC*2. Since it is so small relative to the market, it can produce all it wants and continue to sell at $6. Although we have not drawn in the farm's *MC* curve, you can see that the farm has several output levels from which to choose where price exceeds cost per unit and it can earn economic profit.

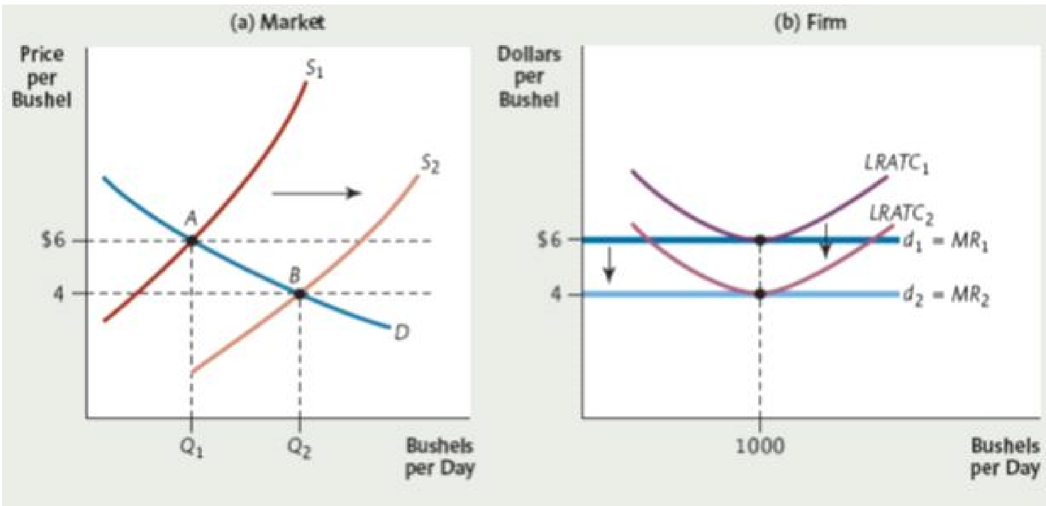
But not for long. In the long run, economic profit at this farm will cause two things to happen.

First, all other farmers in the market will have a powerful incentive to adopt the new technology—to plant the new, genetically engineered seed themselves.

Under perfect competition, they can do so; there are no barriers that prevent any farmer from using the same technology as any other. As these farms adopt the new seed technology, their *LRATC* curves, too, will drop down to *LRATC*2.

Second, outsiders will have an incentive to enter this industry, using the new technology, shifting the market supply curve rightward (from *S*1 to *S*2) and driving down the market price. The process will stop only when the market price has reached the level at which *farms using the new technology* earn zero economic profit. In Figure 2, this occurs at a price of $4 per bushel.

Figure 2: Technological Change in Perfect Competition



Technological change may reduce LRATC. *In panel (b), the first farms that adopt new technology will earn economic profit if they can sell at the old market price of $6 per bushel. That profit will lead its competitors to adopt the same technology and will also attract new entrants. As market supply increases, price falls until each farm is once again earning zero economic profit.*

From this example, we can draw two conclusions about technological change under perfect competition.

First, what will happen to a farmer who is reluctant to use the new technology? As *other* farms make the change, and the market price falls from $6 to $4, the reluctant farmer will suffer an economic loss, since the farm's average cost will remain at $6. Thus, a farmer that refuses to adopt the new technology will be forced to exit the industry. In the end, *all* farms that remain in the market must use the new technology.

Second, who benefits from the new technology in the long run? Not the farmers adopt it. *Some* farmers—the earliest adopters—may enjoy *short-run* profit before the price adjusts completely. But in the long run, all farmers will be right back where they started, earning zero economic profit. The gainers are *consumers* of corn, since they benefit from the lower price.

More generally, we can summarize the impact of technological change as follows:

Conclusion

Under perfect competition, a technological advance leads to a rightward shift of the market supply curve, decreasing market price. In the short run, early adopters may enjoy economic profit, but in the long run, all adopters will earn zero economic profit. Firms that refuse to use the new technology will not survive.

Technological advances in many competitive industries—mining, lumber, communication, entertainment, and others—have indeed spread quickly, shifting market supply curves rapidly and steadily rightward over the past 100 years. Competitive firms in these industries have had to continually adapt to new technologies in order to survive, leading to huge rewards for consumers.

In brief, a technological advance in a perfectly competitive market causes the equilibrium price to fall and equilibrium quantity to rise. Each competitive firm must use the new technology in order to survive, but consumers reap all the benefits by paying a lower price.

Questions

1. How market signals and economy are linked?
2. How change in demand influences the reallocation of resources in competitive markets?
3. Write a short note on technological change in perfect competition.

This lesson will focus on the following:

1. Monopoly and its Causes
2. Monopoly Behaviour: Price or Output Decision
3. Monopoly: Profit and Loss

Monopoly

*A **monopoly** is a market with just one seller.* **Monopoly** is the only seller in market, or a market with just one seller. The term *monopoly* is used for both the market and the firm that operates in that market.

Classifying a real-world firm or market as a monopoly can be tricky, because the number of sellers depends on how broadly a market is defined. Suppose, for example, that you live in a city or town with just one daily newspaper. Is that newspaper a monopoly?

In practice, we usually define a market to include all *close substitutes* for a product. If for most people, Internet news and daily newspapers are close substitutes, we should include them in the same market. If they are more distant substitutes, we should regard them as separate markets.

It makes sense, then, to view monopoly as a spectrum rather than a strict category. On one end of this spectrum is *pure monopoly,* where there is just one seller of a good for which very few buyers could find a substitute. The only doctor, attorney, or food market in a small town comes very close to being a pure monopoly.

How Monopolies Arise

The mere existence of a monopoly means that *something* is causing other firms to stay out of the market rather than enter and compete with the one firm already there. Broadly speaking, there must be some *barrier to entry.*

Why is the market a monopoly?

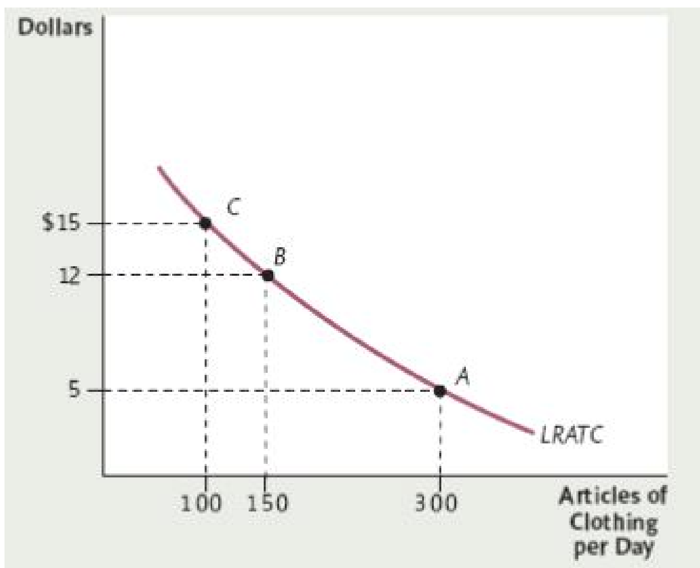
What *barrier* prevents other firms from entering the market?

There are several possible answers.

Economies of Scale

One barrier to entry is economies of scale. Economies of scale in production cause a firm's long-run average cost curve to slope downward. More output the firm produces, the lower will be its cost per unit. If economies of scale persist through a large-enough range of output, then a single firm can produce at lower cost than could two or more firms.

Figure 1: A Natural Monopoly



In the figure 1, the typical firm has an LRATC *curve as shown, with economies of scale through an output level of 300, which is assumed to be the total market quantity.*

A single firm could serve the market at a cost of $5 per unit, operating at point A. *Two firms splitting this market would each produce 150 units, with each operating at point* B *on its* LRATC *curve.*

Cost per unit would be $12, higher than with just one firm. Cost per unit would be even higher with three firms. Each would produce 100 units (point C*), at a cost of*

\$15 per unit. Since a single firm could produce at lower cost than two or more firms, this market tends naturally toward monopoly.

Figure 1 shows an example: the *LRATC* curve for a dry cleaner in a small town.

We'll suppose the entire market for dry cleaning services in this town never exceeds 300 pieces of clothing per day.

In the figure, the *LRATC* curve slopes downward, exhibiting economies of scale.

Why might this be?

A dry cleaning service uses a number of lumpy inputs: a parcel of land for the shop, a store clerk, a small dry cleaning machine if the clothes are cleaned on-site, or daily transportation to an off-site cleaning plant.

In the figure, we assume that cleaning more clothes—by spreading these costs among more units—causes cost per unit to decline all the way to 300 units and beyond.

As a result, one dry cleaner could achieve a lower cost per unit than could two or more dry cleaners.

For example, the *LRATC* curve tells us that one firm could clean 300 pieces of clothing at a cost of \$5 per piece (point *A*). But if two dry cleaners were to split this same output level (150 pieces of clothing each), each would have a higher cost per unit of \$12 at point *B*. For three dry cleaners, cost per article cleaned would be \$15 at point *C*, and so on. The first dry cleaner to locate in the town will have a cost advantage over any potential new entrants. This cost advantage will tend to keep newcomers out of the market.

A monopoly that arises because of economies of scale is called a **natural monopoly.** Local monopolies are often natural monopolies. In a very small town, there might be one gas station, one food market, one doctor, and so on. In all these cases, because there are sizable lumpy inputs and the market is small, the first firm to enter the market will likely be the last.

Legal Barriers

Many monopolies arise because of legal barriers. Of course, since laws are created by human beings, this immediately raises the question: Why would anyone want to create barriers that lead to monopoly? As you'll see, the answer varies depending on the type of barrier being erected. Here, we'll consider two of the most important legal barriers that give rise to monopolies: protection of intellectual property and government franchise.

Protection of Intellectual Property

Intellectual property includes literary, artistic, and musical works, as well as scientific inventions. The market for a specific intellectual property is a monopoly: One firm or individual owns the property and is the sole seller of the rights to use it. There is both good and bad in this.

Prices tend to be higher under monopoly than under perfect competition, and monopolies often earn economic profit as a consequence. A higher price is good for the monopoly and bad for everyone else.

On the other hand, the promise of monopoly profit is what encourages the creation of original products and ideas in the first place. And this benefits the rest of us.

In dealing with intellectual property, government strikes a compromise: It allows the creators of intellectual property to enjoy a monopoly and earn economic profit, *but only for a limited period of time*. Once the time is up, other sellers are allowed to enter the market, and it is hoped that competition among them will, in the end, bring down the price.

The two most important kinds of legal protection for intellectual property are *patents* and *copyrights*.

New scientific discoveries and the products that result from them are protected by a **patent** obtained from the government. The patent prevents anyone else from selling the same discovery or product for about 20 years. If someone uses the discovery without obtaining (and paying for) permission from the patent owner, they can be sued.

Patent A temporary grant of monopoly rights over a new product or scientific discovery.

Literary, musical, and artistic works are protected by a **copyright**, which grants exclusive rights over the material for at least 70 years and often longer. For example, the copyright of a book with a publishing company. No other company or individual can print copies and sell them to the public, and no one can quote from the book at length without obtaining the company's permission.

Copyrights and patents are often sold to another person or firm, but this does not change the monopoly status of the market, since there is still just one seller.

Government Franchise

Some firms have their monopoly status guaranteed through **government franchise**, a grant of exclusive rights over a product. Here, the barrier to entry is quite simple: Any other firm that enters the market will be prosecuted!

Governments often grant franchises when they think the market is a *natural monopoly*. In this case, a single large firm enjoying economies of scale would have a lower cost per unit than multiple smaller firms. Government tries to serve the public interest by *ensuring* that there are no competitors that would cause cost per unit to rise. In exchange for its monopoly status, the seller must submit to either government ownership and control or government regulation over its prices and profits.

Local governments, too, create monopolies by granting exclusive franchises in a variety of industries believed to be natural monopolies. These include utility companies that provide electricity, gas, and water, as well as garbage collection services.

Network Externalities

Imagine that you have created a new, superior operating system for personal computers. Compared to Microsoft Windows, your operating system is less vulnerable to viruses, works 10 percent faster, and uses 10 percent less memory. It even allows the user to turn off the caps-lock key, which most people use only by mistake.

When network externalities are present, joining a large network is more beneficial than joining a small network, even if the product in the larger network is somewhat inferior to the product in the smaller one. Once a network reaches a certain size, additional consumers will want to join just because so many others already have.

And if joining the network requires you to buy a product produced by only one firm, that firm can rapidly become the leading supplier in the market. All of this applies to the market for computer operating systems. When you buy a Windows computer, you benefit from the existence of so many other Windows users in a variety of ways. First, you have access to a large number of other computers—owned by friends and coworkers—that you can easily operate. Second, you have access to more software programs (because software developers know they can reach a bigger market when they write programs for Windows). Finally, there are more people around who can help you when you have a problem, saving you the time and trouble of calling a help desk.

Monopoly Behaviour

The goal of a monopoly, like that of any firm, is to earn the highest profit possible. And, like other firms, a monopolist faces constraints.

A monopolist faces purely *economic* constraints that limit its behavior, constraints that are in some ways similar to those faced by other, non-monopoly firms.

First, there is a constraint on the monopoly's *costs:* For any level of output the monopolist might produce, it must pay some total cost to produce it. This cost constraint is determined by the monopolist's production technology and by the prices it must pay for its inputs. In other words, the constraints on the monopolist's costs are the same as on any other type of firm, such as the perfectly competitive firm we studied in the previous chapter.

The monopolist also faces constraints on the price it can charge. This can be a bit confusing because a monopolist, unlike a competitive firm, is *not* a price taker: It does *not* take the market price as a given. But it does face a given demand curve for its product. Indeed, since a monopoly is the only firm in its market, its demand curve is the *market* demand curve. Thus, the monopoly faces a tradeoff: the more it charges for its product, the fewer units it will be able to sell.

Single Price Versus Price Discrimination

Some firms—including some monopolies—can charge different prices to different consumers, based on differences in the prices they are willing to pay. This kind of pricing is called *price discrimination*.

Other firms— we'll call them *single-price firms*—must charge the same price for every unit they sell, regardless of any differences in willingness to pay among their customers.

Monopoly Price or Output Decision

A monopoly does not make two separate decisions about price and quantity, but rather *one* decision. Once the firm determines its output level, it has also determined its price. The maximum price it can charge and still sell that output level. Similarly, once the firm determines its price, it has also determined its output level. The maximum output the firm can sell at that price.

How does a monopoly determine its profit-maximizing output level (and therefore its profit-maximizing price)?

Single-price monopoly

A monopoly firm that is limited to charging the same price for each unit of output sold.

Table 2 shows some data for a Pool firm that owns and operates the only swimming pool in a small town—a local monopoly. Firm earns revenue by charging an admission fee for using the pool.

The first two columns show various output levels (swimmers per day) and the highest price (admission fee) firm could charge for each output level. These two columns tell us that firm faces a downward-sloping demand curve: The lower the fee, the greater the number of people who will pay to swim each day. This demand curve is graphed in Figure 2 (as the upper curve).

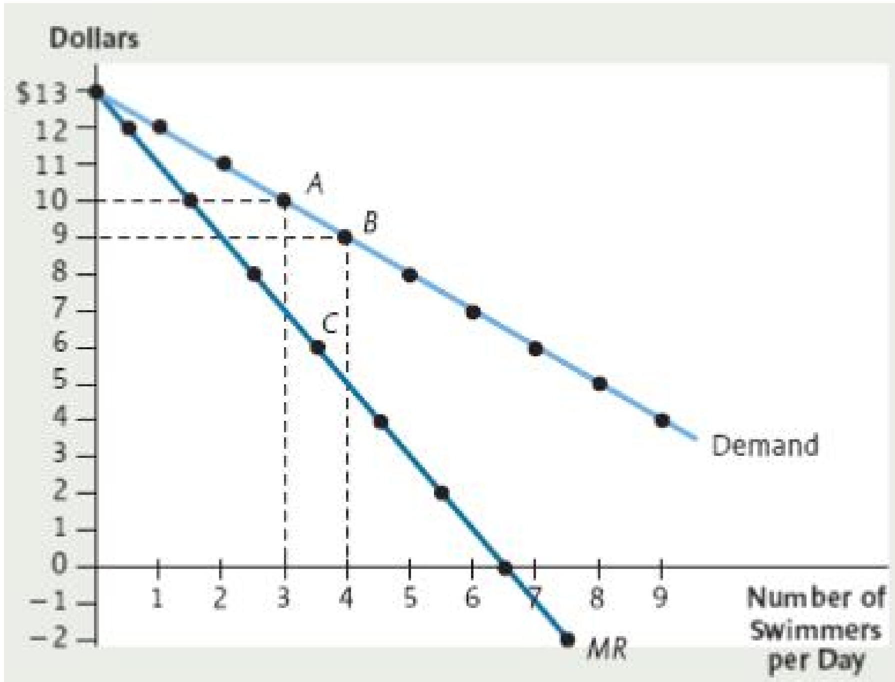
The third column of the table shows firm's total revenue per day (quantity times price) at each output level. For example, at an output level of 3, her daily revenue will be 3 x \$10 =

$30. And the last column shows firm's marginal revenue (*MR*), which is the increase in revenue for a one-unit rise in output. For example, when firm's output rises from 3 to 4, total revenue rises from $30 to $36, so marginal revenue for this change is $36 - $30 = $6.

Figure 1: Demand and Revenue

Q (swimmers per day)	P (admission fee)	TR	MR
0	$13	$ 0	
			$12
1	$12	$12	
			$10
2	$11	$22	
			$ 8
3	$10	$30	
			$ 6
4	$ 9	$36	
			$ 4
5	$ 8	$40	
			$ 2
6	$ 7	$42	
			$ 0
7	$ 6	$42	
			–$ 2
8	$ 5	$40	
			–$ 4
9	$ 4	$36	

Figure 2: Demand and Marginal Revenue for Pool



When a firm faces a downward-sloping demand curve, marginal revenue (MR) *is less than price, and the* MR *curve lies below the demand curve. For example, moving from point* A *to point* B*, output rises from 3 to 4 units, while price falls from $10 to $9. For this move, total revenue rises from $30 to $36, so marginal revenue (plotted at point* C*) is only $6—less than the new price of $9.*

The marginal revenue column is graphed in Figure 2, *below* the demand curve. Why below? Mathematically, this is because when the firm's demand curve slopes downward, marginal revenue is less than the price for all increases in output (except the increase from zero to one unit).

To see this, look at what happens when we move from point *A* to point *B* along the demand curve, and output rises from 3 to 4 units. The new price is $9, but the marginal revenue from producing the fourth unit is $6 (at point *C*), which is less than the new price.

