PART TWO

Microeconomics: Supply, Demand, and Product Markets

CHAPTER

Supply and Demand: Elasticity and Applications

4

You cannot teach a parrot to be an economist simply by teaching it to say "supply" and "demand."

Anonymous

We now move from our introductory survey to a detailed study of microeconomics—of the behavior of individual firms, consumers, and markets. Individual markets contain much of the grand sweep and drama of economic history and the controversies of economic policy. Within the confines of microeconomics we will study the reasons for the vast disparities in earnings between neurosurgeons and textile workers. Microeconomics is crucial to understanding why computer prices have fallen so rapidly and why the use of computers has expanded exponentially. We cannot hope to understand the bitter debates about health care or the minimum wage without applying the tools of supply and demand to these sectors. Even topics such as illegal drugs or crime and punishment are usefully illuminated by considering the way the demand for addictive substances differs from that for other commodities.

But understanding supply and demand requires more than simply parroting the words. A full mastery of microeconomic analysis means understanding the derivation of demand curves and supply curves, learning about different concepts of costs, and understanding how perfect competition differs from monopoly. All these and other key topics will be our subjects as we tour through the fascinating world of microeconomics.

A. PRICE ELASTICITY OF DEMAND AND SUPPLY

Supply and demand can often tell us whether certain forces increase or decrease quantities. But for these tools to be truly useful, we need to know *how much* supply and demand respond to changes in price. Some purchases, like those for vacation travel, are luxuries that are very sensitive to price changes. Others, like food or electricity, are necessities for which consumer quantities respond very little to price changes. The quantitative relationship between price and quantity purchased is analyzed using the crucial concept of *elasticity*. We begin with a careful definition of this term and then use this new concept to analyze the microeconomic impacts of taxes and other types of government intervention.

PRICE ELASTICITY OF DEMAND

Let's look first at the response of consumer demand to price changes:

The **price elasticity of demand** (sometimes simply called **price elasticity**) measures how much the quantity demanded of a good changes when its price

changes. The precise definition of price elasticity is the percentage change in quantity demanded divided by the percentage change in price.

Goods vary enormously in their price elasticity, or sensitivity to price changes. When the price elasticity of a good is high, we say that the good has "elastic" demand, which means that its quantity demanded responds greatly to price changes. When the price elasticity of a good is low, it is "inelastic" and its quantity demanded responds little to price changes.

Goods that have ready substitutes tend to have more elastic demand than those that have no substitutes. If all food or footwear prices were to rise 20 percent tomorrow, you would hardly expect people to stop eating or to go around barefoot, so food and footwear demands are price-inelastic. On the other hand, if mad-cow disease drives up the price of British beef, people can turn to beef from other countries or to lamb or poultry for their meat needs. Therefore, British beef shows a high price elasticity.

The length of time that people have to respond to price changes also plays a role. A good example is that of gasoline. Suppose you are driving across the country when the price of gasoline suddenly increases. Is it likely that you will sell your car and abandon your vacation? Not really. So in the short run, the demand for gasoline may be very inelastic.

In the long run, however, you can adjust your behavior to the higher price of gasoline. You can buy a smaller and more fuel-efficient car, ride a bicycle, take the train, move closer to work, or carpool with other people. The ability to adjust consumption patterns implies that demand elasticities are generally higher in the long run than in the short run.

The price elasticities of demand for individual goods are determined by the economic characteristics of demand. Price elasticities tend to be higher when the goods are luxuries, when substitutes are available, and when consumers have more time to adjust their behavior. By contrast, elasticities are lower for necessities, for goods with few substitutes, and for the short run.

Calculating Elasticities

The precise definition of price elasticity is the percentage change in quantity demanded divided by the percentage change in price. We use the symbol E_D to represent price elasticity, and for convenience we drop the minus signs, so elasticities are all positive.

We can calculate the coefficient of price elasticity numerically according to the following formula:

$$\text{Price elasticity of demand} = E_D = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in price}}$$

Now we can be more precise about the different categories of price elasticity:

- When a 1 percent change in price calls forth more than a 1 percent change in quantity demanded, the good has **price-elastic demand.** For example, if a 1 percent increase in price yields a 5 percent decrease in quantity demanded, the commodity has a highly price-elastic demand.
- When a 1 percent change in price produces less than a 1 percent change in quantity demanded, the good has **price-inelastic demand.** This case occurs, for instance, when a 1 percent increase in price yields only a 0.2 percent decrease in demand.
- One important special case is **unit-elastic demand,** which occurs when the percentage change in quantity is exactly the same as the percentage change in price. In this case, a 1 percent increase in price yields a 1 percent decrease in demand. We will see later that this condition implies that total expenditures on the commodity (which equal $P \times Q$) stay the same even when the price changes.

We illustrate the calculation of elasticities with the example shown in Figure 4-1 and Table 4-1. To begin at point *A*, quantity demanded was 240 units at a price of 90. A price increase to 110 led consumers to reduce their purchases to 160 units, shown as point *B*.

Table 4-1 shows how we calculate price elasticity. The price increase is 20 percent, with the resulting quantity decrease being 40 percent. The price elasticity of demand is evidently $E_D = 40/20 = 2$. The price elasticity is greater than 1, and this good therefore has price-elastic demand in the region from *A* to *B*.

In practice, calculating elasticities is somewhat tricky, and we emphasize three key steps where you have to be especially careful:

1. Recall that we drop the minus signs from the numbers, thereby treating all percentage changes as *positive*. That means all elasticities are written as positive numbers, even though prices and

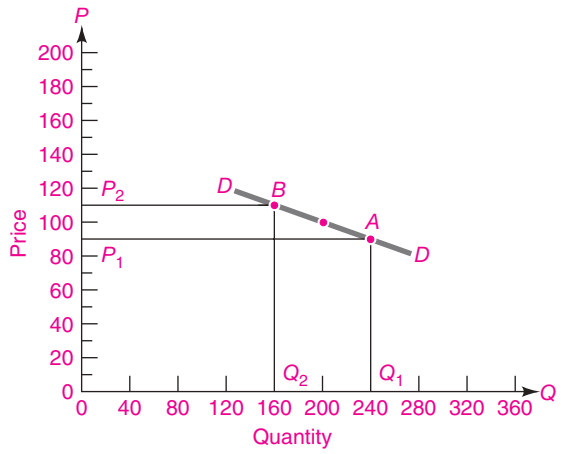


FIGURE 4-1. Elastic Demand Shows Large Quantity Response to Price Change

Market equilibrium is originally at point *A*. In response to a 20 percent price increase, quantity demanded declines 40 percent, to point *B*. Price elasticity is $E_D = 40/20 = 2$. Demand is therefore elastic in the region from *A* to *B*.

Case A: Price = 90 and quantity = 240

Case B: Price = 110 and quantity = 160

Percentage price change = $\Delta P/P = 20/100 = 20\%$

Percentage quantity change = $\Delta Q/Q = -80/200 = -40\%$

Price elasticity = $E_D = 40/20 = 2$

TABLE 4-1. Example of Good with Elastic Demand

Consider the situation where price is raised from 90 to 110. According to the demand curve, quantity demanded falls from 240 to 160. Price elasticity is the ratio of percentage change in quantity divided by percentage change in price. We drop the minus signs from the numbers so that all elasticities are positive.

quantities demanded move in opposite directions for downward-sloping demand curves.

2. Note that the definition of elasticity uses *percentage changes* in price and demand rather than absolute changes. This has the neat effect that a change in the units of measurement does not affect the elasticity. So whether we measure price in pennies or dollars, the price elasticity stays the same.

3. Note the use of *averaging* to calculate percentage changes in price and quantity. The formula for a percentage change is $\Delta P/P$. The value of ΔP in Table 4-1 is clearly $20 = 110 - 90$. But it's not immediately clear what value we should use for P in the denominator. Is it the original value of 90, the final value of 110, or something in between?

   For very small percentage changes, such as from 100 to 99, it does not much matter whether we use 99 or 100 as the denominator. But for larger changes, the difference is significant. To avoid ambiguity, we will take the average price to be the base price for calculating price changes. In Table 4-1, we used the average of the two prices [$P = (90 + 110)/2 = 100$] as the base or denominator in the elasticity formula. Similarly, we used the average quantity [$Q = (160 + 240)/2 = 200$] as the base for measuring the percentage change in quantity. The exact formula for calculating elasticity is therefore

$$E_D = \frac{\Delta Q}{(Q_1 + Q_2)/2} \div \frac{\Delta P}{(P_1 + P_2)/2}$$

   where P_1 and Q_1 represent the original price and quantity and P_2 and Q_2 stand for the new price and quantity.

Price Elasticity in Diagrams

It's possible to determine price elasticities in diagrams as well. Figure 4-2 illustrates the three cases of elasticities. In each case, price is cut in half and consumers change their quantity demanded from *A* to *B*.

In Figure 4-2(*a*), a halving of price has tripled quantity demanded. Like the example in Figure 4-1, this case shows price-elastic demand. In Figure 4-2(*c*), cutting price in half led to only a 50 percent increase in quantity demanded, so this is the case of price-inelastic demand. The borderline case of unit-elastic demand is shown in Figure 4-2(*b*); in this example, the doubling of quantity demanded exactly matches the halving of price.

Figure 4-3 displays the important polar extremes where the price elasticities are infinite and zero, or completely elastic and completely inelastic. Completely inelastic demands, or ones with zero elasticity, are ones where the quantity demanded responds not at all to price changes; such demand is seen to be a vertical demand curve. By contrast, when demand is infinitely elastic, a tiny change in price will lead to an

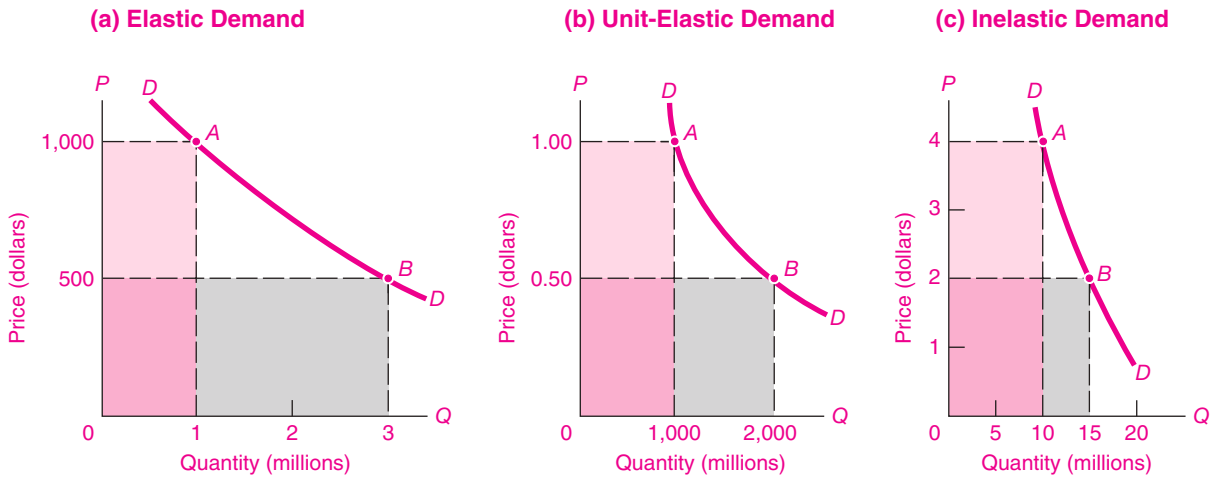


FIGURE 4-2. Price Elasticity of Demand Falls into Three Categories

indefinitely large change in quantity demanded, as in the horizontal demand curve in Figure 4-3.

A Shortcut for Calculating Elasticities

There is a simple rule for calculating the price elasticity of a demand curve:

> The elasticity of a straight line at a point is given by the ratio of the length of the line segment below the point to the length of the line segment above the point.

The procedure is shown in Figure 4-4. At the top of the line, a very small percentage price change induces a very large percentage quantity change, and the elasticity is therefore extremely large. Price

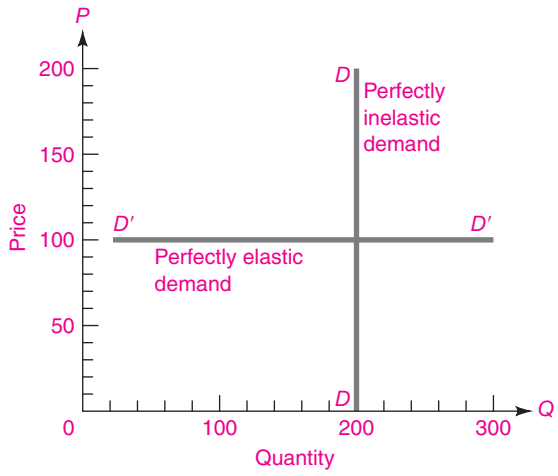


FIGURE 4-3. Perfectly Elastic and Inelastic Demands

Polar extremes of demand are vertical demand curves, which represent perfectly inelastic demand ($E_D = 0$), and horizontal demand curves, which show perfectly elastic demand ($E_D = \infty$).

Elasticity of Straight Line

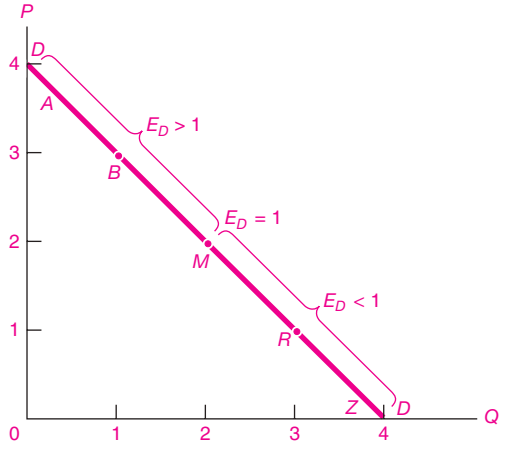


FIGURE 4-4. A Simple Rule for Calculating the Demand Elasticity

We can calculate the elasticity as the ratio of the lower segment to the upper segment at the demand point. For example, at point *B*, the lower segment is 3 times as long as the upper segment, so the price elasticity is 3.

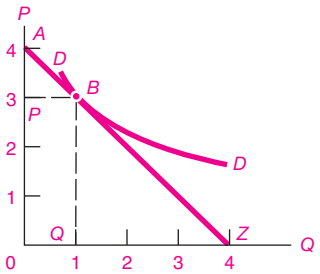


FIGURE 4-5. Calculating the Demand Elasticity for Curved Demand

To calculate the demand elasticity for a nonlinear demand curve, first draw a tangent line at the point. Then take the ratio of the length of the straight-line segment below the point to the length of the line segment above the point. Hence, at point *B* the elasticity can be calculated to be 3.

elasticity is relatively large when we are high up the linear *DD* curve. We use the rule to calculate the elasticity at point *B* in Figure 4-4. Calculate the ratio of the line segment *BZ* to the segment *AB*. Looking at the axes, we see that the ratio is 3. Therefore, price elasticity at point *B* is 3.

A similar calculation at point *R* shows that demand at that point is inelastic, with an elasticity of ⅓.

Finally, calculate elasticity at point *M*. Here, the ratio of the two line segments is one, so demand is unit-elastic at the midpoint *M*.

We can also use the rule to calculate the elasticity of a curved demand curve, as shown in Figure 4-5. For this case, you begin by drawing a line that is tangent to the point, and you then calculate the ratio of segments for the tangent line. This will provide the correct calculation of elasticity for the curved line. Use as an example point *B* in Figure 4-5. We have drawn a tangent straight line. A careful inspection will show that the ratio of the lower to upper segments of the straight line is 3. Therefore, the curved demand has an elasticity of 3 at point *B*.

The Algebra of Elasticities

For the mathematically inclined, we can show the algebra of elasticities for straight-line (linear) demand curves. We begin with a demand curve, which is written as $Q = a - bP$. The demand elasticity at point (P_0, Q_0) is defined as $E_D = (\%\,\Delta Q)/(\%\,\Delta P) = (\Delta Q/Q_0)/(\Delta P/P_0) = (\Delta Q/\Delta P)(P_0/Q_0)$. This implies that the elasticity at point (P_0, Q_0) is

$$E_D = b(P_0/Q_0)$$

Note that the elasticity depends upon the slope of the demand curve, but it also depends upon the specific price and quantity pair. Question 11 at the end of this chapter provides examples that allow you to apply this formula.

Elasticity Is Not the Same as Slope

We must always remember not to confuse the elasticity of a curve with its slope. This distinction is easily seen when we examine the straight-line demand curves that are often found in illustrative examples.

What is the price elasticity of a straight-line demand curve? Surprisingly, along a straight-line demand curve, the price elasticity varies from zero to infinity! Table 4-2 gives a detailed set of elasticity calculations using the same technique as that in Table 4-1. This table shows that linear demand curves start out with high price elasticity, where price is high and quantity is low, and end up with low elasticity, where price is low and quantity is high.

This illustrates an important point. When you see a demand curve in a diagram, it is not true that a steep slope for the demand curve means inelastic demand or that a flat slope signifies elastic demand. The slope is not the same as the elasticity because the demand curve's slope depends upon the *changes* in *P* and *Q*, whereas the elasticity depends upon the *percentage changes* in *P* and *Q*. The only exceptions are the polar cases of completely elastic and inelastic demands.

We also illustrate the point in Figure 4-4. This straight-line demand curve has elastic demand in the top region and inelastic demand in the bottom region.

Finally, look at Figure 4-2(*b*). This demand curve is clearly not a straight line with constant slope. Yet it has a constant demand elasticity of $E_D = 1$ because the percentage change in price is equal everywhere to the percentage change in quantity.

Elasticities cannot be inferred by slope alone. The general rule for elasticities is that the elasticity can be calculated as the ratio of the length of the straight-line or tangent segment below the demand point to the length of the segment above the point.

Numerical Calculation of Elasticity Coefficient

Q	ΔQ	P	ΔP	$\frac{Q_1 + Q_2}{2}$	$\frac{P_1 + P_2}{2}$	$E_D = \frac{\Delta Q}{(Q_1 + Q_2)/2} \div \frac{\Delta P}{(P_1 + P_2)/2}$	
0		6					
	10		2	5	5	$\frac{10}{5} \div \frac{2}{5} = 5$	(elastic)
10		4					
	10		2	15	3	$\frac{10}{15} \div \frac{2}{3} = 1$	(unit-elastic)
20		2					
	10		2	25	1	$\frac{10}{25} \div \frac{2}{1} = 0.2$	(inelastic)
30		0					

TABLE 4-2. Calculation of Price Elasticity along a Linear Demand Curve

ΔP denotes the change in price, i.e., $\Delta P = P_2 - P_1$, while $\Delta Q = Q_2 - Q_1$. To calculate numerical elasticity, the percentage change of price equals price change, ΔP, divided by average price $[(P_1 + P_2)/2]$; the percentage change in output is calculated as ΔQ divided by average quantity, $[(Q_1 + Q_2)/2]$. Treating all figures as positive numbers, the resulting ratio gives numerical price elasticity of demand, E_D. Note that for a straight line, elasticity is high at the top, low at the bottom, and exactly 1 in the middle.

ELASTICITY AND REVENUE

Many businesses want to know whether raising prices will raise or lower revenues. This question is of strategic importance for businesses like airlines, baseball teams, and magazines, which must decide whether it is worthwhile to raise prices and whether the higher prices make up for lower demand. Let's look at the relationship between price elasticity and total revenue.

Total revenue is by definition equal to price times quantity (or $P \times Q$). If consumers buy 5 units at \$3 each, total revenue is \$15. If you know the price elasticity of demand, you know what will happen to total revenue when price changes:

1. When demand is price-inelastic, a price decrease reduces total revenue.
2. When demand is price-elastic, a price decrease increases total revenue.
3. In the borderline case of unit-elastic demand, a price decrease leads to no change in total revenue.

The concept of price elasticity is widely used today as businesses attempt to separate customers into groups with different elasticities. This technique has been extensively pioneered by the airlines (see the box that follows). Another example is software companies, which have a wide range of different prices for their products in an attempt to exploit different elasticities. For example, if you are desperate about buying a new operating system immediately, your elasticity is low and the seller will profit from charging you a relatively high price. On the other hand, if you are not in a hurry for an upgrade, you can search around for the best price and your elasticity is high. In this case, the seller will try to find a way to make the sale by charging a relatively low price.

Fly the Financial Skies of "Elasticity Air"

Understanding demand elasticities is worth billions of dollars each year to U.S. airlines. Ideally, airlines would like to charge a relatively high price to business travelers, while charging leisure

passengers a low-enough price to fill up all their empty seats. That is a strategy for raising revenues and maximizing profits.

But if they charge low-elasticity business travelers one price and high-elasticity leisure passengers a lower price, the airlines have a big problem—keeping the two classes of passengers separate. How can they stop the low-elasticity business travelers from buying up the cheap tickets meant for the leisure travelers and not let high-elasticity leisure flyers take up seats that business passengers would have been willing to buy?

The airlines have solved their problem by engaging in "price discrimination" among their different customers in a way that exploits different price elasticities. **Price discrimination** is the practice of charging different prices for the same service to different customers. Airlines offer discount fares for travelers who plan ahead and who tend to stay longer. One way of separating the two groups is to offer discounted fares to people who stay over a Saturday night—a rule that discourages business travelers who want to get home for the weekend. Also, discounts are often unavailable at the last minute because many business trips are unplanned expeditions to handle an unforeseen crisis—another case of price-inelastic demand. Airlines have devised extremely sophisticated computer programs to manage their seat availability as a way of ensuring that their low-elasticity passengers cannot benefit from discount fares.

The Paradox of the Bumper Harvest

We can use elasticities to illustrate one of the most famous paradoxes of all economics: the paradox of the bumper harvest. Imagine that in a particular year nature smiles on farming. A cold winter kills off the pests; spring comes early for planting; there are no killing frosts; rains nurture the growing shoots; and a sunny October allows a record crop to come to market. At the end of the year, family Jones happily settles down to calculate its income for the year. The Joneses are in for a major surprise: *The good weather and bumper crop have lowered their and other farmers' incomes.*

How can this be? The answer lies in the elasticity of demand for foodstuffs. The demands for basic food products such as wheat and corn tend to be inelastic; for these necessities, consumption changes very little in response to price. But this means farmers as a whole receive less total revenue when the harvest is good than when it is bad. The increase in supply arising from an abundant harvest tends to lower the price. But the lower price doesn't increase quantity demanded very much. The implication is that a low price elasticity of food means that large harvests (high Q) tend to be associated with low revenue (low $P \times Q$).

These ideas can be illustrated by referring back to Figure 4-2. We begin by showing how to measure revenue in the diagram itself. Total revenue is the product of price times quantity, $P \times Q$. Further, the area of a rectangle is always equal to the product of its base times its height. Therefore, total revenue at any point on a demand curve can be found by examining the area of the rectangle determined by the P and Q at that point.

Next, we can check the relationship between elasticity and revenue for the unit-elastic case in Figure 4-2(b). Note that the shaded revenue region ($P \times Q$) is \$1000 million for both points A and B. The shaded areas representing total revenue are the same because of offsetting changes in the Q base and the P height. This is what we would expect for the borderline case of unit-elastic demand.

We can also see that Figure 4-2(a) corresponds to elastic demand. In this figure, the revenue rectangle expands from \$1000 million to \$1500 million when price is halved. Since total revenue goes up when price is cut, demand is elastic.

In Figure 4-2(c) the revenue rectangle falls from \$40 million to \$30 million when price is halved, so demand is inelastic.

Which diagram illustrates the case of agriculture, where a bumper harvest means lower total revenues for farmers? Clearly it is Figure 4-2(c). Which represents the case of vacation travel, where a lower price could mean higher revenues? Surely Figure 4-2(a).

Table 4-3 shows the major points to remember about price elasticities.

Cigarette Taxes and Smoking

What is the impact of cigarette taxes on smoking? Some people say, "Cigarettes are so addictive that people will pay anything for their daily habit." Implicitly, when you say that the quantity demanded does not respond to price, you are saying

Value of demand elasticity	Description	Definition	Impact on revenues
Greater than one ($E_D > 1$)	Elastic demand	Percentage change in quantity demanded *greater* than percentage change in price	Revenues *increase* when price decreases
Equal to one ($E_D = 1$)	Unit-elastic demand	Percentage change in quantity demanded *equal* to percentage change in price	Revenues *unchanged* when price decreases
Less than one ($E_D < 1$)	Inelastic demand	Percentage change in quantity demanded *less* than percentage change in price	Revenues *decrease* when price decreases

TABLE 4-3. Elasticities: Summary of Crucial Concepts

that the price elasticity is zero. What does the evidence say about the price elasticity of cigarette consumption?

We can use a historical example to illustrate the issue. New Jersey doubled its cigarette tax from 40 cents to 80 cents per pack. The tax increased the average price of cigarettes from $2.40 to $2.80 per pack. Economists estimate that the effect of the price increase alone was a decrease in New Jersey's cigarette consumption from 52 million to 47.5 million packs.

Using the elasticity formula, you can calculate that the short-run price elasticity is 0.59. (Make sure you can get the same number.) Similar estimates come from more detailed statistical studies. The evidence indicates that the price elasticity of cigarettes is definitely not zero.

PRICE ELASTICITY OF SUPPLY

Of course, consumption is not the only thing that changes when prices go up or down. Businesses also respond to price in their decisions about how much to produce. Economists define the price elasticity of supply as the responsiveness of the quantity supplied of a good to its market price.

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

As with demand elasticities, there are polar extremes of high and low elasticities of supply. Suppose the amount supplied is completely fixed, as in the case of perishable fish brought to market to be sold at whatever price they will fetch. This is the limiting case of zero elasticity, or completely inelastic supply, which is a vertical supply curve.

At the other extreme, say that a tiny cut in price will cause the amount supplied to fall to zero, while the slightest rise in price will coax out an indefinitely large supply. Here, the ratio of the percentage change in quantity supplied to percentage change in price is extremely large and gives rise to a horizontal supply curve. This is the polar case of infinitely elastic supply.

Between these extremes, we call supply elastic or inelastic depending upon whether the percentage change in quantity is larger or smaller than the percentage change in price. In the borderline unit-elastic case, where price elasticity of supply equals 1, the percentage increase of quantity supplied is exactly equal to the percentage increase in price.

You can readily see that the definitions of price elasticities of supply are exactly the same as those for price elasticities of demand. The only difference is that for supply the quantity response to price is positive, while for demand the response is negative.