When any firm, including a monopoly, faces a downward-sloping demand curve, marginal revenue is less than the price of output. Therefore, the marginal revenue curve will lie below the demand curve.

Why must marginal revenue be less than the price? Because when a firm faces a downward-sloping demand curve, it must lower the price in order to sell a greater quantity. The new, lower price applies to *all* units it sells, including those it was *previously* selling at some higher price.

For example, suppose firm initially has 3 swimmers per day at $10 each. If firm wants 4 swimmers, Table 1 tells us that firm must lower price to $9. Firm would *gain* $9 in revenue by admitting one more swimmer at that price. But firm would also *lose* some revenue, because each of the first three swimmers that firm *used* to charge $10 will now be charged $9—a *loss* of $3 in revenue. If we add the $9 gained on the fourth swimmer and subtract the $3 lost from lowering the price to the other three, the net impact on revenue is an increase of $6—less than the $9 price she is now charging.

Notice, too, that for increases in output beyond 7, marginal revenue turns negative. For these changes in output, firm loses more in revenue from dropping the price on previous units than she gains by selling one new unit. No firm would ever want to operate where marginal revenue is negative, because it could then increase its revenue *and* have lower costs by *decreasing* output.

Profit-Maximizing Output Level

To maximize profit, a monopoly—like any firm—should produce the quantity where MC = MR *and the* MC *curve crosses the* MR *curve from below.*

In Figure 3, we've plotted the firm's demand curve, showing the number of cable subscribers at each monthly price. Firm's marginal revenue curve lies below its demand curve. The figure also shows firm's marginal cost curve.

The greatest profit possible occurs at an output level of 16,000, where the *MC* curve crosses the *MR* curve from below. In order to sell this level of output, the firm will charge a price of $90, located at point *E* on its demand curve.

For a monopoly, *price and output are not independent decisions, but different ways of expressing the same decision*. Once firm determines its profit maximizing output level (16,000 units), it has also determined its profit-maximizing price ($90), and vice versa.

Monopoly and Market Power

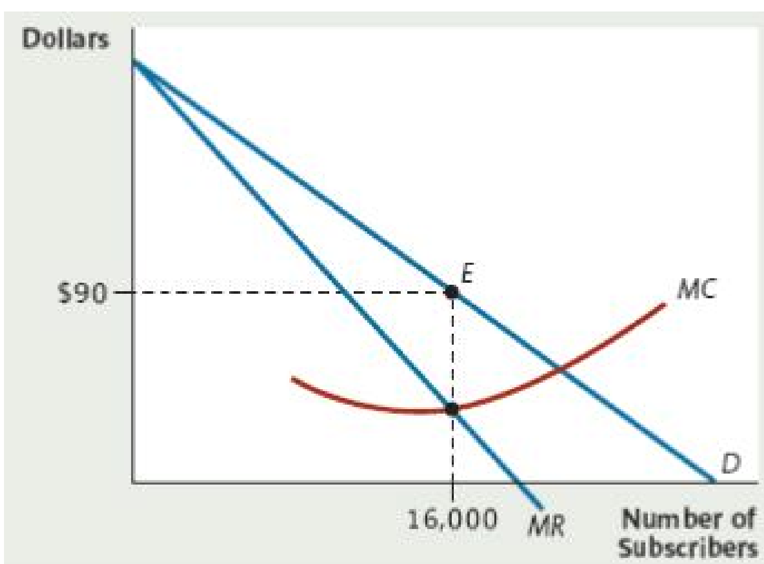
A monopoly is an example of a firm with **market power**—the ability to raise price without causing quantity demanded to go to zero. Any firm facing a downward sloping demand curve has market power: As it raises its price, quantity demanded falls, but some customers who value the firm's product will continue to buy it at the higher price.

Only perfectly competitive firms, which face horizontal demand curves, have no market power at all. For a competitive firm, raising price even a tiny bit above the market price reduces quantity demanded to zero. A competitive firm is a *price taker:* It must accept the market price as a given, so there is no decision about price.

By contrast, when a firm has market power, it is a **price setter**—it makes a choice about what price to charge. The choice is limited by constraints (such as the demand curve itself), but it is still a choice. Monopolies are one example of price-setting firms, but they are not the only example. In the next chapter, you'll learn about other market structures besides monopoly in which firms have market power and are therefore price setters.

Market power The ability of a seller to raise price without losing all demand for the product being sold.

Figure 3: Monopoly Price and Output Determination



Like any firm, the monopolist maximizes profit by producing where MC *equals* MR. *Here, that quantity is 16,000 units. The price charged ($90) is read off the demand curve. It is the highest price at which the monopolist can sell the profit-maximizing level of output.*

Profit and Loss

In Figure 3, we've illustrated firm's price and output level, but we cannot yet see whether the firm is making an economic profit or loss. This will require one more addition to the diagramme—the average cost curve. Remember that Profit per unit = *P* - *ATC*.

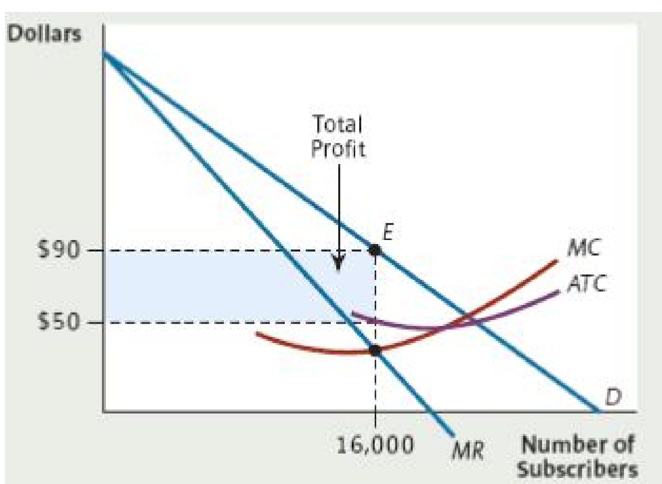
At any output level, the price is read off the demand curve. Profit per unit, then, is just the vertical distance between the firm's demand curve and its *ATC* curve. Figure 4 is just like Figure 3 but adds firm's *ATC* curve.

At the profit-maximizing output level of 16,000, price is $90 and average total cost is $50, so profit per unit is $40.

The height of shaded rectangle is profit per unit ($40), and the width is the number of units produced (16,000). The *area* of the rectangle = height x width = firm's total profit, or $40 x 16,000 = $640,000.

Price setter A firm (with market power) that selects its price, rather than accepting the market price as a given.

Figure 4: A Monopoly Earning Profit

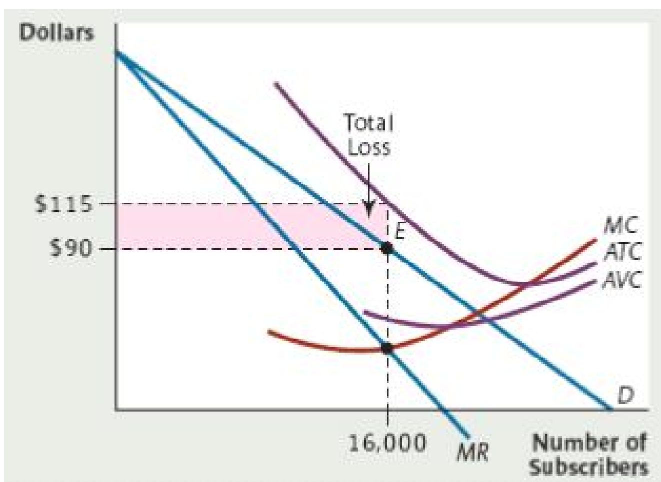


The monopoly in this figure is earning a profit. At the profit maximizing output level (16,000), profit per unit is equal to the difference between price ($90) and ATC ($50). Total profit is equal to profit per unit multiplied by the number of units, or $40 x *16,000 = $640,000, represented by the shaded rectangle.*

A monopoly earns a profit whenever P > ATC. *Its total profit at the best output level equals the area of a rectangle with height equal to the distance between* P *and* ATC *and width equal to the level of output.*

Figure 5 illustrates the case of a monopoly suffering a loss. Here, costs are higher than in Figure 4, and the *ATC* curve lies everywhere above the demand curve. As a result, the firm will suffer a loss at any level of output. At the best output level (where *MC* = *MR*), the loss will be smallest. In the figure, this occurs at 16,000 units, with *ATC* = $115 and price = $90, so the loss per unit is $25. The total loss ($400,000) is the area of the pink rectangle, whose height is the loss per unit ($25) and width is the best output level (16,000).

Figure 5: A Monopoly Suffering a Loss



The monopoly in this figure is suffering a loss. At the profit maximizing output level (16,000), the loss per unit is equal to the difference between price ($90) and ATC ($115). The total loss is equal to loss per unit multiplied by the number of units, or $25 x *16,000 = $400,000, represented by the shaded rectangle.*

Conclusion

Being a monopolist is no guarantee of profit. If costs are too high, or demand is insufficient, a monopolist may break even or suffer a loss. A monopoly suffers a loss whenever $P < ATC$. Its total loss at the best output level equals the area of a rectangle with height equal to the distance between ATC and P and width equal to the level of output.

Questions

1. What is monopoly?
2. What are causes of monopoly?
3. Explain price or output decision of a monopoly.
4. Explain profit and loss of a monopoly using suitable diagramme.

This lesson will focus on the following:

1. Monopoly Markets: Short-Run and Long-Run Equilibrium
2. Monopoly Vs Perfect Competition
3. Monopoly Decisions and Changes in Demand and Cost-Saving Technology

Equilibrium in Monopoly Markets

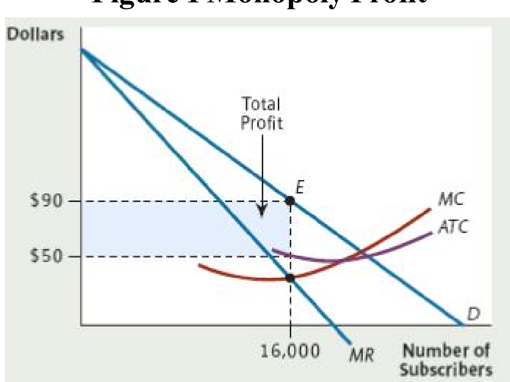
A monopoly market is in equilibrium when the only firm in the market, the monopoly firm, is maximizing its profit. After all, once the firm is producing the profit maximizing quantity— and charging the highest price that will enable it to sell that quantity—it has no incentive to change either price or quantity, unless something in the market changes.

Short-Run Equilibrium

In the short run, a monopoly may earn an economic profit or suffer an economic loss. It may, of course, break even as well.

A monopoly that is earning an economic profit will, of course, continue to operate in the short run, charging the price and producing the output level at which *MR* = *MC*, as in Figure 1.

Figure 1 Monopoly Profit



The monopoly in this figure is earning a profit. At the profit maximizing output level (16,000), profit per unit is equal to the difference between price ($90) and ATC ($50). Total profit is equal to profit per unit multiplied by the number of units, or $40 x *16,000 = $640,000, represented by the shaded rectangle.*

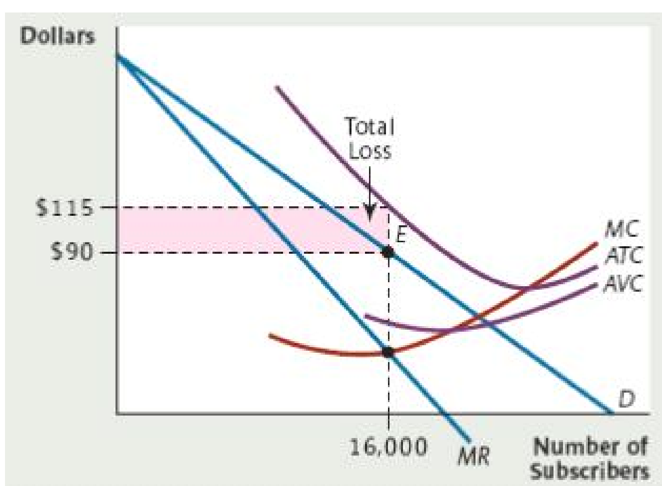
A monopoly earns a profit whenever P > ATC. *Its total profit at the best output level equals the area of a rectangle with height equal to the distance between* P *and* ATC *and width equal to the level of output.*

Figure 2 illustrates the case of a monopoly suffering a loss. Here, costs are higher than in Figure 1, and the *ATC* curve lies everywhere above the demand curve. As a result, the firm will suffer a loss at any level of output. At the best output level (where *MC* = *MR*), the loss will be smallest. In the figure, this occurs at 16,000 units, with *ATC* = $115 and price = $90, so the loss per unit is $25. The total loss ($400,000) is the area of the pink rectangle, whose height is the loss per unit ($25) and width is the best output level (16,000).

Being a monopolist is no guarantee of profit. If costs are too high, or demand is insufficient, a monopolist may break even or suffer a loss.

A monopoly suffers a loss whenever P < ATC. *Its total loss at the best output level equals the area of a rectangle with height equal to the distance between* ATC *and* P *and width equal to the level of output.*

Figure 2: Monopoly Loss



The monopoly in this figure is suffering a loss. At the profit maximizing output level (16,000), the loss per unit is equal to the difference between price ($90) and ATC ($115). The total loss is equal to loss per unit multiplied by the number of units, or $25 x *16,000 = $400,000, represented by the shaded rectangle.*

But what if a monopoly suffers a loss in the short run?

Then it will have to make the same decision as any other firm: to shut down or not to shut down. A firm should shut down if *TR* < *TVC* at the output level where marginal revenue and marginal cost are equal—applies to any firm, including a monopoly.

Therefore, *any firm—including a monopoly—should shut down if* P < AVC *at the output level where* MR = MC.

In Figure 1, firm is suffering a loss. But since *P* = $90 and *AVC* is less than $90 at an output of 16,000, we have *P* > *AVC*: The firm should keep operating.

The shutdown rule should accurately predict the behaviour of most privately owned and operated monopolies. But if a monopoly operates under a government franchise or regulation and produces a vital service such as transportation, mail delivery, or mass transit, the government may not allow it to shut down.

If, for example, the monopoly suddenly finds that *P* < *AVC* at every output level, the government might order the firm to continue operating, and use tax revenue to cover the loss.

Long-Run Equilibrium

The perfectly competitive firms will *not* earn a profit in long-run equilibrium. Profit attracts new firms into the market, and market production increases. This, in turn, causes the market price to fall, eliminating any temporary profit earned by a competitive firm.

But there is no such process at work in a monopoly market, where barriers *prevent* the entry of other firms into the market. Outsiders will *want* to enter an industry when a monopoly is earning positive economic profit, but they will be *unable to do so.* Thus, the market provides no mechanism to eliminate monopoly profit.

Unlike perfectly competitive firms, monopolies may earn economic profit in the long run.

What about economic loss?

If a monopoly is franchised or regulated by the government, and it faces the prospect of long-run loss, the government may decide to subsidize it in order to keep it running. But if the monopoly is privately owned and controlled, it will not tolerate long-run losses. A monopoly suffering an economic loss that it expects to continue indefinitely should always exit the industry, just like any other firm.

A privately owned, unregulated monopoly suffering an economic loss in the long run will exit the industry, just as would any other business firm. In the long run, therefore, we should not find such monopolies suffering economic losses.

Comparing Monopoly to Perfect Competition

In perfect competition, economic profit is relentlessly reduced to zero by the entry of other firms; in monopoly, economic profit can continue indefinitely.

But monopoly also differs from perfect competition in another way:

All else equal, a monopoly market will have a higher price and lower output than a perfectly competitive market.

To see why this is so, let's explore what would happen if a single firm took over a perfectly competitive market, changing the market to a monopoly.

Panel (a) of Figure 3 illustrates a competitive market consisting of 100 identical firms. The market is in long-run equilibrium at point *E,* with a market price of $10 and market output of 100,000 units. In panel (b), the typical firm faces a horizontal demand curve at $10, produces output of 1,000 units, and earns zero economic profit.

Now, imagine that a single company buys all 100 firms, to form a monopoly. The new monopoly market is illustrated in panel (c). Under monopoly, the horizontal demand curve facing each firm becomes irrelevant. Now, the demand curve facing the monopoly is the downward-sloping *market* demand curve *D*—the same as the market demand curve in panel (a). Since the demand curve slopes downward, marginal revenue will be less than price, and the *MR* curve will lie everywhere below the demand curve. To maximize profit, the monopoly will want to find the output level at which *MC* = *MR*. But what is the new monopoly's *MC* curve?

The monopoly's marginal cost curve will be the same as the market supply curve in panel (a). Why?

First, remember that in a perfectly competitive industry the market supply curve is obtained by adding up each individual firm's supply curve, that is, each individual firm's marginal cost curve. Therefore, the market supply curve tells us the marginal cost—at each firm—of producing another unit of output for the market. When the mono poly takes over each of these individual firms, the market supply curve tells us how much it will cost the monopoly to produce another unit of output at each of its plants.

For example, point *E* on the market supply curve tells us that, when total supply is 100,000, with each plant producing 1,000 units, increasing output by one more unit will cost the monopoly $10 because that is the marginal cost at each of its plants. The same is true at every other point along the old competitive market supply curve: It will always tell us the new monopoly's cost of producing one more unit at each of the plants it now owns. In other words, the upward-sloping curve in panel (c), which is the market supply curve when the market is competitive, becomes the marginal cost curve for a single firm when the market is monopolized.

Now we have all the information we need to find the monopoly's choice of price and quantity. In panel (c), the monopoly's *MC* curve crosses the *MR* curve from below at 60,000 units of output. This will be the monopoly's profit-maximizing output level. To sell this much output, the monopoly will charge $15 per unit— point *F* on its demand curve.

After the monopoly takes over, the price rises from $10 to $15, and market quantity drops from 100,000 to 60,000. The monopoly, compared to a competitive market, *charges more and produces less.*

Why does this happen?

When the market was perfectly competitive, each firm could sell all the output it wanted at the given market price of $10, and each firm knew it could earn an additional $10 in revenue for each additional unit it sold. The best option for the firm was to increase output until marginal cost rose to $10.

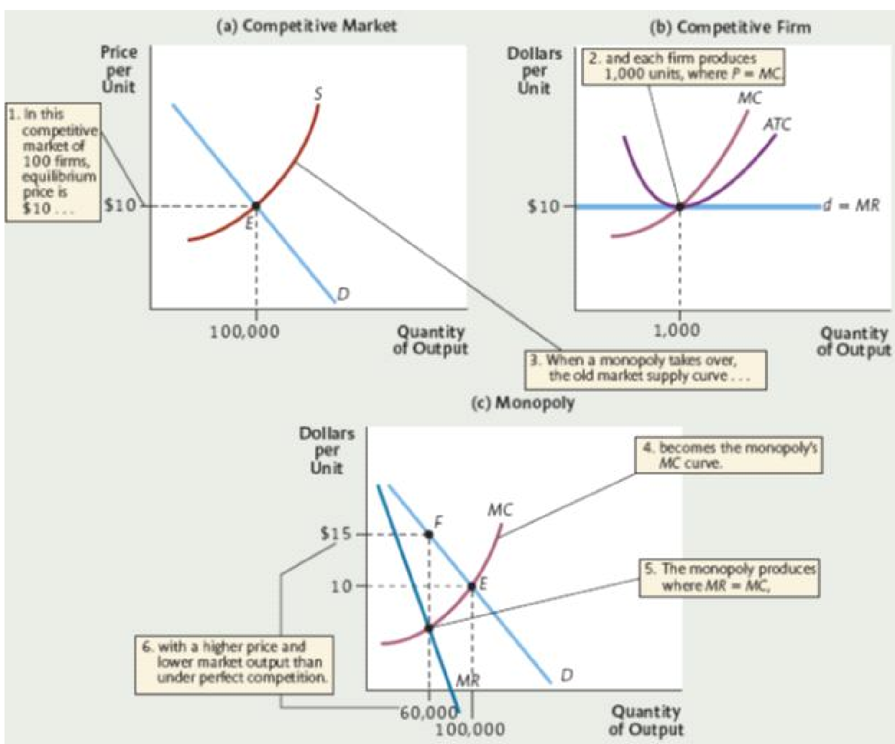
But the new monopoly does *not* treat price, or marginal revenue, as given values. Instead, it knows that raising its own output *lowers* the market price. So if the monopoly goes all the way to the competitive output level (100,000 units in the figure), it will be producing units for which $MR < MC$ (all units beyond 60,000).

This will reduce its profit. To maximize profit, the monopoly has to stop short of the competitive output—producing 60,000 rather than 100,000. Of course, since the monopoly sells a lower market quantity, it will charge a higher market price.

Now let's see who gains and who loses from the takeover. By raising price and restricting output, the new monopoly earns economic profit. We know this because if the firm were to charge $10—the old competitive price—each of its plants would break even, giving it zero economic profit. But we've just seen that $10 is *not* the profit-maximizing price—$15 is. So, the firm must make higher profit at $15 than at $10, ensuring it will earn more than zero economic profit.

Consumers, however, lose in two ways: They pay more for the output they buy and, due to higher prices, they buy less output. The changeover from perfect competition to monopoly thus benefits the owners of the monopoly and harms consumer of the product.

Figure 3: Comparing Monopoly and Perfect Competition



In conclusion, price is higher and output is lower under monopoly *if all else is equal.* In particular, we have assumed that after the market is monopolized, the technology of production remains unchanged at each of the monopoly's "plants",

But a monopoly may be able to *change* the technology of production, so that all else would *not* remain equal. For example, a monopoly may have each of its new plants *specialize* in some part of the production process, or it may be able to achieve efficiencies in product planning, employee supervision, bookkeeping, or customer relations. These cost savings might shift down the monopoly's marginal cost curve.

If you draw a new, lower *MC* curve in panel (c), you'll see that this works to *decrease* the monopoly's price and *increase* its output level—exactly the reverse of the effects discussed earlier. If the cost savings are great enough, and the *MC* curve drops low enough, a profit-maximizing monopoly could even charge a lower price and produce more output than would a competitive market.

Government and Monopoly Profit

Monopolies often exist with government permission. When we bring the government into our analysis, the monopoly's total profit may be less than that predicted by the analysis we've done so far. Government involvement reduces monopoly profit in two ways.

In many cases of natural monopoly, a firm is granted a government franchise to be the sole seller in a market. This has been true of monopolies that provide water service, electricity, and natural gas. In exchange for its franchise, the monopoly must accept government regulation, often including the requirement that it submit its prices to a public commission for approval. The government will often want to keep prices high enough to keep the monopoly in business, but no higher. Since the monopoly will stay in business unless it suffers a long-run loss, the ideal pricing strategy for the regulatory commission would be to keep the monopoly's economic profit at zero.

Economic profit includes the opportunity cost of the funds invested by the monopoly's owners. If the public commission succeeds, the monopoly's *accounting* profit will be just enough to match what the owners could earn by investing their funds elsewhere—that is, the monopoly will earn zero economic profit.

Another factor that reduces a monopoly's profit comes from the interplay between politics and economics. Many monopolies achieve and maintain their monopoly status due to legal barriers to entry. And many of these monopolies are completely unregulated. For example, a movie theater may enjoy a monopoly in an area because zoning regulations prevent entry by competitors.

In less developed countries, a single firm may be granted the exclusive right to sell or produce a particular good even though it is not a natural monopoly. In all of these cases, the monopoly is left free to set its price as it wishes.

But legal barriers to entry—for example, zoning laws—are often controversial. A monopoly may charge a higher price and produce less output than would a competitive market. Thus, government will be tempted to pull the plug on a monopoly's exclusive status and allow competitors into the market. The monopoly, in turn, will often take action to *preserve* legal barriers to entry. Economists call such actions *rent-seeking activity.*

Any costly action a firm undertakes to establish or maintain its monopoly status is called ***rent-seeking activity.***

In economics, the term *economic rent* refers to any earnings beyond the minimum needed in order for a good or service to be produced. For example, the minimum price to get *land* "produced" is zero, since it's a gift of nature. This is why all the earnings of landowners are called "rent." A monopoly's economic profit is another example of rent, since it represents earnings above the minimum needed to keep the monopoly in business.

In countries with the most corrupt bureaucracies, rent-seeking activity typically takes the form of outright bribes to government officials. But rent seeking occurs in virtually all countries. It includes the time and money spent lobbying legislators and the public for favourable policies. The costs of such activities can reduce a monopoly's profit below what the simple monopoly model would suggest.

What Happens When Things Change?

Once a monopoly is maximizing profit, it has no incentive to change its price or its level of output . We'll consider two such events: a change in demand for the monopolist's product, and a change in its costs.

Change in Demand

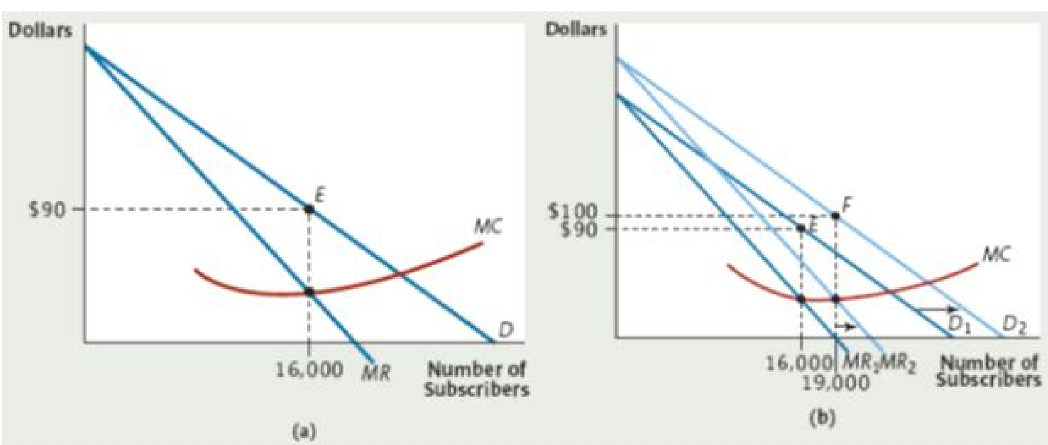
In a competitive market, an increase in demand caused an increase in both market price and market quantity.

Does the same general conclusion hold for a monopolist?

Let's see. Panel (a) of Figure 4 shows firm earning a positive profit in the short run. As before, it is producing 16,000 units per month, charging \$90 per unit, and earning a monthly profit of \$640,000 (not shown). The fact that firm is a monopolist, however, does not mean that it is immune to shifts in demand. What might cause a monopolist to experience a shift in demand? For example, an increase in consumer tastes for the monopolist's good will shift its demand curve rightward, and a decrease in consumer incomes can shift it leftward.

In panel (b) of Figure 4, this is shown as a rightward shift of the demand curve from *D*1 to *D*2. Marginal revenue curve shifts as well, from *MR*1 to *MR*2. Why is this? A rightward shift in demand is also an *upward* shift in demand. At each quantity, the firm can charge a greater price than before. With a higher price, the rise in revenue (*MR*) for each increase in quantity will be greater as well. So the *MR* curve shifts upward (rightward), just like the demand curve.

Figure 4: Change in Demand



Panel (a) shows firm in equilibrium. It is providing 16,000 units of cable TV service at a price of $90 per month. Panel (b) shows the same firm following an increase in demand from D*1 to* D*2. With the increased demand,* MR *is higher at each level of output. In the new equilibrium, firm is charging a higher price ($100), providing more service (19,000 units), and earning a larger profit.*

With an unchanged cost structure, the new short-run equilibrium will occur where *MR*2 intersects the unchanged *MC* curve. As you can see, the result is an increase in quantity from 16,000 to 19,000 and a higher price: $100 per month rather than the original $90. In this sense, monopoly markets behave very much like competitive markets.

What about the monopolist's profit? With both price and quantity now higher, total revenue has clearly increased. But total cost is higher as well. So it seems as if profit could either rise or fall.

Profit *must* be higher in the new equilibrium at point *F*. We know that because firm has the option of continuing to sell its original quantity, 16,000, at a price higher than before. If, as we assume, it started out earning a profit at that output level, then the higher price would certainly give it an eve *higher* profit. But the logic of *MR* = *MC* tells us that the greatest profit of all occurs at 19,000 units. So profit is certainly greater after the increase in demand.

We can conclude that:

A monopolist will generally react to an increase in demand by producing more output, charging a higher price, and earning a larger profit. It will react to a decrease in demand by reducing output, lowering price, and suffering a reduction in profit.

Cost-Saving Technology

In a perfectly competitive market, all cost savings from a technological advance are passed along to consumers in the form of lower prices. Is the same true of monopoly?

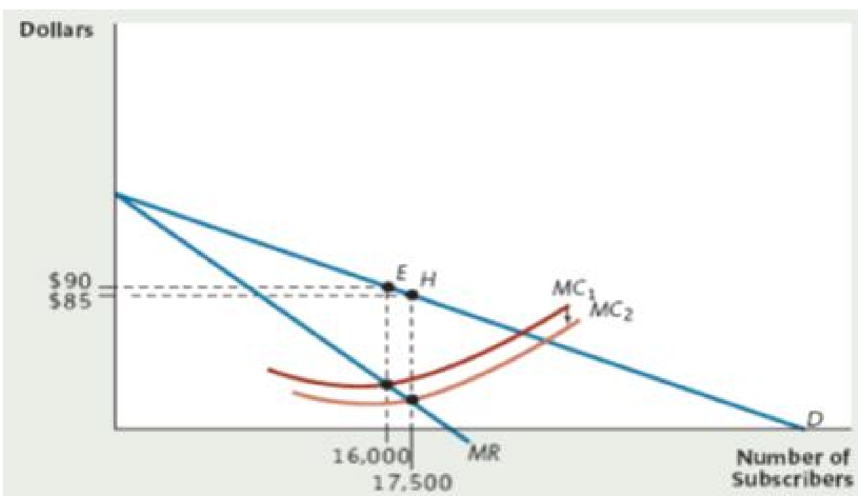
Suppose a new type of cable box becomes available that breaks down less often, requiring fewer service calls. When firm begins using this equipment, it finds that it gets fewer service calls, so its labour costs decrease by $15 per customer.

Figure 5 shows the result. Before the new equipment is used, firm is charging $90 and producing output of 16,000, where its *MR* and *MC* curves cross. The technological advance, when it's distributed to all of firm's customers, will lower not only the monthly cost per *current* customer (shifting the *ATC* curve down by \$15, which isn't shown), but also the monthly cost of servicing each *additional* subscriber. That is, firm's *marginal cost* curve will shift down by \$15, from *MC*1 to *MC*2.

Firm will now want to add subscribers. After the downward shift in the *MC* curve, *MR* exceeds *MC* at the original output of 16,000. An opportunity to raise profit by increasing output has been created. In the figure, the new intersection point between *MC* and *MR* occurs at an output level of 17,500, so that's firm's new profit-maximizing output level. The demand curve tells us that when output is 17,500, firm will charge a price of \$85.

Furthermore, we know that firm's profits have increased. How? If firm had left its output unchanged, the downward shift in its *ATC* curve (not shown) would have raised its profit. Increasing output from 16,000 to 17,500 increased profit further (because *MR* > *MC* for that move). Thus, profit must be greater than before.

Figure 5: Cost Saving Technological Change



A cost-saving technological advance shifts the monopolist's marginal cost curve down, from MC*1 to* MC*2. Consumers gain because th e price falls, but the drop in price is less than the drop in marginal cost. The monopoly gains because its profit is greater.*

In general, a monopoly will pass to consumers only part *of the benefits from a cost-saving technological change. After the change in technology, the monopoly's profits will be higher.*

This stands in sharp contrast to the impact of technological change in perfectly competitive markets, where—as stated earlier—all of the cost saving is passed along to consumers in the long run.

Questions

1. Draw demand, *MR,* and *ATC* curves that show a monopoly that is just breaking even.
2. Explain short-run and long-run equilibrium of monopoly.
3. Make a comparison of monopoly and perfect competition.
3. Explain monopoly decisions and changes in demand and cost-saving technology.

This lesson will focus on the following:

1. Price Discrimination: Conditions and Effects
2. Types of Price Discrimination
3. Price Discrimination and Multiple Prices

Price Discrimination

A single-price monopoly is one that charges the same price on every unit that it sells. But not all monopolies operate this way. For example, local utilities typically charge different rates per kilowatt-hour, depending on whether the energy is used in a home or business. Telephone companies charge different rates for calls made by people on different calling plans. Nor is this multi price policy limited to monopolies: Movie theaters charge lower prices to senior citizens, airlines charge lower prices to those who book their flights in advance, and supermarkets and food companies charge lower prices to customers who clip coupons from their local newspaper.

In some cases, the different prices are due to differences in the firm's costs of production. For example, it may be more expensive to deliver a product a great distance from the factory, so a firm may charge a higher price to customers in outlying areas. But in other cases, the different prices arise not from cost differences but from the firm's recognition that *some customers are willing to pay more than others:*

The term *discrimination* in this context requires some getting used to. In everyday language, *discrimination* carries a negative connotation: We think immediately of discrimination against someone because of his or firm race, sex, or age. But a price discriminating monopoly does not discriminate based on prejudice, stereotypes, or ill will toward any person or group; rather, it divides its customers into different categories based on their *willingness to pay* for the good—nothing more and nothing less. By doing so, a monopoly can squeeze even more profit out of the market. Why, then, doesn't *every* firm practice price discrimination?

Conditions for Price Discrimination

Every firm would *like* to practice price discrimination, not all of them can. To successfully price discriminate, three conditions must be satisfied:

Market Power

To price discriminate, a firm must have *market power.* That is, the firm must face a downward-sloping demand curve so that it behaves as a price setter. To see why, think about a perfectly competitive firm that faces a horizontal demand curve and has no market power. If such a firm tried to charge some customers a higher price than others, the high-price customers would simply buy from other firms that are selling the same product at the market price. By contrast, all monopolies face a downward sloping demand curve, so they meet the market power requirements.

Identifying Willingness to Pay

In order to determine which prices to charge to which customers, a firm must be able to identify how much different customers or groups of customers are willing to pay. But this is often difficult. Suppose your barber or hairstylist wanted to price discriminate. How would he determine how much you are willing to pay for a haircut? He could *ask* you, but . . . let's be real: You wouldn't tell him the truth, since you know he would only use the information to charge you more than you've been paying. Price-discriminating firms—in most cases—must be a bit sneaky, relying on more indirect methods to gauge their customers' willingness to pay.

For example, airlines know that business travelers, who must get to their destination, are willing to pay a higher price for air travel than are tourists or vacationers, who can more easily travel by train, bus, or car. Of course, if airlines merely *announced* a higher price for business travel, then no one would admit to being a business traveler when buying a ticket. So the airlines must find some way to identify business travelers without actually asking.

Their method is crude but reasonably effective: Business travelers typically plan their trips at the last minute and don't stay over Saturday night, while tourists and vacationers generally plan long in advance and do stay over Saturday. Thus, the airlines give a discount to any customer who books a flight several weeks in advance and stays over, and they charge a higfirm price to those who book at the last minute and don't stay over.

Prevention of Resale

To price discriminate, a firm must be able to prevent low price customers from reselling its product to high-price customers. This can be a vexing problem for many would-be discriminators. For example, when airlines began price discriminating, a resale market developed: Business travelers could buy tickets at the last minute from intermediaries, who had booked in advance at the lower price and then advertised their tickets for sale. To counter this, the airlines imposed the additional requirement of a Saturday stay over for the lower price. By adding this restriction, the airlines were able to substantially reduce the reselling of low-price tickets to business travelers.

It is often easy to prevent resale of a *service* because of its personal nature. A hairstylist can charge different prices to different customers without fearing that one customer will sell firm haircut to another. The same is true of the services provided by physicians, attorneys, and music teachers. Resale of *goods,* however, is much harder to prevent, since goods can be easily transferred from person to person without losing their usefulness.