The exact definition of the price elasticity of supply, E_S, is as follows:

$$E_S = \frac{\text{percentage change in quantity supplied}}{\text{percentage change in price}}$$

Figure 4-6 displays three important cases of supply elasticity: (a) the vertical supply curve, showing completely inelastic supply; (c), the horizontal supply curve, displaying completely elastic supply; and (b), an intermediate case of a straight line, going

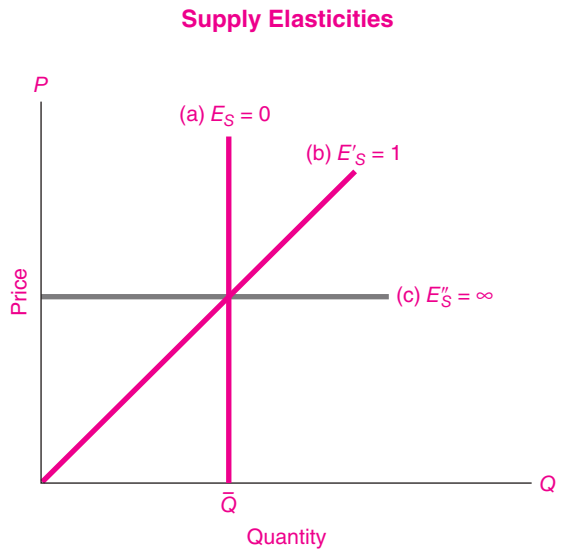


FIGURE 4-6. Supply Elasticity Depends upon Producer Response to Price

When supply is fixed, supply elasticity is zero, as in curve (a). Curve (c) displays an indefinitely large quantity response to price changes. Intermediate case (b) arises when the percentage quantity and price changes are equal.

through the origin, illustrating the borderline case of unit elasticity.[1]

What factors determine supply elasticity? The major factor influencing supply elasticity is the ease with which production in the industry can be increased. If all inputs can be readily found at going market prices, as is the case for the textile industry, then output can be greatly increased with little increase in price. This would indicate that supply elasticity is relatively large. On the other hand, if production capacity is severely limited, as is the case for gold mining, then even sharp increases in the price of gold will call forth but a small response in gold production; this would be inelastic supply.

Another important factor in supply elasticities is the time period under consideration. A given change in price tends to have a larger effect on amount supplied as the time for suppliers to respond increases. For very brief periods after a price increase, firms may be unable to increase their inputs of labor, materials, and capital, so supply may be very price-inelastic. However, as time passes and businesses can hire more labor, build new factories, and expand capacity, supply elasticities will become larger.

We can use Figure 4-6 to illustrate how supply may change over time for the fishing case. Supply curve (a) might hold for fish on the day they are brought to market, where they are simply auctioned off for whatever they will bring. Curve (b) might hold for the intermediate run of a year or so, with the given stock of fishing boats and before new labor is attracted to the industry. Over the very long run, as new fishing boats are built, new labor is attracted, and new fish farms are constructed, the supply of fish might be very price-elastic, as in case (c) in Figure 4-6.

[1] You can determine the elasticity of a supply curve that is not a straight line as follows: (*a*) Draw the straight line that lies tangent to the curve at a point, and (*b*) then measure the elasticity of that tangential straight line.

B. APPLICATIONS TO MAJOR ECONOMIC ISSUES

Having laid the groundwork with our study of elasticities, we now show how these tools can assist our understanding of many of the basic economic trends and policy issues. We begin with one of the major transformations since the Industrial Revolution, the decline of agriculture. Next, we examine the implications of taxes on an industry, using the example of a gasoline tax. We then analyze the consequences of various types of government intervention in markets.

THE ECONOMICS OF AGRICULTURE

Our first application of supply-and-demand analysis comes from agriculture. The first part of this section lays out some of the economic fundamentals of the farm sector. Then we will use the theory of supply and demand to study the effects of government intervention in agricultural markets.

Long-Run Relative Decline of Farming

Farming was once our largest single industry. A hundred years ago, half the American population lived and worked on farms, but that number has declined to less

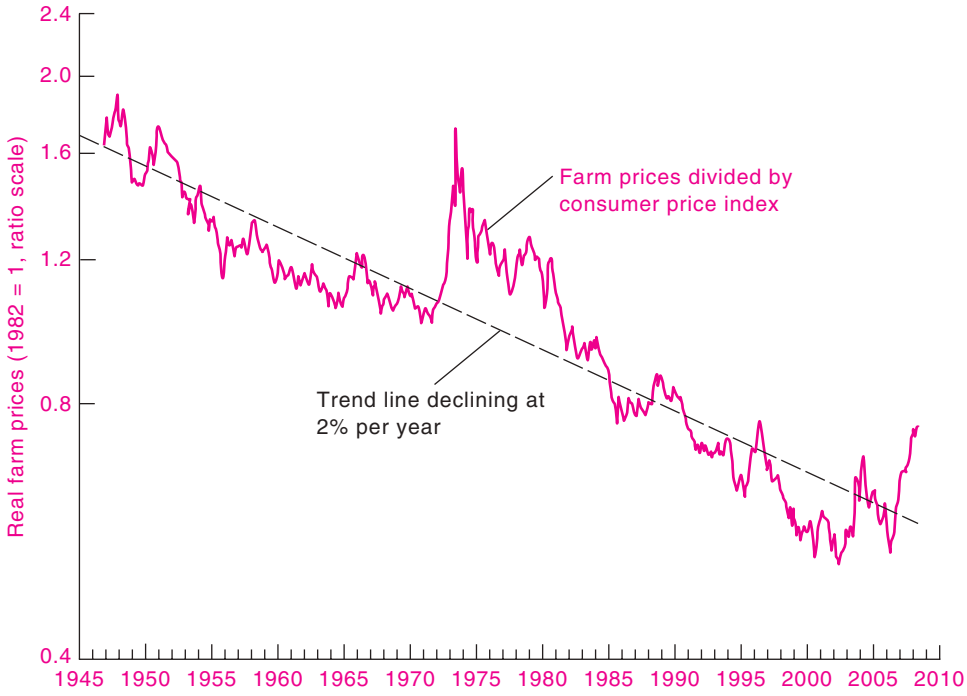


FIGURE 4-7. Prices of Basic Farm Products Have Declined Sharply

One of the major forces affecting the U.S. economy has been the decline in the relative prices of basic farm products—wheat, corn, soybeans, and the like. Over the past decades, farm prices have declined 2 percent per year relative to the general price level. The grain shortages since 2005 have slowed but not reversed the long slide in relative food prices. However, the recent upturn in food prices has contributed to inflation in most countries, and even to food riots in poor countries.

Source: Bureau of Labor Statistics.

than 3 percent of the workforce today. At the same time, prices for farm products have fallen relative to incomes and other prices in the economy. Figure 4-7 shows the steady decline of farm prices over the last half-century. While median family income has more than doubled, farm incomes have stagnated. Farm-state senators fret about the decline of the family farm.

A single diagram can explain the cause of the sagging trend in farm prices better than libraries of books and editorials. Figure 4-8 shows an initial equilibrium with high prices at point *E*. Observe what happens to agriculture as the years go by. Demand for food increases slowly because basic foods are necessities; the demand shift is consequently modest in comparison to growing average incomes.

What about supply? Although many people mistakenly think that farming is a backward business, statistical studies show that productivity (output per unit of input) has grown more rapidly in agriculture than in most other industries. Important advances include mechanization through tractors, combines, and cotton pickers; fertilization and irrigation; selective breeding; and development of genetically modified crops. All these innovations have vastly increased the productivity of agricultural inputs. Rapid productivity growth has increased supply greatly, as shown by the supply curve's shift from *SS* to *S′S′* in Figure 4-8.

What must happen at the new competitive equilibrium? Sharp increases in supply outpaced modest increases in demand, producing a downward trend in farm prices relative to other prices in the economy. And this is precisely what has happened in recent decades, as is seen in Figure 4-7.

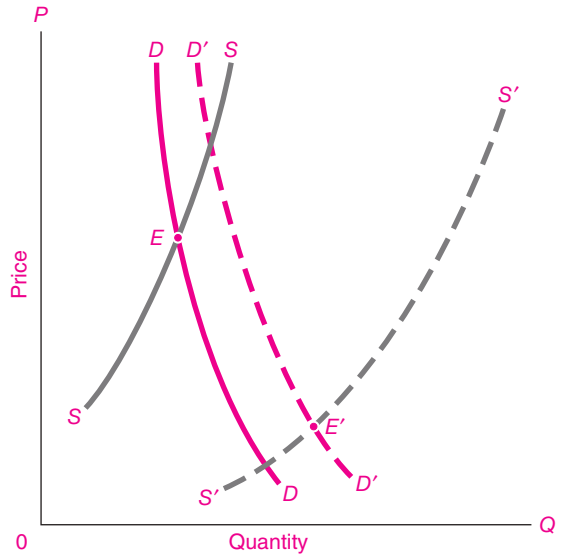


FIGURE 4-8. Agricultural Distress Results from Expanding Supply and Price-Inelastic Demand

Equilibrium at *E* represents conditions in the farm sector decades ago. Demand for farm products tends to grow more slowly than the impressive increase in supply generated by technological progress. Hence, competitive farm prices tend to fall. Moreover, with price-inelastic demand, farm incomes decline with increases in supply.

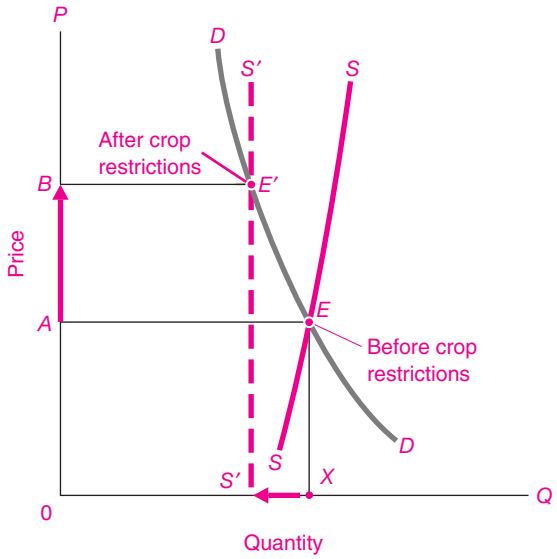


FIGURE 4-9. Crop-Restriction Programs Raise Both Price and Farm Income

Before the crop restriction, the competitive market produces an equilibrium with low price at *E*. When government restricts production, the supply curve is shifted leftward to *S′S′*, moving the equilibrium to *E′* and raising price to *B*. Confirm that new revenue rectangle 0*BE′S′* is larger than original revenue rectangle 0*AEX*—higher revenue being the result of inelastic demand.

Crop Restrictions. In response to falling incomes, farmers have often lobbied the federal government for economic assistance. Over the years, governments at home and abroad have taken many steps to help farmers. They have raised prices through price supports; they have curbed imports through tariffs and quotas; and they sometimes simply sent checks to farmers who agreed *not* to produce on their land.

How can *reducing production* actually *help* farmers? We can use the paradox of the bumper harvest to explain this result. Suppose the government requires every farmer to reduce production. As Figure 4-9 shows, this has the effect of shifting the supply curve up and to the left. Because the demand for food is inelastic, crop restrictions not only raise the price of crops but also tend to raise farmers' total revenues. Just as bumper harvests hurt farmers, crop restrictions raise farm incomes. Of course, consumers are hurt by the crop restrictions and higher prices—just as they would be if a flood or drought created a scarcity of food.

Restrictions on production are a typical example of government interference in individual markets. They often raise the income of one group at the expense of consumers. These policies are generally inefficient: the gain to farmers is less than the harm to consumers.

IMPACT OF A TAX ON PRICE AND QUANTITY

Governments tax a wide variety of commodities—cigarettes, alcohol, imported goods, telephone services, and so on. We are often interested in determining who actually bears the burden of the tax, and here is where supply and demand are essential.

Take the example of gasoline taxes. In 2008, the average tax on gasoline in the United States was around 50 cents per gallon. Many economists and environmentalists advocate much higher gasoline

taxes for the United States. They point out that higher taxes would curb consumption, and thereby reduce global warming as well as lower our dependence on insecure foreign sources of oil. Some advocate raising gasoline taxes by $1 or $2 per gallon. What would be the impact of such a change?

For concreteness, suppose that the government decides to discourage oil consumption by levying a gasoline tax of $2 per gallon. Prudent legislators would of course be reluctant to raise gasoline taxes so sharply without a firm understanding of the consequences of such a move. They would want to know the incidence of the tax. *By* ***incidence*** *we mean the ultimate economic effect of a tax on the real incomes of producers and consumers.* Just because oil companies write a check for the taxes does not mean that the taxes in fact reduce their profits. By using supply and demand, we can analyze the exact incidence of the tax.

It could be that the burden of the tax is shifted forward to the consumers, which would occur if the retail price of gasoline goes up by the full $2 of the tax. Or perhaps consumers cut back so sharply on gasoline purchases that the burden of the tax is shifted back completely onto the oil companies. Where the actual impact lies between these extremes can be determined only from supply-and-demand analysis.

Figure 4-10 provides the answer. It shows the original pretax equilibrium at *E,* the intersection of the original *SS* and *DD* curves, at a gasoline price of $2 a gallon and total consumption of 100 billion gallons per year. We portray the imposition of a $2 tax in the retail market for gasoline as an upward shift of the supply curve, with the demand curve remaining unchanged. The demand curve does not shift because the quantity demanded at each retail price is unchanged by the gasoline-tax increase. Note that the demand curve for gasoline is relatively inelastic.

By contrast, the supply curve definitely does shift upward by $2. The reason is that producers are willing to sell a given quantity (say, 100 billion gallons) only if they receive the same *net* price as before. That is, at each quantity supplied, the market price must rise by exactly the amount of the tax. If producers had originally been willing to sell 80 billion gallons at $1.80 per gallon, they would still be willing to sell the same amount at a retail price of $3.80 (which, after subtracting the tax, yields the producers the same $1.80 per gallon).

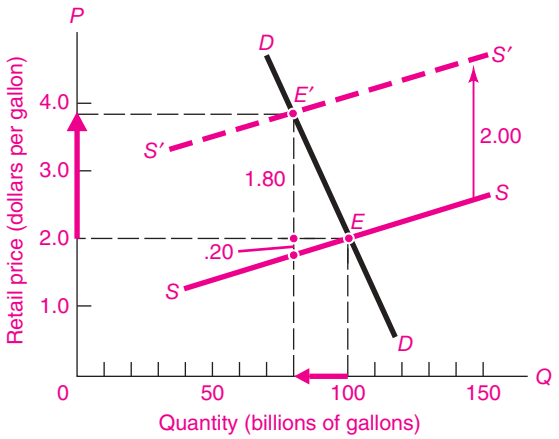


FIGURE 4-10. Gasoline Tax Falls on Both Consumer and Producer

What is the incidence of a tax? A $2 tax on gasoline shifts the supply curve up $2 everywhere, giving a new supply curve, *S′S′*, parallel to the original supply curve, *SS*. This new supply curve intersects *DD* at the new equilibrium, *E′*, where the price to consumers has risen 180 cents and the producers' price has fallen 20 cents. The green arrows show changes in *P* and *Q*. Note that consumers bear most of the burden of the tax.

What is the new equilibrium price? The answer is found at the intersection of the new supply and demand curves at *E′*, where *S′S′* and *DD* meet. Because of the supply shift, the retail price is higher. Also, the quantity supplied and demanded is reduced. If we read the graph carefully, we find that the new equilibrium price has risen from $2 to about $3.80. The new equilibrium output, at which supply and demand are in equilibrium, has fallen from 100 billion to about 80 billion gallons.

Who ultimately pays the tax? What is its incidence? Clearly the oil industry pays a small fraction, for it receives only $1.80 ($3.80 less the $2 tax) rather than $2. But the consumer bears most of the burden, with the retail price rising $1.80, because supply is relatively price-elastic whereas demand is relatively price-inelastic.

Subsidies. If taxes are used to discourage consumption of a commodity, subsidies are used to encourage

production. One pervasive example of subsidies comes in agriculture. You can examine the impact of a subsidy in a market by shifting *down* the supply curve. The general rules for subsidies are exactly parallel to those for taxes.

General Rules on Tax Shifting. Gasoline is just a single example of how to analyze tax shifting. Using this apparatus, we can understand how cigarette taxes affect both the prices and the consumption of cigarettes; how taxes or tariffs on imports affect foreign trade; and how property taxes, social security taxes, and corporate-profit taxes affect land prices, wages, and interest rates.

The key issue in determining the incidence of a tax is the relative elasticities of supply and demand. If demand is inelastic relative to supply, as in the case of gasoline, most of the cost is shifted to consumers. By contrast, if supply is inelastic relative to demand, as is the case for land, then most of the tax is shifted to the suppliers. Here is the general rule for determining the incidence of a tax:

The incidence of a tax denotes the impact of the tax on the incomes of producers and consumers. In general, the incidence depends upon the relative elasticities of demand and supply. (1) A tax is shifted *forward* to consumers if the *demand is inelastic* relative to supply. (2) A tax is shifted *backward* to producers if *supply is inelastic* relative to demand.

MINIMUM FLOORS AND MAXIMUM CEILINGS

Sometimes, rather than taxing or subsidizing a commodity, the government legislates maximum or minimum prices. History is full of examples. From biblical days, governments have limited the interest rates that lenders can charge (so-called usury laws). In wartime, governments often impose wage and price controls to prevent spiraling inflation. During the energy crisis of the 1970s, there were controls on gasoline prices. A few large cities, including New York, have rent controls on apartments.[2] Today, there are increasingly stringent limitations on the prices that doctors or hospitals can charge under federal health programs such as Medicare. Sometimes there are price floors, as in the case of the minimum wage.

These kinds of interferences with the laws of supply and demand are genuinely different from those in which the government imposes a tax and then lets the market act through supply and demand. Although political pressures always exist to keep prices down and wages up, experience has taught that sector-by-sector price and wage controls tend to create major economic distortions. Nevertheless, as Adam Smith well knew when he protested against mercantilist policies of an earlier age, most economic systems are plagued by inefficiencies stemming from well-meaning but inexpert interferences with the mechanisms of supply and demand. Setting maximum or minimum prices in a market tends to produce surprising and sometimes perverse economic effects. Let's see why.

Two important examples of government intervention are the minimum wage and price controls on gasoline. These will illustrate the surprising side effects that can arise when governments interfere with market determination of price and quantity.

The Minimum-Wage Controversy

The minimum wage sets a minimum hourly rate that employers are allowed to pay workers. In the United States, the federal minimum wage began in 1938 when the government required that covered workers be paid at least 25 cents an hour. By 1947, the minimum wage was fully 65 percent of the average rate paid to manufacturing workers (see Figure 4-11). The most recent law increased the minimum wage to $7.25 per hour in 2009.

This is an issue that divides even the most eminent economists. For example, Nobel laureate Gary Becker stated flatly, "Hike the minimum wage, and you put people out of work." Another group of Nobel Prize winners countered, "We believe that the federal minimum wage can be increased by a moderate amount without significantly jeopardizing employment opportunities."

How can nonspecialists sort through the issues when the experts are so divided? How can we resolve these apparently contradictory statements? To begin with, we should recognize that statements on the

[2] See question 9 at the end of this chapter for an examination of rent controls.

FIGURE 4-11. The Minimum Wage and Teenage Unemployment, 1947–2009

The green line shows the level of the minimum wage relative to average hourly earnings in manufacturing. Note how the minimum wage declined slowly relative to other wages over the last half-century. Additionally, the blue line shows the ratio of teenage unemployment to overall unemployment. Do you see any relationship between the two lines? What does this tell you about the minimum-wage controversy?

Source: Data are from the U.S. Department of Labor. Background on the minimum wage can be found at the Labor Department's website at *www.dol.gov/esa/minwage/q-a.htm.*

desirability of raising the minimum wage contain personal value judgments. Such statements might be informed by the best positive economics and still make different recommendations on important policy issues.

A cool-headed analysis indicates that the minimum-wage debate centers primarily on issues of interpretation rather than fundamental disagreements on empirical findings. Begin by looking at Figure 4-12, which depicts the market for unskilled workers. The figure shows how a minimum wage rate sets a floor for most jobs. As the minimum wage rises above the market-clearing equilibrium at *M,* the total number of jobs moves up the demand curve to *E,* so employment falls. The gap between labor supplied and labor demanded is shown as *U.* This represents the amount of unemployment.

Using supply and demand, we see that there is likely to be a rise in unemployment and a decrease in employment of low-skilled workers. But how large will these magnitudes be? And what will be the impact on the wage income of low-income workers? On these questions, we can look at the empirical evidence.

Most studies indicate that a 10 percent increase in the minimum wage would reduce employment of teenagers by between 1 and 3 percent. The impact on adult employment is even smaller. Some recent studies put the adult employment effects very close to zero, and one set of studies suggests that employment might even increase. So a careful reading of the

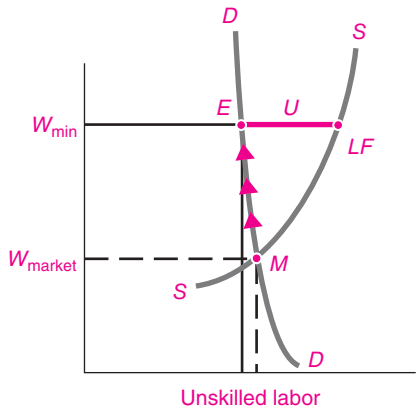


FIGURE 4-12. Effects of a Minimum Wage

Setting the minimum-wage floor at W_{min}, high above the free-market equilibrium rate at W_{market}, results in employment at *E*. Employment is reduced, as the arrows show, from *M* to *E*. Additionally, unemployment is *U*, which is the difference between labor supplied at *LF* and employment at *E*. If the demand curve is inelastic, increasing the minimum wage will increase the income of low-wage workers. To see this, shade in the rectangle of total wages before and after the minimum-wage increase.

quotations from the eminent economists indicates that some economists consider small to be "insignificant" while others emphasize the existence of at least some job losses. Our example in Figure 4-12 shows a case where the *employment* decline (shown as the difference between *M* and *E*) is very small while the *unemployment* caused by the minimum wage (shown by the *U* line) is relatively large.

Figure 4-11 on page 78 shows the history of the minimum wage and teenage unemployment over the last half-century. With the declining power of the labor movement, the ratio of the minimum wage to the manufacturing wage declined from two-thirds in 1947 to around one-third in 2008. There was a slight upward trend in the relative unemployment rate of teenagers over this period. It is worth examining the pattern of changes to see whether you can detect an impact of the minimum wage on teenage unemployment.

Another factor in the debate relates to the impact of the minimum wage on incomes. Virtually every study concludes that the demand for low-wage workers is price-inelastic. The results we just cited indicate that the price elasticity is between 0.1 and 0.3. Given the elasticities just cited, a 10 percent increase in the minimum wage will increase the incomes of the affected groups by 7 to 9 percent. Figure 4-12 shows how the *incomes* of low-income workers rise despite the decline in their *total employment*. This can be seen by comparing the income rectangles under the equilibrium points *E* and *M*. (See question 8*e* at the end of this chapter.)

The impact on incomes is yet another reason why people may disagree about the minimum wage. Those who are particularly concerned about the welfare of low-income groups may feel that modest inefficiencies are a small price to pay for higher incomes. Others—who worry more about the cumulative costs of market interferences or about the impact of higher costs upon prices, profits, and international competitiveness—may hold that the inefficiencies are too high a price. Still others might believe that the minimum wage is an inefficient way to transfer buying power to low-income groups; they would prefer using direct income transfers or government wage subsidies rather than gumming up the wage system. How important are each of these three concerns to you? Depending upon your priorities, you might reach quite different conclusions on the advisability of increasing the minimum wage.

Energy Price Controls

Another example of government interference comes when the government legislates a maximum price ceiling. This occurred in the United States in the 1970s, and the results were sobering. We return to our analysis of the gasoline market to see how price ceilings function.

Let's set the scene. Suppose there is suddenly a sharp rise in oil prices. This has occurred because of reduced cartel supply and booming demand, but it might also come about because of political disturbances in the Middle East due to war or revolution. Figure 3-1 on p. 46 showed the results of the interaction of supply and demand in oil markets.

Politicians, seeing the sudden jump in prices, rise to denounce the situation. They claim that consumers are being "gouged" by profiteering oil companies. They worry that the rising prices threaten to ignite an inflationary spiral in the cost of living. They fret about the impact of rising prices on the poor and the elderly. They call upon the government to "do something." In the face of rising prices, the U.S.

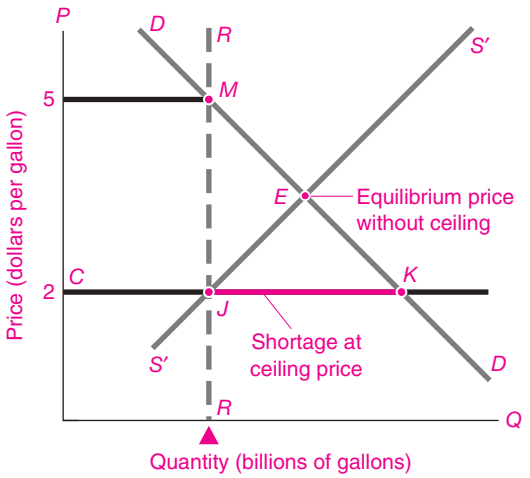


FIGURE 4-13. Price Controls Produce Shortages

Without a legal price ceiling, price would rise to *E*. At the ceiling price of $2, supply and demand do not balance, and shortages break out. Some method of formal or informal rationing is needed to allocate the short supply and bring the actual demand down to supply at *RR*. If *CJ* ration coupons become marketable, this would imply a new supply curve of *RR*. At the ceiling price of $2, coupons would sell for $3, and the total price (coupons plus cash) would be $5.

government might be inclined to listen to these arguments and place a ceiling on oil prices, as it did from 1973 to 1981.

What are the effects of such a ceiling? Suppose the initial price of gasoline is $2 a gallon. Then, because of a drastic cut in oil supply, the market price of gasoline rises sharply. Now consider the gasoline market after the supply shock. In Figure 4-13, the post-shock equilibrium is given at point *E*. If the free market were allowed to operate, the market would clear with a price of perhaps $3.50. Consumers would complain but would willingly pay the higher price rather than go without fuel.

Rationing by the Queue, by Coupons, or by the Purse?

Enter the government, which passes a law setting the maximum price for gasoline at the old level of $2 a gallon. We can picture this legal maximum price as the ceiling-price line *CJK* in Figure 4-13.

At the legal ceiling price, quantities supplied and demanded do not match. The market does not "clear" because it is against the law for sellers to charge the equilibrium price. Consumers want more gasoline than producers are willing to supply at the controlled price. This is shown by the gap between *J* and *K*. There follows a period of frustration and shortage—a game of musical chairs in which somebody is left without gasoline when the pump runs dry.

The inadequate supply of gasoline must somehow be rationed. Initially, this may be done through a first-come, first-served approach. People wait in line—this is rationing by the "queue." Because people's time is valuable, the length of the line will serve as a kind of price that limits demand. We see rationing by the queue today in markets like health care, where the price of medical care is subsidized. This is a wasteful system because much valuable time is spent waiting in line just as a way of preventing prices from reaching equilibrium.

Sometimes, particularly during large wars such as World War II, governments design a more efficient system of nonprice rationing based on formal allocation or coupon rationing. Perhaps people get a gasoline ration that is distributed on the basis of the number of automobiles. Under coupon rationing, each customer must have a coupon as well as money to buy the goods—in effect, there are two kinds of money. When rationing is adopted, shortages disappear because demand is limited by the allocation of the coupons.

Just how do ration coupons change the supply-and-demand picture? In Figure 4-13, suppose the government hands out coupons corresponding to quantity *CJ*. Then, supply and the new demand balance at the ceiling price of $2.

Sometimes, the ration coupons will be marketable. Figure 4-13 shows a supply of coupons of *RR*. With this supply curve, the equilibrium price of gasoline is $5 per gallon, and the price of coupons is given by *JM*, or $3 per gallon. At this point, gasoline is once again a market commodity, where you pay $2 for the gasoline and $3 for a coupon. The price has indeed risen, but in an indirect way. Additionally, people with coupons have been given a new form of income in coupons. Note that because of the price control, quantity supplied is still at the old level, but the total price including coupons ($5) is actually

higher than the original equilibrium price without rationing ($3.50).

All of this sounds complicated, and it is. History has shown that legal and illegal evasions of price controls grow over time. The inefficiencies eventually overwhelm whatever favorable impacts the controls might have on consumers. Particularly when there is room for ample substitution (i.e., when elasticities of supply or demand are high), price controls are costly, difficult to administer, and ineffective. Consequently, price controls on most goods are rarely used in most market economies.

There is a profound lesson here: Goods are always scarce. Society can never fulfill everyone's desires. In normal times, price itself rations the scarce supplies. When governments step in to interfere with supply and demand, prices no longer fill the role of rationers. Waste, inefficiency, and aggravation are likely companions of such interferences.

SUMMARY

A. Price Elasticity of Demand and Supply

1. Price elasticity of demand measures the quantitative response of demand to a change in price. Price elasticity of demand (E_D) is defined as the percentage change in quantity demanded divided by the percentage change in price. That is,

$$\text{Price elasticity of demand} = E_D = \frac{\text{percentage change in quantity demanded}}{\text{percentage change in price}}$$

   In this calculation, the sign is taken to be positive, and P and Q are averages of old and new values.
2. We divide price elasticities into three categories: (*a*) Demand is elastic when the percentage change in quantity demanded exceeds the percentage change in price; that is, $E_D > 1$. (*b*) Demand is inelastic when the percentage change in quantity demanded is less than the percentage change in price; here, $E_D < 1$. (*c*) When the percentage change in quantity demanded exactly equals the percentage change in price, we have the borderline case of unit-elastic demand, where $E_D = 1$.
3. Price elasticity is a pure number, involving percentages; it should not be confused with slope.
4. The demand elasticity tells us about the impact of a price change on total revenue. A price reduction increases total revenue if demand is elastic; a price reduction decreases total revenue if demand is inelastic; in the unit-elastic case, a price change has no effect on total revenue.
5. Price elasticity of demand tends to be low for necessities like food and shelter and high for luxuries like snowmobiles and vacation air travel. Other factors affecting price elasticity are the extent to which a good has ready substitutes and the length of time that consumers have to adjust to price changes.
6. Price elasticity of supply measures the percentage change of output supplied by producers when the market price changes by a given percentage.

B. Applications to Major Economic Issues

7. One of the most fruitful arenas for application of supply-and-demand analysis is agriculture. Improvements in agricultural technology mean that supply increases greatly, while demand for food rises less than proportionately with income. Hence free-market prices for foodstuffs tend to fall. No wonder governments have adopted a variety of programs, like crop restrictions, to prop up farm incomes.
8. A commodity tax shifts the supply-and-demand equilibrium. The tax's incidence (or impact on incomes) will fall more heavily on consumers than on producers to the degree that the demand is inelastic relative to supply.
9. Governments occasionally interfere with the workings of competitive markets by setting maximum ceilings or minimum floors on prices. In such situations, quantity supplied need no longer equal quantity demanded; ceilings lead to excess demand, while floors lead to excess supply. Sometimes, the interference may raise the incomes of a particular group, as in the case of farmers or low-skilled workers. Often, distortions and inefficiencies result.

CONCEPTS FOR REVIEW

Elasticity Concepts

price elasticity of demand, supply
elastic, inelastic, unit-elastic demand
E_D = % change in Q/% change in P
determinants of elasticity
total revenue = $P \times Q$
relationship of elasticity and revenue change

Applications of Supply and Demand

incidence of a tax
distortions from price controls
rationing by price vs. rationing by the queue

FURTHER READING AND INTERNET WEBSITES

Further Reading

If you have a particular concept you want to review, such as elasticity, you can often look in an encyclopedia of economics, such as John Black, *Oxford Dictionary of Economics,* 2d ed. (Oxford, New York, 2002), or David W. Pearce, ed., *The MIT Dictionary of Modern Economics* (MIT Press, Cambridge, Mass., 1992). The most comprehensive encyclopedia, covering many advanced topics in seven volumes, is Steven N. Durlauf and Lawrence E. Blume, eds., *The New Palgrave Dictionary of Economics* (Macmillan, London, 2008), available in most libraries.

The minimum wage has generated a fierce debate among economists. A recent book by two labor economists presents evidence that the minimum wage has little effect on employment: David Card and Alan Krueger, *Myth and Measurement: The New Economics of the Minimum Wage* (Princeton University Press, Princeton, N.J., 1997).

Websites

There are currently no reliable online dictionaries for terms in economics. There are few good websites for understanding fundamental economic concepts like supply and demand or elasticities. The concise online encyclopedia of economics at *www.econlib.org/library/CEE.html* is generally reliable but covers only a small number of topics. Sometimes, the free site of the *Encyclopaedia Britannica* at *www.britannica.com* provides background or historical material. When all else fails, you can go to the online encyclopedia at *en.wikipedia.org/wiki/Main_Page*, but be warned that it is often unreliable. (For example, the 2008 definition of "price elasticity of demand" is close to incomprehensible.)

Current issues such as the minimum wage are often discussed in policy papers at the website of the Economic Policy Institute, a think tank focusing on economic issues of workers, at *www.epinet.org*.

QUESTIONS FOR DISCUSSION

1. "A good harvest will generally lower the income of farmers." Illustrate this proposition using a supply-and-demand diagram.
2. For each pair of commodities, state which you think is the more price-elastic and give your reasons: perfume and salt; penicillin and ice cream; automobiles and automobile tires; ice cream and chocolate ice cream.
3. "The price drops by 1 percent, causing the quantity demanded to rise by 2 percent. Demand is therefore elastic, with $E_D > 1$." If you change 2 to ½ in the first sentence, what two other changes will be required in the quotation?
4. Consider a competitive market for apartments. What would be the effect on the equilibrium output and price after the following changes (other things held equal)? In each case, explain your answer using supply and demand.
   a. A rise in the income of consumers
   b. A $10-per-month tax on apartment rentals
   c. A government edict saying apartments cannot rent for more than $200 per month
   d. A new construction technique allowing apartments to be built at half the cost
   e. A 20 percent increase in the wages of construction workers
5. Consider a proposal to raise the minimum wage by 10 percent. After reviewing the arguments in the chapter, estimate the impact upon employment and upon

the incomes of affected workers. Using the numbers you have derived, write a short essay explaining how *you* would decide if you had to make a recommendation on the minimum wage.

6. A conservative critic of government programs has written, "Governments know how to do one thing well. They know how to create shortages and surpluses." Explain this quotation using examples like the minimum wage or interest-rate ceilings. Show graphically that if the demand for unskilled workers is price-elastic, a minimum wage will decrease the total earnings (wage times quantity demanded of labor) of unskilled workers.
7. Consider what would happen if a tariff of $2000 were imposed on imported automobiles. Show the impact of this tariff on the supply and the demand, and on the equilibrium price and quantity, of American automobiles. Explain why American auto companies and autoworkers often support import restraints on automobiles.
8. Elasticity problems:
   a. The world demand for crude oil is estimated to have a short-run price elasticity of 0.05. If the initial price of oil were $100 per barrel, what would be the effect on oil price and quantity of an embargo that curbed world oil supply by 5 percent? (For this problem, assume that the oil-supply curve is completely inelastic.)
   b. To show that elasticities are independent of units, refer to Table 3-1. Calculate the elasticities between each demand pair. Change the price units from dollars to pennies; change the quantity units from millions of boxes to tons, using the conversion factor of 10,000 boxes to 1 ton. Then recalculate the elasticities in the first two rows. Explain why you get the same answer.
   c. Jack and Jill went up the hill to a gas station that does not display the prices. Jack says, "Give me $10 worth of gas." Jill says, "Give me 10 gallons of gas." What are the price elasticities of demand for gasoline of Jack and of Jill? Explain.
   d. Can you explain why farmers during a depression might approve of a government program requiring that pigs be killed and buried under the ground?
   e. Look at the impact of the minimum wage shown in Figure 4-12. Draw in the rectangles of total income with and without the minimum wage. Which is larger? Relate the impact of the minimum wage to the price elasticity of demand for unskilled workers.
9. No one likes to pay rent. Yet scarcities of land and urban housing often cause rents to soar in cities. In response to rising rents and hostility toward landlords, governments sometimes impose *rent controls*. These generally limit the increases on rent to a small year-to-year increase and can leave controlled rents far below free-market rents.
   a. Redraw Figure 4-13 to illustrate the impact of rent controls for apartments.
   b. What will be the effect of rent controls on the vacancy rate of apartments?
   c. What nonrent options might arise as a substitute for the higher rents?
   d. Explain the words of a European critic of rent controls: "Except for bombing, nothing is as efficient at destroying a city as rent controls." (*Hint:* What would happen to maintenance?)
10. Review the example of the New Jersey cigarette tax (p. 71). Using graph paper or a computer, draw supply and demand curves that will yield the prices and quantities before and after the tax. (Figure 4-10 shows the example for a gasoline tax.) For this example, assume that the supply curve is perfectly elastic. [*Extra credit:* A demand curve with constant price elasticity takes the form $Y = AP^{-e}$, where Y is quantity demanded, P is price, A is a scaling constant, and e is the (absolute value) of the price elasticity. Solve for the values of A and e which will give the correct demand curve for the prices and quantities in the New Jersey example.]
11. Review the algebra of demand elasticities on p. 69. Then assume that the demand curve takes the following form: $Q = 100 - 2P$.
    a. Calculate the elasticities at $P = 1$, 25, and 49.
    b. Explain why elasticity is different from slope using the formula.

CHAPTER

5 Demand and Consumer Behavior

O, reason not the need: our basest beggars
Are in the poorest thing superfluous.

Shakespeare,
King Lear

We make countless decisions every day about how to allocate our scarce money and time. Should we buy a pizza or a hamburger? Buy a new car or fix our old one? Spend our income today or save for future consumption? Should we eat breakfast or sleep late? As we balance competing demands and desires, we make the choices that define our lives.

The results of these individual choices are what underlie the demand curves and price elasticities that we met in earlier chapters. This chapter explores the basic principles of consumer choice and behavior. We shall see how patterns of market demand can be explained by the process of individuals' pursuing their most preferred bundle of consumption goods. We also will learn how to measure the benefits that each of us receives from participating in a market economy.

CHOICE AND UTILITY THEORY

In explaining consumer behavior, economics relies on the fundamental premise that people choose those goods and services they value most highly. To describe the way consumers choose among different consumption possibilities, economists a century ago developed the notion of *utility*. From the notion of utility, they were able to derive the demand curve and explain its properties.

What do we mean by "utility"? In a word, **utility** denotes satisfaction. More precisely, it refers to how consumers rank different goods and services. If basket A has higher utility than basket B for Smith, this ranking indicates that Smith prefers A over B. Often, it is convenient to think of utility as the subjective pleasure or usefulness that a person derives from consuming a good or service. But you should definitely resist the idea that utility is a psychological function or feeling that can be observed or measured. Rather, utility is a scientific construct that economists use to understand how rational consumers make decisions. We derive consumer demand functions from the assumption that people make decisions that give them the greatest satisfaction or utility.

In the theory of demand, we assume that people maximize their utility, which means that they choose the bundle of consumption goods that they most prefer.

Marginal Utility and the Law of Diminishing Marginal Utility

How does utility apply to the theory of demand? Say that consuming the first unit of ice cream gives you a certain level of satisfaction or utility. Now imagine consuming a second unit. Your total utility goes up

because the second unit of the good gives you some additional utility. What about adding a third and fourth unit of the same good? Eventually, if you eat enough ice cream, instead of adding to your satisfaction or utility, it makes you sick!

This leads us to the fundamental economic concept of marginal utility. When you eat an additional unit of ice cream, you will get some additional satisfaction or utility. The increment to your utility is called **marginal utility.**

The expression "marginal" is a key term in economics and always means "additional" or "extra." Marginal utility denotes the additional utility you get from the consumption of an additional unit of a commodity.

One of the fundamental ideas behind demand theory is the **law of diminishing marginal utility.** This law states that the amount of extra or marginal utility declines as a person consumes more and more of a good.

To understand this law, first remember that utility tends to increase as you consume more of a good. However, as you consume more and more, your total utility will grow at a slower and slower rate. This is the same thing as saying that your marginal utility (the extra utility added by the last unit consumed of a good) diminishes as more of a good is consumed.

The law of diminishing marginal utility states that, as the amount of a good consumed increases, the marginal utility of that good tends to decline.

A Numerical Example

We can illustrate utility numerically as in Table 5-1. The table shows in column (2) that total utility (U) enjoyed increases as consumption (Q) grows, but it increases at a decreasing rate. Column (3) measures marginal utility as the extra utility gained when 1 extra unit of the good is consumed. Thus when the individual consumes 2 units, the marginal utility is $7 - 4 = 3$ units of utility (call these units "utils").

Focus next on column (3). The fact that marginal utility declines with higher consumption illustrates the law of diminishing marginal utility.

Figure 5-1 on page 86 shows graphically the data on total utility and marginal utility from Table 5-1. In part (a), the blue blocks add up to the total utility at each level of consumption. In addition, the smooth blue curve shows the smoothed utility level

(1) Quantity of a good consumed Q	(2) Total utility U	(3) Marginal utility MU
0	0	
		4
1	4	
		3
2	7	
		2
3	9	
		1
4	10	
		0
5	10	

TABLE 5-1. Utility Rises with Consumption

As we consume more of a good or service like pizza or concerts, total utility increases. The increment of utility from one unit to the next is the "marginal utility"—the extra utility added by the last extra unit consumed. By the law of diminishing marginal utility, the marginal utility falls with increasing levels of consumption.

for fractional units of consumption. It shows utility increasing, but at a decreasing rate. Figure 5-1(b) depicts marginal utilities. Each of the blue blocks of marginal utility is the same size as the corresponding block of total utility in (a). The straight blue line in (b) is the smoothed curve of marginal utility.

The law of diminishing marginal utility implies that the marginal utility (MU) curve in Figure 5-1(b) must slope downward. This is exactly equivalent to saying that the total utility curve in Figure 5-1(a) must look concave, like a dome.

Relationship of Total and Marginal Utility. Using Figure 5-1, we can easily see that the total utility of consuming a certain amount is equal to the sum of the marginal utilities up to that point. For example, assume that 3 units are consumed. Column (2) of Table 5-1 shows that the total utility is 9 units. In column (3) we see that the sum of the marginal utilities of the first 3 units is also $4 + 3 + 2 = 9$ units.

Examining Figure 5-1(b), we see that the total area under the marginal utility curve at a particular level of consumption—as measured either by blocks or by the area under the smooth MU curve—must

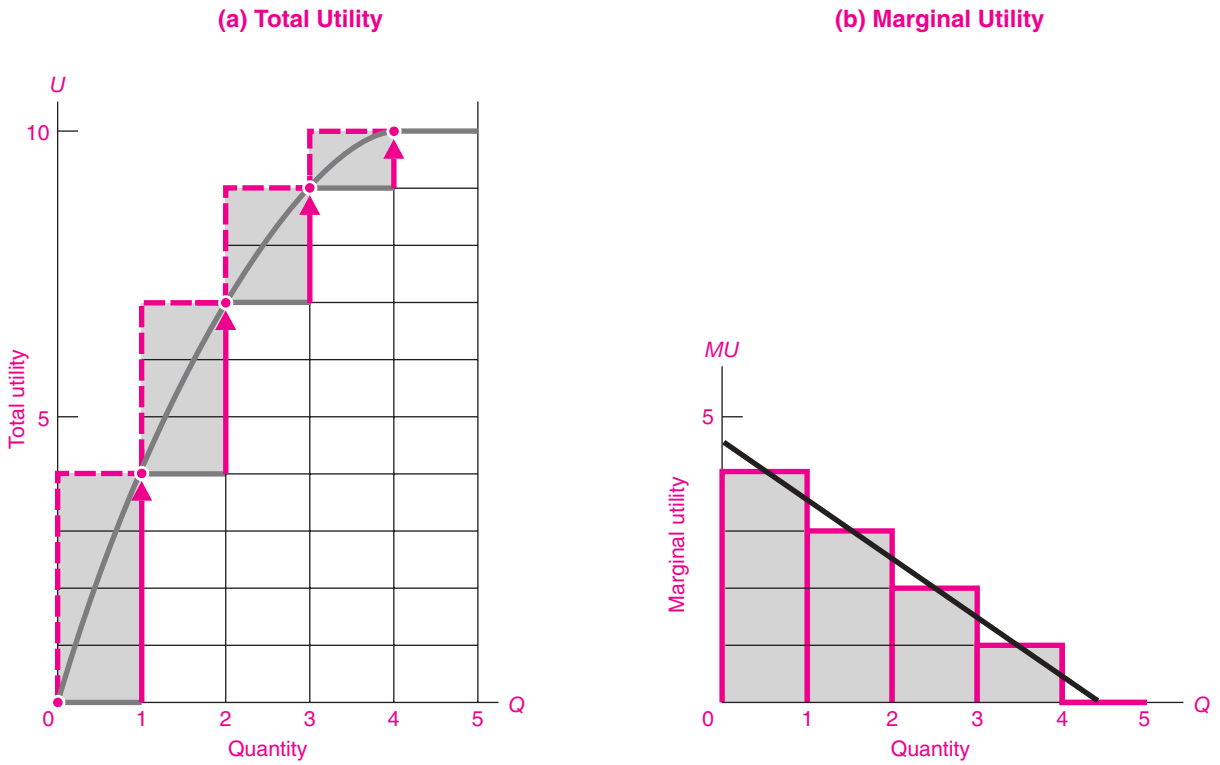


FIGURE 5-1. The Law of Diminishing Marginal Utility

Total utility in **(a)** rises with consumption, but it rises at a decreasing rate, showing diminishing marginal utility. This observation led early economists to formulate the law of downward-sloping demand.

The blue blocks show the extra utility added by each new unit. The fact that total utility increases at a decreasing rate is shown in **(b)** by the declining steps of marginal utility. If we make our units smaller, the steps in total utility are smoothed out and total utility becomes the smooth blue curve in **(a)**. Moreover, smoothed marginal utility, shown in **(b)** by the blue downward-sloping smooth curve, becomes indistinguishable from the slope of the smooth curve in **(a)**.

equal the height of the total utility curve shown for the same number of units in Figure 5-1(*a*).

Whether we examine this relationship using tables or graphs, we see that total utility is the sum of all the marginal utilities that were added from the beginning.

History of Utility Theory

Modern utility theory stems from *utilitarianism,* which has been one of the major currents of Western intellectual thought of the last two centuries. The notion of utility arose soon after 1700, as the basic ideas of mathematical probability were being developed. Thus Daniel Bernoulli, a member of a brilliant Swiss family of mathematicians, observed in 1738 that people act as if the dollar they stand to gain in a fair bet is worth less to them than the dollar they stand to lose. This means that they are averse to risk and that successive new dollars of wealth bring them smaller and smaller increments of true utility.

An early introduction of the utility notion into the social sciences was accomplished by the English philosopher Jeremy Bentham (1748–1832). After studying legal theory, and under the influence of Adam Smith's doctrines, Bentham turned to the study of the principles necessary

for drawing up social legislation. He proposed that society should be organized on the "principle of utility," which he defined as the "property in any object . . . to produce pleasure, good or happiness or to prevent . . . pain, evil or unhappiness." All legislation, according to Bentham, should be designed on utilitarian principles, to promote "the greatest happiness of the greatest number." Among his other legislative proposals were quite modern-sounding ideas about crime and punishment in which he suggested that raising the "pain" to criminals by harsh punishments would deter crimes.

Bentham's views about utility seem familiar to many people today. But they were revolutionary 200 years ago because they emphasized that social and economic policies should be designed to achieve certain practical results, whereas legitimacy at that time was usually based on tradition, the divine right of kings, or religious doctrines. Today, many political thinkers defend their legislative proposals with utilitarian notions of what will make the largest number of people best off.

The next step in the development of utility theory came when the neoclassical economists—such as William Stanley Jevons (1835–1882)—extended Bentham's utility concept to explain consumer behavior. Jevons thought economic theory was a "calculus of pleasure and pain," and he developed the theory that rational people would base their consumption decisions on the extra or marginal utility of each good.

The ideas of Jevons and his coworkers led directly to the modern theories of ordinal utility and indifference curves developed by Vilfredo Pareto, John Hicks, R. G. D. Allen, Paul Samuelson, and others in which the Benthamite ideas of measurable cardinal utility are no longer needed.

DERIVATION OF DEMAND CURVES

The Equimarginal Principle

Having explained utility theory, we now apply that theory to explain consumer demand and to understand the nature of demand curves.

We assume that each consumer maximizes utility, which means that the consumer chooses the most preferred bundle of goods from what is available. We also assume that consumers have a certain income and face given market prices for goods.

What would be a sensible rule for choosing the preference bundle of goods in this situation? Certainly, I would not expect that the last egg brings the same marginal utility as the last pair of shoes, for shoes cost much more per unit than eggs. A satisfactory rule would be: If good A costs twice as much as good B, then buy good A only when its marginal utility is at least twice as great as good B's marginal utility.

This leads to the *equimarginal principle* that I should arrange my consumption so that the last dollar spent on each good is bringing me the same marginal utility.

Equimarginal principle: The fundamental condition of maximum satisfaction or utility is the equimarginal principle. It states that a consumer will achieve maximum satisfaction or utility when the marginal utility of the last dollar spent on a good is exactly the same as the marginal utility of the last dollar spent on any other good.

Why must this condition hold? If any one good gave more marginal utility per dollar, I would increase my utility by taking money away from other goods and spending more on that good—until the law of diminishing marginal utility drove its marginal utility per dollar down to equality with that of other goods. If any good gave less marginal utility per dollar than the common level, I would buy less of it until the marginal utility of the last dollar spent on it had risen back to the common level. The common marginal utility per dollar of all commodities in consumer equilibrium is called the *marginal utility of income*. It measures the additional utility that would be gained if the consumer could enjoy an extra dollar's worth of consumption.

This fundamental condition of consumer equilibrium can be written in terms of the marginal utilities (*MU*s) and prices (*P*s) of the different goods in the following compact way:

$$\frac{MU_{\text{good 1}}}{P_1} = \frac{MU_{\text{good 2}}}{P_2} = \frac{MU_{\text{good 3}}}{P_3} = \cdots = MU \text{ per \$ of income}$$

Why Demand Curves Slope Downward

Using the fundamental rule for consumer behavior, we can easily see why demand curves slope downward. For simplicity, hold the common marginal

utility per dollar of income constant. Then increase the price of good 1. With no change in quantity consumed, the first ratio (i.e., $MU_{\text{good 1}}/P_1$) will be below the *MU* per dollar of all other goods. The consumer will therefore have to readjust the consumption of good 1. The consumer will do this by (*a*) lowering the consumption of good 1, thereby (*b*) raising the *MU* of good 1, until (*c*) at the new, reduced level of consumption of good 1, the new marginal utility per dollar spent on good 1 is again equal to the *MU* per dollar spent on other goods.

A higher price for a good reduces the consumer's desired consumption of that commodity; this shows why demand curves slope downward.

Leisure and the Optimal Allocation of Time

A Spanish toast to a friend wishes "health, wealth, and the time to enjoy them." This saying captures the idea that we must allocate our time budgets in much the same way as we do our dollar budgets. Time is the great equalizer, for even the richest person has but 24 hours a day to "spend." Let's see how our earlier analysis of allocating scarce dollars applies to time.

Consider leisure, often defined as "time which one can spend as one pleases." Leisure brings out our personal eccentricities. The seventeenth-century philosopher Francis Bacon held that the purest of human pleasures was gardening. The British statesman Winston Churchill wrote of his holiday: "I have had a delightful month building a cottage and dictating a book: 200 bricks and 2000 words a day."

We can apply utility theory to the allocation of time as well as money. Suppose that, after satisfying all your obligations, you have 3 hours a day of free time and can devote it to gardening, laying bricks, or writing history. What is the best way to allocate your time? Let's ignore the possibility that time spent on some of these activities might be an investment that will enhance your earning power in the future. Rather, assume that these are all pure consumption or utility-yielding pursuits. The principles of consumer choice suggest that you will make the best use of your time when you equalize the marginal utilities of the last minute spent on each activity.

To take another example, suppose you want to maximize your knowledge in your courses but you have only a limited amount of time available. Should you study each subject for the same amount of time? Surely not. You may find that an equal study time for economics, history, and chemistry will not yield the same amount of knowledge in the last minute. If the last minute produces a greater marginal knowledge in chemistry than in history, you would raise your total knowledge by shifting additional minutes from history to chemistry, and so on, until the last minute yields the same incremental knowledge in each subject.

The same rule of maximum utility per hour can be applied to many different areas of life, including engaging in charitable activities, improving the environment, or losing weight. It is not merely a law of economics. It is a law of rational choice.

Are Consumers Wizards? The View from Behavioral Economics

All of this discussion makes it sound as if consumers are mathematical wizards who routinely make calculations of marginal utility to the tenth decimal place and solve complicated systems of equations in their everyday lives.

This unrealistic view is definitely not what we assume in economics. We know that most decisions are made in a routine and intuitive way. We may have Cheerios and yogurt for breakfast every day because they are not too expensive, are easy to find in the store, and slake our morning hunger.