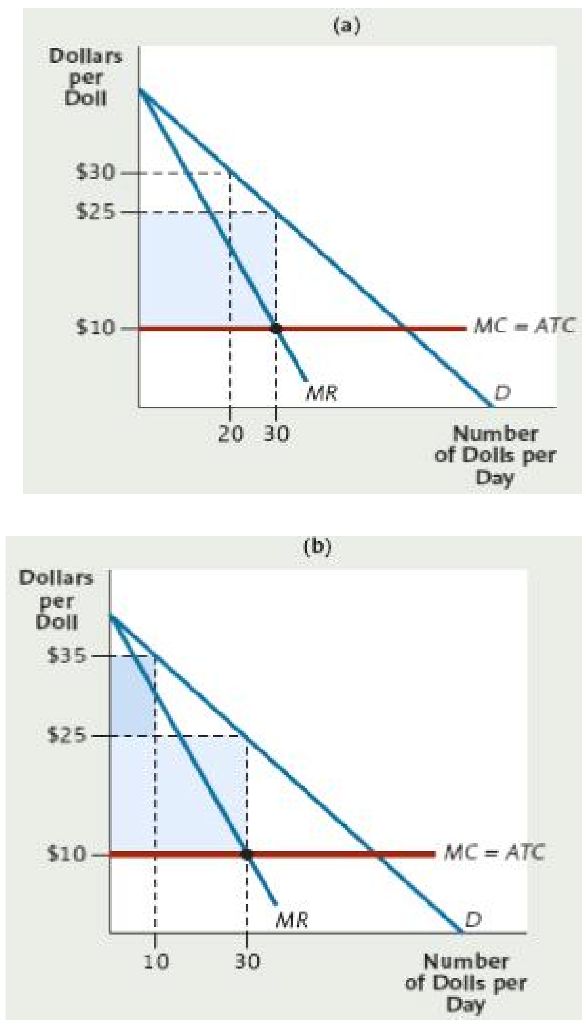
Effects of Price Discrimination

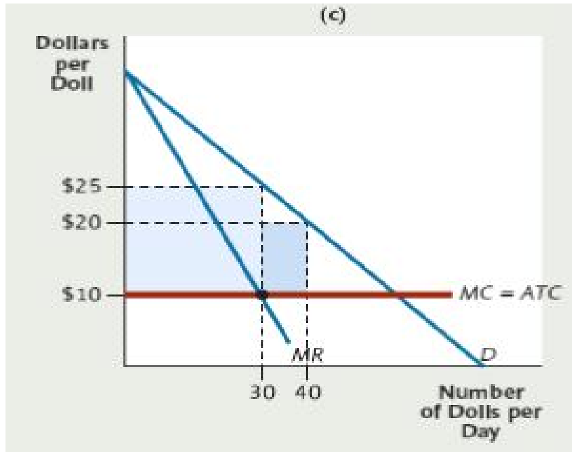
Price discrimination always benefits the owners of a firm: When the firm can charge different prices to different consumers, it can use this ability to increase its profit. But the effects on consumers can vary. To understand how price discrimination affects the firm and the consumers of its product, consider, a monopolist who produces and sells dolls at flea-markets.

To keep our analysis as simple as possible, we'll assume that Firm has no fixed costs, and that each doll costs \$10 to make no matter how many firm produces. Thus, cost per doll (*ATC*) is \$10 at every output level. Furthermore, because each additional doll costs \$10 to make, firm marginal cost (*MC*) is also \$10 at any output level. This is why, in Figure 1(a), both the *MC* and *ATC* curves are represented by the same horizontal line at \$10.

Let's first suppose that firm is a single-price monopolist, charging a pre - announced price on every doll firm sells. The figure shows the demand and marginal revenue curves firm would face on a typical day. Using the $MR = MC$ rule, Firm would earn maximum profit by selling 30 dolls per day, and charging \$25 per doll. Firm profit per unit would be \$25 - \$10 = \$15, which is the vertical distance between the ATC curve and the demand curve at 30 units. Firm total profit would be \$15 x 30 dolls = \$450 per day, which is equal to the area of the shaded rectangle.

Figure 1: Two Kinds of Price Discrimination





Panel (a) shows a single-price monopolist, selling 30 dolls per day at $30 each and earning a profit of $450 per day, as shown by the blue shaded rectangle. In panel (b), firm price discriminates by charging a high firm price of $35 for 10 dolls per day, while still charging $10 for the remaining 20 dolls.

Profit increases by $100 per day, the area of the dark shaded rectangle. Panel (c) shows a different type of price discrimination: charging the original $25 on the first 30 dolls, and a lower price on just 10 additional dolls, bringing firm total output to 40. Compared to panel (a), firm profit in panel (c) rises by $100 per day—the area of the dark-shaded rectangle.

Price Discrimination That Harms Consumers

Now suppose that firm discovers that on a typical day, some of firm dolls are sold to particularly eager customers who show up first thing in the morning to buy them. Firm has found a way to identify those willing to pay more, so firm charges the early-morning buyers $35 each and continues charging $25 to everyone else. The result, seen in panel (b), is that the first 10 dolls are now sold to those willing to pay $35 or more for them, and the remaining 20 dolls continue to sell for $25.

What will happen to firm's profit?

Because firm's selling the same total number of dolls each day, firm costs are unchanged. In effect, firm has merely raised the price of the first 10 dolls by $10 each (from $25 to

$35), increasing firm revenue by $10 x 10 dolls = $100. Thus, Firm's profit must rise by $100. This *increase* in profit is represented by the darker shaded rectangle in Figure 1(b).

Firm's total profit is now equal to the areas of *both* shaded rectangles to get firm—the lightly-shaded one (firm original profit of $450) and the darker one (firm additional profit of $100). Thus, firm total profit has risen to $550.

What about firm customers?

Those who pay a high firm price are harmed by firm price discrimination (compared to the single-price outcome). Firm customers, who continue to pay $25, are unaffected. Thus, the increase in firm's profit is equal to the additional payments by the customer who pay more.

More generally, *Price discrimination can raise the price for some consumers above the price they would pay under a single-price policy. The additional profit for the firm comes at the expense of the consumers who pay more.*

Price Discrimination That Benefits Consumers

Let's go back to the initial, single-price policy and suppose that firm had discovered a different way to price discriminate. Observing that those who come late in the day and tend not to buy any dolls at $25, firm decides to lower the price to $20 during last hour of business. Sure enough, firm sells an additional 10 dolls that way. The result is shown in panel (c), firm charges $25 for the first 30 dolls, and $20 for an additional 10 beyond those. Firm total output is now 40 dolls per day.

By pushing firm output all the way to 40 dolls per day, isn't firm violating the $MC = MR$ rule and decreasing firm profit?

The *MR* curve in the figure was drawn under the assumption that firm would have to lower firm price on *all* dolls in order to sell more of them. But this is no longer the case. With price discrimination, the *MR* curve no longer tells us what will happen to firm's revenue when firm increases firm output.

In fact, we know that each doll firm sells for $20 will now add a full $20 to firm revenue. At the same time, each one adds only $10 to firm cost. So firm earns profit of $20 - $10

= \$10 on each additional doll. Firm should sell as many of these additional dolls at \$20 as people will buy.

According to the demand curve, that is 10 dolls beyond the previous 30. Selling these 10 additional dolls increases firm total profit by \$10 x 10 = \$100. This *increase* in profit is represented by the darker shaded rectangle in Figure 1(c).

Firm's total profit is once again equal to the areas of *both* shaded rectangles to get firm— the lightly shaded one (firm original profit of \$450) and the darker one (firm additional profit of \$100). Firm total profit has risen to \$550.

In this case, firm's customers are better off, too. The first 30 customers are unaffected— they continue to pay \$25. But with price discrimination, an additional 10 people are able to buy dolls at a price they are willing to pay.

Price discrimination can lower the price for some consumers below the price they would pay under a single-price policy. Those consumers benefit, while the firm earns high firm profit.

Of course, it is possible for a firm to combine *both* types of price discrimination. That is, it could raise the price above what it would charge as a single-price monopoly for some consumers, and lower it for firms. This would increase the firm's profit, while benefiting some consumers and harming firms.

Perfect Price Discrimination

Suppose a firm could somehow find out the maximum price customers would be willing to pay for *each* unit of output it sells. Then it could increase its profits even further by practicing *perfect price discrimination:*

Under ***perfect price discrimination,*** *a firm charges each customer the most the customer would be willing to pay for each unit he or firm buys.*

Perfect price discrimination is very difficult to practice in the real world, since it would require the firm to read its customers' minds. However, many real-world situations come rather close to perfect price discrimination. Used-car dealers routinely post a sticker price far higher than the price they think they can actually get. They then size up each customer to determine the discount needed to complete the sale.

The dealer may look at the customer's clothes and the car the customer is currently driving, inquire about the customer's job, and observe how sophisticated the customer is about cars. The aim is to determine the maximum price he or firm would be willing to pay. A similar sizing up takes place in flea markets, yard sales, and many other situations in which the final price is *negotiated* rather than fixed in advance.

In effect, firm knows exactly the demand curve each customer. With firm new skills, firm can increase firm profit by becoming a *perfect price discriminator:*

For each unit along the horizontal axis, firm will charge the price indicated by the vertical height of the demand curve.

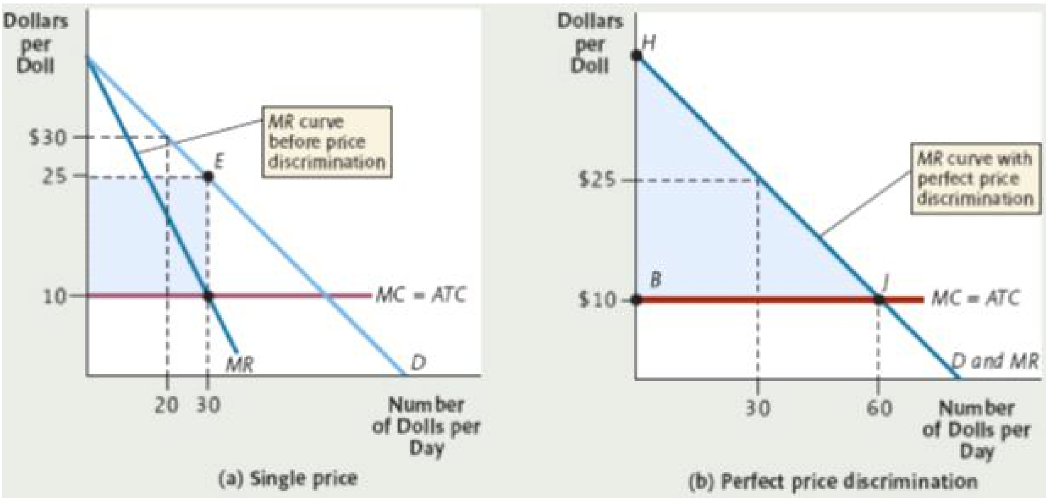
How many dolls should Firm sell now? To answer this question, we need to find the new output level at which $MR = MC$. But once again, the MR curve in the figure is no longer valid: It was based on the assumption that Firm had to lower the price on *all* units each time firm wanted to sell another one.

Now, as a perfect price discriminator, firm needs to lower the price only on the *additional* unit firm sells, and firm revenue will rise by the price of that additional unit. For example, if firm is currently selling 30 dolls and wants to sell 31, firm would lower the price just on the additional doll by a tiny bit—say, to \$24.50—and in that case, firm revenue would rise by \$24.50.

Perfect price discrimination

Charging each customer the most he or firm would be willing to pay for each unit purchased.

Figure 2: Perfect Price Discrimination



The single-price monopolist sells 30 dolls per day at \$25 each. With a constant ATC *of \$10, firm earns a profit of \$450 per day, as shown by the blue rectangle in panel (a). However, if firm can charge each customer the maximum the customer is willing to pay, shown by the height of the demand curve, then firm* MR *curve is the demand curve firm faces. In panel (b), firm would sell 60 dolls,* MC = P *at point* J. *Firm profit would increase to the area of triangle* HBJ.

For a perfect price discriminator, marginal revenue is equal to the price of the additional unit sold. Thus, the firm's MR *curve is the same as its demand curve.*

Now it is easy to see what Firm should do: Since our requirement for profit maximization is that $MC = MR$, and for a perfect price discriminator, MR is the same as price (P), Firm should produce at $MC = P$. In Figure 2(b), this occurs at point J, the MC curve intersects the demand curve—at 60 units of output.

At that point, the only way to increase sales would be to lower the price on an additional doll below \$10, but since the marginal cost of a doll is always \$10, we would have $P < MC$, and Firm's profit would decline.

What is firm's profit-maximizing price?

What about firm's total profit?

On each unit of output, firm charges a price given by the demand curve and bears a cost of $10. Adding up the profit on *all* units gives us the area under the demand curve and above $10, or the area of triangle *HBJ* (not shaded).

Now we can determine who gains and who loses when firm transforms from a single-price monopolist to a perfect price discriminator. Firm clearly gains: Firm profit increases, from the rectangle in Figure 2(a) to the larger, shaded triangle in Figure 2(b). Consumers of the product are the clear losers: Since they all pay the most they would willingly pay, no one gets to buy a doll at a price he or firm would regard as a "good deal."

A perfect price discriminator *increases profit at the expense of consumers, charging each customer the most he or firm would willingly pay for the product.*

How Firms Choose Multiple Prices

Consider Air India Airlines, the only airline that flies direct between two small cities. Air India offers one flight each day, and serves two kinds of customers: business travelers and college students. Business travelers want to minimize their travel time, so are less sensitive to price. Students, by contrast, are more willing to travel by train or take a road trip and are generally more price-sensitive.

Suppose Air India has determined that it can separate these two markets by requiring a Saturday stay-over and advanced booking for a special "student price," while charging everyone else (for example, business travelers) the "normal price." It will charge a single (different) price in *each* market. But what should those prices be?

The guiding rule is:

A price discriminating firm facing separate market demand curves in different markets (A, B, C, etc. . .) *should choose its prices and output levels so that marginal revenue in each market is equal to its marginal cost of production:*

$$MRA = MRB = MRC = \ldots = MC.$$

Look at Figure 3, the marginal cost is constant at $40. (For Air India, marginal cost would include the additional fuel for an additional passenger, additional in flight snacks, and perhaps the additional labor hours of ticket takers and baggage handlers). Panel (a) shows the demand and marginal revenue curves of the business travelers, while panel (b) shows the same curves for students.

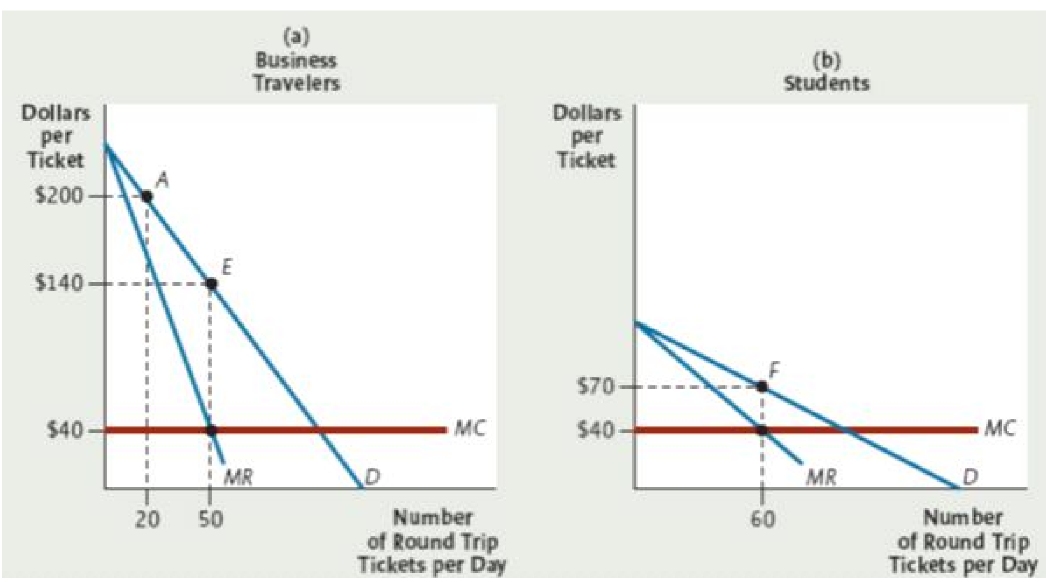
Imagine, first, that Air India was charging \$200 for business travelers, and selling 20 tickets, at point *A*.

Could this be the best price to charge?

The answer is no.

If you draw a vertical line from point *A* down to the *MR* curve, you will find that marginal revenue at 20 tickets is about \$160 substantially above the \$40 marginal cost. So the airline could increase profit by lowering its price for business travelers and selling more tickets to them. In fact, the airline should continue lowering its price to business travelers until its marginal revenue decreases to \$40, which occurs at 50 tickets per day. At that number of tickets, the *MR* curve intersects the *MC* curve, and the demand curve (at point *E*) tells us that price will be \$140. Any further increase in ticket sales in the business market— say, to 60 tickets per day—would cause *MR* to drop below *MC*, and profit would decrease.

Figure 3: How a Price-Discriminating Monopoly sets Prices in Multiple Markets



No Choice Airlines is able to separate travelers into two different types: Business travelers in panel (a) and students in panel (b). In each market, it sells the profit-maximizing number of tickets at which marginal revenue is equal to its marginal cost, and charges the price on the demand curve for that number of tickets. Because the demand curves are different, so are the marginal revenue curves, so price and output will differ in each market. In panel (a), for business travelers, MR = MC *at 50 tickets, and No Choice charges $140. In panel (b), for students,* MR = MC *at 60 tickets, and No Choice charges $70.*

Air India should follow the same procedure in the student market in panel (b). The *MR* curve intersects the *MC* curve at 60 tickets, and the demand curve (at point *F*) gives us a price of $70. So Air India will charge $70 in the student market.

We cannot use this diagram to determine Air India's total profit or loss, because Air India—like any airline—has other costs that are not part of its horizontal *MC* curve. (This includes the cost for its air terminal, salaries for pilot and copilot, and plane servicing costs.) But we do know that, assuming Air India operates at all, it maximizes profit by selling 50 tickets to business travelers for $140 each, and 60 tickets to students at $70 each. At any other prices (or output levels), profit would be smaller.

Price Discrimination in Everyday Life

Price discrimination is not limited to monopolies. It can be practiced by *any* firm that satisfies the three requirements discussed earlier. As a result, price discrimination is more prevalent.

Rebates on electronic goods are an example. If you've recently purchased a printer or computer, chances are your receipt included a coupon for a rebate from the store or the manufacturer—in effect, offering you a lower price. But to pay this lower price, you must go through the time and trouble to read all the directions, cut the UPC code from the box, mail it in, wait several weeks or months for your check to arrive, and then deposit the check.

Many people complain about all this time and trouble, and wonder why the manufacturer or store doesn't just lower the sticker price. The answer, in large part, is price discrimination. By adding time, trouble, and delay for the discount, the store can separate those who are very price sensitive from those who are not. In effect, each group is charged a different price.

Discount coupons for the supermarket or drugstore work much the same way. You only get the discount if you happen to have the coupon with you at the store. Only the most price-sensitive customers will go through the trouble of clipping, saving, and organizing their coupons so that they have them when they need them.

When retailers put items “on sale” (for a reduced price) after a delay of weeks or months, it is in part an effort to price discriminate. Those who feel they must have the latest fashions, video games, or DVDs immediately after they arrive at the store, and have the income to buy them at higher prices, will make their purchases soon after the goods arrive. Weeks or months later, when the goods go on sale, everyone else pays a lower price.

Conclusion

A *monopoly firm* is the only seller of a good or service in a market. The market is defined broadly enough to include any close substitutes. Monopoly arises because of some barrier to entry: economies of scale, legal barriers, or network externalities. As the only seller, the monopoly faces the market demand curve and must decide what price (or prices) to charge in order to maximize profit.