Rather, what we assume in consumer demand theory is that consumers are reasonably consistent in their tastes and actions. We expect that people do not flail around and make themselves miserable by constantly making mistakes. If most people act consistently most of the time, avoiding erratic changes in buying behavior and generally choosing their most preferred bundles, our theory of demand will provide a reasonably good approximation to the facts.

As always, however, we must be alert to situations where irrational or inconsistent behavior crops up. We know that people make mistakes. People sometimes buy useless gadgets or are bilked by unscrupulous sales pitches. A new area of research is *behavioral economics,* which recognizes that people have limited time and memory, that information is incomplete, and that patterns of irrational-looking behavior are persistent. This approach allows for the possibility that imperfect information, psychological biases, and costly decision making may lead to poor decisions.

Behavioral economics explains why households save too little for retirement, why stock market bubbles occur, and how used-car markets behave when people's information is limited. A significant recent example illustrating behavioral principles came when millions of people took out "subprime mortgages" to buy homes in the 2000s. They did not read or could not understand the fine print, and as a result many people defaulted on their mortgages and lost their homes, triggering a major financial crisis and an economic downturn. It turns out that poor consumers were not the only people who could not read the fine print, however, for they were joined by banks, hedge-fund managers, bond-rating firms, and thousands of investors who bought assets that they did not understand.

Behavioral economics joined the mainstream in 2001 and 2002 when Nobel Prizes were awarded for economic research in this area. George Akerlof (University of California at Berkeley) was cited for developing a better understanding of the role of asymmetric information and the market for "lemons." Daniel Kahneman (Princeton University) and Vernon L. Smith (George Mason University) received the prize for "the analysis of human judgment and decision-making . . . and the empirical testing of predictions from economic theory by experimental economists."

Analytical Developments in Utility Theory

We pause here to provide an elaboration of some of the advanced issues behind the concept of utility and its application to demand theory. Economists today generally reject the notion of a cardinal (or measurable) utility that people feel or experience when consuming goods and services. Utility does not ring up like numbers on a gasoline pump.

Rather, what counts for modern demand theory is the principle of **ordinal utility.** Under this approach, consumers need to determine only their preference ranking of bundles of commodities. Ordinal utility asks, "Do I prefer a pastrami sandwich to a chocolate milk shake?" A statement such as "Bundle A is preferred to bundle B"—which does not require that we know how much A is preferred to B—is called ordinal, or dimensionless. Ordinal variables are ones that we can rank in order, but for which there is no measure of the quantitative difference between the situations. We might rank pictures in an exhibition by order of beauty without having a quantitative measure of beauty. Using only such ordinal preference rankings, we can establish firmly the general properties of market demand curves described in this chapter and in its appendix.

The discerning reader will wonder whether the equimarginal principle describing consumer equilibrium behavior implies cardinal utility. In fact, it does not; only ordinal measures are needed. An ordinal utility measure is one that we can stretch while always maintaining the same greater-than or less-than relationship (like measuring with a rubber band). Examine the marginal condition for consumer equilibrium. If the utility scale is stretched (say, by doubling or multiplying times 3.1415), you can see that all the numerators in the condition are changed by exactly the same amount, so the consumer equilibrium condition still holds.

For certain special situations the concept of *cardinal,* or dimensional, utility is useful. An example of a cardinal measure comes when we say that the speed of a plane is six times that of a car. People's behavior under conditions of uncertainty is today analyzed using a cardinal concept of utility. This topic will be examined further when we review the economics of risk, uncertainty, and gambling in Chapter 11.

Our treatment of utility in the equimarginal principle assumed that goods can be divided into indefinitely small units. However, sometimes indivisibility of units is important and cannot be glossed over. Thus, Hondas cannot be divided into arbitrarily small portions the way juice can. Suppose I buy one Honda, but definitely not two. Then the additional utility of the first car is enough larger than the additional utility of the same number of dollars spent elsewhere to induce me to buy this first unit. The additional utility that the second Honda would bring is enough less to ensure I do not buy it. When indivisibility matters, our equality rule for equilibrium can be restated as an inequality rule.

AN ALTERNATIVE APPROACH: SUBSTITUTION EFFECT AND INCOME EFFECT

The concept of marginal utility has helped explain the fundamental law of downward-sloping demand. But over the last few decades, economists have developed an alternative approach to the analysis of demand—one that makes no mention of marginal utility. This alternative approach uses "indifference

curves," which are explained in the appendix to this chapter, to rigorously and consistently produce the major propositions about consumer behavior. This approach also helps explain the factors that tend to make the responsiveness of quantity demanded to price—the price elasticity of demand—large or small.

Indifference analysis asks about the substitution effect and the income effect of a change in price. By looking at these, we can see why the quantity demanded of a good declines as its price rises.

Substitution Effect

The substitution effect is the most obvious factor for explaining downward-sloping demand curves. If the price of coffee goes up while other prices do not, then coffee has become relatively more expensive. When coffee becomes a more expensive beverage, less coffee and more tea or cola will be bought. Similarly, because sending electronic mail is cheaper and quicker than sending letters through the regular mail, people are increasingly relying on electronic mail for correspondence. More generally, the **substitution effect** says that when the price of a good rises, consumers will tend to substitute other goods for the more expensive good in order to satisfy their desires more inexpensively.

Consumers, then, behave the way businesses do when the rise in price of an input causes firms to substitute low-priced inputs for high-priced inputs. By this process of substitution, businesses can produce a given amount of output at the least total cost. Similarly, when consumers substitute less expensive goods for more expensive ones, they are buying a given amount of satisfaction at lower cost.

Income Effect

A second impact of a price change comes through its effect on real income. The term *real income* means the actual quantity of goods that your money income can buy. When a price rises and money income is fixed, real income falls because the consumer cannot afford to buy the same quantity of goods as before. This produces the **income effect,** which is the change in the quantity demanded that arises because a price change lowers consumer real incomes. Most goods respond positively to higher incomes, so the income effect will normally reinforce the substitution effect in producing a downward-sloping demand curve.

We can obtain a quantitative measure of the income effect using a new concept, **income elasticity.** This term denotes the percentage change in quantity demanded divided by the percentage change in income, holding other things, such as prices, constant.

$$\text{Income elasticity} = \frac{\%\text{ change in quantity demanded}}{\%\text{ change in income}}$$

High income elasticities, such as those for airline travel or yachts, indicate that the demand for these goods rises rapidly as income increases. Low income elasticities, such as for potatoes or used furniture, denote a weak response of demand to increases in income.

Calculation of Income Elasticity

Suppose you are a city planner for Santa Fe, New Mexico, and you are concerned about the growth in the demand for water consumption by households in that arid region. You make inquiries and find the following data for 2000: The population is 62,000; the projected growth rate of the population is 20 percent per decade; per capita annual water consumption in 2000 was 1000 gallons; per capita incomes are projected to grow by 25 percent over the next decade; and the income elasticity of water use per capita is 0.50. You then estimate the water needs for 2010 (with unchanged prices) as

Water consumption in 2010

= population in 2000 × population growth factor
× per capita water use
× [1 + (income growth × income elasticity)]
= 62,000 × 1.2 × 1000 × (1 + 0.25 × 0.50)
= 83,700,000

From these data, you project a growth in total household water use of 35 percent from 2000 to 2010.

Income and substitution effects combine to determine the major characteristics of demand curves of different commodities. Under some circumstances the resulting demand curve is very price-elastic, as where the consumer has been spending a good deal on the commodity and ready substitutes are available. In this case both the income and the

substitution effects are strong and the quantity demanded responds strongly to a price increase.

But consider a commodity like salt, which requires only a small fraction of the consumer's budget. Salt is not easily replaceable by other items and is needed in small amounts to complement more important items. For salt, both income and substitution effects are small, and demand will tend to be price-inelastic.

FROM INDIVIDUAL TO MARKET DEMAND

Having analyzed the principles underlying a single individual's demand for coffee or electronic mail, we next examine how the entire market demand derives from the individual demand. *The demand curve for a good for the entire market is obtained by summing up the quantities demanded by all the consumers.* Each consumer has a demand curve along which the quantity demanded can be plotted against the price; it generally slopes downward and to the right. If all consumers were exactly alike in their demands and if there were 1 million consumers, we could think of the market demand curve as a millionfold enlargement of each consumer's demand curve.

In fact, of course, people differ in their tastes. Some have high incomes, some low. Some greatly desire coffee; others prefer tea. To obtain the total market demand curve, we calculate the sum total of what all the different consumers consume at each price. We then plot that total amount as a point on the market demand curve. Alternatively, we might construct a numerical demand table by summing the quantities demanded by all individuals at each market price.

As a matter of convention, we label *individual* demand and supply curves with lowercase letters (*dd* and *ss*), while using uppercase letters (*DD* and *SS*) for the *market* demand and supply curves.

The market demand curve is the sum of individual demands at each price. Figure 5-2 shows how to add individual *dd* demand curves horizontally to get the market *DD* demand curve.

FIGURE 5-2. Market Demand Derived from Individual Demands

We add all individual consumers' demand curves to get the market demand curve. At each price, such as $5, we add quantities demanded by each person to get the market quantity demanded. The figure shows how, at a price of $5, we add horizontally Smith's 1 unit demanded to Brown's 2 units to get the market demand of 3 units.

Demand Shifts

We know that changes in the price of coffee affect the quantity of coffee demanded. We know this from budget studies, from historical experience, and from examining our own behavior. We discussed briefly in Chapter 3 some of the important nonprice determinants of demand. We now review the earlier discussion in light of our analysis of consumer behavior.

An increase in income tends to increase the amount we are willing to buy of most goods. Necessities tend to be less responsive than most goods to income changes, while luxuries tend to be more responsive to income. And there are a few anomalous goods, known as inferior goods, for which purchases may shrink as incomes increase because people can afford to replace them with other, more desirable goods. Soup bones, intercity bus travel, and black-and-white TVs are examples of inferior goods for many Americans today.

What does all this mean in terms of the demand curve? The demand curve shows how the quantity of a good demanded responds to a change in its own price. But the demand is also affected by the prices of other goods, by consumer incomes, and by special influences. The demand curve was drawn on the assumption that these other things were held constant. But what if these other things change? Then the whole demand curve will shift to the right or to the left.

Figure 5-3 illustrates changes in factors affecting demand. Given people's incomes and the prices for other goods, we can draw the demand curve for coffee as *DD*. Assume that price and quantity are at point *A*. Suppose that incomes rise while the prices of coffee and other goods are unchanged. Because coffee is a normal good with a positive income elasticity, people will increase their purchases of coffee. Hence the demand curve for coffee will shift to the right, say, to *D′D′*, with *A′* indicating the new quantity demanded of coffee. If incomes should fall, then we would expect a reduction in demand and in quantity bought. This downward shift we illustrate by *D″D″* and by *A″*.

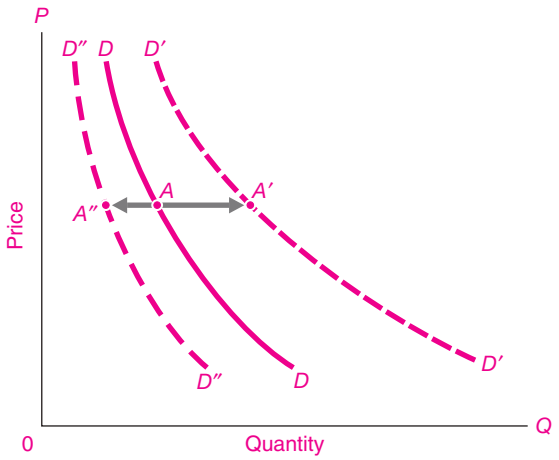


FIGURE 5-3. Demand Curve Shifts with Changes in Income or in Other Goods' Prices

As incomes increase, consumers generally want more of a good, thus increasing demand or shifting demand outward (explain why higher incomes shift *DD* to *D′D′*). Similarly, a rise in the price of a substitute good increases or shifts out the demand curve (e.g., from *DD* to *D′D′*). Explain why a decrease in income would generally shift demand to *D″D″*. Why would a decrease in chicken prices shift hamburger demand to *D″D″*?

Substitutes and Complements

Everyone knows that raising the price of beef will decrease the amount of beef demanded. We have seen that it will also affect the demand for other commodities. For example, a higher price for beef will increase the demand for substitutes like chicken. A higher beef price may lower the demand for goods like hamburger buns and ketchup that are used along with beef hamburgers. It will probably have little effect on the demand for economics textbooks.

We say, therefore, that beef and chicken are substitute products. Goods A and B are **substitutes** if an increase in the price of good A will increase the demand for substitute good B. Hamburgers and hamburger buns, or cars and gasoline, on the other hand, are complementary products; they are called **complements** because an increase in the price of good A causes a decrease in the demand for its complementary good B. In between are **independent goods,** such as beef and textbooks, for which a price change for one has no effect on the demand for the other. Try classifying the pairs turkey and cranberry sauce, oil and coal, college and textbooks, shoes and shoelaces, salt and shoelaces.

Say Figure 5-3 represented the demand for beef. A fall in the price of chickens may well cause consumers to buy less beef; the beef demand curve would therefore shift to the left, say, to *D″D″*. But what if the price of hamburger buns were to fall? The resulting change on *DD,* if there is one, will be in the direction of increased beef purchases, a rightward shift of the demand curve.

Why do we see this difference in response? Because chicken is a substitute product for beef, while hamburger buns are complements to beef.

Review of key concepts:

- The **substitution effect** occurs when a higher price leads to substitution of other goods for the good whose price has risen.
- The **income effect** is the change in the quantity demanded of a good because the change in its price has the effect of changing a consumer's real income.
- **Income elasticity** is the percentage change in quantity demanded of a good divided by the percentage change in income.
- Goods are **substitutes** if an increase in the price of one increases the demand for the other.
- Goods are **complements** if an increase in the price of one decreases the demand for the other.
- Goods are **independent** if a price change for one has no effect on the demand for the other.

Empirical Estimates of Price and Income Elasticities

For many economic applications, it is essential to have numerical estimates of price elasticities. For example, an automobile manufacturer will want to know the impact on sales of the higher car prices that result from installation of costly pollution-control equipment; a college needs to know the impact of higher tuition rates on student applications; and a publisher will calculate the impact of higher textbook prices on its sales. All these applications require a numerical estimate of price elasticity.

Similar decisions depend on income elasticities. A government planning its road or rail network will estimate the impact of rising incomes on automobile travel; the federal government must calculate the effect of higher incomes on energy consumption in designing policies for air pollution or global warming; in determining the necessary investments for generating capacity, electrical utilities require income elasticities for estimating electricity consumption.

Economists have developed useful statistical techniques for estimating price and income elasticities. The quantitative estimates are derived from market data on quantities demanded, prices, incomes, and other variables. Tables 5-2 and 5-3 show selected estimates of elasticities.

Commodity	Price elasticity
Tomatoes	4.60
Green peas	2.80
Legal gambling	1.90
Taxi service	1.24
Furniture	1.00
Movies	0.87
Shoes	0.70
Legal services	0.61
Medical insurance	0.31
Bus travel	0.20
Residential electricity	0.13

TABLE 5-2. Selected Estimates of Price Elasticities of Demand

Estimates of price elasticities of demand show a wide range of variation. Elasticities are generally high for goods for which ready substitutes are available, like tomatoes or peas. Low price elasticities exist for those goods like electricity which are essential to daily life and which have no close substitutes.

Source: Heinz Kohler, *Microeconomics: Theory and Applications* (Heath, Lexington, Mass., 1992).

Commodity	Income elasticity
Automobiles	2.46
Owner-occupied housing	1.49
Furniture	1.48
Books	1.44
Restaurant meals	1.40
Clothing	1.02
Physicians' services	0.75
Tobacco	0.64
Eggs	0.37
Margarine	–0.20
Pig products	–0.20
Flour	–0.36

TABLE 5-3. Income Elasticities for Selected Products

Income elasticities are high for luxuries, whose consumption grows rapidly relative to income. Negative income elasticities are found for "inferior goods," whose demand falls as income rises. Demand for many staple commodities, like clothing, grows proportionally with income.

Source: Heinz Kohler, *Microeconomics: Theory and Applications* (Heath, Lexington, Mass., 1992).

THE ECONOMICS OF ADDICTION

In a free-market economy, the government generally lets people decide what to buy with their money. If some want to buy expensive cars while others want to buy expensive houses, we assume that they know what is best for them and that in the interests of personal freedom the government should respect their preferences.

In some cases, but sparingly and with great hesitation, the government decides to overrule private adult decisions. These are cases of *merit goods,* whose consumption is thought intrinsically worthwhile, and the opposite, which are *demerit goods,* whose consumption is deemed harmful. For these goods, we recognize that some consumption activities have such serious effects that overriding individuals' private decisions may be desirable. Today, most societies provide for free public education and emergency health care; on the other hand, society also penalizes or forbids consumption of such harmful substances as cigarettes, alcohol, and heroin.

Among the most controversial areas of social policy are demerit goods involving addictions. An addiction is a pattern of compulsive and uncontrolled use of a substance. The heavy smoker or the heroin user may bitterly regret the acquired habit, but such habits are extremely difficult to break after they have become established. A regular user of cigarettes or heroin is much more likely to desire these substances than is a nonuser. Moreover, the demands for addictive substances are quite price-inelastic.

The markets for addictive substances are big business. Consumer expenditures on tobacco products in 2007 were $95 billion, while total expenditures on alcoholic beverages were $155 billion. Numbers for illegal drugs involve guesswork, but recent estimates of spending on illegal drugs place the total at around $75 billion annually.

Consumption of these substances raises major public policy issues because addictive substances may injure users and often impose costs and harms on society. The harms to users include around 450,000 premature deaths annually, along with a wide variety of medical problems attributable to smoking; 10,000 highway fatalities a year attributed to alcohol; and failures in school, job, and family, along with high levels of AIDS, from intravenous heroin use. Harms to society include the predatory crime that addicts of high-price drugs engage in; the costs of providing subsidized medical care to those who consume drugs, cigarettes, or tobacco; the rapid spread of communicable diseases, especially AIDS and pneumonia; and the tendency of existing users to recruit new users.

One policy approach, often followed in the United States, is to prohibit the sale and use of addictive substances and to enforce prohibition with criminal sanctions. Economically, prohibition can be interpreted as a sharp upward shift in the supply curve. After the upward shift, the price of the addictive substance is much higher. During Prohibition (1920–1933), alcohol prices were approximately 3 times higher than before. Estimates are that cocaine currently sells for at least 20 times its free-market price.

What is the effect of supply restrictions on the consumption of addictive substances? And how does the prohibition affect the injuries to self and to society? To answer these questions, we need to consider the nature of the demand for addictive substances. The evidence indicates that casual consumers of illegal drugs have cheap substitutes like alcohol and tobacco and thus will have relatively high price elasticity of demand. By contrast, hard-core users are often addicted to particular substances and have price-inelastic demands.

We can illustrate the market for addictive substances in Figure 5-4. The demand curve *DD* is extremely price-inelastic for established users. Now consider a policy of discouraging drug use. One approach, used for cigarettes, is to impose a large tax. As we saw in the previous chapter, this can be analyzed as an upward shift in the supply curve. A policy of prohibition such as is used for illegal substances has the same effect of shifting the supply curve from *SS* to *S′S′*.

Because demand is price-inelastic, quantity demanded will decline very little. At the higher price, total spending on drugs increases sharply. For illegal drugs, the required outlays may be so great that the user engages in predatory crime. The results, in the view of two economists who have studied the subject, are that "the market in illegal drugs promotes crime, destroys inner cities, spreads AIDS, corrupts law enforcement officials and politicians, produces and exacerbates poverty, and erodes the moral fabric of society."

A different case would arise for highly price-sensitive consumers such as casual users. For example,

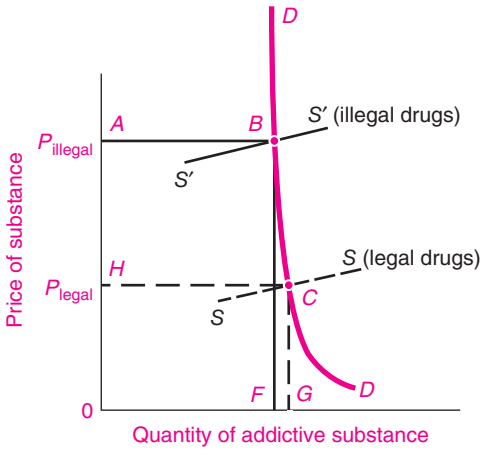


FIGURE 5-4. Market for Addictive Substances

The demand for addictive substances is price-inelastic for heavy smokers or hard-core users of drugs like heroin or cocaine. As a result, if prohibition or heavy taxation shifts supply from *SS* to *S′S′*, total spending on drugs will rise from 0*HCG* to 0*ABF*. For drugs that are highly price-inelastic, this implies that spending on drugs will rise when supply is restrained. What will happen to criminal activity after prohibition if a substantial fraction of the income of addicts is obtained by theft? Can you see why some people would argue for reduced drug enforcement or even decriminalization for addictive drugs?

a teenager might experiment with an addictive substance if it is affordable, while a high price (accompanied by low availability) would reduce the number of people who start down the road to addiction. In this case, supply restraints are likely to lower use sharply and to reduce spending on addictive substances. (See question 10 at the end of this chapter for further discussion.)

One of the major difficulties with regulating addictive substances comes because of the patterns of substitution among them. Many drugs appear to be close substitutes rather than complements. As a result, experts caution, raising the price of one substance may drive users to other harmful substances. For example, states that have criminal penalties for marijuana use tend to have higher teenage consumption of alcohol and tobacco.

Clearly, social policy toward addictive substances raises extremely complex issues. But the economic theory of demand provides some important insights into the impacts of alternative approaches. First, it suggests that raising the prices of harmful addictive substances can reduce the number of casual users who will be attracted into the market. Second, it cautions us that many of the negative consequences of illegal drugs result from the prohibition of addictive substances rather than from their consumption per se. Many thoughtful observers conclude with the paradoxical observation that the overall costs of addictive substances—to users, to other people, and to the ravaged inner cities in which the drug trade thrives—would be lower if government prohibitions were relaxed and the resources currently devoted to supply restrictions were instead put into treatment and counseling.

THE PARADOX OF VALUE

More than two centuries ago, in *The Wealth of Nations,* Adam Smith posed the paradox of value:

> Nothing is more useful than water; but it will scarce purchase anything. A diamond, on the contrary, has scarce any value in use; but a very great quantity of other goods may frequently be had in exchange for it.

In other words, how is it that water, which is essential to life, has little value, while diamonds, which are generally used for conspicuous consumption, command an exalted price?

Although this paradox troubled Adam Smith 200 years ago, we can imagine a dialogue between a probing student and a modern-day Adam Smith as follows:

Student: How can we resolve the paradox of value?

Modern Smith: The simplest answer is that the supply and demand curves for water intersect at a very low price, while the supply and demand for diamonds yield a very high equilibrium price.

Student: But you have always taught me to go behind the curves. Why do supply and demand for water intersect at such a low price and for diamonds at a high price?

Modern Smith: The answer is that diamonds are very scarce and the cost of getting extra ones is high, while water is relatively abundant and costs little in many areas of the world.

Student: But where is utility in this picture?

Modern Smith: You are right that this answer still does not reconcile the cost information with the equally valid fact that the world's water is vastly more critical than the world's supply of diamonds. So, we need to add a second truth: The total utility from water consumption does not determine its price or demand. Rather, water's price is determined by its *marginal* utility, by the usefulness of the *last* glass of water. Because there is so much water, the last glass sells for very little. Even though the first few drops are worth life itself, the last few are needed only for watering the lawn or washing the car.

Student: Now I get it. The theory of economic value is easy to understand if you just remember that in economics the tail wags the dog. It is the tail of marginal utility that wags the dog of prices.

Modern Smith: Precisely! An immensely valuable commodity like water sells for next to nothing because its last drop is worth next to nothing.

We can restate this dialogue as follows: The more there is of a commodity, the less is the relative desirability of its last little unit. It is therefore clear why water has a low price and why an absolute necessity like air can become a free good. In both cases, it is the large quantities that pull the marginal utilities so far down and thus reduce the prices of these vital commodities.

CONSUMER SURPLUS

The paradox of value emphasizes that the recorded monetary value of a good (measured by price times quantity) may be a misleading indicator of the total economic value of that good. The measured economic value of the air we breathe is zero, yet air's contribution to welfare is immeasurably large.

The gap between the total utility of a good and its total market value is called **consumer surplus.** The surplus arises because we "receive more than we pay for" as a result of the law of diminishing marginal utility.

We have consumer surplus basically because we pay the same amount for each unit of a commodity that we buy, from the first to the last. We pay the same price for each egg or glass of water. Thus we pay for *each* unit what the *last* unit is worth. But by our fundamental law of diminishing marginal utility, the earlier units are worth more to us than the last. Thus, we enjoy a surplus of utility on each of these earlier units.

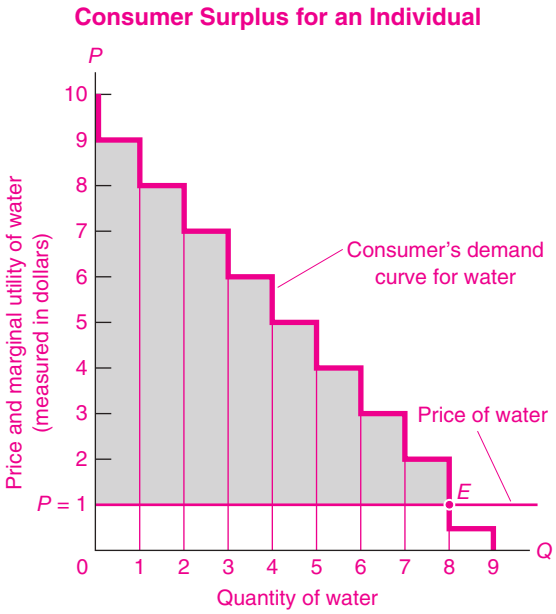


FIGURE 5-5. Because of Diminishing Marginal Utility, Consumer's Satisfaction Exceeds What Is Paid

The downward-sloping demand for water reflects the diminishing marginal utility of water. Note how much excess or surplus satisfaction occurs from the earlier units. Adding up all the blue surpluses ($8 of surplus on unit 1 + $7 of surplus on unit 2 + · · · + $1 of surplus on unit 8), we obtain the total consumer surplus of $36 on water purchases.

In the simplified case seen here, the area between the demand curve and the price line is the total consumer surplus.

Figure 5-5 illustrates the concept of consumer surplus in the case where money provides a firm measuring rod for utility. Here, an individual consumes water, which has a price of $1 per gallon. This is shown by the horizontal green line at $1 in Figure 5-5. The consumer considers how many gallon jugs to buy at that price. The first gallon is highly valuable, slaking extreme thirst, and the consumer is willing to pay $9 for it. But this first gallon costs only the market price of $1, so the consumer has gained a surplus of $8.