Like other firms, a single-price monopolist will produce at $MR = MC$ and set the maximum price consumers are willing to pay for that quantity. Monopoly profit (P - ATC multiplied by the quantity produced) can persist in the long run because of barriers to entry.

However, government regulation and rent-seeking activity can reduce monopoly profit. All else equal, a monopoly charges a higher price and produces less output than a perfectly competitive market. When demand for a monopoly’s product increases, it will raise prices and increase production. When a monopoly’s marginal costs decrease, it will pass only part of the cost savings on to consumers.

Some monopolies can practice *price discrimination* by charging different prices to different customers. Doing so requires the ability to identify customers who are willing to pay more and to prevent low-price customers from reselling to high-price customers. Price discrimination always benefits the monopolist (otherwise, it would charge a single price), but it can either benefit or harm consumers, depending on whether they face higher or lower prices after the discrimination. With *perfect* price discrimination, every consumer is charged the highest price they are willing to pay.

When a price discriminating firm faces more than one market, it maximizes its profits by equating the marginal revenue in each market to its marginal cost of production. This leads to a higher price in markets, buyers are less price-sensitive, and lower prices in markets, buyers are more price sensitive.

Questions

1. Define price discrimination.
2. What are conditions of price discrimination?
3. Discuss effects of price discrimination.
4. Explain the types of price discrimination.
5. How multiple prices are charged under price discrimination?

This lesson will focus on the following:

1. Characteristics of monopolistic competition
2. Monopolistic Competition in Short Run
3. Monopolistic Competition in Long Run

Concept of Imperfect Competition

When thinking of market structure, perfect competition and monopoly can be viewed as the two extremes.

In perfect competition, there are so many firms producing the same product that each takes the market price as given. In monopoly, there is only one firm in the market, producing a product without close substitutes. It can set its price without worrying about other firms that are selling a similar product.

Most goods and services, however, are sold in markets that are neither perfectly competitive nor monopolies. Instead, they lie somewhere *between* these two extremes. In these markets, there is more than one firm, but each firm has some market power—some ability to set price. Consider, for example, the market for wireless phone service in the United States.

It is certainly not a monopoly, because there is more than one firm. But neither does it resemble perfect competition.

Monopolistic Competition and Oligopoly

In terms of the number of firms and differences in the product, the market for wireless phone service falls somewhere between the extremes of monopoly and perfect competition.

Consider restaurants. Even a modest-size city such as Jammu has hundreds of different restaurants. This is certainly a large number of competitors. But they are not *perfect* competitors, because each one sells a product that is differentiated in important ways—in the type of food served, the recipes used, the location, and even the friendliness of the staff. The markets for wireless phone service and restaurant meals in most cities are examples of **imperfect competition.**

Imperfectly competitive markets have more than one firm (so they are not monopolies), but they violate one or more of the requirements of perfect competition.

Monopolistic Competition

The monopolistic competition is a hybrid of perfect competition and monopoly, sharing some of the features of each. Specifically, *a monopolistically competitive market has three fundamental characteristics:*

1. many buyers and sellers;

2. sellers offer a differentiated product; and

3. sellers can easily enter or exit the market.

Many Buyers and Sellers

In *perfect* competition, the existence of many buyers and sellers played an important role: ensuring that no individual buyer or seller could influence the market price.

In monopolistic competition, the "many buyers and sellers" assumption plays the same

Imperfect competition A market structure in which there is more than one firm but one or more of the requirements of perfect competition is violated.

Monopolistic competition A market structure in which there are many firms selling products that are differentiated, and in which there is easy entry and exit. 1 Imperfect competition is sometimes defined as *any* market structure other than perfect competition, which would include monopoly as well.

But an individual seller, in spite of having many competitors, has market power and acts as a *price setter*.

Our assumption of many sellers, however, has another purpose: it rules out strategic interaction among firms in the market. That is, when a firm under monopolistic competition makes a decision (about price, advertising, product guarantees, etc.), it does not take into account how it will affect other firms, or how they might respond. There are simply too many other firms, each supplying such a small part of the market, that no one of them can have much impact on the other.

Sellers Offer a Differentiated Product

In perfect competition, sellers offer a standardized product. In *monopolistic competition,* by contrast, each seller produces a somewhat different product from the others. No two coffee houses, photocopy shops, or food markets are exactly the same.

For this reason, a monopolistic competitor can raise its price (up to a point) and lose only *some* of its customers. The others will stay with the firm because they like its product, even when it charges somewhat more than its competitors.

Thus, a monopolistic competitor faces a *downward-sloping demand curve,* so it has market power. In this sense, it is more like a monopolist than a perfect competitor: *Because it produces a differentiated product, a monopolistic competitor faces a downward-sloping demand curve: It can sell more by charging less, or raise its price without losing all of its customers.*

What makes a product differentiated?

Sometimes, it is the *quality* of the product. By many objective standards—longevity, performance, frequency of repair—a Mercedes is a better car than a Hyundai. Similarly, based on room size and service, the Hilton has better hotel rooms than Motel.

But the difference can also be a matter of taste. Objectively speaking, Colgate toothpaste may be neither better nor worse than Patanjali. But each has its own flavor and texture, and each appeals to different people.

Another type of differentiation arises from differences in *location.* Two bookstores may be identical in every respect—range of selection, service— but you will often prefer the one closer to your home or office. Ultimately, though, product differentiation is subjective: A product is different whenever people *think* that it is, whether their perception is accurate or not.

Because a monopolistic competitor faces a downward-sloping demand curve, the firm *chooses* its price. Like a monopoly, it is a *price setter*. Monopolistic competitors do imitate the successful practices of others in the market. But a monopolistic competitor does not take into account the *potential* for imitation when making a decision. In the second half of this chapter, when we study oligopoly, we'll examine what happens when firms *do* take into account the potential reactions of their rivals.

Easy Entry and Exit

This feature is shared by monopolistic competition and perfect competition. It plays the same role in both: ensuring that firms earn zero economic profit in the long run. "Easy entry" does *not* mean that entry is effortless or inexpensive.

In monopolistic competition, however, easy entry extends to business practices as well: Any firm can copy the successful practices of other firms. If one movie theater finds that offering lower prices for afternoon showings generates economic profit, any other movie theater can do the same.

Monopolistic Competition in the Short Run

The individual monopolistic competitor behaves very much like a monopoly. Its goal is to maximize profit by producing therefore $MR = MC$. The result may be economic profit or loss in the short run.

The key difference is this:

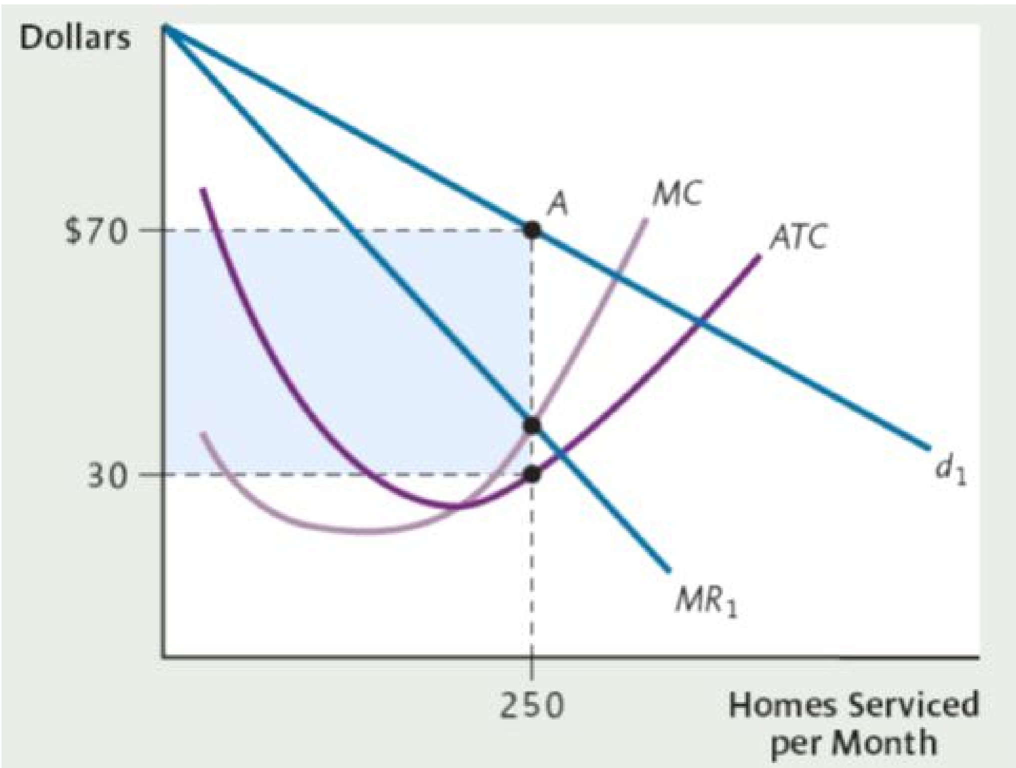
While a monopoly is the *only* seller in its market, a monopolistic competitor is one of many sellers. When a *monopoly* raises its price, its customers must pay up or buy less of the product. When a *monopolistic competitor* raises its price, its customers have one additional option: They can buy a similar (though not identical) good from some other firm. Thus, all else equal, the demand curve facing a firm will be more elastic under monopolistic competition than under monopoly.

Figure 1 illustrates the situation of a monopolistic competitor. The figure shows the demand curve, $d1$, that the firm faces, as well as the marginal revenue, marginal cost, and average total cost curves. A monopolistic competitor competes with many other firms in its local area. Thus, if it raises its price, it will lose some of its customers to the competition.

Like any other firm, this will produce where $MR = MC$. In Figure 1, when firm faces demand curve $d1$ and the associated marginal revenue curve $MR1$, its profit-maximizing output level is 250 per month, and its profit-maximizing price is \$70 per unit. In the short run, the firm may earn an economic profit or an economic loss, or it may break even. In the figure, firm is earning an economic profit:

Profit per unit is P - ATC = \$70 - \$30 = \$40, and total monthly profit—the area of the rectangle—is \$40 x 250 = \$10,000.

Figure 1: A Monopolistically Competitive Firm in the Short Run



Like any other firm, a monopolistic competitor maximizes profit by producing the level of output where its MR *and* MC *curves intersect. Firm maximizes profit by servicing 250 homes per month. The profit-maximizing price (\$70) is found on the demand curve at an output level of 250 (point* A*). Profit per unit of \$40 is the difference between the price (\$70) and average total cost (\$30) at output of 250. Total profit is profit per unit times output (\$40* x *250 = \$10,000), equal to the area of the shaded rectangle.*

Monopolistic Competition in the Long Run

If firm is a monopoly, the firm could continue to earn economic profit forever, since barriers to entry would keep out any potential competitors.

But under monopolistic competition—in which there are no barriers to entry and exit—the firm will not enjoy its profit for long. New sellers will enter the market, attracted by the profits that can be earned. Firm will lose some of its customers to the new entrants. At any given price, firm will find itself servicing fewer homes than before, so the demand curve it faces will shift leftward. Entry will continue to occur, and the demand curve will continue to shift leftward, until firm and other firms are earning zero economic profit.

This process of adjustment is shown in Figure 2. The demand curve shifts leftward (from *d*1 to *d*2). The marginal revenue curve shifts left as well (from *MR*1 to *MR*2). Kafka's new profit-maximizing output level, 100, is found at the intersection point between its marginal cost curve and its *new* marginal revenue curve *MR*2.

Firm's new price—found on its demand curve *d*2 at 100 units—is $40. Finally, since *ATC* is also $40 at that output level, firm is earning zero economic profit—the best it can do in the long run. In long-run equilibrium, the profit-maximizing price, $40, will always equal the average total cost of production.

We can also reverse these steps. If the typical firm is suffering an economic loss, *exit* will occur. With fewer competitors, those firms that remain in the market will gain customers, so their demand curves will shift *rightward.*

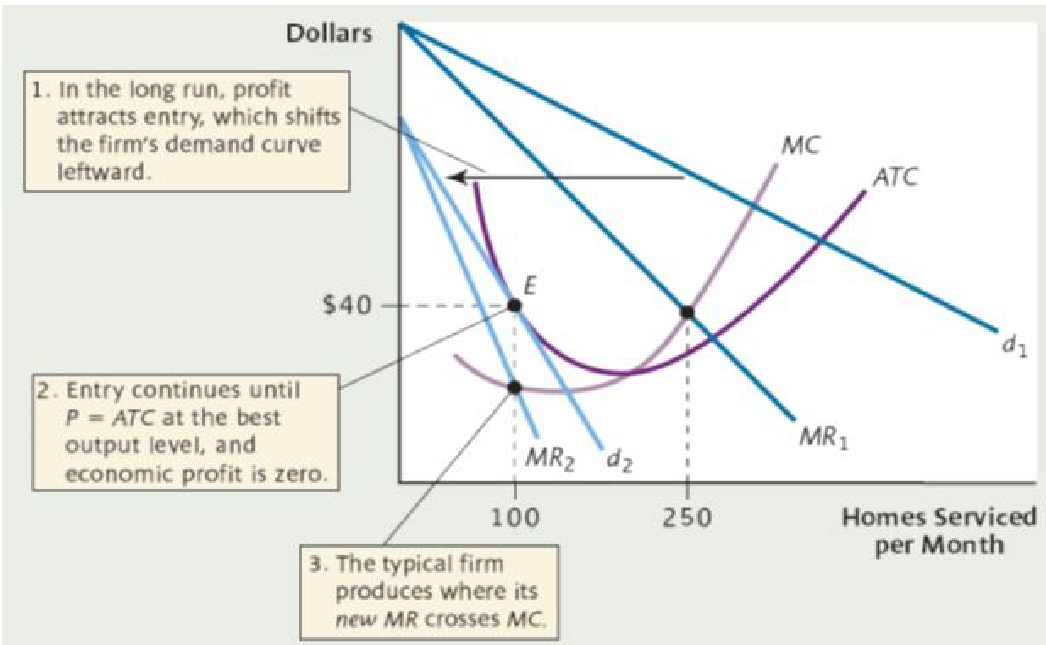
Exit will cease only when the typical firm is earning zero economic profit, the demand curve just touches the *ATC* curve point like *E* in Figure 2. Thus, the typical firm will earn zero economic profit in the long run, whether we start from a position of economic profit or economic loss:

Under monopolistic competition, firms can earn positive or negative economic profit in the short run. But in the long run, free entry and exit ensure that each firm earns zero economic profit, just as under perfect competition.

Is this prediction of our model realistic?

Indeed it is: In the real world, monopolistic competitors often earn economic profit or loss in the short run. But, given enough time, profits attract new entrants while losses result in an "industry shakeout" as firms exit. In the long run, restaurants, retail stores, hair salons, and other monopolistically competitive firms earn zero economic profit for their owners. That is, there is just enough accounting profit to cover the implicit costs of doing business—just enough to keep the owners from shifting their time and money to some alternative enterprise.

Figure 2: A Monopolistically Competitive Firm in the Long Run



Excess Capacity Under Monopolistic Competition

Look again at Figure 2, which shows firm's long-run equilibrium, at point *E,* after entry has eliminated its profits.

At that point, the *ATC* curve has the same slope as the demand curve, a *negative* slope. Thus, in the long run, a monopolistic competitor always produces on the *downward-sloping* portion of its *ATC* curve and therefore *never produces at minimum average cost.*

Indeed, its output level is always *too small* to minimize cost per unit. The firm operates with *excess capacity.* The output level at which cost per unit is minimized is often called capacity output.

In Figure 2, firm *would* reach minimum cost per unit by servicing about 200 homes per month (the firm's capacity output), but in the long run, it will service only 100 homes per month.

In the long run, a monopolistic competitor will operate with excess capacity— that is, it will not sell enough output to achieve minimum cost per unit.

To see why a monopolistic competitor *cannot* minimize average cost in the long run, imagine that firm wanted to do so, by servicing 200 homes per month. With its current demand curve, it would suffer a loss, since $P < ATC$ at that output level. It would quickly return to its profit-maximizing output of 100 homes, at least it breaks even.

Excess capacity suggests that monopolistic competition is costly to consumers, and indeed it is. Recall that under perfect competition, P = minimum ATC in long run equilibrium.

But under monopolistic competition, $P >$ minimum ATC in the long run. Thus, if the ATC curves were the same, price would always be greater under monopolistic competition.

This reasoning may tempt you to leap to a conclusion:

Consumers are better off under perfect competition. But don't leap so fast.

Remember that in order to get the beneficial results of perfect competition, all firms must produce identical output. It is precisely because monopolistic competitors produce *differentiated* output—and therefore have downward-sloping demand curves—that $P >$ minimum ATC in the long run.

Non-Price Competition

If a monopolistic competitor wants to increase its output, one way is to cut its price. That is, it can move *along* its demand curve. But a price cut is not the only way to increase output. Since the firm produces a differentiated product, it can also sell more by convincing people that its own output is better than that of competing firms. Such efforts, if successful, will *shift* the firm's demand curve rightward.

*Any action a firm takes to shift the demand curve for its output to the right is called **non-price competition.***

Better service, product guarantees, free home delivery, more attractive packaging, better locations, as well as advertising to inform customers about these things, are all examples of non-price competition.

This type of competition is another reason why monopolistic competitors earn zero economic profit in the long run. If an innovative firm discovers a way to shift its demand curve rightward—say, by offering better service or more clever advertising—then in the *short run,* it may be able to earn a profit.

But not for long. Remember that in monopolistic competition, the "free entry" assumption includes the ability of new entrants, as well as existing firms, to replicate the successful business practices of others. If product guarantees are enabling some firms to earn economic profit, then *all* firms will offer product guarantees. If advertising is doing the trick, then *all* firms will start ad campaigns. In the long run, imitation by others will reverse any advantage that the initiators hoped to achieve, and will begin shifting their demand curves back again. At the same time, the costs of the non-price competition shifts up each firm's *ATC* curve. After all, firms have to*pay* for advertising, product guarantees, or better staff training.

Even if non-price competition leads to profits for the early adopters in the short run, we can identify *two* forces that shrink profit back to zero in the long run:

(1) imitation by others reverses the initial rightward shift in demand; and

(2) the costs of non price competition shift the *ATC* curve upward.

In the end, each firm will once again earn zero economic profit, with its demand curve tangent to the new, higher *ATC* curve.

Questions

1. Define monopolistic competition.
2. List the characteristics of monopolistic competition.
3. Explain short-run equilibrium of firm under monopolistic competition.
4. Explain long-run equilibrium of firm under monopolistic competition.

This lesson will focus on the following:

1. Concept of oligopoly
2. Causes of oligopoly
3. Cournot Model
4. Stackelberg Model, and
5. Bertrand Model

Concept of Oligopoly

A monopolistic competitor enjoys a certain amount of independence. There are so many *other* firms selling in the market—each one such a small fish in such a large pond—that each of them can make decisions without worrying about how they will react. For example, if a single pharmacy in a large city cuts its prices or begins advertising, it can safely assume that any other pharmacy that could benefit from price cutting or advertising has already done so, or will shortly do so, *regardless of its own actions*. Thus, there is no reason for the first pharmacy to take the reactions of other pharmacies into account when making its own decisions.

But in some markets, most of the output is sold by just a few firms. These markets are not monopolies (there is more than one seller), but they are not monopolistically competitive either. There are so few firms that the actions taken by any one will *very much* affect the others and will likely generate a direct response.

Before the management team makes a decision, it must reason as follows: "If we take action *A,* our competitors will do *B,* and then we would do *C,* and they would respond with *D*. . . ," and so on. This kind of strategic interaction among firms is the hallmark of the market structure we call *oligopoly:*

An ***oligopoly*** *is a market dominated by a small number of strategically interacting firms.*

There are many different types of oligopolies. The products may be more or less identical among firms, such as copper wire, or differentiated, such as laptop computers.

Oligopoly A market structure with a small number of strategically interacting firms.

Oligopoly markets can have different characteristics, but in all cases, *a small number of strategically interacting firms produce the dominant share of output in the market.*

Oligopoly in Real World

While defining an oligopoly in theory is straightforward, *applying* the definition to real-world markets is sometimes difficult. The extent to which a market follows the oligopoly model—with market dominance by a few firms—is at the heart of public policy toward market structure.

Whether we view a market as an oligopoly depends on how the market is defined. With a narrow-enough definition, we can find oligopoly everywhere. For example, in a large city there will be thousands of restaurants, so we would properly consider the market for "restaurant meals" in that city to be monopolistically competitive. But if we define the market as "Thai restaurants within a half-mile from the civic center," there may be only two or three such firms, we have an oligopoly!

Causes of Oligopoly

If a market has just a few sellers, we should naturally ask: Why aren't there more? Especially because in some oligopoly markets firms earn economic profit year after year. Why doesn't such profit attract entry, as it does in perfect competition and monopolistic competition?

What *barriers to entry* keep out new competitors, leaving just a few firms with the market all to themselves?

Economies of Scale

One familiar barrier to entry is economies of scale. When a firm has economies of scale over a wide range of output, a large firm will have lower cost per unit than would a small firm. This can create a natural monopoly if a firm's minimum efficient scale (MES) occurs when it produces for the entire market.

Or it can create a **natural oligopoly** if the MES occurs when a firm produces for a large fraction of the market. Since small firms can't compete, only a few large firms survive, and the market becomes an oligopoly. Airlines, college textbook publishers, and passenger jet manufacturers are all examples of oligopolies in which economies of scale play a large role.

Reputation as a Barrier

A new entrant may suffer just from being new. Established oligopolists are likely to have favourable reputations. In many oligopolies—like the markets for soft drinks—heavy advertising expenditure has also helped to build and maintain brand loyalty. A new entrant might be able to catch up to those already in the industry, but this may require a substantial period of high advertising costs and low revenues. This puts new entrants at a disadvantage compared to the firms already in the industry.

Strategic Barriers

Oligopoly firms often pursue strategies designed to keep out potential competitors. They can maintain excess production capacity as a signal to a potential entrant that, with little advance notice, they could easily saturate the market and leave the new entrant with little or no revenue. They can make special deals with distributors to receive the best space in retail stores or make long-term arrangements with customers to ensure that their products are not displaced quickly by those of a new entrant.

Legal Barriers

Patents and copyrights, which can be responsible for monopoly, can also create oligopolies. For example, only four medications have received government approval for treatment of mild to moderate Alzheimer's disease, and all four are still protected by patents. Until these patents expire, or several new drugs are developed, this market will continue to be an oligopoly in which just four large pharmaceutical companies are the sellers.

Like monopolies, oligopolies are not shy about lobbying the government to preserve their market domination. One of the easiest targets is foreign competition. U.S. steel companies are relentless in their efforts to limit the amount of foreign— especially Japanese—steel sold in the U.S. market. In the past, they have succeeded in getting special taxes on imported steel and financial penalties imposed upon successful foreign steel companies.

In local markets, zoning regulations may prohibit the building of a new supermarket, movie theater, or auto repair shop, thereby preserving the oligopoly status of the few firms already established firms. Lobbying by established firms is often the source of these restrictive practices.

Cournot Model

The duopoly model to analyze competition between a small number of competitors is given by French economist Augustin Cournot. Consequently it has come to be known as the Cournot duopoly model. While the maximizing behaviour that is incorporated in this model can apply to a situation with several firms rather than two, we will develop the model with two firms. This differs slightly from the preceding section, where each firm has simply a choice between a high or low output.

The critical element of the Cournot approach is that the firms each determine their optimal strategy – one that maximizes profit – by reacting optimally to their opponent's strategy, which in this case involves their choice of output.

Cournot behaviour involves each firm reacting optimally in their choice of output to their competitors' output decisions.

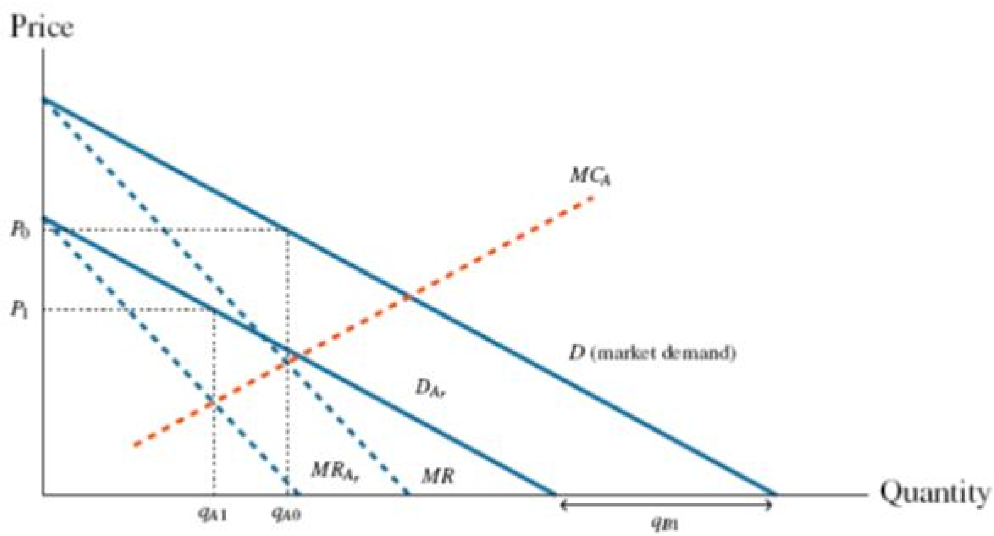
A central element here is the reaction function of each firm, which defines the optimal output choice conditional upon their opponent's choice.

Reaction functions define the optimal choice of output conditional upon a rival's output choice.

We can develop an optimal strategy with the help of Figure 1. D is the market demand, and two firms supply this market. If B supplies a zero output, then A would face the whole demand, and would maximize profit where $MC = MR$. Let this output be defined by $qA0$. We transfer this output combination to Figure 2, where the output of each firm is on one of the axes—A on the vertical axis and B on the horizontal. This particular combination of zero output for B and $qA0$ for A is represented on the vertical axis as the point $qA0$.

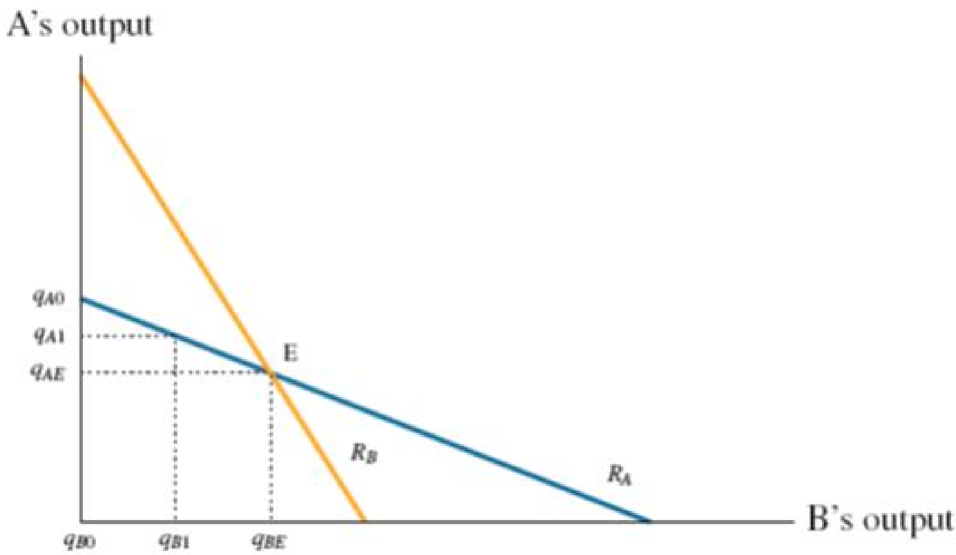
Instead, suppose that B produces a quantity $qB1$ in Figure 11.3. This reduces the demand curve facing A correspondingly from D to DAr , which we call A's *residual demand*. When subject to such a choice by B, firm A maximizes profit by producing where $MRAr = MC$, where $MRAr$ is the marginal revenue corresponding to the residual demand DAr . The optimum for A is now $qA1$, and this pair of outputs is represented by the combination $(qA1,qB1)$ in Figure 2.

Figure 1: Duopoly Behaviour



When one firm, B, chooses a specific output, e.g. $qB1$, then A's residual demand DAr is the difference between the market demand and $qB1$. A's profit is maximized at $qA1$ – where$MC = MRAr$. This is an optimal reaction by A to B's choice. For all possible choices by B, A can form a similar optimal response. The combination of these responses forms A's reaction function.

Figure 2: Reaction Functions and Equilibrium



The reaction function for A (*RA*) defines the optimal output response for A to any output choice by B. The reaction function for B is defined similarly. The equilibrium occurs at the intersection of *RA* and *RB*. Any other combination will induce one firm to change its output, and therefore could not be an equilibrium.

Firm *A* forms a similar optimal response for every possible output level that *B* could choose, and these responses define A's reaction function. The reaction function illustrated for *A* in Figure 2.

It is thus the locus of all optimal response outputs on the part of *A*. The downward-sloping function makes sense: The more *B* produces, the smaller is the residual market for *A*, and therefore the less *A* will produce.

But *A* is just one of the players in the game. If *B* acts in the same optimizing fashion, *B* too can formulate a series of optimal reactions to A's output choices. The combination of such choices would yield a reaction function for *B*. This is plotted as *RB* in Figure 2.

An equilibrium is defined by the intersection of the two reaction functions, in this case by the point E. At this output level *each firm is making an optimal decision, conditional upon the choice of its opponent*. Consequently, neither firm has an incentive to change its output; therefore it can be called the Nash equilibrium.

Any other combination of outputs on either reaction function would lead one of the players to change its output choice, and therefore could not constitute an equilibrium. To see this, suppose that B produces an output greater than qBE; how will A react? A's reaction function indicates that it should choose a quantity to supply less than qAE. If so, how will B respond in turn to that optimal choice? It responds with a quantity read from its reaction function, and this will be less than the amount chosen at the previous stage. By tracing out such a sequence of reactions it is clear that the output of each firm will move to the equilibrium qE.

The Cournot model yields an outcome that lies between monopoly (or collusion/cartel) and competitive market models. It does not necessarily assume that the firms are identical in terms of their cost structure, although the lower-cost producer will end up with a larger share of the market.

The next question that arises is whether this duopoly market will be sustained as a duopoly, or if entry may take place. In particular, if economic profits accrue to the participants will such profits be competed away by the arrival of new producers, or might there be barriers of either a 'natural' or 'constructed' type that operate against new entrants?

Entry, exit & potential competition

At this point we inquire about the potential entry and impact of new firms – firms who might enter the industry if conditions were sufficiently enticing, meaning the presence of economic profits. One way of examining entry in this oligopolistic world is to envisage potential entry barriers as being either intended or unintended, though the difference between the two can be blurred. Broadly, an unintended or 'natural' barrier is one related to cost conditions and the size of the market. An intended barrier involves a strategic decision on the part of the firm to prevent entry.

Unintended entry barriers

Oligopolists tend to have substantial fixed costs, accompanied by declining average costs up to very high output levels. Such a cost structure 'naturally' gives rise to a supply side with a small number of suppliers. For examples, given demand and cost structures, could Vancouver support two professional soccer teams; could Calgary support two professional hockey teams; could Jammu sustain two professional football teams? The answer to each of these questions is likely 'no'.

Because given the cost structure of these markets, it would not be possible to induce twice as many spectators without reducing the price per game ticket to such a degree that revenue would be insufficient to cover costs. (We will neglect for the moment that the governing bodies of these sports also have the power to limit entry.) Fixed costs include stadium costs, staff payrolls and player payrolls. In fact most costs in these markets are relatively fixed. Market size relative to fixed and variable costs is not large enough to sustain two teams in most cities. Exceptions in reality are huge urban areas such as Jalandhar and Patiala.

Accordingly, it is possible that the existing team, or teams, may earn economic profit from their present operation; but such profit does not entice further entry, because the market structure is such that the entry of an additional team could lead to all teams making losses.

Intended entry barriers

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.

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.

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'.

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).

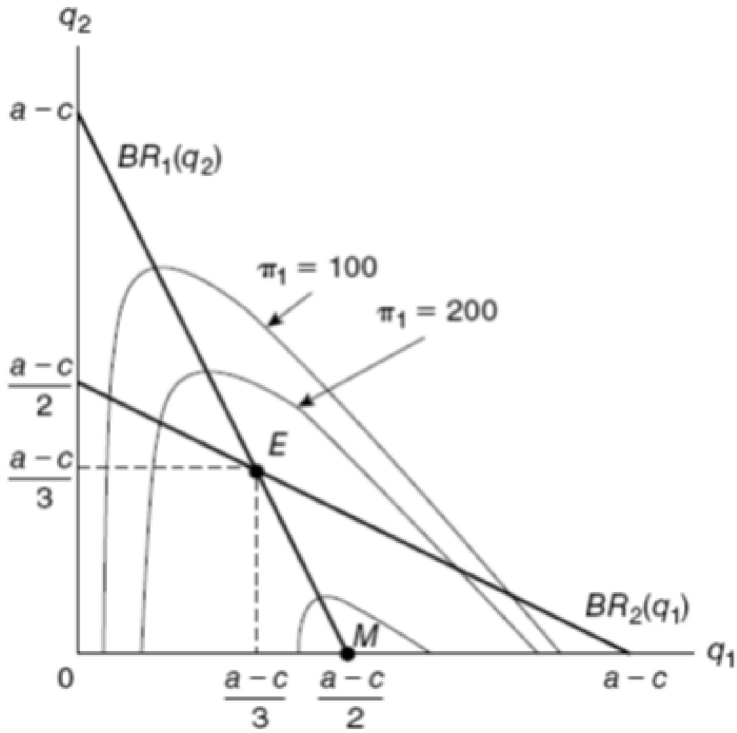
An ***over-investment*** **strategy** means that an existing supplier generates additional production capacity through investment in new plant or capital. This is costly to the incumbent and is intended as a signal to any potential entrant that this capacity could be brought on-line immediately should a potential competitor contemplate entry. For example, a ski-resort owner may invest in a new chair-lift, even if she does not use it. The existence of the additional capacity may scare potential entrants. A key component of this strategy is that the incumbent firm invests ahead of time – and inflicts a cost on itself. The incumbent does not simply say "I will build another chair-lift if you decide to develop a nearby mountain into a ski hill." That policy does not carry the same degree of credibility as actually incurring the cost of construction ahead of time.

However, such a strategy may not always be feasible: It might be just too costly to pre-empt entry by putting spare capacity in place. Spare capacity is not so different from brand development through advertising; both are types of sunk cost. The threats associated with the incumbent's behaviour become a credible threat because the incumbent incurs costs up front. A credible threat is one that is effective in deterring specific behaviours; a competitor must believe that the threat will be implemented if the competitor behaves in a certain way.

Best-Response Diagram for Cournot Duopoly

Firms' best responses are drawn as thick lines in figure 3; their intersection (E) is the Nash equilibrium of the Cournot game. Isoprofit curves for firm 1 increase until point M is reached, which is the monopoly outcome for firm 1.

Figure 3: Best-Response for Cournot Duopoly

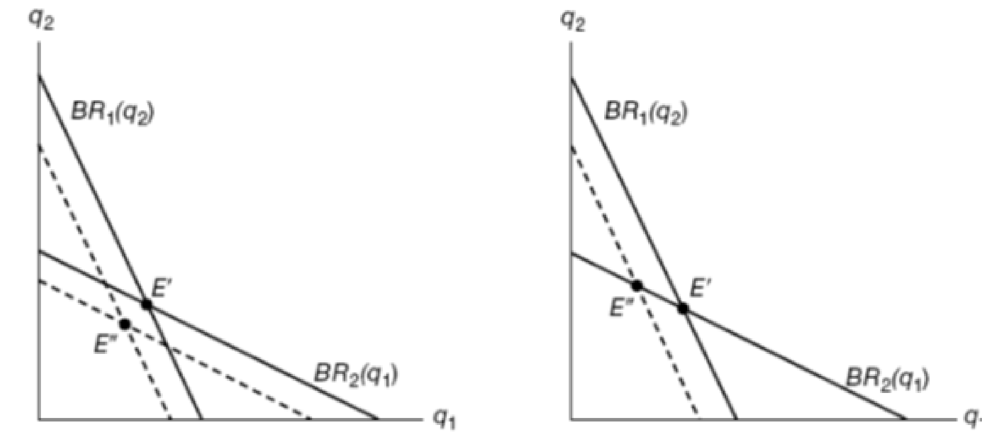


Several example isoprofits for firm 1 are shown in the figure. As profit increases from 100 to 200 to yet higher levels, the associated isoprofits shrink down to the monopoly point, which is the highest isoprofit on the diagram.

Shifting Cournot Best Responses

Firms' initial best responses are drawn as solid lines, resulting in a Nash equilibrium at point E0. Panel (a) depicts an increase in both firms' marginal costs, shifting their best responses—now given by the dashed lines—inward. The new intersection point, and thus the new equilibrium, is point E00. Panel (b) depicts an increase in just firm 1's marginal cost.

Figure 4: Shifting Cournot Best Responses



(a) Increase in both firms' marginal costs (b) Increase in firm 1's marginal cost

Prices or quantities?