Consider the second gallon. This is worth $8 to the consumer, but again costs only $1, so the surplus is $7. And so on down to the ninth gallon, which is worth only 50 cents to the consumer, and so it is not bought. The consumer equilibrium comes at point *E*, where 8 gallons of water are bought at a price of $1 each.

But here we make an important discovery: Even though the consumer has paid only $8, the total

Consumer Surplus for a Market

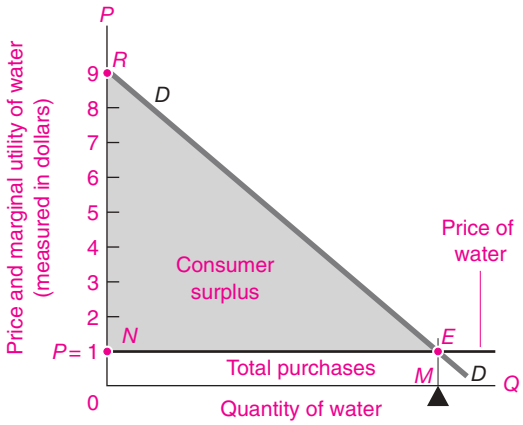


FIGURE 5-6. Total Consumer Surplus Is the Area under the Demand Curve and above the Price Line

The demand curve measures the amount consumers would pay for each unit consumed. Thus the total area under the demand curve (0*REM*) shows the total utility attached to the consumption of water. By subtracting the market cost of water to consumers (equal to 0*NEM*), we obtain the consumer surplus from water consumption as the blue triangle *NER*. This device is useful for measuring the benefits of public goods and the losses from monopolies and import tariffs.

value of the water is $44. This is obtained by adding up each of the marginal utility columns (= $9 + $8 + · · · + $2). Thus the consumer has gained a surplus of $36 over the amount paid.

Figure 5-5 examines the case of a single consumer purchasing water. We can also apply the concept of consumer surplus to a market as a whole. The market demand curve in Figure 5-6 is the horizontal summation of the individual demand curves. The logic of the individual consumer surplus carries over to the market as a whole. The area of the market demand curve above the price line, shown as *NER* in Figure 5-6, represents the total consumer surplus.

Because consumers pay the price of the last unit for all units consumed, they enjoy a surplus of utility over cost. Consumer surplus measures the extra value that consumers receive above what they pay for a commodity.

Applications of Consumer Surplus

The concept of consumer surplus is useful in helping evaluate many government decisions. For example, how can the government decide on the value of building a new highway or of preserving a recreation site? Suppose a new highway has been proposed. Being free to all, it will bring in no revenue. The value to users will be found in time saved or in safer trips and can be measured by the individual consumer surplus. To avoid difficult issues of interpersonal utility comparisons, we assume that there are 10,000 users, all identical in every respect.

Suppose that each individual's consumer surplus is $350 for the highway. The highway will raise consumer economic welfare if its total cost is less than $3.5 million (10,000 × $350). Economists use consumer surplus when they are performing a *cost-benefit analysis,* which attempts to determine the costs and benefits of a government program. Generally, an economist would recommend that a free road should be built if its total consumer surplus exceeds its costs. Similar analyses have been used for environmental questions such as whether to preserve wilderness areas for recreation or whether to require new pollution-abatement equipment.

The concept of consumer surplus also points to the enormous privilege enjoyed by citizens of modern societies. Each of us enjoys a vast array of enormously valuable goods that can be bought at low prices. This is a humbling thought. If you know someone who is bragging about his economic productivity, or explaining how high her real wages are, suggest a moment of reflection. If such people were transported with their specialized skills to an uninhabited desert island, how much would their wages buy? Indeed, without capital machinery, without the cooperation of others, and without the technological knowledge which each generation inherits from the past, how much could any of us produce? It is only too clear that all of us reap the benefits of an economic world we never made. As the great British sociologist L. T. Hobhouse said:

> The organizer of industry who thinks that he has "made" himself and his business has found a whole social system ready to his hand in skilled workers, machinery, a market, peace and order—a vast apparatus and a pervasive atmosphere, the joint creation of millions of men and scores of generations. Take away the whole social factor and we [are] but . . . savages living on roots, berries, and vermin.

Now that we have surveyed the essentials of demand, we move on to costs and supply.

SUMMARY

1. Market demands or demand curves are explained as stemming from the process of individuals' choosing their most preferred bundle of consumption goods and services.
2. Economists explain consumer demand by the concept of utility, which denotes the relative satisfaction that a consumer obtains from using different commodities. The additional satisfaction obtained from consuming an additional unit of a good is given the name *marginal utility,* where "marginal" means the extra or incremental utility. The law of diminishing marginal utility states that as the amount of a commodity consumed increases, the marginal utility of the last unit consumed tends to decrease.
3. Economists assume that consumers allocate their limited incomes so as to obtain the greatest satisfaction or utility. To maximize utility, a consumer must satisfy the *equimarginal principle* that the marginal utilities of the last dollar spent on each and every good must be equal.

   Only when the marginal utility per dollar is equal for apples, bacon, coffee, and everything else will the consumer attain the greatest satisfaction from a limited dollar income. But be careful to note that the marginal utility of a $50-per-ounce bottle of perfume is not equal to the marginal utility of a 50-cent glass of cola. Rather, their marginal utilities divided by price per unit are all equal in the consumer's optimal allocation. That is, their marginal utilities per last dollar, MU/P, are equalized.
4. Equal marginal utility or benefit per unit of resource is a fundamental rule of choice. Take any scarce resource, such as time. If you want to maximize the value or utility of that resource, make sure that the marginal benefit per unit of the resource is equalized in all uses.
5. The market demand curve for all consumers is derived by adding horizontally the separate demand curves of each consumer. A demand curve can shift for many reasons. For example, a rise in income will normally shift DD rightward, thus increasing demand; a rise in the price of a substitute good (e.g., chicken for beef) will also create a similar upward shift in demand; a rise in the price of a complementary good (e.g., hamburger buns for beef) will in turn cause the DD curve to shift downward and leftward. Still other factors—changing tastes, population, or expectations—can affect demand.
6. We can gain added insight into the factors that cause downward-sloping demand by separating the effect of a price rise into substitution and income effects. (*a*) The substitution effect occurs when a higher price leads to substitution of other goods to meet satisfactions; (*b*) the income effect means that a price increase lowers real income and thereby reduces the desired consumption of most commodities. For most goods, substitution and income effects of a price increase reinforce one another and lead to the law of downward-sloping demand. We measure the quantitative responsiveness of demand to income by the income elasticity, which is the percentage change in quantity demanded divided by the percentage change in income.
7. Remember that it is the tail of marginal utility that wags the market dog of prices. This point is emphasized by the concept of *consumer surplus.* We pay the same price for the last quart of milk as for the first. But, because of the law of diminishing marginal utility, marginal utilities of earlier units are greater than that of the last unit. This means that we would have been willing to pay more than the market price for each of the earlier units. The excess of total value over market value is called consumer surplus. Consumer surplus reflects the benefit we gain from being able to buy all units at the same low price. In simplified cases, we can measure consumer surplus as the area between the demand curve and the price line. It is a concept relevant for many public decisions—such as deciding when the community should incur the heavy expenses of a road or bridge or set aside land for a wilderness area.

CONCEPTS FOR REVIEW

utility, marginal utility
utilitarianism
law of diminishing marginal utility
demand shifts from income and other sources
ordinal utility
equimarginal principle: $MU_1/P_1 = MU_2/P_2 = \cdots = MU$ per \$ of income
market demand vs. individual demand
income elasticity
substitutes, complements, independent goods
substitution effect and income effect
merit goods, demerit goods
paradox of value
consumer surplus

FURTHER READING AND INTERNET WEBSITES

Further Reading

An advanced treatment of consumer theory can be found in intermediate textbooks; see the Further Reading section in Chapter 3 for some good sources.

Utilitarianism was introduced in Jeremy Bentham, *An Introduction to the Principles of Morals* (1789).

An interesting survey of psychology and economics is contained in Matthew Rabin, "Psychology and Economics," *Journal of Economic Literature,* March 1998, while serious students of the subject may want to read Colin Camerer, George Loewenstein, and Matthew Rabin, eds., *Advances in Behavioral Economics* (Princeton University Press, Princeton, N.J., 2003).

Consumers often need help in judging the utility of different products. Look at *Consumer Reports* for articles that attempt to rate products. They sometimes rank products as "Best Buys," which might mean the most utility per dollar of expenditure.

Jeffrey A. Miron and Jeffrey Zwiebel, "The Economic Case against Drug Prohibition," *Journal of Economic Perspectives,* Fall 1995, pp. 175–192, is an excellent nontechnical survey of the economics of drug prohibition.

Websites

Data on total personal consumption expenditures for the United States are provided at the website of the Bureau of Economic Analysis, *www.bea.doc.gov.* Data on family budgets are contained in Bureau of Labor Statistics, *Consumer Expenditures,* available at *www.bls.gov.*

Practical guides for consumers are provided at the government site *www.consumer.gov.* The organization Public Citizen lobbies in Washington "for safer drugs and medical devices, cleaner and safer energy sources, a cleaner environment, fair trade, and a more open and democratic government." Its website at *www.citizen.org* contains articles on many consumer, labor, and environmental issues.

You can read the Nobel lectures of laureates Akerlof, Kahneman, and Smith, with their views on behavioral economics, at *nobelprize.org/nobel_prizes/economics/laureates/.*

QUESTIONS FOR DISCUSSION

1. Explain the meaning of utility. What is the difference between total utility and marginal utility? Explain the law of diminishing marginal utility and give a numerical example.
2. Each week, Tom Wu buys two hamburgers at \$2 each, eight cokes at \$0.50 each, and eight slices of pizza at \$1 each, but he buys no hot dogs at \$1.50 each. What can you deduce about Tom's marginal utility for each of the four goods?
3. Which pairs of the following goods would you classify as complementary, substitute, or independent goods: beef, ketchup, lamb, cigarettes, gum, pork, radio, television, air travel, bus travel, taxis, and paperbacks? Illustrate the resulting shift in the demand curve for one good when the price of another good goes up. How would a change in income affect the demand curve for air travel? The demand curve for bus travel?

4. Why is it wrong to say, "Utility is maximized when the marginal utilities of all goods are exactly equal"? Correct the statement and explain.
5. Here is a way to think about consumer surplus as it applies to movies:
   a. How many movies did you watch last year?
   b. How much in total did you pay to watch movies last year?
   c. What is the *maximum* you would pay to see the movies you watched last year?
   d. Calculate **c** minus **b.** That is your consumer surplus from movies.
6. Consider the following table showing the utility of different numbers of days skied each year:

Number of days skied	Total utility ($)
0	0
1	70
2	110
3	146
4	176
5	196
6	196

   Construct a table showing the marginal utility for each day of skiing. Assuming that there are 1 million people with preferences shown in the table, draw the market demand curve for ski days. If lift tickets cost $40 per day, what are the equilibrium price and quantity of days skied?
7. For each of the commodities in Table 5-2, calculate the impact of a doubling of price on quantity demanded. Similarly, for the goods in Table 5-3, what would be the impact of a 50 percent increase in consumer incomes?
8. As you add together the identical demand curves of more and more people (in a way similar to the procedure in Figure 5-2), the market demand curve becomes flatter and flatter on the same scale. Does this fact indicate that the elasticity of demand is becoming larger and larger? Explain your answer carefully.
9. An interesting application of supply and demand to addictive substances compares alternative techniques for supply restriction. For this problem, assume that the demand for addictive substances is inelastic.
   a. One approach (used today for heroin and cocaine and for alcohol during Prohibition) is to reduce supply at the nation's borders. Show how this raises price and increases the total income of the suppliers in the drug industry.
   b. An alternative approach (followed today for tobacco and alcohol) is to tax the goods heavily. Using the tax apparatus developed in Chapter 4, show how this reduces the total income of the suppliers in the drug industry.
   c. Comment on the difference between the two approaches.
10. Demand may be price-elastic for casual users of drugs—ones who are not addicted or for whom substitute products are readily available. In this case, restrictions or price increases will have a significant impact on use. Draw a supply and demand diagram like Figure 5-4 where the demand curve is price-elastic. Show the effect of a steep tax on quantity demanded. Show that, because demand is price-elastic, total spending on drugs with restrictions will fall. Explain why this analysis would support the argument of those who would severely limit the availability of addictive substances.
11. Suppose you are very rich and very fat. Your doctor has advised you to limit your food intake to 2000 calories per day. What is your consumer equilibrium for food consumption?
12. *Numerical problem on consumer surplus:* Assume that the demand for travel over a bridge takes the form $Y = 1{,}000{,}000 - 50{,}000P$, where Y is the number of trips over the bridge and P is the bridge toll (in dollars).
   a. Calculate the consumer surplus if the bridge toll is $0, $1, and $20.
   b. Assume that the cost of the bridge is $1,800,000. Calculate the toll at which the bridge owner breaks even. What is the consumer surplus at the break-even toll?
   c. Assume that the cost of the bridge is $8 million. Explain why the bridge should be built even though there is no toll that will cover the cost.

Appendix 5

GEOMETRICAL ANALYSIS OF CONSUMER EQUILIBRIUM

An alternative and more advanced approach to deriving demand curves uses the approach called indifference curves. This appendix derives the major conclusions of consumer behavior with this new tool.

THE INDIFFERENCE CURVE

Start by assuming that you are a consumer who buys different combinations of two commodities, say, food and clothing, at a given set of prices. For each combination of the two goods, assume that you prefer one to the other or are indifferent between the pair. For example, when asked to choose between combination A of 1 unit of food and 6 units of clothing and combination B of 2 units of food and 3 of clothing, you might (1) prefer A to B, (2) prefer B to A, or (3) be indifferent between A and B.

Now suppose that A and B are equally good in your eyes—that you are indifferent as to which of them you receive. Let us consider some other combinations of goods about which you are likewise indifferent, as listed in the table for Figure 5A-1.

Figure 5A-1 shows these combinations diagrammatically. We measure units of clothing on one axis and units of food on the other. Each of our four combinations of goods is represented by its point, *A, B, C, D*. But these four are by no means the only combinations among which you are indifferent. Another batch, such as 1½ units of food and 4 of clothing, might be ranked as equal to *A, B, C,* or *D,* and there are many others not shown. The curved contour of Figure 5A-1, linking up the four points, is an **indifference curve.** The points on the curve represent consumption bundles among which the consumer is indifferent; all are equally desirable.

Law of Substitution

Indifference curves are drawn as bowl-shaped, or convex to the origin. Hence, as you move downward and to the right along the curve—a movement that implies increasing the quantity of food and reducing the units of clothing—the curve becomes flatter. The curve is drawn in this way to illustrate a property that seems most often to hold true in reality and which we call the law of substitution:

The scarcer a good, the greater its relative substitution value; its marginal utility rises relative to the marginal utility of the good that has become plentiful.

Thus, in going from *A* to *B* in Figure 5A-1, you would swap 3 of your 6 clothing units for 1 extra food unit. But from *B* to *C,* you would sacrifice only 1 unit of your remaining clothing supply to obtain a third food unit—a 1-for-1 swap. For a fourth unit of food, you would sacrifice only ½ unit from your dwindling supply of clothing.

If we join the points *A* and *B* of Figure 5A-1, we find that the slope of the resulting line (neglecting its negative sign) has a value of 3. Join *B* and *C,* and the slope is 1; join *C* and *D,* and the slope is ½. These figures—3, 1, ½—are the *substitution ratios* (sometimes called the *marginal rates of substitution*) between the two goods. As the size of the movement along the curve becomes very small, the closer the substitution ratio comes to the actual slope of the indifference curve.

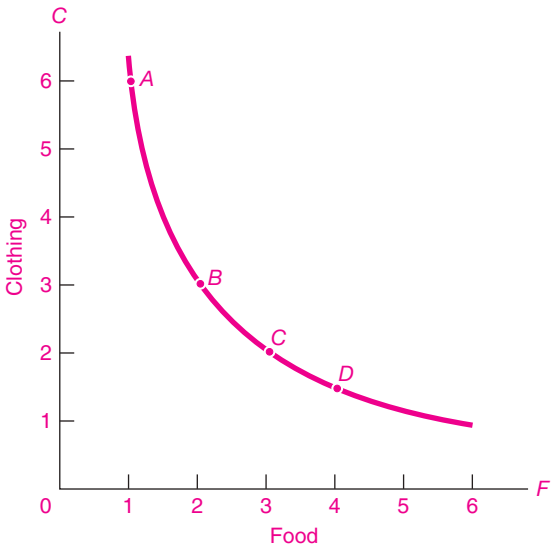
The slope of the indifference curve is the measure of the goods' relative marginal utilities, or of the substitution terms on which—for very small changes—the consumer would be willing to exchange a little less of one good in return for a little more of the other.

An indifference curve that is convex in the manner of Figure 5A-1 conforms to the law of substitution. As the amount of food you consume goes up—and the quantity of clothing goes down—food must become relatively cheaper in order for you to be persuaded to take a little extra food in exchange for a little sacrifice of clothing. The precise shape and slope of an indifference curve will, of course, vary from one consumer to the next, but the typical shape will take the form shown in Figures 5A-1 and 5A-2.

The Indifference Map

The table in Figure 5A-1 is one of an infinite number of possible tables. We could start with a more preferred consumption situation and list some of the different combinations that would bring the consumer this higher level of satisfaction. One such table might have begun with 2 food units and 7 clothing units;

A Consumer's Indifference Curve



Indifference Combinations

	Food	Clothing
A	1	6
B	2	3
C	3	2
D	4	$1\frac{1}{2}$

FIGURE 5A-1. Indifference Curve for a Pair of Goods

Getting more of one good compensates for giving up some of the other. The consumer likes situation *A* exactly as much as *B, C,* or *D*. The food-clothing combinations that yield equal satisfaction are plotted as a smooth indifference curve. This is convex from below in accord with the law of substitution, which says that as you get more of a good, its substitution ratio, or the indifference curve's slope, diminishes.

another with 3 food units, 8 clothing units. Each table could be portrayed graphically, each with a corresponding indifference curve.

Figure 5A-2 shows four such curves; the curve from Figure 5A-1 is labeled U_3. This diagram is analogous to a geographic contour map. A person who walks along the path indicated by a particular height contour on such a map is neither climbing nor descending; similarly, the consumer who moves from one position to another along a single indifference curve enjoys neither increasing nor decreasing satisfaction from the change in consumption. Only a few of the possible indifference curves are shown in Figure 5A-2.

Note that as we increase both goods and thus move in a northeasterly direction across this map, we are crossing successive indifference curves; hence, we are reaching higher and higher levels of satisfaction (assuming that the consumer gets greater satisfaction from receiving increased quantities of both goods). Curve U_3 stands for a higher level of satisfaction than

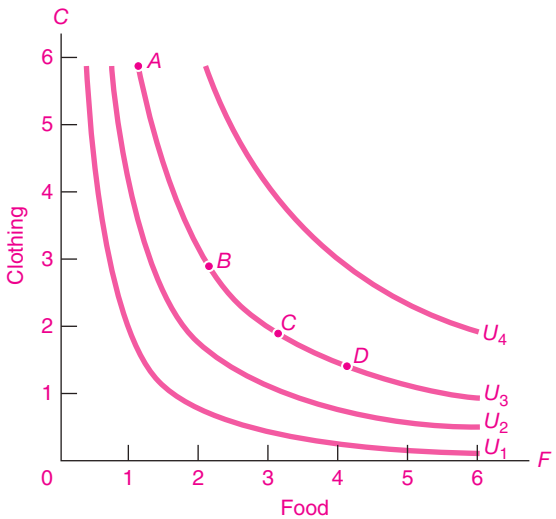
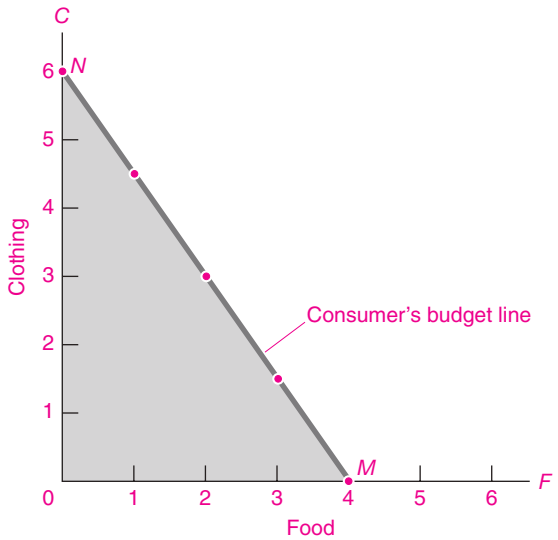


FIGURE 5A-2. A Family of Indifference Curves

The curves labeled U_1, U_2, U_3, and U_4 represent indifference curves. Which indifference curve is most preferred by the consumer?

A Consumer's Budget Line



Alternative Consumption Possibilities

	Food	Clothing
M	4	0
	3	1½
	2	3
	1	4½
N	0	6

FIGURE 5A-3. Income Constrains Consumer Spending

The budget limit on expenditures can be seen in a numerical table. The total cost of each budget (reckoned as $\$1.50F + \$1C$) adds up to exactly \$6 of income. We can plot the budget constraint as a straight line whose absolute slope equals the P_F/P_C ratio. *NM* is the consumer's budget line. When income is \$6, with food and clothing prices \$1.50 and \$1, the consumer can choose any point on this budget line. (Why is its slope \$1.50/\$1 = ½?)

U_2; U_4, for a higher level of satisfaction than U_3; and so forth.

BUDGET LINE OR BUDGET CONSTRAINT

Now let us set a particular consumer's indifference map aside for a moment and give the consumer a fixed income. He has, say, \$6 per day to spend, and he is confronted with fixed prices for each food and clothing unit—\$1.50 for food, \$1 for clothing. It is clear that he could spend his money on any one of a variety of alternative combinations of food and clothing. At one extreme, he could buy 4 food units and no clothing; at the other, 6 clothing units and no food. The table with Figure 5A-3 illustrates some of the possible ways in which he could allocate his \$6.

Figure 5A-3 plots five of these possibilities. Note that all the points lie on a straight line, labeled *NM*. Moreover, any other attainable point, such as 3⅓ food units and 1 clothing unit, lies on *NM*. The straight budget line *NM* sums up all the possible combinations of the two goods that would just exhaust the consumer's income.[1] The slope of *NM* (neglecting its sign) is ⅔, which is the ratio of the food price to the clothing price. The meaning of the slope is that, given these prices, every time our consumer gives up 3 clothing units (thereby dropping down 3 vertical units on the diagram), he can gain 2 units of food (i.e., move right 2 horizontal units).

We call *NM* the consumer's **budget line** or **budget constraint.**

[1] This is so because, if we designate quantities of food and clothing bought as *F* and *C*, respectively, total expenditure on food must be $\$1.50F$ and total expenditure on clothing, $\$1C$. If daily income and expenditure are \$6, the following equation must hold: $\$6 = \$1.50F + \$1C$. This is a linear equation, the equation of the budget line *NM*. Note:

Arithmetic slope of *NM* = \$1.50 ÷ \$1
= price of food ÷ price of clothing

THE EQUILIBRIUM POSITION OF TANGENCY

Now we are ready to put our two parts together. The axes of Figure 5A-3 are the same as those of Figures 5A-1 and 5A-2. We can superimpose the blue budget line *NM* upon this green consumer indifference map, as shown in Figure 5A-4. The consumer is free to move anywhere along *NM*. Positions to the right and above *NM* are not allowed because they require more than $6 of income; positions to the left and below *NM* are irrelevant because the consumer is assumed to spend the full $6.

Where will the consumer move? Obviously, to that point which yields the greatest satisfaction—that is, to the highest possible indifference curve—which in this case must be at the green point *B*. At *B*, the budget line just touches, but does not cross, the indifference curve U_3. At this point of tangency, where the budget line just kisses but does not cross an indifference contour, is found the highest utility contour the consumer can reach.

Geometrically, the consumer is at equilibrium where the slope of the budget line (which is equal to the ratio of food to clothing prices) is exactly equal to the slope of the indifference curve (which is equal to the ratio of the marginal utilities of the two goods).

Consumer equilibrium is attained at the point where the budget line is tangent to the highest indifference curve. At that point, the consumer's substitution ratio is just equal to the slope of the budget line.

Put differently, the substitution ratio, or the slope of the indifference curve, is the ratio of the marginal utility of food to the marginal utility of clothing. So our tangency condition is just another way of stating that the ratio of prices must be equal to the ratio of marginal utilities; in equilibrium, the consumer is getting the same marginal utility from the last penny spent on food as from the last penny spent on clothing. Therefore, we can derive the following equilibrium condition:

$$\frac{P_F}{P_C} = \text{substitution ratio} = \frac{MU_F}{MU_C}$$

This is exactly the same condition as we derived for utility theory in the main part of this chapter.

Consumer's Equilibrium

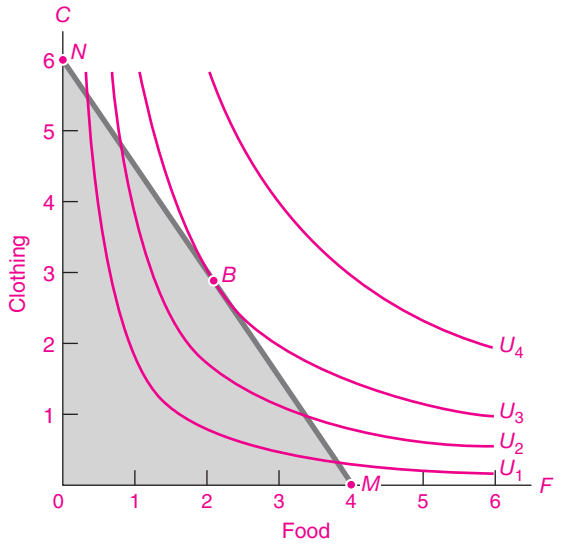


FIGURE 5A-4. Consumer's Most Preferred and Feasible Consumption Bundle Is Attained at *B*

We now combine the budget line and indifference contours in one diagram. The consumer reaches the highest indifference curve attainable with fixed income at point *B*, which is the tangency of the budget line with the highest indifference curve. At tangency point *B*, substitution ratio equals price ratio P_F/P_C. This means that all goods' marginal utilities are proportional to their prices, with the marginal utility of the last dollar spent on every good being equalized.

CHANGES IN INCOME AND PRICE

Two important applications of indifference curves are frequently used to consider the effects of (*a*) a change in money income and (*b*) a change in the price of one of the two goods.

Income Change

Assume, first, that the consumer's daily income is halved while the two prices remain unchanged. We could prepare another table, similar to the table for Figure 5A-3, showing the new consumption possibilities. Plotting these points on a diagram such as Figure 5A-5, we should find that the new budget line occupies the position $N'M'$ in Figure 5A-5. The line has made a parallel shift inward.[2] The consumer is

[2] The equation of the new $N'M'$ budget line is now $\$3 = \$1.50F + \$1C$.

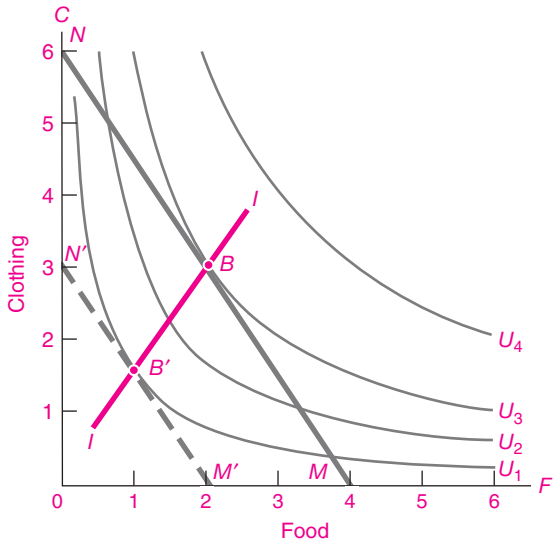


FIGURE 5A-5. Effect of Income Change on Equilibrium

An income change shifts the budget line in a parallel way. Thus, halving income to \$3 shifts NM to $N'M'$, moving equilibrium to B'. (Show what raising income to \$8 would do to equilibrium. Estimate where the new tangency point would come.)

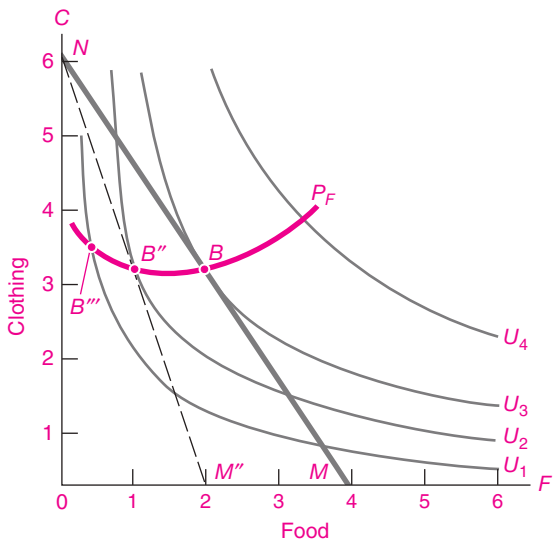


FIGURE 5A-6. Effect of Price Change on Equilibrium

A rise in the price of food makes the budget line pivot on N, rotating from NM to NM''. The new tangency equilibrium is at B'', where there is definitely less food consumed but clothing consumption may either go up or down.

now free to move only along this new (and lower) budget line; to maximize satisfaction, he will move to the highest attainable indifference curve, or to point B'. A tangency condition for consumer equilibrium applies here as before.

Single Price Change

Now return our consumer to his previous daily income of \$6, but assume that the price of food rises from \$1.50 to \$3 while the price of clothing is unchanged. Again we must examine the change in the budget line. This time we find that it has pivoted on point N and is now NM'', as illustrated in Figure 5A-6.[3]

The common sense of such a shift is clear. Since the price of clothing is unchanged, point N is just as available as it was before. But since the price of food has risen, point M (which represents 4 food units) is no longer attainable. With food costing \$3 per unit, only 2 units can now be bought with a daily income of \$6. So the new budget line still passes through N, but it must pivot at N and pass through M'', which is to the left of M.

Equilibrium is now at B'', and we have a new tangency point. Higher food price has definitely reduced food consumption, but clothing consumption may move in either direction. To clinch your understanding, work out the cases of an increase in income and a fall in the price of clothing or food.

DERIVING THE DEMAND CURVE

We are now in a position to derive the demand curve. Look carefully at Figure 5A-6. Note that as we increased the price of food from \$1.50 per unit to \$3 per unit, we kept other things constant. Tastes as represented by the indifference curves did not change, and money income and the price of clothing stayed constant. Therefore, we are in the ideal position to trace the demand curve for food. At a price of \$1.50, the consumer buys 2 units of food, shown as equilibrium point B. When the price rises to \$3 per unit, the food purchased is 1 unit, at equilibrium point

[3] The budget equation of NM'' is now $\$6 = \$3F + \$1C$.

B''. If you draw in the budget line corresponding to a price of \$6 per unit of food, the equilibrium occurs at point B''', and food purchases are 0.45 unit.

Now plot the price of food against the purchases of food, again holding other things constant. You will have derived a neat downward-sloping demand curve from indifference curves. Note that we have done this without ever needing to mention the term "utility"—basing the derivation solely on measurable indifference curves.

SUMMARY TO APPENDIX

1. An indifference curve depicts the points of equally desirable consumption bundles. The indifference contour is usually drawn convex (or bowl-shaped) in accordance with the law of diminishing relative marginal utilities.
2. When a consumer has a fixed money income, all of which she spends, and is confronted with market prices of two goods, she is constrained to move along a straight line called the budget line or budget constraint. The line's slope will depend on the ratio of the two market prices; how far out it lies will depend on the size of her income.
3. The consumer will move along this budget line until reaching the highest attainable indifference curve. At this point, the budget line will touch, but not cross, an indifference curve. Hence, equilibrium is at the point of tangency, where the slope of the budget line (the ratio of the prices) exactly equals the slope of the indifference curve (the substitution ratio or the ratio of the marginal utilities of the two goods). This gives additional proof that, in equilibrium, marginal utilities are proportional to prices.
4. A fall in income will move the budget line inward in a parallel fashion, usually causing less of both goods to be bought. A change in the price of one good alone will, other things being constant, cause the budget line to pivot so as to change its slope. After a price or income change, the consumer will again attain a new tangency point of highest satisfaction. At every point of tangency, the marginal utility per dollar is equal for every good. By comparing the new and old equilibrium points, we trace the usual downward-sloping demand curve.

CONCEPTS FOR REVIEW

indifference curves
slope or substitution ratio
budget line or budget constraint
convexity of indifference curves and law of diminishing relative marginal utilities
optimal tangency condition: P_F/P_C = substitution ratio = MU_F/MU_C

QUESTIONS FOR DISCUSSION

1. Draw the indifference curves (*a*) between complementary goods like left shoes and right shoes and (*b*) between perfect substitutes like two bottles of cola sitting next to each other in a store.
2. Consider noodles and yachts. Draw a set of indifference curves and budget lines like those in Figure 5A-5 which show noodles as an inferior good and yachts as a "luxury" with an income elasticity greater than 1.

CHAPTER

Production and Business Organization

6

The business of America is business.

Calvin Coolidge

Before we can eat our daily bread, someone must bake it. Similarly, the economy's ability to build cars, generate electricity, write computer programs, and deliver the multitude of goods and services that are in our gross domestic product depends upon our productive capacity. Productive capacity is determined by the size and quality of the labor force, by the quantity and quality of the capital stock, by the nation's technical knowledge along with the ability to use that knowledge, and by the nature of public and private institutions. Why are living standards high in North America? Low in tropical Africa? For answers, we should look to how well the machine of production is running.

Our goal is to understand how market forces determine the supply of goods and services. Over the next three chapters we will lay out the essential concepts of production, cost, and supply and show how they are linked. We first explore the fundamentals of production theory, showing how firms transform inputs into desirable outputs. Production theory also helps us understand why productivity and living standards have risen over time and how firms manage their internal activities.

A. THEORY OF PRODUCTION AND MARGINAL PRODUCTS

BASIC CONCEPTS

A modern economy has an enormously varied set of productive activities. A farm takes fertilizer, seed, land, and labor and turns them into wheat or corn. Modern factories take inputs such as energy, raw materials, computerized machinery, and labor and use them to produce tractors, DVDs, or tubes of toothpaste. An airline takes airplanes, fuel, labor, and computerized reservation systems and provides passengers with the ability to travel quickly through its network of routes.

The Production Function

We have spoken of inputs like land and labor and outputs like wheat and toothpaste. But if you have a fixed amount of inputs, how much output can you get? On any day, given the available technical knowledge, land, machinery, and so on, only a certain quantity of tractors or toothpaste can be obtained from a

given amount of labor. The relationship between the amount of input required and the amount of output that can be obtained is called the *production function*.

The **production function** specifies the maximum output that can be produced with a given quantity of inputs. It is defined for a given state of engineering and technical knowledge.

An important example is the production function for generating electricity. Visualize it as a book with technical specifications for different kinds of plants. One page is for gas turbines, showing their inputs (initial capital cost, fuel consumption, and the amount of labor needed to run the turbine) and their outputs (amount of electricity generated). The next page shows inputs and outputs of coal-fired generating plants. Yet other pages describe nuclear power plants, solar power stations, and so forth. Taken together, they constitute the production function for electricity generation.

Note that our definition assumes that firms always strive to produce efficiently. In other words, they always attempt to produce the maximum level of output for a given dose of inputs.

Consider the humble task of ditchdigging. Outside our windows in America, we see a large and expensive tractor, driven by one person with another to supervise. This team can easily dig a trench 5 feet deep and 50 feet long in 2 hours. When we visit Africa, we see 50 laborers armed only with picks. The same trench might take an entire day. These two techniques—one capital-intensive and the other labor-intensive—are part of the production function for ditchdigging.

There are literally millions of different production functions—one for each and every product or service. Most of them are not written down but are in people's minds. In areas of the economy where technology is changing rapidly, like computer software and biotechnology, production functions may become obsolete soon after they are used. And some, like the blueprints of a medical laboratory or cliff house, are specially designed for a specific location and purpose and would be useless anywhere else. Nevertheless, the concept of a production function is a useful way of describing the productive capabilities of a firm.

Total, Average, and Marginal Product

Starting with a firm's production function, we can calculate three important production concepts: total, average, and marginal product. We begin by computing the total physical product, or **total product,** which designates the total amount of output produced, in physical units such as bushels of wheat or number of sneakers. Figure 6-1(*a*) on page 109 and column (2) of Table 6-1 on page 110 illustrate the concept of total product. For this example, they show how total product responds as the amount of labor applied is increased. The total product starts at zero for zero labor and then increases as additional units of labor are applied, reaching a maximum of 3900 units when 5 units of labor are used.

Once we know the total product, it is easy to derive an equally important concept, the marginal product. Recall that the term "marginal" means "extra."

The **marginal product** of an input is the extra output produced by 1 additional unit of that input while other inputs are held constant.

For example, assume that we are holding land, machinery, and all other inputs constant. Then labor's marginal product is the extra output obtained by adding 1 unit of labor. The third column of Table 6-1 calculates the marginal product. The marginal product of labor starts at 2000 for the first unit of labor and then falls to only 100 units for the fifth unit. Marginal product calculations such as this are crucial for understanding how wages and other factor prices are determined.

The final concept is the **average product,** which equals total output divided by total units of input. The fourth column of Table 6-1 shows the average product of labor as 2000 units per worker with one worker, 1500 units per worker with two workers, and so forth. In this example, average product falls through the entire range of increasing labor input.

Figure 6-1 plots the total and marginal products from Table 6-1. Study this figure to make sure you understand that the blocks of marginal products in (*b*) are related to the changes in the total product curve in (*a*).

The Law of Diminishing Returns

Using production functions, we can understand one of the most famous laws in all economics, the law of diminishing returns:

Under the **law of diminishing returns,** a firm will get less and less extra output when it adds additional

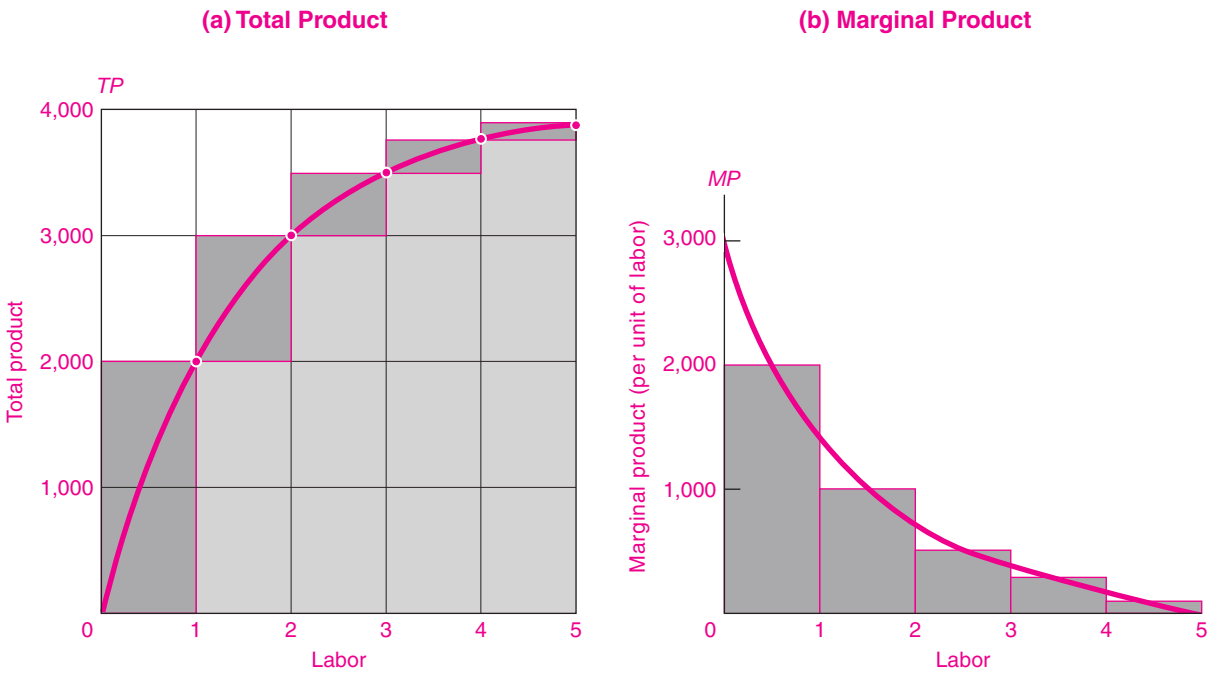


FIGURE 6-1. Marginal Product Is Derived from Total Product

Diagram **(a)** shows the total product curve rising as additional inputs of labor are added, holding other things constant. However, total product rises by smaller and smaller increments as additional units of labor are added (compare the increments of the first and the fifth worker). By smoothing between points, we get the green-colored total product curve.

Diagram **(b)** shows the declining steps of marginal product. Make sure you understand why each dark rectangle in **(b)** is equal to the equivalent dark rectangle in **(a)**. The area in **(b)** under the green-colored marginal product curve (or the sum of the dark rectangles) adds up to the total product in **(a)**.

units of an input while holding other inputs fixed. In other words, the marginal product of each unit of input will decline as the amount of that input increases, holding all other inputs constant.

The law of diminishing returns expresses a very basic relationship. As more of an input such as labor is added to a fixed amount of land, machinery, and other inputs, the labor has less and less of the other factors to work with. The land gets more crowded, the machinery is overworked, and the marginal product of labor declines.

Table 6-1 illustrates the law of diminishing returns. Given fixed land and other inputs, we see that there is zero total output of corn with zero inputs of labor. When we add our first unit of labor to the same fixed amount of land, we observe that 2000 bushels of corn are produced.

In our next stage, with 2 units of labor and fixed land, output goes to 3000 bushels. Hence, the second unit of labor adds only 1000 bushels of additional output. The third unit of labor has an even lower marginal product than does the second, and the fourth unit adds even less. Table 6-1 thus illustrates the law of diminishing returns.

Figure 6-1 also illustrates the law of diminishing returns for labor. Here we see that the marginal product curve in (*b*) declines as labor inputs increase, which is the precise meaning of diminishing returns. In Figure 6-1(*a*), diminishing returns are seen as a concave or dome-shaped total product curve.

(1) Units of labor input	(2) Total product	(3) Marginal product	(4) Average product
0	0		
		2,000	
1	2,000		2,000
		1,000	
2	3,000		1,500
		500	
3	3,500		1,167
		300	
4	3,800		950
		100	
5	3,900		780

TABLE 6-1. Total, Marginal, and Average Product

The table shows the total product that can be produced for different inputs of labor when other inputs (capital, land, etc.) and the state of technical knowledge are unchanged. From total product, we can derive important concepts of marginal and average products.

What is true for labor is also true for any other input. We can interchange land and labor, now holding labor constant and varying land. We can calculate the marginal product of each input (labor, land, machinery, water, fertilizer, etc.), and the marginal product would apply to any output (wheat, corn, steel, soybeans, and so forth). We would find that other inputs also tend to show the law of diminishing returns.

Diminishing Returns in Farm Experiments

The law of diminishing returns is often observed in agriculture. As Farmer Tilly adds more labor, the fields will be more thoroughly seeded and weeded, irrigation ditches will be neater, and scarecrows better oiled. At some point, however, the additional labor becomes less and less productive. The third hoeing of the field or the fourth oiling of the machinery adds little to output. Eventually, output grows very little as more people crowd onto the farm; too many tillers spoil the crop.

Agricultural experiments are one of the most important kinds of technological research. These techniques have been used for over a century to test different seeds, fertilizers, and other combinations of inputs in a successful effort to raise agricultural productivity. Figure 6-2 shows the results of an experiment in which different doses of phosphorus fertilizer were applied on two different plots, holding constant land area, nitrogen fertilizer, labor, and other inputs. Real-world experiments are complicated by "random errors"—in this case, due primarily to differences in soils. You can see that diminishing returns set in quickly after about 100 pounds of phosphorus per acre. Indeed, beyond an input level of around 300 pounds per acre, the marginal product of additional phosphorus fertilizer is negative.

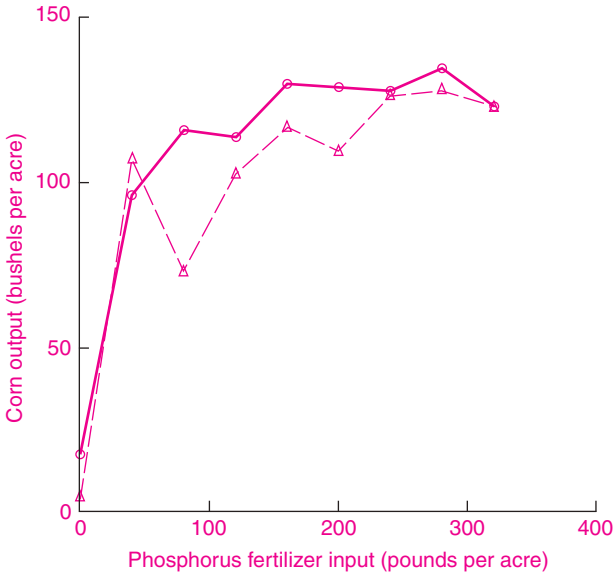


FIGURE 6-2. Diminishing Returns in Corn Production

Agricultural researchers experimented with different doses of phosphorus fertilizer on two different plots to estimate the production function for corn in western Iowa. In conducting the experiment, they were careful to hold constant other things such as nitrogen fertilizer, water, and labor inputs. Because of variations in soils and microclimate, even the most careful scientist cannot prevent some random variation, which accounts for the jagged nature of the lines. If you fit a smooth curve to the data, you will see that the relationship displays diminishing returns for every dose and that marginal product becomes negative for a phosphate input of around 300.

Source: Earl O. Heady, John T. Pesek, and William G. Brown, *Crop Response Surfaces and Economic Optima in Fertilizer Use* (Agricultural Experiment Station, Iowa State College, Ames, Iowa, 1955), table A-15.

Diminishing returns are a key factor in explaining why many countries in Asia are so poor. Living standards in crowded Rwanda or Bangladesh are low because there are so many workers per acre of land and not because farmers are ignorant or fail to respond to economic incentives.

We can also use the example of studying to illustrate the law of diminishing returns. You might find that the first hour of studying economics on a given day is productive—you learn new laws and facts, insights and history. The second hour might find your attention wandering a bit, with less learned. The third hour might show that diminishing returns have set in with a vengeance, and by the next day the third hour is a blank in your memory. Does the law of diminishing returns suggest why the hours devoted to studying should be spread out rather than crammed into the day before exams?

The law of diminishing returns is a widely observed empirical regularity rather than a universal truth like the law of gravity. It has been found in numerous empirical studies, but exceptions have also been uncovered. Moreover, diminishing returns might not hold for all levels of production. The very first inputs of labor might actually show increasing marginal products, since a minimum amount of labor may be needed just to walk to the field and pick up a shovel. Notwithstanding these reservations, diminishing returns will prevail in most situations.

RETURNS TO SCALE

Diminishing returns and marginal products refer to the response of output to an increase of a *single* input when all other inputs are held constant. We saw that increasing labor while holding land constant would increase food output by ever-smaller increments.

But sometimes we are interested in the effect of increasing *all* inputs. For example, what would happen to wheat production if land, labor, water, and other inputs were increased by the same proportion? Or what would happen to the production of tractors if the quantities of labor, computers, robots, steel, and factory space were all doubled? These questions refer to the *returns to scale,* or the effects of scale increases of inputs on the quantity produced. Three important cases should be distinguished:

- **Constant returns to scale** denote a case where a change in all inputs leads to a proportional change in output. For example, if labor, land, capital, and other inputs are doubled, then under constant returns to scale output would also double. Many handicraft industries (such as haircutting in America or handloom operation in a developing country) show constant returns.
- **Increasing returns to scale** (also called **economies of scale**) arise when an increase in all inputs leads to a more-than-proportional increase in the level of output. For example, an engineer planning a small-scale chemical plant will generally find that increasing the inputs of labor, capital, and materials by 10 percent will increase the total output by more than 10 percent. Engineering studies have determined that many manufacturing processes enjoy modestly increasing returns to scale for plants up to the largest size used today.
- **Decreasing returns to scale** occur when a balanced increase of all inputs leads to a less-than-proportional increase in total output. In many processes, scaling up may eventually reach a point beyond which inefficiencies set in. These might arise because the costs of management or control become large. One case has occurred in electricity generation, where firms found that when plants grew too large, risks of plant failure grew too large. Many productive activities involving natural resources, such as growing wine grapes or providing clean drinking water to a city, show decreasing returns to scale.

Production shows increasing, decreasing, or constant returns to scale when a balanced increase in all inputs leads to a more-than-proportional, less-than-proportional, or just-proportional increase in output.

One of the common findings of engineers is that modern mass-production techniques require that factories be a certain minimum size. Chapter 2 explained that as output increases, firms may divide production into smaller steps, taking advantage of specialization and division of labor. In addition, large-scale production allows intensive use of specialized capital equipment, automation, and computerized

Production concept	Definition
Diminishing returns	Declining marginal product of an input, holding all other inputs constant
Returns to scale	Increase in output for balanced increase in all inputs is
Decreasing	. . . less than proportional
Constant	. . . proportional
Increasing	. . . more than proportional

TABLE 6-2. Important Production Concepts

This table shows succinctly the important production concepts.

design and manufacturing to perform simple and repetitive tasks quickly.

Information technologies often display strong economies of scale. A good example is Microsoft's Windows Vista operating system. Developing this program reportedly required $10 billion in research, development, beta-testing, and promotion. Yet the cost of adding Windows Vista to a new computer is very close to zero because doing so simply requires a few seconds of computer time. We will see that strong economies of scale often lead to firms with significant market power and sometimes pose major problems of public policy.

Table 6-2 summarizes the important concepts from this section.

SHORT RUN AND LONG RUN

Production requires not only labor and land but also time. Pipelines cannot be built overnight, and once built they last for decades. Farmers cannot change crops in midseason. It often takes a decade to plan, construct, test, and commission a large power plant. Moreover, once capital equipment has been put in the concrete form of a giant automobile assembly plant, the capital cannot be economically dismantled and moved to another location or transferred to another use.

To account for the role of time in production and costs, we distinguish between two different time periods. We define the **short run** as a period in which firms can adjust production by changing variable factors such as materials and labor but cannot change fixed factors such as capital. The **long run** is a period sufficiently long that all factors including capital can be adjusted.

To understand these concepts more clearly, consider the way the production of steel might respond to changes in demand. Say that Nippon Steel is operating its furnaces at 70 percent of capacity when an unexpected increase in the demand for steel occurs because of the need to rebuild from an earthquake in Japan or California. To adjust to the higher demand for steel, the firm can increase production by increasing worker overtime, hiring more workers, and operating its plants and machinery more intensively. The factors which are increased in the short run are called *variable* factors.

Suppose that the increase in steel demand persisted for an extended period of time, say, several years. Nippon Steel would examine its capital needs and decide that it should increase its productive capacity. More generally, it might examine all its *fixed* factors, those that cannot be changed in the short run because of physical conditions or legal contracts. The period of time over which all inputs, fixed and variable, can be adjusted is called the long run. In the long run, Nippon might add new and more efficient production processes, install a rail link or new computerized control system, or build a plant in Mexico. When all factors can be adjusted, the total amount of steel will be higher and the level of efficiency can increase.

Efficient production requires time as well as conventional inputs like labor. We therefore distinguish between two different time periods in production and cost analysis. The short run is the period of time in which only some inputs, the variable inputs, can be adjusted. In the short run, fixed factors, such as plant and equipment, cannot be fully modified or adjusted. The long run is the period in which all factors employed by the firm, including capital, can be changed.

That Smells So Good!

The production processes of a modern market economy are extraordinarily complex. We can illustrate this with the lowly hamburger.

As Americans spend more time in the workplace and less in the kitchen, their demand for prepared food has risen dramatically. TV dinners have replaced store-bought carrots and peas, while hamburgers bought at McDonald's now number in the billions. The move to processed foods has the undesirable property that the food—after being washed, sorted, sliced, blanched, frozen, thawed, and reheated—often loses most of its flavor. You want a hamburger to smell and taste like a hamburger, not like cooked cardboard.

This is where the "production of tastes and smells" enters. Companies like International Flavors and Fragrances (IFF) synthesize the flavor of potato chips, breakfast cereals, ice cream, cookies, and just about every other kind of processed food, along with the fragrance of many fine perfumes, soaps, and shampoos. If you read most food labels, you will discover that the food contains "natural ingredients" or "artificial ingredients"—such compounds as amyl acetate (banana flavor) or benzaldehyde (almond flavor).