Why price and quantity are such different strategic variables. Starting from equal prices, a small reduction in one firm's price allows it to steal all of market demand from its competitors. This sharp benefit from undercutting makes price competition extremely "tough." Quantity competition is "softer." Starting from equal quantities, a small increase in one firm's quantity has only a marginal effect on the revenue that other firms receive from their existing output. Firms have less of an incentive to out produce each other with quantity competition than to undercut each other with price competition.

An advantage of the Cournot model is its realistic implication that the industry grows more competitive as the number n of firms entering the market increases from monopoly to perfect competition. In the Bertrand model there is a discontinuous jump from monopoly to perfect competition if just two firms enter, and additional entry beyond two has no additional effect on the market outcome.

An apparent disadvantage of the Cournot model is that firms in real-world markets tend to set prices rather than quantities, contrary to the Cournot assumption that firms choose quantities. For example, grocers advertise prices for orange juice, say \$2.50 a container, in newpaper circulars rather than the number of containers it stocks.

Stackelberg Model: Quantity Leadership

In the case of quantity leadership, one firm makes a choice before the other firm. This is sometimes called the **Stackelberg model** in honor of the first economist who systematically studied leader-follower interactions. The Stackelberg model is often used to describe industries in which there is a dominant firm, or a natural leader. For example, IBM is often considered to be a dominant firm in the computer industry. A commonly observed pattern of behaviour is for smaller firms in the computer industry to wait for IBM's announcements of new products and then adjust their own product decisions accordingly. In this case we might want to model the computer industry with IBM playing the role of a Stackelberg leader, and the other firms in the industry being Stackelberg followers.

Let us turn now to the details of the theoretical model. Suppose that firm 1 is the leader and that it chooses to produce a quantity $y1$. Firm 2 responds by choosing a quantity $y2$. Each firm knows that the equilibrium price in the market depends on the total output produced. We use the inverse demand function $p(Y)$ to indicate the equilibrium price as a function of industry output, $Y = y1 + y2$.

What output should the leader choose to maximize its profits? The answer depends on how the leader thinks that the follower will react to its choice. Presumably the leader should expect that the follower will attempt to maximize profits as well, given the choice made by the leader. In order for the leader to make a sensible decision about its own production, it has to consider the follower's profit-maximization problem.

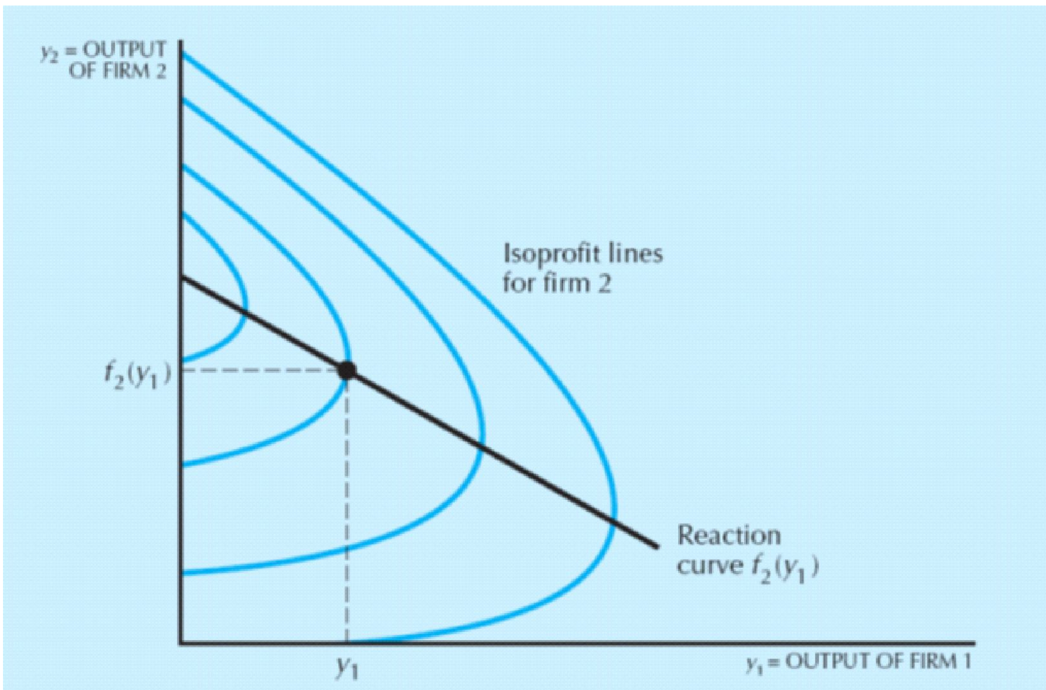
The Follower's Problem

The follower's profit depends on the output choice of the leader, but from the viewpoint of the follower the leader's output is predetermined—the production by the leader has already been made, and the follower simply views it as a constant.

The marginal revenue has the usual interpretation. When the follower increases its output, it increases its revenue by selling more output at the market price. But it also pushes the price down by Δp, and this lowers its profits on all the units that were previously sold at the higher price. The important thing to observe is that the profit-maximizing choice of the follower will depend on the choice made by the leader.

The **reaction function** since it tells us how the follower will react to the leader's choice of output. Note that profits to firm 2 will increase as we move to isoprofit lines that are further to the left. This is true since if we fix the output of firm 2 at some level, firm 2's profits will increase as firm 1's output decreases. Firm 2 will make its maximum possible profits when it is a monopolist; that is, when firm 1 chooses to produce zero units of output.

Figure 1: Reaction Function



Derivation of a reaction curve

This reaction curve gives the profit-maximizing output for the follower, firm 2, for each output choice of the leader, firm 1. For each choice of $y1$ the follower chooses the output level $f2(y1)$ associated with the isoprofit line farthest to the left.

For each possible choice of firm 1's output, firm 2 wants to choose its own output to make its profits as large as possible. This means that for each choice of $y1$, firm 2 will pick the value of $y2$ that puts it on the isoprofit line furthest to the left, as illustrated in Figure. This point will satisfy the usual sort of tangency condition: the slope of the isoprofit line must be vertical at the optimal choice. The locus of these tangencies describes firm 2's reaction curve, $f2(y1)$.

To see this result algebraically, we need an expression for the marginal revenue associated with the profit function for firm 2. It turns out that this expression is given by

$$MR2(y1, y2) = a \text{ “ } by1 \text{ “ } 2by2.$$

The Leader's Problem

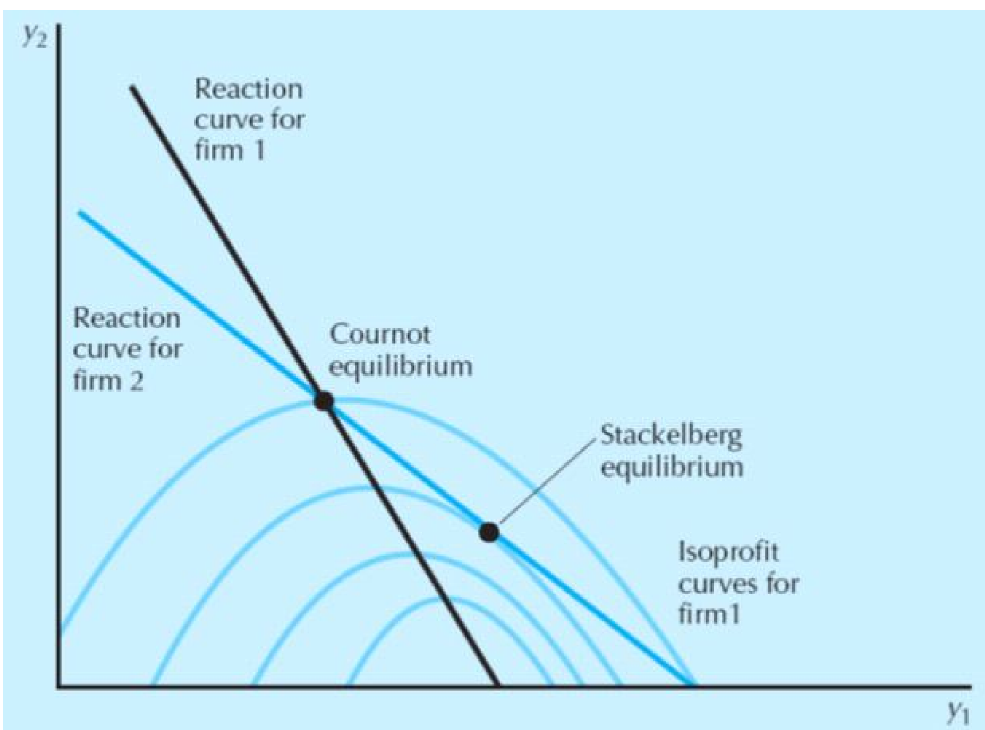
We have now examined how the follower will choose its output *given* the choice of the leader. We turn now to the leader's profit-maximization problem.

Presumably, the leader is also aware that its actions influence the output choice of the follower. This relationship is summarized by the reaction function $f2(y1)$. Hence when making its output choice it should recognize the influence that it exerts on the follower.

When the leader contemplates changing its output it has to recognize the influence it exerts on the follower.

Firm 1, the leader, chooses the point on firm 2's reaction curve that touches firm 1's lowest possible isoprofit line, thus yielding the highest possible profits for firm 1.

Figure 2: Stackelberg Equilibrium



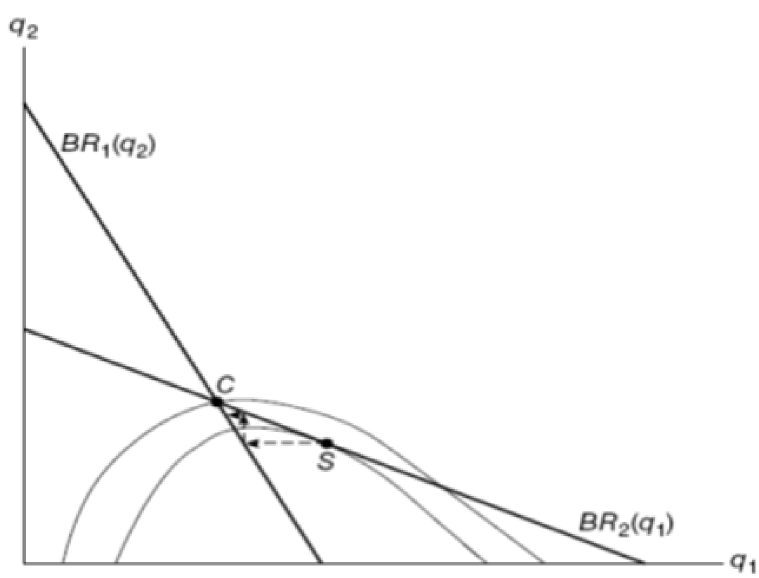
In brief, the simplest setting to illustrate the first-mover advantage is in the Stackelberg model. The model is similar to a duopoly version of the Cournot model except that—rather than simultaneously choosing the quantities of their identical outputs—firms move sequentially, with firm 1 (the leader) choosing its output first and then firm 2 (the follower) choosing after observing firm 1's output.

We use backward induction to solve for the sub game-perfect equilibrium of this sequential game. Begin with the follower's output choice. Firm 2 chooses the output q2 that maximizes its own profit, taking firm 1's output q1 as given. In other words, firm 2 best responds to firm 1's output. This results in the same best-response function for firm 2 as we computed in the Cournot game from the first-order condition.

Stackelberg Game

Best-response functions from the Cournot game are drawn as thick lines. Frown-shaped curves are firm 1's isoprofits. Point C is the Nash equilibrium of the Cournot game (involving simultaneous output choices). The Stackelberg equilibrium is point S, the point at which the highest isoprofit for firm 1 is reached on 2's best-response function. At S, 1's isoprofit is tangent to 2's best-response function. If firm 1 cannot commit to its output then the outcome unravels, following the dotted line from S back to C.

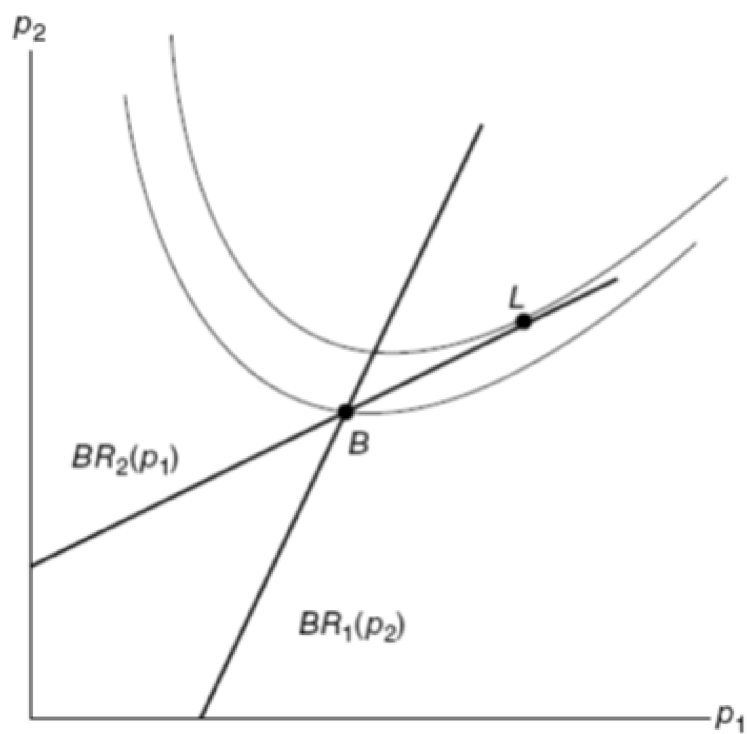
Figure 3: Stackelberg Game



Price-Leadership Game

Thick lines are best-response functions from the game in which firms choose prices for differentiated products. U-shaped curves are firm 1's isoprofits. Point B is the Nash equilibrium of the simultaneous game, and L is the sub game-perfect equilibrium of the sequential game in which firm 1 moves first. At L, 1's isoprofit is tangent to 2's best response.

Figure 4: Price-Leadership Game



Conclusion

Both firms' prices and profits are higher in this sequential game than in the simultaneous one, but now the follower earns even more than the leader. As illustrated in the best-response function diagram in above figure, firm 1 commits to a high price in order to induce firm 2 to raise its price also, essentially "softening" the competition between them.

Bertrand Model

Simultaneous Price Setting

In the Cournot model, it is assumed that firms were choosing their quantities and letting the market determine the price. Another approach is to think of firms as setting their prices and letting the market determine the quantity sold. This model is known as **Bertrand competition**.

When a firm chooses its price, it has to forecast the price set by the other firm in the industry. Just as in the case of Cournot equilibrium we want to find a pair of prices such that each price is a profit-maximizing choice given the choice made by the other firm.

What does a Bertrand equilibrium look like?

When firms are selling identical products, as we have been assuming, the Bertrand equilibrium has a very simple structure indeed. It turns out to be the competitive equilibrium, where price equals marginal cost!

How can we get a competitive price if there are only two firms in the market?

If we think of the Bertrand model as a model of competitive bidding it makes more sense. Suppose that one firm "bids" for the consumers' business by quoting a price above marginal cost. Then the other firm can always make a profit by undercutting this price with a lower price. It follows that the only price that each firm cannot rationally expect to be undercut is a price equal to marginal cost.

It is often observed that competitive bidding among firms that are unable to collude can result in prices that are much lower than can be achieved by other means. This phenomenon is simply an example of the logic of Bertrand competition.

Collusion

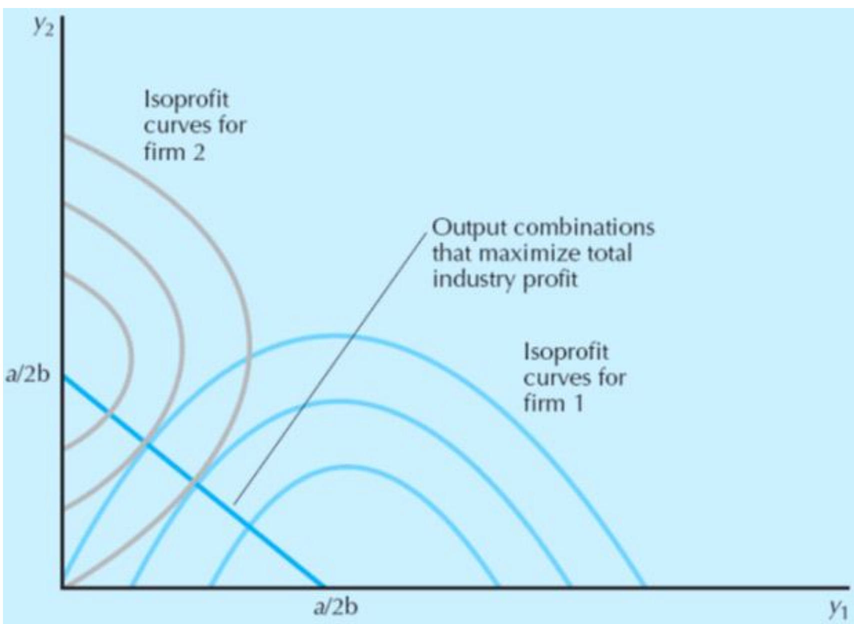
In the models we have examined up until now the firms have operated independently. But if the firms collude so as to jointly determine their output, these models are not very reasonable. If collusion is possible, the firms would do better to choose the output that maximizes total industry profits and then divide up the profits among themselves. When firms get together and attempt to set prices and outputs so as to maximize total industry profits, they are known as a **cartel**.

A cartel is simply a group of firms that jointly collude to behave like a single monopolist and maximize the sum of their profits. Thus the profit-maximization problem facing the two firms is to choose their outputs $y1$ and $y2$ so as to maximize total industry profits:

Punishment Strategies

We have seen that a cartel is fundamentally unstable in the sense that it is always in the interest of each of the firms to increase their production above that which maximizes aggregate profit. If the cartel is to operate successfully, some way must be found to "stabilize" the behavior. One way to do this is for firms to threaten to punish each other for cheating on the cartel agreement. In figure below we investigate the size of punishments necessary to stabilize a cartel.

Figure 1: Punishment Strategies



Consider a duopoly composed of two identical firms. If each firm produces half the monopoly amount of output, total profits will be maximized and each firm will get a payoff of, say, πm. In an effort to make this outcome stable, one firm announces to the other: "If you stay at the production level that maximizes joint industry projects, fine. But if I discover you cheating by producing more than this amount, I will punish you by producing the Cournot level of output forever." This is known as a **punishment strategy**.

When will this sort of threat be adequate to stabilize the cartel? We have to look at the benefits and costs of cheating as compared to those of cooperating. Suppose that cheating occurs, and the punishment is carried out. Since the optimal response to Cournot behavior is Cournot behaviour (by definition), this results in each firm receiving a per-period profit of, say, πc. Of course, the Cournot payoff, πc is less than the cartel payoff, πm.