But these unfamiliar chemicals can do amazing things. A food researcher recounts the following experience in the laboratories of IFF:

> [After dipping a paper fragrance-testing filter into each bottle from the lab,] I closed my eyes. Then I inhaled deeply, and one food after another was conjured from the glass bottles. I smelled fresh cherries, black olives, sautéed onions, and shrimp. [The] most remarkable creation took me by surprise. After closing my eyes, I suddenly smelled a grilled hamburger. The aroma was uncanny, almost miraculous. It smelled like someone in the room was flipping burgers on a hot grill. But when I opened my eyes, there was just a narrow strip of white paper.[1]

This story reminds us that "production" in a modern economy is much more than planting potatoes and casting steel. It sometimes involves disassembling things like chickens and potatoes into their tiny constituents, and then reconstituting them along with new synthesized tastes halfway around the world. Such complex production processes can be found in every sector, from pharmaceuticals that change our mood or help our blood flow more smoothly to financial instruments that take apart, repackage, and sell the streams of mortgage payments. And most of the time, we don't even know what exotic substances lie inside the simple (recycled) paper that wraps our $2 hamburger.

[1] Eric Schlosser, *Fast Food Nation* (Perennial Press, New York, 2002), p. 129.

TECHNOLOGICAL CHANGE

Economic history records that total output in the United States has grown more than tenfold over the last century. Part of that gain has come from increased inputs, such as labor and machinery. But much of the increase in output has come from technological change, which improves productivity and raises living standards.

Some examples of technological change are dramatic: wide-body jets that increased the number of passenger-miles per unit of input by almost 50 percent; fiber optics that have lowered cost and improved reliability in telecommunications; and improvements in computer technologies that have increased computational power by more than 1000 times in three decades. Other forms of technological change are more subtle, as is the case when a firm adjusts its production process to reduce waste and increase output.

We distinguish *process innovation,* which occurs when new engineering knowledge improves production techniques for existing products, from *product innovation,* whereby new or improved products are introduced in the marketplace. For example, a process innovation allows firms to produce more output with the same inputs or to produce the same output with fewer inputs. In other words, a process innovation is equivalent to a shift in the production function.

Figure 6-3 illustrates how technological change, in the form of a process innovation, would shift the total product curve. The lower line represents the feasible output, or production function, for a particular industry in the year 1995. Suppose that productivity, or output per unit of input, in this industry is rising at 4 percent per year. If we return to the same industry a decade later, we would likely see that changes in technical and engineering knowledge have led to a 48 percent improvement in output per unit of input $[(1 + .04)^{10} = 1.48]$.

Next, consider product innovations, which involve new and improved products. It is much more difficult to quantify the importance of product innovations, but they may be even more important in raising living standards than process innovations. Many of today's goods and services did not even exist 50 years ago. In producing this textbook, the authors used computer software, microprocessors, Internet

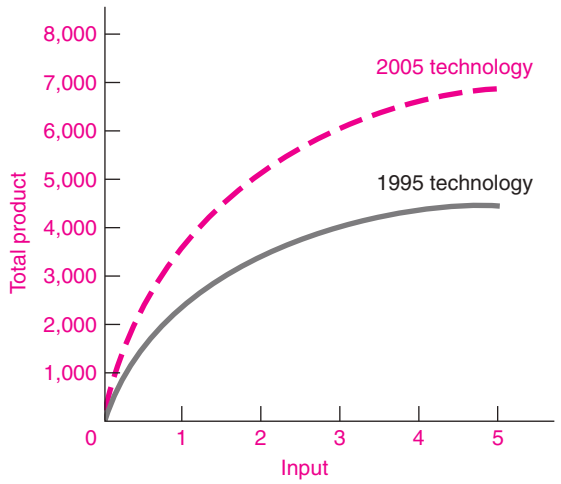


FIGURE 6-3. Technological Change Shifts Production Function Upward

The solid line represents maximum producible output for each level of inputs given the state of technical knowledge in 1995. As a result of improvements in computer technology and management practices, technological change shifts the production function upward, allowing much more output to be produced in 2005 for each level of inputs.

sites, and databases that were not available a decade ago. Medicine, communications, and entertainment are other sectors where product innovations have been critical. The whole arena of the Internet, from e-commerce to e-mail, was not found even in science fiction literature 30 years ago. For fun, and to see this point, try to find any commodity or production process that has not changed since your grandparents were your age!

Figure 6-3 shows the happy case of a technological advance. Is the opposite case—technological regress—possible? The answer is no for a well-functioning market economy. Inferior technologies are unprofitable and tend to be discarded in a market economy, while more productive technologies are introduced because they increase the profits of the innovating firms. To see this, suppose that someone invents an expensive new mousetrap that will never catch a mouse. No profit-oriented firm would produce such a device; and if a poorly managed firm decided to produce it, rational consumers who lived in mouse-infested houses would decline to buy it. Well-functioning markets innovate with better, not inferior, mousetraps.

When there are market failures, however, technological regress might occur. An unregulated company might introduce a socially wasteful process, say, dumping toxic wastes into a stream, because the wasteful process is more profitable. *But the economic advantage of inferior technologies comes only because the social costs of pollution are not included in the firm's calculations of the costs of production.* If pollution costs were included in a firm's decisions, say by pollution taxes, the regressive process would no longer be profitable. In competitive markets, inferior products follow Neanderthals into extinction.

Networks

Many products have little use by themselves and generate value only when they are used in combination with other products. Such products are strongly complementary. An important case is a *network,* where different people are linked together through a particular medium. Types of networks include both those defined by physical linkages, such as telecommunication systems, electricity transmission networks, computer clusters, pipelines, and roads, and the indirect networks that occur when people use compatible systems (such as Windows operating systems) or speak the same language (such as English).

To understand the nature of networks, consider how far you could drive your car without a network of gas stations or how valuable your telephone or e-mail would be if no one else had telephones or computers.

Network markets are special because consumers derive benefits not simply from their own use of a good but also from the number of other consumers who adopt the good. This is known as an *adoption externality.* When I get a phone, everyone else with a phone can now communicate with me. Therefore, my joining this network leads to positive external effects for others. The network externality is the reason why many colleges provide universal e-mail for all their students and faculty—the value of e-mail is much higher when everyone participates. Figure 6-4 on page 115 illustrates how one individual's joining a network has an external benefit to others.

Economists have discovered many important features of network markets. First, network markets are "tippy," meaning that the equilibrium tips toward one or only a

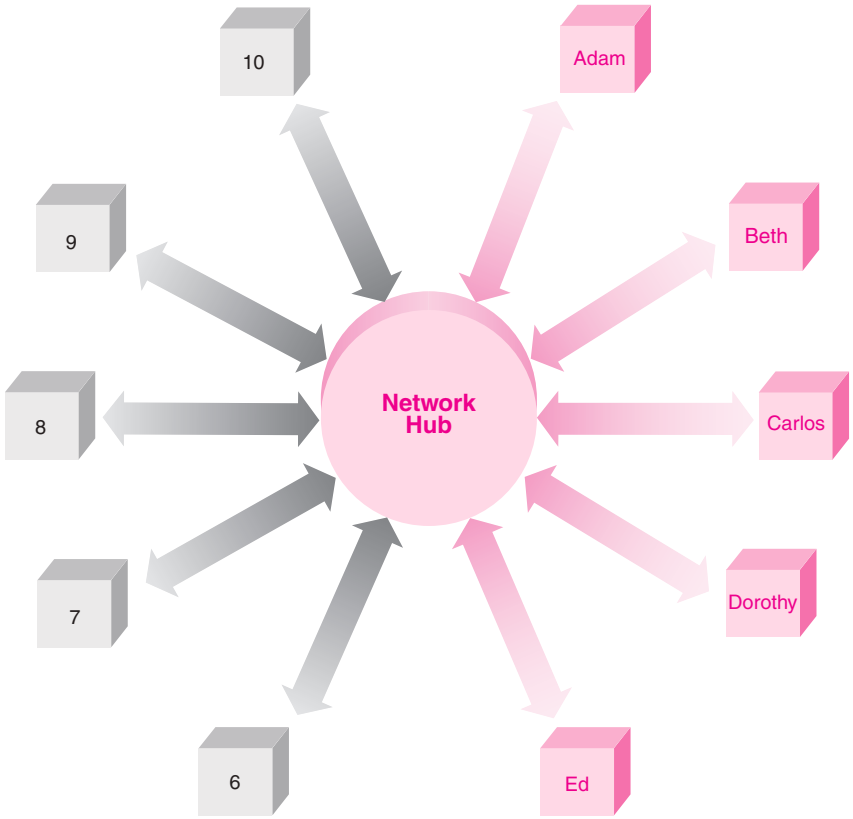


FIGURE 6-4. Value of Networking Increases as Membership Rises

Assume that each person derives a value of $1 for each additional person who is connected to a telephone or e-mail network. If Ed decides to join, he will get $4 of value from being connected to Adam, Beth, Carlos, and Dorothy. But there is an "adoption externality" because each of the four people already in the network gets $1 of additional value when Ed joins, for a total of $4 of external additional value.

These network effects make it difficult for networks to get started. To see this point, note that the second or third person who joins the network gets little value from joining. But when many people are in the network, each new member has a high value of joining because they are networked with a large number of people. (As an exercise, calculate the value of joining for the second and for the tenth people who join the network.)

few products. Because consumers dislike buying products that may turn out to be incompatible with dominant technologies, the equilibrium tends to gravitate to a single product which wins out over its rivals. One of the best-known examples is computer operating systems, where Microsoft Windows became the dominant system in part because consumers wanted to make sure that their computers could operate all the available software. (The important antitrust case involving Microsoft is discussed in Chapter 10.)

A second interesting feature is that "history matters" in network markets. A famous example is the QWERTY keyboard used with your computer. You might wonder why this particular configuration of keys, with its awkward placement of the letters, became the standard. The design of the QWERTY keyboard in the nineteenth century was based on the concept of keeping frequently used keys (like "e" and "o") physically separated in order to prevent manual typewriters from jamming. By the time the technology for electronic typing evolved, tens of millions of people had

already learned to type on millions of typewriters. Replacing the QWERTY keyboard with a more efficient design would have been both expensive and difficult to coordinate. Thus, the placement of the letters remains unchanged on today's keyboards.

This example shows how an embedded network technology can be extremely stable. A similar example that worries many environmentalists is America's "wasteful" automobile culture, where the existing network of cars, roads, gasoline stations, and residential locations will be difficult to dislodge in favor of more environmentally friendly alternatives, like improved mass transit.

Third, because networks involve a complicated interplay of economies of scale, expectations, dynamics, and tipping, they lead to a fascinating array of business strategies. The tippy nature of networks means that they tend to be "winner-take-all" markets with intense rivalry in the early stages and but a few competitors once the winning technology has emerged. In addition, network markets are often inertial, so once a product has a substantial lead, it may be very difficult for other products to catch up. These characteristics mean that companies often want to get an early lead on their rivals.

Suppose you are producing a network product. In order to build on your early lead, you might persuade users that you are number one by puffing up your sales; use "penetration pricing" by offering very low prices to early adopters; bundle your product with another popular product; or raise questions about your competitors' quality or staying power. Above all, you would probably invest heavily in advertising to shift out the demand curve for the product. If you are the fortunate winner, you will benefit from the economies of scale in the network and enjoy your monopoly profits. But don't take your dominant position for granted. Once your commanding lead is questioned, the virtuous cycle of market dominance can easily turn into the vicious cycle of market decline.

Networks raise important issues for public policy. Should government set standards to ensure competition? Should government regulate network industries? How should government antitrust policy treat monopolists like Microsoft that have been the fortunate winners in the network race but use anticompetitive tactics? These questions are on the minds of many public policymakers today.[2]

[2] See the Further Reading section at the end of this chapter.

PRODUCTIVITY AND THE AGGREGATE PRODUCTION FUNCTION

Productivity

One of the most important measures of economic performance is productivity. **Productivity** is a concept measuring the ratio of total output to a weighted average of inputs. Two important variants are **labor productivity,** which calculates the amount of output per unit of labor, and **total factor productivity,** which measures output per unit of total inputs (typically of capital and labor).

Productivity Growth from Economies of Scale and Scope

A central concept in economics is *productivity,* a term denoting the ratio of output to inputs. Economists typically look at two measures of productivity. Total factor productivity is output divided by an index of all inputs (labor, capital, materials, . . .), while labor productivity measures output per unit of labor (such as hours worked). When output is growing faster than inputs, this represents **productivity growth.**

Productivity grows because of technological advances such as the process and product innovations described above. Additionally, productivity grows because of economies of scale and scope.

Economies of scale and mass production have been important elements of productivity growth since the Industrial Revolution. Most production processes are many times larger than they were during the nineteenth century. A large ship in the mid-nineteenth century could carry 2000 tons of goods, while the largest supertankers today carry over 1 million tons of oil.

If increasing returns to scale prevail, the larger scale of inputs and production would lead to greater productivity. Suppose that, with no change in technology, the typical firm's inputs increased by 10 percent and that, because of economies of scale, output increased by 11 percent. Economies of scale would be responsible for a growth in total factor productivity of 1 percent.

A different kind of efficiency arises when there are **economies of scope,** which occur when a number of different products can be produced more efficiently together than apart. A prominent example is seen for computer software. Software programs often incorporate additional features as they evolve. For

example, when consumers buy software to prepare their federal income taxes, the CD-ROM usually contains several other modules, including a link to a Web page, government documents, and a tax preparation manual. This shows economies of scope because the different modules can be more inexpensively produced, packaged, and used together than separately. Economies of scope are like the specialization and division of labor that increase productivity as economies become larger and more diversified.

While increasing returns to scale and scope are potentially large in many sectors, at some point decreasing returns to scale and scope may take hold. As firms become larger and larger, the problems of management and coordination become increasingly difficult. In relentless pursuit of greater profits, a firm may find itself expanding into more geographic markets or product lines than it can effectively manage. A firm can have only one chief executive officer, one chief financial officer, and one board of directors. With less time to study each market and spend on each decision, top managers may become insulated from day-to-day production and begin to make mistakes. Like empires that have been stretched too thin, such firms find themselves vulnerable to invasion by smaller and more agile rivals.

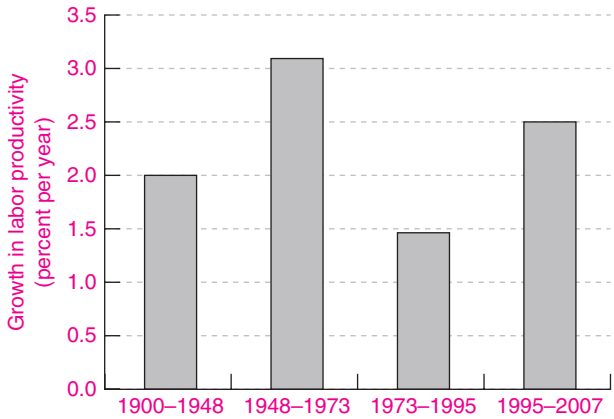


FIGURE 6-5. Growth in Labor Productivity

We see here the average growth in total productivity per hour worked during different periods. The last half-century had rapid growth after World War II, then a slowdown during the troubled 1970s and 1980s, and rapid growth during the period of rapid penetration of information technologies since 1995.

Source: Bureau of Labor Statistics and private scholars.

Empirical Estimates of the Aggregate Production Function

Now that we have examined the principles of production theory, we can apply these theories to measure how well the whole U.S. economy has performed in recent years. To do this, we need to look at *aggregate production functions,* which relate total output to the quantity of inputs (like labor, capital, and land). What have economic studies found? Here are a few of the important results:

- Total factor productivity has been increasing over the last century because of technological progress and higher levels of worker education and skill.
- The average rate of total productivity growth was slightly under 1½ percent per year since 1900.
- Over the twentieth century, labor productivity (output per hour worked) grew at an average rate of slightly more than 2 percent per year. From the early 1970s to the mid-1990s, however, all measures of productivity showed a marked slowdown in growth, and real wages and living standards consequently stagnated over this period. Since the mid-1990s, fueled largely by information technologies, there has been a marked upturn in productivity growth, with rates above the historical norm. (Figure 6-5 shows the historical trends.)
- The capital stock has been growing faster than the number of worker-hours. As a result, labor has a growing quantity of capital goods to work with; hence, labor productivity and wages have tended to rise even faster than the 1½ percent per year attributable to total factor productivity growth alone.

We end with a final word on the difficulties of measuring productivity growth accurately. Recent empirical studies suggest that we have seriously underestimated productivity growth in some areas. Studies of medical care, capital goods, consumer electronics, computers, and computer software indicate that our measuring rod for productivity is distorted. One particularly important shortcoming is the failure to account for the economic value of new and improved products. For example, when compact discs replaced "long-playing records," our measures of productivity did not include the improvement in

durability and sound quality. Similarly, our economic accounts cannot accurately measure the contribution of the Internet to consumer economic welfare.

B. BUSINESS ORGANIZATIONS

THE NATURE OF THE FIRM

So far we have talked about production functions as if they were machines that could be operated by anyone: put a pig in one end and a sausage comes out the other. In reality, almost all production is done by specialized organizations—the small, medium, and large businesses that dominate the landscape of modern economies. Why does production generally take place in firms rather than in our basements?

Firms or business enterprises exist for many reasons, but the most important is that *business firms are specialized organizations devoted to managing the process of production*. Among their important functions are exploiting economies of mass production, raising funds, and organizing factors of production.

In the first place, production is organized in firms because of *economies of specialization*. Efficient production requires specialized labor and machinery, coordinated production, and the division of production into many small operations. Consider a service such as a college education. This activity requires specialized personnel to teach economics and mathematics and Spanish, to produce the meals and housing services, to keep records, collect tuition, and pay the bills. We could hardly expect that a student could organize all these activities by herself. If there were no need for specialization and division of labor, we could each produce our own college education, surgical operations, electricity, and compact discs in our own backyard or buy them on the Internet. We obviously cannot perform such feats; efficiency generally requires large-scale production in businesses.

A second function of firms is *raising resources* for large-scale production. Developing a new commercial aircraft costs billions of dollars or Euros; the research and development expenses for a new computer microprocessor are just as high. Where are such funds to come from? In the nineteenth century, businesses could often be financed by wealthy, risk-taking individuals. Today, in a private-enterprise economy, most funds for production must come from company profits or from money borrowed in financial markets. Indeed, efficient production by private enterprise would be virtually unthinkable if corporations could not raise billions of dollars each year for new projects.

A third reason for the existence of firms is to *manage and coordinate the production process*. Once all the factors of production are engaged, someone has to monitor their daily activities to ensure that the job is being done effectively and honestly. The manager is the person who organizes production, introduces new ideas, products, or processes, makes the business decisions, and is held accountable for success or failure. Production cannot, after all, organize itself. Someone has to supervise the construction of a new factory, negotiate with labor unions, and purchase materials and supplies.

Take the case of a baseball team. How likely is it that 25 people would organize themselves into just the right combination of pitchers, catchers, and hitters, all in the right order and using the best strategy? If you were to purchase the franchise for a baseball team, you would have to rent a stadium, hire baseball players, negotiate with people for concessions, hire ushers, deal with unions, and sell tickets. This is the role of firms, to manage the production process, purchasing or renting land, capital, labor, and materials.

Business firms are specialized organizations devoted to managing the process of production. Production is organized in firms because efficiency generally requires large-scale production, the raising of significant financial resources, and careful management and coordination of ongoing activities.

Production in the Firm or the Market?

If markets are such a powerful mechanism for efficiency, why does so much production take place within large organizations? A related question is, Why do some firms decide on an integrated production structure while others contract out a large fraction of their sales? For example, before 1982 AT&T was vertically and horizontally integrated, doing its own research

and development, designing and producing its own equipment, installing and renting telephones, and providing telephone service. By contrast, most personal computers are "produced" by assemblers who purchase the hard drives, circuits, monitors, and keyboards from outside vendors and package and sell them.

These central issues of industrial organization were first raised by Ronald Coase in a pathbreaking study for which he was awarded the 1991 Nobel Prize.[3] This exciting area analyzes the comparative advantage of organizing production through the hierarchical control of firms as compared to the contractual relationships of the market.

Why might organizing through large firms be efficient? Perhaps the most important reason is the difficulty of designing "complete contracts" that cover all contingencies. For example, suppose Snoozer Inc. thinks it has discovered a hot new drug to cure laziness. Should it do the research in its own laboratories or contract out to another company, WilyLabs, Inc.? The problem with contracting out is that there are all kinds of unforeseen contingencies that could affect the profitability of the drug. What would happen if the drug proves useful for other conditions? What if the patent, tax, or international-trade laws change? What if there is a patent-infringement suit?

Because of the contractual incompleteness, the company runs the risk of the *holdup problem*. Suppose that WilyLabs discovers that the antilaziness drug works only when taken with another drug that WilyLabs owns. WilyLabs goes to Snoozer and says, "Sorry, pal, but to get both drugs will cost you another $100 million." This is holdup with a vengeance. Fear of being held up in situations which involve relationship-specific investments and contractual incompleteness will lead Snoozer to do the research internally so that it can control the outcomes of its research.

The recent trend in many industries has been to move away from highly integrated firms by "outsourcing" or contracting out production. This has definitely been the trend in the computer industry since the days when IBM was almost as integrated as AT&T. Contracting out can function well in situations where, as in the PC industry, the components are standardized or "commoditized." Another example is Nike, which contracts out much of its production because the production process is standard and Nike's real value is tied to its design and trademark. In addition, new contractual forms, such as long-term contracts based on reputations, attempt to minimize holdup problems.

[3] See the Further Reading section at the end of this chapter for examples of Coase's and related writings.

Those who study organizations point to the vital importance of large firms in promoting innovation and increasing productivity. In the nineteenth century, railroads not only brought wheat from farm to market but also introduced time zones. Indeed, the very notion of being "on time" first became crucial when being off schedule produced train wrecks. As the tragic story of centrally planned economies so clearly shows, without the organizational genius of the modern private-enterprise firm, all the land, labor, and capital can work for naught.

BIG, SMALL, AND INFINITESIMAL BUSINESSES

Production in a market economy takes place in a wide variety of business organizations—from the tiniest individual proprietorships to the giant corporations that dominate economic life in a capitalist economy. There are currently around 30 million different businesses in America. The majority of these are tiny units owned by a single person—the individual proprietorship. Others are partnerships, owned by two or perhaps two hundred partners. The largest businesses tend to be corporations.

Tiny businesses predominate in numbers. But in sales and assets, in political and economic power, and in size of payroll and employment, the few hundred largest corporations dominate the economy.

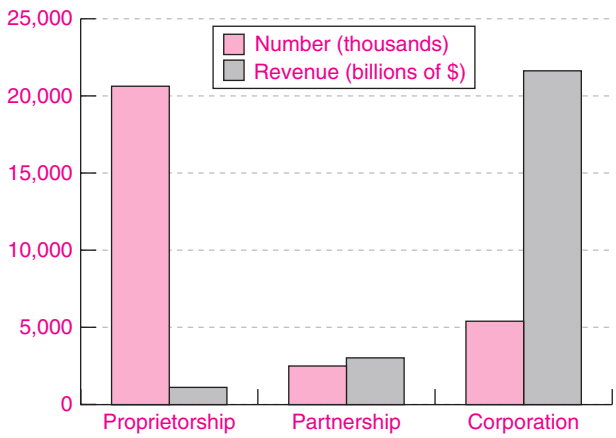


FIGURE 6-6. Number and Size of Different Business Forms, 2004

Corporations are fewer in number but dominate the economy.

Source: Internal Revenue Service.

Figure 6-6 shows the number and total revenue of the three major forms of economic organization in the United States.

The Individual Proprietorship

At one end of the spectrum are the individual proprietorships, the classic small businesses often called "mom-and-pop" stores. A small store might do a few hundred dollars of business per day and barely provide a minimum wage for the owners' efforts.

These businesses are large in number but small in total sales. For most small businesses, a tremendous amount of personal effort is required. The self-employed often work 50 or 60 hours per week and take no vacations, yet the average lifetime of a small business is only a year. Still, some people will always want to start out on their own. Theirs may be the successful venture that gets bought out for millions of dollars.

The Partnership

Often a business requires a combination of talents—say, lawyers or doctors specializing in different areas. Any two or more people can get together and form a partnership. Each agrees to provide a fraction of the work and capital and to share a percentage of the profits and losses.

Today, partnerships account for only a small fraction of total economic activity, as Figure 6-6 shows. Up to recently, partnerships were unattractive because they imposed *unlimited liability*. Under unlimited liability, partners are liable without limit for all debts contracted by the partnership. If you own 1 percent of the partnership and the business fails, you will be called upon to pay 1 percent of the bills. However, if your partners cannot pay, you may be called upon to pay all the debts, even if you must sell off your prized possessions to do so. Some states in the United States allow limited-liability partnerships for certain professions like law and architecture.

Except for a few sectors involving real estate and professionals, partnerships are cumbersome to administer and are less important than the corporate form of organization for most businesses.

The Corporation

The bulk of economic activity in an advanced market economy takes place in private corporations. Centuries ago, corporate charters were awarded by special acts of the monarch or legislature. The British East India Company was a privileged corporation and as such it practically ruled India for more than a century. In the nineteenth century, railroads often had to spend as much money on getting a charter through the legislature as on preparing their roadbeds. Over the past century, laws have been passed that allow almost anyone the privilege of forming a corporation for almost any purpose.

Today, a **corporation** is a form of business organization chartered in one of the 50 states or abroad and owned by a number of individual stockholders. The corporation has a separate legal identity, and indeed is a legal "person" that may on its own behalf buy, sell, borrow money, produce goods and services, and enter into contracts. In addition, the corporation enjoys the right of *limited liability,* whereby each owner's investment and financial exposure in the corporation is strictly limited to a specified amount.

The central features of a modern corporation are the following:

- The ownership of a corporation is determined by the ownership of the company's common stock. If you own 10 percent of a corporation's shares, you have 10 percent of the ownership. Publicly owned corporations are valued on stock exchanges, like the New York Stock Exchange. It is in such stock markets that the titles to the largest corporations are traded and that much of the nation's risk capital is raised and invested.
- In principle, the shareholders control the companies they own. They collect dividends in proportion to the fraction of the shares they own, and they elect directors and vote on many important issues. But don't think that the shareholders have a significant role in running giant corporations. In practice, shareholders of giant corporations exercise virtually no control because they are too dispersed to overrule the entrenched managers.
- The corporation's managers and directors have the legal power to make decisions for the corporation. They decide what to produce and how to produce it. They negotiate with labor unions and decide whether to sell the firm if another firm wishes to take it over. When the newspaper announces that a firm has laid off 20,000 workers, this decision was made by the managers. The shareholders own the corporation, but the managers run it.

Advantages and Disadvantages of Corporations. Corporations are the dominant form of organization in a market economy because they are an extremely efficient way to engage in business. A corporation is a legal person that can conduct business. Also, the corporation may have perpetual succession or existence, regardless of how many times the shares of stock change hands. Corporations are hierarchical, with the chief executive officer (CEO) exercising such power that they are sometimes called "autocratic" organizations. Managers can make decisions quickly, and often ruthlessly, which is in stark contrast to the way economic decisions are made by legislatures.