Let us suppose that the two firms are each producing at the collusive, monopoly level of production. Put yourself in the place of one of the firms trying to decide whether to continue to produce at your quota. If you produce more output, deviating from your quota, you make profit πd, where $\pi d > \pi m$. This is the standard temptation facing a cartel member described above: if each firm restricts output and pushes the price up, then each firm has an incentive to capitalize on the high price by increasing its production.

Conclusion

The weakness of this model is that the threat to revert to Cournot behavior forever is not very believable. One firm certainly may believe that the other will punish it for deviating, but "forever" is a long time. A more realistic model would consider shorter periods of retaliation, but the analysis then becomes much more complex. In the next chapter, we discuss some models of "repeated games" that illustrate some of the possible behaviours.

Questions

1. Discuss the concept of oligopoly.
2. What are causes of oligopoly?
3. Elaborate competition and cartel.
4. Explain Cournot Model.
5. Examine Stackelberg Model.
6. Discuss Bertrand Model.

This lesson will focus on the following:

1. Oligopoly versus Other market structures
2. Game Theory Approach
3. Simple Oligopoly Games
4. Cooperative Behaviour in Oligopoly
5. Advertising in Monopolistic Competition and Oligopoly

Oligopoly and Other Market Structures

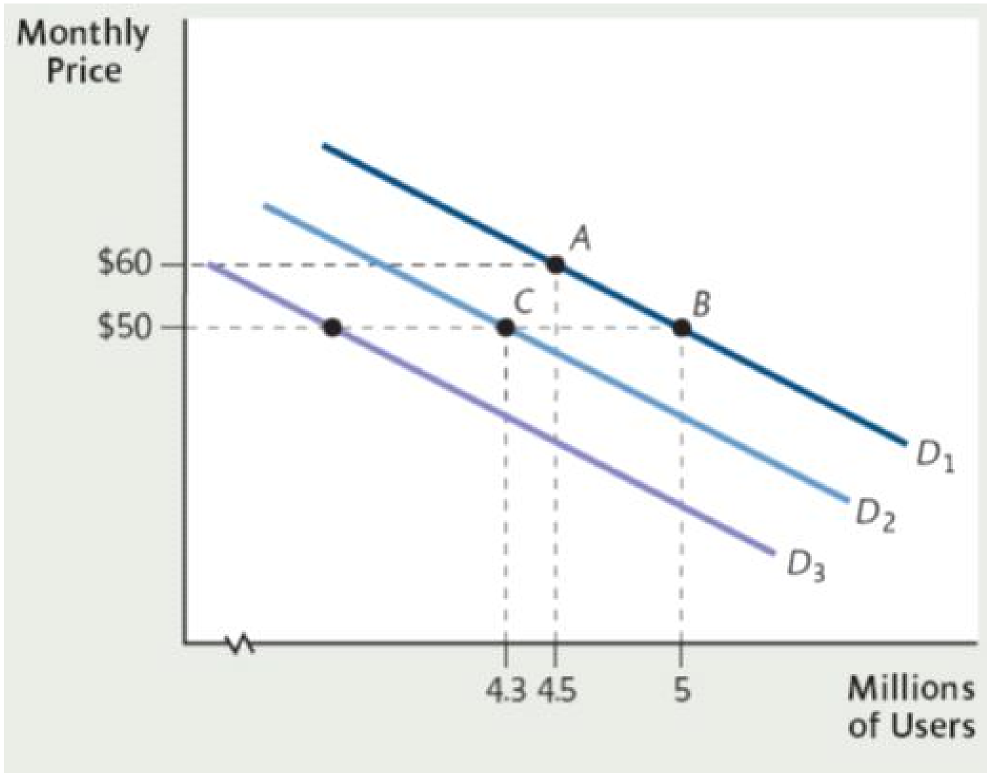
Of the market structures, oligopoly presents the greatest challenge to economists. To see why, look at Figure 1. Let's first consider the demand curve labeled $D1$. It shows (hypothetically) the number of customers the firm would have at each price—*if* the other three major carriers continued to charge $99 for their own, similar plans.

For example, with a price of $60, the diagramme shows that firm would have 4.5 million customers (point *A*). With a lower price of $50—and the other carriers continuing to charge $99—firm's service would attract 5 million customers (point *B*).

If firm was a monopolistic competitor, it could do the usual: find the marginal revenue curve associated with demand curve $D1$, find its marginal cost curve, and then find the profit maximizing number of customers where $MR = MC$.

Finally, it would find the price for that output level on its demand curve. It wouldn't have to worry about how the other carriers would respond to that price, because Sprint would be one of many small firms—too small for its decisions to influence the others and elicit a reaction from them.

Figure 1: An Oligopolist's Demand Curve Depends on its Competitors' Responses



For an oligopoly firm, such as Sprint, the quantity demanded at each price depends on the response of its competitors. Demand curve D1 *is drawn assuming that firm's competitors maintain their current prices (say, $99). In that case, firm can charge $60 and attract 4.5 million users (point* A*) or $50 and attract 5 million users (point* B*).*

However, if firm decides to charge $50, and one of its competitors matches firm's price, firm's demand curve will shift leftward, perhaps to D2*. In that case, at a price of $50, firm will attract only 4.3 million users (point* C*). If more than one competitor drops its price to match firm's $50 price, the demand curve will shift further leftward, as in the move to* D3.

But because this market is an oligopoly, firm could *not* assume that the other wireless carriers would continue to charge $99. On the contrary, firm would have to *anticipate* how the other carriers would respond after it priced its own service. Suppose, for example, that firm decided to charge $50, and T-Mobile lowered its own price to $60 in response.

Then firm's demand curve would shift leftward, to a curve like *D*2. After all, *D*1 was drawn under the assumption that all the other carriers would continue to charge $99; if one of them (T-Mobile) lowers its price, firm will have fewer customers at each price than before. Along the new demand curve *D*2, firm would now have 4.3 million customers at $50 (point *C*), rather than the 5 million customers we found earlier (point *B*).

Cooperative Behavour in Oligopoly

In the real world, oligopolists will usually get more than one chance to choose their prices. Pepsi and Coca-Cola have been rivals in the soft drink market for decades. These firms can change their strategies after observing their rivals' strategies.

The equilibrium in a game with repeated plays may be very different from the equilibrium in a game played only once. Often, firms will evolve some form of *cooperation* in the long run. For example, look again at Figure 2. If this game were played only once, we would expect each player to pursue its dominant strategy, select a low price, and end up with $25,000 in yearly profit. But there is a better outcome for both players. If each were to charge a high price, each would make a profit of $50,000 per year.

Explicit Collusion

The simplest form of cooperation is **explicit collusion,** in which managers meet face to-face to decide how to set prices. These arrangements are commonly called price fixing agreements.

The most extreme form of explicit collusion is the creation of a **cartel**—a group of firms that tries to maximize the total profits of the group as a whole. To do this, the group of firms behaves as if it were a monopoly, treating the market demand curve as the "monopoly's" demand curve. Then, it finds the point on the demand curve—the price and quantity of output—that maximizes total profit.

Each member is instructed to charge the agreed-upon price (cartels are often called *price-fixing* agreements), and each is allotted a share of the cartel's total output. This last step is crucial: If any member produces and sells more than its allotted portion, then the group's total *output* rises. The extra output would cause the price to fall below the agreed-upon profit-maximizing price.

The most famous cartel in recent years has been OPEC—the Organization of Petroleum Exporting Countries—which meets periodically to influence the price of oil by setting the amount that each of its members can produce.

If explicit collusion to raise prices is such a good thing for oligopolists, why don't they all do it? A major reason is that it's usually *illegal*. OPEC was not considered illegal by any of the oil-producing nations, but cartels are against the law in the United States, the European Union, and most of the developed nations.

In these countries, explicit collusion must be conducted with the utmost secrecy. And the penalties, if the oligopolists are caught, can be severe.

Interestingly, authorities in both the United States and Europe now use a strategy based on the prisoner's dilemma game to uncover price-fixing agreements:

The chances of getting caught, and the severe penalties at stake, often lead oligopolists to other forms of collusion that are harder to detect.

Tacit Collusion

Any time firms cooperate *without* an explicit agreement, they are engaging in **tacit collusion.** Typically, players adopt strategies along the following lines: "In general, I will set a high price. If my rival also sets a high price, I will go on setting a high price. If my rival sets a low price this time, I will punish him by setting a low price next time." You can see that if both players stick to this strategy, they will both likely set the high price.

Each is waiting for the other to go first in setting a low price, so it may never happen. This type of strategy is often called **tit-for-tat**, defined as doing to the other player what he has just done to you.

Tit-for-tat strategies are prominent in the airline industry. When one major airline announces special discounted fares, its rivals almost always announce identical fares the next day. The response from the rivals not only helps them remain competitive, but also provides a signal to the price-cutting airline that it will not be able to offer discounts that are unmatched by its rivals

Another form of tacit collusion is **price leadership**, in which one firm, the *price leader,* sets its price, and other sellers copy that price. The leader may be the dominant firm in the industry (the one with the greatest market share, for example), or the position of leader may rotate from firm to firm.

The Limits to Collusion

Collusion—whether explicit or tacit—gives oligopolies absolute power over their markets, leaving them free to jack up prices and exploit the public without limit. But oligopoly power, even with collusion, has its limits.

First, even colluding firms are constrained by the market demand curve: A rise in price will always reduce the quantity demanded from all firms together. There is one price—the cartel monopoly price—that maximizes the total profits of all firms in the market, and it will never serve the group's interest to charge any price higher than this.

Second, oligopolies are often weakened—and sometimes destroyed—by new technologies. This is especially true of local oligopolies. A small town, for example, might be able to support only a few stores selling luggage, office equipment, or books. But the Internet has enabled residents in small towns everywhere to choose among dozens or more online sellers of the same merchandise.

Third, collusion is limited by powerful incentives to cheat on any agreement. Perhaps not, because each player has an incentive to cheat by switching back to the low price. The other player may punish the cheater by lowering his own price, and cooperation may be restored. But periodic cheating often plagues oligopolies.

Anti-Trust Legislation and Enforcement

Explicit price-fixing agreements among firms violate the law in most countries. But even tacit collusion can attract the watchful eye of government. Antitrust policies—which are designed to protect the interests of consumers and preserve adequate competition—in the United States and many other countries often prevent oligopolies from forming, or police them when they do.

In practice, antitrust enforcement has focused on three types of actions:

(1) preventing collusive agreements among firms, such as price-fixing agreements;

(2) breaking up large firms or limiting their activities when market dominance harms consumers; and

(3) preventing mergers that would lead to harmful market domination.

The impact of these antitrust actions goes far beyond the specific companies called into the courtroom. Managers of firms even considering anticompetitive moves have to think long and hard about the consequences of acts that might violate the antitrust laws.

Advertising in Monopolistic Competition and Oligopoly

Perfect competitors never advertise and monopolies advertise relatively little. But advertising is almost always found under monopolistic competition and very often in oligopoly. Why?

All monopolistic competitors, and many oligopolists, produce differentiated products. In these types of markets, the firm gains customers by convincing them that its product is different and better in some way than that of its competitors.

Advertising, whether it merely informs customers about the product or attempts to influence them more subtly and psychologically, is one way to sharply differentiate a product in the minds of consumers. Since other firms will take advantage of the opportunity to advertise, any firm that *doesn't* advertise will be lost in the shuffle.

Advertising and Market Equilibrium Under Monopolistic Competition

A monopolistic competitor advertises for two reasons: to shift its demand curve rightward (greater quantity demanded at each price) and to make demand for its output *less* elastic (so it can raise price and suffer a smaller decrease in quantity demanded).

Advertising costs money, so in addition to its impact on the demand curve, it will also affect the firm's *ATC* curve. What is the ultimate impact of advertising on the typical firm?

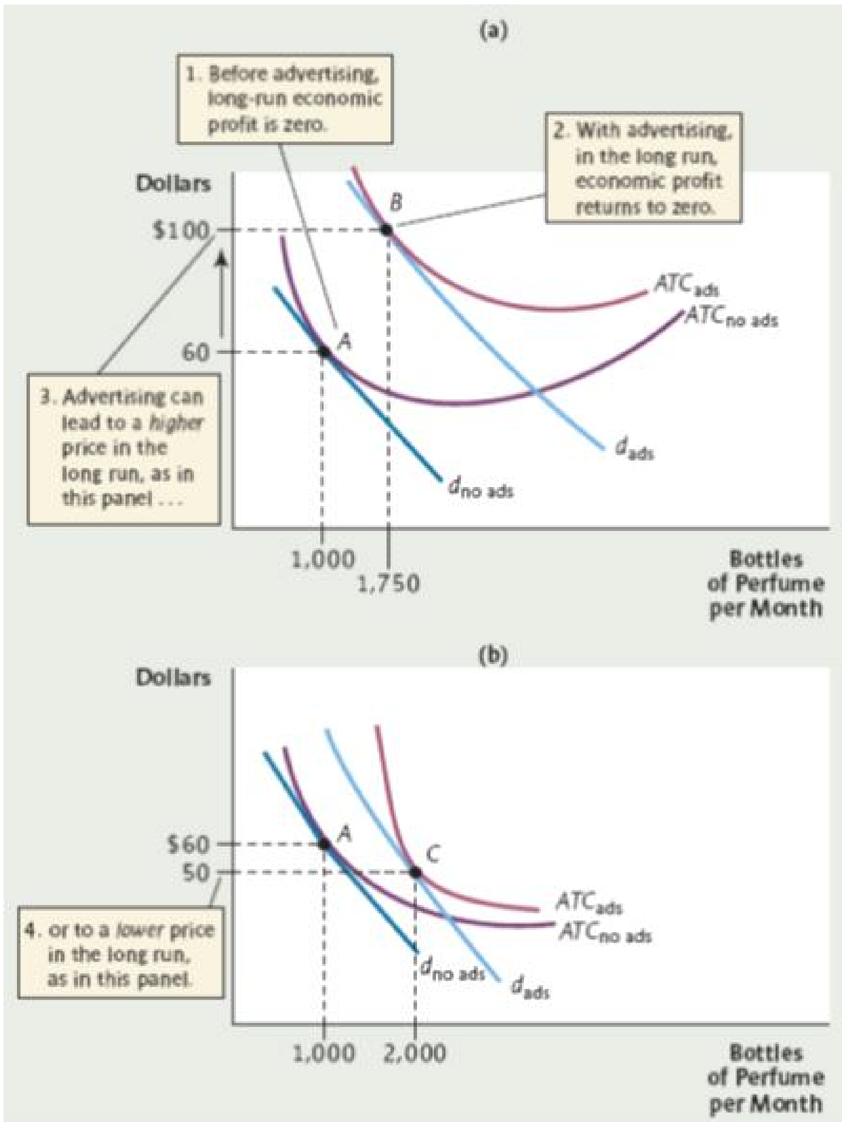
Figure 2(a) shows demand and *ATC* curves for a company that manufactures and sells perfume. Initially, there is no advertising at all in the industry. Firm is in long-run equilibrium at point *A,* in panel (a), where its demand curve and *ATC* curve touch, so economic profit is zero.

The firm charges \$60 per bottle and sells at the profit-maximizing output level of 1,000 bottles per month. This is the output level where its marginal revenue and marginal cost curves (not shown) intersect.

Now suppose that we introduce advertising into this market. Initially, the first few firms that discover advertising may have a temporary advantage over firms that don't advertise. But remember that in monopolistic competition any successful form of non-price competition will be automatically replicated by *all* firms (otherwise they would be at a competitive disadvantage). So let's skip over the temporary situation in which only some firms advertise, and examine our new long-run equilibrium when *all* firms advertise. In the long run, how will advertising change the situation of a typical monopolistic competitor in this market?

One change is that, with each firm paying additional costs for advertising, cost per unit will be greater at every output level. So the typical firm's *ATC* curve will shift upward. In panel (a), we show that firm's *ATC* curve shifts upward to *ATC*ads. Notice, however, that the upward shift is smaller at higher output levels, where the cost of any given ad campaign is spread over a larger number of units.

Figure 2: Advertising in Monopolistic Competition



In addition to the shift in *ATC*, we can expect that with all firms advertising, the demand for the product *in general* will increase. More people are aware of the product, or have had their appetites stimulated. And this, in turn, means that *each* firm should be able to sell more units at any given price than before:

The demand curve facing each firm shifts rightward. How much will the typical firm's demand curve shift? We know that, in the long run, the combination of a rightward shift in demand and an upward shift in *ATC* must eventually lead to a new equilibrium in which economic profit is zero. To see why, remember that if advertising creates economic profit in the short run, entry will occur, and every firm's demand curve will then shift leftward. If advertising creates economic loss in the short run, exit will occur, and the remaining firms' demand curves will shift rightward. In the end, long-run equilibrium (a situation of neither entry nor exit) requires that the typical firm earn zero economic profit. And in monopolistic competition, as you've learned, this can only occur when the demand curve touches but does not cross the *ATC* curve, with $P = ATC$ at the profit maximizing output level.

In panel (a), the new long-run equilibrium for our typical firm, occurs at point *B*. Firm sells 1,750 bottles of perfume and charges consumers a higher price ($100) than before. But because it has to pay for advertising, it is breaking even, just as it was in the initial long-run equilibrium without advertising. In panel (a), the impact of advertising is to *raise* prices for consumers. When consumers buy perfume, they are now paying for the advertising as well as all the inputs they paid for before. But you may be surprised that advertising can also have the opposite result: It can actually *lower* prices for consumers. Panel (b) illustrates this case. As before, we begin in a long-run equilibrium with no advertising in the market, and Firm operating at point *A*. When we introduce advertising to all firms, each firm (including Firm) sees its *ATC* curve shift upward, to *ATC*ads. But this time, when long-run equilibrium is restored with zero economic profit (point *C*), Firm is charging only $50—less than the initial $60. Advertising has brought down the price of perfume. How can this be? By advertising, each firm is able to produce and sell more output.

This remains true even when *all* firms advertise because total market demand has increased. Since the firm was originally on the downward-sloping portion of its *ATC* curve, we know that its *non*-advertising costs per unit will decline as output expands. If this decline is great enough—as in panel (b)—then costs per unit will drop, even when the cost of advertising is included.

Conclusion

Under monopolistic competition, advertising may increase the size of the market so that more units are sold. But in the long run, each firm earns zero economic profit, *just as it would if no firm were advertising. The price to the consumer may either rise or fall.*

Questions

1. How might advertising reduce economic wellbeing?
2. How might advertising increase economic well-being?
3. How might advertising with no apparent informational content in fact convey information to consumers?
4. Explain two benefits that might arise from the existence of brand names.