In addition, corporate stockholders enjoy limited liability, which protects them from incurring the debts or losses of the corporation beyond their initial contribution. If we buy $1000 worth of stock, we cannot lose more than our original investment.

Corporations face one major disadvantage: The government levies an extra tax on corporate profits. For an unincorporated business, any income after expenses is taxed as ordinary personal income. The large corporation is treated differently in that some of its income is doubly taxed—first as corporate profits and then as individual income on dividends.

Economists have criticized the corporation income tax as "double taxation" and have sometimes proposed integrating the corporate tax with the individual tax system. Under tax integration, corporate income is allocated to individuals and then taxed as individual income.

Sometimes, corporations undertake actions that provoke public outrage and government actions. In the late nineteenth century, corporations engaged in fraud, price fixing, and bribery, which led to enactment of antitrust and securities-fraud legislation. In the last few years, corporate scandals erupted when it was discovered that some companies engaged in massive accounting fraud and many corporate executives feathered their nests with huge bonuses and stock options. In private as in public life, power sometimes corrupts.

Efficient production often requires large-scale enterprises, which need billions of dollars of invested capital. Corporations, with limited liability and a convenient management structure, can attract large supplies of private capital, produce a variety of related products, and pool investor risks.

Ownership, Control, and Executive Compensation

The operation of large corporations raises important issues of public policy. They control much of a market economy, yet they are not controlled by the public. Indeed, scholars have come to recognize that they are not really controlled by their owners. Let us review some of the issues here.

The first step in understanding large corporations is to realize that most large corporations are "publicly owned." Corporate shares can be bought by anyone, and ownership is spread among many investors. Take a company like IBM, which was worth about $170 billion in 2008. Tens of millions of people have a financial interest in IBM through their mutual funds or pension accounts. However, no single person owned even 0.1 percent of the total. Such dispersed ownership is typical of our large publicly owned corporations.

Because the stock of large companies is so widely dispersed, *ownership is typically divorced from control.* Individual owners cannot easily affect the actions of large corporations. And while the stockholders of a company do in principle elect its board of directors—a group of insiders and knowledgeable outsiders—it is the management that makes the major decisions about corporate strategy and day-to-day operations.

In some situations, there is no conflict of interest between management and stockholders. Higher profits benefit everyone. But one important potential conflict between managers and stockholders has caught people's attention—the question of executive compensation. Top managers are able to extract from their boards large salaries, stock options, expense accounts, bonuses, free apartments, expensive artwork, and generous retirement pensions at the stockholders' expense. Nobody is arguing that managers should work for the minimum wage, but executive pay in U.S. corporations has risen very rapidly in recent years. Some top executives at poorly performing companies—or even at companies like WorldCom or Enron which later went bankrupt—received salaries and bonuses totaling $100 million or more.

Figure 6-7 shows an arresting graph: the ratio of the average pay of the top executives in the largest firms to that of the average worker. That ratio rose

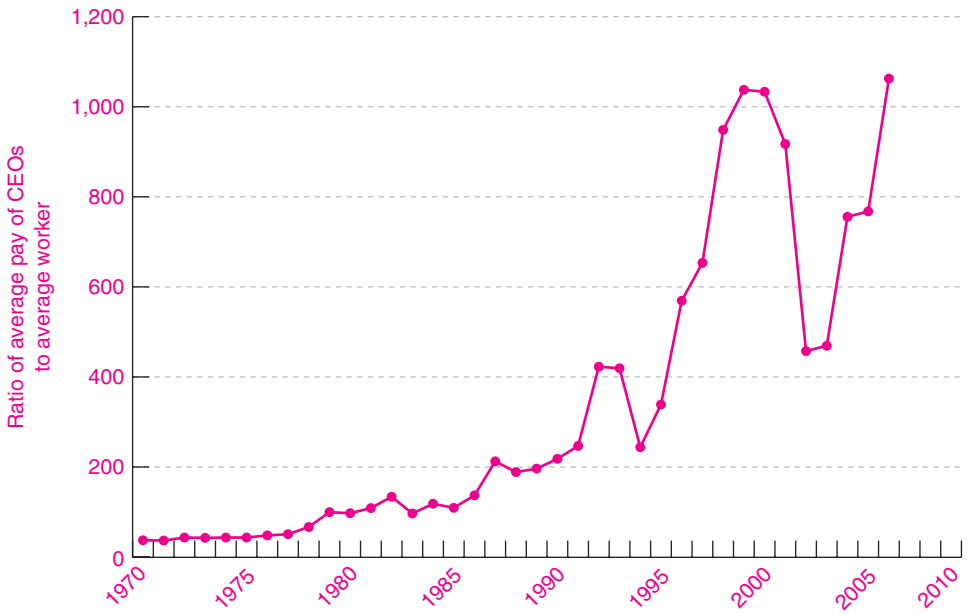


FIGURE 6-7. The Explosion in Executive Compensation

The figure shows the ratio of the average pay of the top 100 chief executive officers (CEOs) of U.S. corporations to the pay of the average U.S. worker. The ratio has risen from around 40 in 1970 to over 1000 in the mid-2000s. Many factors lie behind this explosive growth, but the most important is probably the ability of CEOs to manage the compensation process.

Source: Thomas Piketty and Emmanuel Saez, data from their website at *elsa.berkeley.edu/~saez/*.

from a historical average of around 40 to more than 1000 in recent years. The rise in executive compensation has been part of the reason for the growth in income inequality in the United States. What is the reason for this increase? Why, economists ask, are American executives often paid 10 or 20 times more than are executives in comparable firms of other countries?

Research in this area has pointed to several reasons for the dramatic change. Defenders point to the great importance of managers in efficient capitalism, but this overlooks the role of marginal productivity in competitive markets. Defenders also argue that stock options, which have been the major source of increased executive pay, are efficient devices because they tie compensation to performance through stock prices.

Critics answer that the most important reason for the trend is the divorce of ownership from control. This is the symptom of a malady known as the *principal-agent problem,* wherein the incentives of the agents (the managers) are not appropriately aligned with the interests of the principal (the owners). Moreover, managers tend to hide the compensation procedures from stockholders, and so the owners never really have a vote on managerial compensation. Additionally, stock options may give incentives for management to distort the financial accounts as well as to produce sound profits.

The rising tide of executive compensation raises important questions about public policy. What are effective means of ensuring that compensation is efficient? Most economists are reluctant to have the government set any kind of pay standards. They would argue that a system of progressive taxation is the most evenhanded way to deal with income inequalities. Most agree that better information and more power to owners can also wring out the largest excesses.

SUMMARY

A. Theory of Production and Marginal Products

1. The relationship between the quantity of output (such as wheat, steel, or automobiles) and the quantities of inputs (of labor, land, and capital) is called the production function. Total product is the total output produced. Average product equals total output divided by the total quantity of inputs. We can calculate the marginal product of a factor as the extra output added for each additional unit of input while holding all other inputs constant.
2. According to the law of diminishing returns, the marginal product of each input will generally decline as the amount of that input increases, when all other inputs are held constant.
3. The returns to scale reflect the impact on output of a balanced increase in all inputs. A technology in which doubling all inputs leads to an exact doubling of outputs displays constant returns to scale. When doubling inputs leads to less than double (more than double) the quantity of output, the situation is one of decreasing (increasing) returns to scale.
4. Because decisions take time to implement, and because capital and other factors are often very long-lived, the reaction of production may change over different time periods. The short run is a period in which variable factors, such as labor or material inputs, can be easily changed but fixed factors cannot. In the long run, the capital stock (a firm's machinery and factories) can depreciate and be replaced. In the long run, all inputs, fixed and variable, can be adjusted.
5. Technological change refers to a change in the underlying techniques of production, as occurs when a new product or process of production is invented or an old product or process is improved. In such situations, the same output is produced with fewer inputs or more output is produced with the same inputs. Technological change shifts the production function upward.
6. Attempts to measure an aggregate production function for the American economy tend to corroborate theories of production and marginal products. In the twentieth century, technological change increased the productivity of both labor and capital. Total factor productivity (measuring the ratio of total output to total inputs) grew at around 1½ percent per year over the twentieth century, although from the 1970s to the mid-1990s the rate of productivity growth slowed markedly and real wages stopped growing. But underestimating the importance of new and improved products may lead to a significant underestimate of productivity growth.

B. Business Organizations

7. Business firms are specialized organizations devoted to managing the process of production.
8. Firms come in many shapes and sizes—with some economic activity in tiny one-person proprietorships, some in partnerships, and the bulk in corporations. Each kind of enterprise has advantages and disadvantages. Small businesses are flexible, can market new products, and can disappear quickly. But they suffer from the fundamental disadvantage of being unable to accumulate large amounts of capital from a dispersed group of investors. Today's large corporation, granted limited liability by the state, is able to amass billions of dollars of capital by borrowing from banks, bondholders, and stock markets.
9. In a modern economy, business corporations produce most goods and services because economies of mass production necessitate that output be produced at high volumes, the technology of production requires much more capital than a single individual would willingly put at risk, and efficient production requires careful management and coordination of tasks by a centrally directed entity.
10. The modern corporation may involve divided incentives because of the divorce of ownership from control, which has produced the vast gulf between executive compensation and average wages.

CONCEPTS FOR REVIEW

inputs, outputs, production function
total, average, and marginal product
diminishing marginal product and the law of diminishing returns
constant, increasing, and decreasing returns to scale
short run vs. long run
technological change: process innovation, product innovation
Productivity:
  defined as output/input
  two versions: labor productivity, total factor productivity
aggregate production function
reasons for firms: scale economies, financial needs, management
major business forms: individual proprietorship, partnership, corporation
unlimited and limited liability
firm vs. market and the holdup problem
Divorce of ownership from control: principal-agent problem

FURTHER READING AND INTERNET WEBSITES

Further Reading

Ronald Coase's classic work is "The Nature of the Firm," *Economica,* November 1937. Students may enjoy a recent nontechnical survey of the field in the symposium "The Firm and Its Boundaries," *Journal of Economic Perspectives,* Fall 1998. For a thoughtful analysis of network effects, see the symposium in *Journal of Economic Perspectives,* Spring 1994. A fascinating study of networks and the new economy is contained in Chapter 7 in Carl Shapiro and Hal R. Varian, *Information Rules: A Strategic Guide to the Network Economy* (Harvard Business School Press, Cambridge, Mass., 1997).

For a recent survey of the issues and policies concerning executive compensation, see Gary Shorter and Marc Labonte, *The Economics of Corporate Executive Pay,* March 22, 2007, available at *digitalcommons.ilr.cornell.edu/crs/36/*. A discussion of the economic background on this subject is contained in a symposium in *The Journal of Economic Perspectives,* Fall 2003, particularly the article by Kevin Murphy and Brian Hall.

Trends in the income of top executives are shown in Thomas Piketty and Emmanuel Saez, "Income Inequality in the United States, 1913–1998," *Quarterly Journal of Economics,* 2003, pp. 1–39; that article and an updated version are available at *elsa.berkeley.edu/~saez/*.

Websites

One of the most interesting websites about networks is compiled by Hal R. Varian, dean of the School of Information Management and Systems at the University of California at Berkeley. This site, called "The Economics of the Internet, Information Goods, Intellectual Property and Related Issues," is at *www.sims.berkeley.edu/resources/infoecon*.

A specialized site on network economics maintained by Nicholas Economides of New York University is found at *raven.stern.nyu.edu/networks/site.html*.

QUESTIONS FOR DISCUSSION

1. Explain the concept of a production function. Describe the production function for hamburgers, computers, concerts, haircuts, and a college education.
2. Consider a production function of the following form: $X = 100L^{1/2}$, where X = output and L = input of labor (assuming other inputs are fixed).
   a. Construct a figure like Figure 6-1 and a table like Table 6-1 for inputs of $L = 0, 1, 2, 3$, and 4.
   b. Explain whether this production function shows diminishing returns to labor. What values would the exponent need to take for this production function to exhibit increasing returns to labor?
3. The following table describes the actual production function for oil pipelines. Fill in the missing values for marginal products and average products:

(1)	(2)	(3)	(4)
	18-Inch Pipe		
Pumping horsepower	Total product (barrels per day)	Marginal product (barrels per day per hp)	Average product (barrels per day per hp)
10,000	86,000		
20,000	114,000		
30,000	134,000		
40,000	150,000		
50,000	164,000		

4. Using the data in question 3, plot the production function of output against horsepower. On the same graph, plot the curves for average product and marginal product.
5. Suppose you are running the food concession at the athletic events for your college. You sell hot dogs, colas, and potato chips. What are your inputs of capital, labor, and materials? If the demand for hot dogs declines, what steps could you take to reduce output in the short run? In the long run?
6. An important distinction in economics is between shifts of the production function and movements along the production function. For the food concession in question 5, give an example of both a shift of and a movement along the hot-dog production function. Illustrate each with a graph of the relation between hot-dog production and labor employed.
7. Substitution occurs when firms replace one input for another, as when a farmer uses tractors rather than labor when wages rise. Consider the following changes in a firm's behavior. Which represent substitution of one factor for another with an unchanged technology, and which represent technological change? Illustrate each with a graphical production function.
   a. When the price of oil increases, a firm replaces an oil-fired plant with a gas-fired plant.
   b. A bookseller reduces its sales staff by 60 percent after it sets up an Internet outlet.
   c. Over the period 1970–2000, a typesetting firm decreases its employment of typesetters by 200 workers and increases its employment of computer operators by 100 workers.
   d. After a successful unionization drive for clerical workers, a college buys personal computers for its faculty and reduces its secretarial workforce.
8. Consider a firm that produces pizzas with capital and labor inputs. Define and contrast diminishing returns and decreasing returns to scale. Explain why it is possible to have diminishing returns for one input and constant returns to scale for both inputs.
9. Show that if the marginal product is always decreasing, the average product is always above the marginal product.
10. Review the example of a network shown in Figure 6-4. Assume that only one person can join the network each month, starting with Adam and proceeding clockwise.
    a. Construct a table showing the value to the joining person as well as the external value to others (i.e., the value to all others in the network) when an additional person joins. (*Hint:* The entries for Ed are $4 and $4.) Then calculate the total social value for each level of membership. Graph the relationship between the size of the network and the total social value. Explain why this shows increasing returns rather than diminishing returns.
    b. Assume that the cost of joining is $4.50. Draw a graph which shows how membership changes over time if six people are in the network to begin with. Draw another one which shows what happens if there are initially three people in the network. What is the point at which the equilibrium "tips" toward universal membership?
    c. Suppose you are the sponsor of the network shown in Figure 6-4. What kind of pricing could you use to get the network started when there are only one or two members?

CHAPTER

7

Analysis of Costs

Costs merely register competing attractions.

Frank Knight

Risk, Uncertainty, and Profit (1921)

Everywhere that production goes, costs follow close behind like a shadow. Firms must pay for their inputs: screws, solvents, software, sponges, secretaries, and statisticians. Profitable businesses are acutely aware of this simple fact as they determine their production strategies, since every dollar of unnecessary costs reduces the firm's profits by that same dollar.

But the role of costs goes far beyond influencing production and profits. Costs affect input choices, investment decisions, and even the decision of whether to stay in business. Is it cheaper to hire a new worker or to pay overtime? To open a new factory or expand an old one? To invest in new machinery domestically or to outsource production abroad? Businesses want to choose those methods of production that are most efficient and produce output at the lowest cost.

This chapter is devoted to a thorough analysis of cost. First we consider the full array of economic costs, including the central notion of marginal costs. Then we examine how business accountants measure cost in practice. Finally, we look at the notion of opportunity cost, a broad concept that can be applied to a wide range of decisions. This comprehensive study of cost will lay the foundation for understanding the supply decisions of business firms.

A. ECONOMIC ANALYSIS OF COSTS

TOTAL COST: FIXED AND VARIABLE

Consider a firm that produces a quantity of output (denoted by q) using inputs of capital, labor, and materials. The firm's accountants have the task of calculating the total dollar costs incurred to produce output level q.

Table 7-1 on page 127 shows the total cost (TC) for each different level of output q. Looking at columns (1) and (4), we see that TC goes up as q goes up. This makes sense because it takes more labor and other inputs to produce more of a good; extra factors involve an extra money cost. It costs \$110 in all to produce 2 units, \$130 to produce 3 units, and so forth. In our discussion, we assume that the firm always produces output at the lowest possible cost.

Fixed Cost

Columns (2) and (3) of Table 7-1 separate total cost into two components: total fixed cost (FC) and total variable cost (VC).

(1) Quantity q	(2) Fixed cost FC ($)	(3) Variable cost VC ($)	(4) Total cost TC ($)
0	55	0	55
1	55	30	85
2	55	55	110
3	55	75	130
4	55	105	160
5	55	155	210
6	55	225	280

TABLE 7-1. Fixed, Variable, and Total Costs

The major elements of a firm's costs are its fixed costs (which do not vary at all when output changes) and its variable costs (which increase as output increases). Total costs are equal to fixed plus variable costs: $TC = FC + VC$.

Fixed costs are expenses that must be paid even if the firm produces zero output. Sometimes called "overhead" or "sunk costs," they consist of items such as rent for factory or office space, interest payments on debts, salaries of tenured faculty, and so forth. They are fixed because they do not change if output changes. For example, a law firm might have an office lease which runs for 10 years and remains an obligation even if the firm shrinks to half its previous size. Because FC is the amount that must be paid regardless of the level of output, it remains constant at $55 in column (2).

Variable Cost

Column (3) of Table 7-1 shows variable cost (VC). **Variable costs** do vary as output changes. Examples include materials required to produce output (such as steel to produce automobiles), production workers to staff the assembly lines, power to operate factories, and so on. In a supermarket, checkout clerks are a variable cost, since managers can adjust the clerks' hours worked to match the number of shoppers coming through the store.

By definition, VC begins at zero when q is zero. VC is the part of TC that grows with output; indeed, the jump in TC between any two outputs is the same as the jump in VC.

Let us summarize these cost concepts:

Total cost represents the lowest total dollar expense needed to produce each level of output q. TC rises as q rises.

Fixed cost represents the total dollar expense that is paid out even when no output is produced; fixed cost is unaffected by any variation in the quantity of output.

Variable cost represents expenses that vary with the level of output—such as raw materials, wages, and fuel—and includes all costs that are not fixed.

Always, by definition,

$$TC = FC + VC$$

Minimum Attainable Costs

Anyone who has managed a business knows that when we write down a cost schedule like the one in Table 7-1, we make the firm's job look altogether too simple. Much hard work lies behind Table 7-1. To attain the lowest level of costs, the firm's managers have to make sure that they are paying the least possible amount for necessary materials, that the lowest-cost engineering techniques are incorporated into the factory layout, that employees are being honest, and that countless other decisions are made in the most economical fashion.

For example, suppose you are the owner of a baseball team. You have to negotiate salaries with players, choose managers, bargain with vendors, worry about electricity and other utility bills, consider how much insurance to buy, and deal with the 1001 other issues that are involved in running the team with minimum cost.

The total costs shown in Table 7-1 are the minimum costs that result from all these hours of managerial work.

DEFINITION OF MARGINAL COST

Marginal cost is one of the most important concepts in all of economics. **Marginal cost** (MC) denotes the extra or additional cost of producing 1 extra unit of output. Say a firm is producing 1000 compact discs for a total cost of $10,000. If the total cost of

(1) Output q	(2) Total cost TC ($)	(3) Marginal cost MC ($)
0	55	
		30
1	85	
		25
2	110	
		20
3	130	
		—
4	160	
		50
5	210	

TABLE 7-2. Calculation of Marginal Cost

Once we know total cost, it is easy to calculate marginal cost. To calculate the *MC* of the fifth unit, we subtract the total cost of the 4 units from the total cost of the 5 units, i.e., *MC* = \$210 − \$160 = \$50. Fill in the blank for the marginal cost of the fourth unit.

producing 1001 discs is \$10,006, then the marginal cost of production is \$6 for the 1001st disc.

Sometimes, the marginal cost of producing an extra unit of output can be quite low. For an airline flying planes with empty seats, the added cost of another passenger is literally peanuts; no additional capital (planes) or labor (pilots and flight attendants) is necessary. In other cases, the marginal cost of another unit of output can be quite high. Consider an electric utility. Under normal circumstances, it can generate enough power using only its lowest-cost, most efficient plants. But on a hot summer day, when everyone's air conditioners are running and demand for electricity is high, the utility may be forced to turn on its old, high-cost, inefficient generators. This added electric power comes at a high marginal cost to the utility.

Table 7-2 uses the data from Table 7-1 to illustrate how we calculate marginal costs. The green-colored *MC* numbers in column (3) of Table 7-2 come from subtracting the *TC* in column (2) from the *TC* of the subsequent quantity. Thus the *MC* of the first unit is \$30(= \$85 − \$55); the marginal cost of the second unit is \$25(= \$110 − \$85); and so on.

Instead of getting *MC* from the *TC* column, we could get the *MC* figures by subtracting each *VC* number in column (3) of Table 7-1 from the *VC* in the row below it. Variable cost always grows exactly like total cost, the only difference being that *VC* must (by definition) start out from 0 rather than from the constant *FC* level. (Check that \$30 − \$0 = \$85 − \$55, and \$55 − \$30 = \$110 − \$85, and so on.)

The marginal cost of production is the additional cost incurred in producing 1 extra unit of output.

Marginal Cost in Diagrams. Figure 7-1 illustrates total cost and marginal cost. It shows that *TC* is related to *MC* in the same way that total product is related to marginal product or that total utility is related to marginal utility.

The Marginal Cost of Distributing Software

When the software company Microsoft decided to enter the market for Internet browsers, it did so by giving away its Internet Explorer browser, either as a stand-alone product or in combination with the Windows operating system. Its competitors complained that Microsoft was engaged in "predatory behavior." How could it give the browser software away and not lose money?

The answer lies in the unusual property of information technology (IT). According to IT specialist Hal Varian, IT "typically has the property that it is very costly to produce the first copy and very cheap to produce subsequent copies." In this case, while it cost Microsoft a great deal to develop Internet Explorer, the marginal cost of distributing an extra unit of the software was close to zero. That is, the cost to Microsoft of delivering 1,000,001 units was no more than the cost of 1,000,000 units. As long as the marginal cost was zero, Microsoft was not losing money by giving Internet Explorer away.

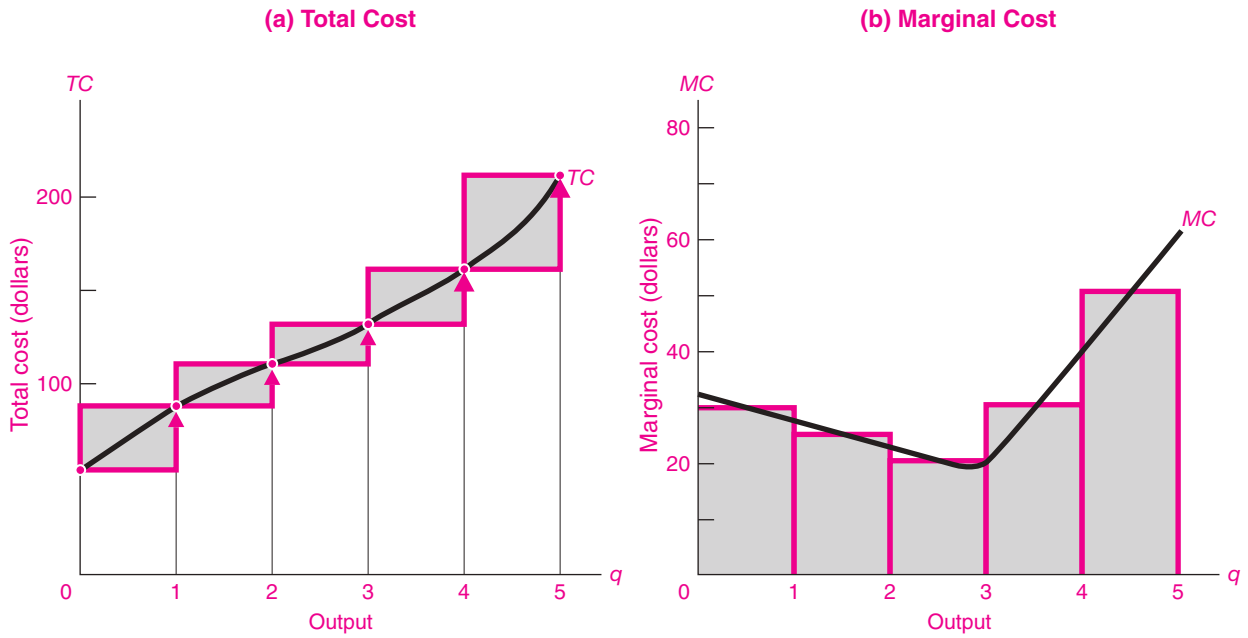


FIGURE 7-1. The Relationship between Total Cost and Marginal Cost

These graphs show the data from Table 7-2. Marginal cost in **(b)** is found by calculating the extra cost added in **(a)** for each unit increase in output. Thus to find the *MC* of producing the fifth unit, we subtract \$160 from \$210 to get *MC* of \$50. A smooth blue curve has been drawn through the points of *TC* in **(a),** and the smooth blue *MC* curve in **(b)** links the discrete steps of *MC*.

AVERAGE COST

We complete our catalog of the cost concepts with a discussion of different kinds of average or unit cost. Table 7-3 on page 130 expands the data of Tables 7-1 and 7-2 to include three new measures: average cost, average fixed cost, and average variable cost.

Average or Unit Cost

Average cost (AC) is a concept widely used in business; by comparing average cost with price or average revenue, businesses can determine whether or not they are making a profit. **Average cost** is the total cost divided by the total number of units produced, as shown in column (6) of Table 7-3. That is,

$$\text{Average cost} = \frac{\text{total cost}}{\text{output}} = \frac{TC}{q} = AC$$

In column (6), when only 1 unit is produced, average cost has to be the same as total cost, or \$85/1 = \$85. But for $q = 2$, $AC = TC/2 = \$110/2 = \$55$, as shown. Note that average cost falls lower and lower at first. (We shall see why in a moment.) AC reaches a minimum of \$40 at $q = 4$, and then slowly rises.

Figure 7-2 on page 131 plots the cost data shown in Table 7-3. Figure 7-2(a) depicts the total, fixed, and variable costs at different levels of output. Figure 7-2(b) shows the different average cost concepts, along with a smoothed marginal cost curve. Graph (a) shows how total cost moves with variable cost while fixed cost remains unchanged.

Now turn to graph (b). This plots the U-shaped AC curve and aligns AC right below the TC curve from which it is derived.

Average Fixed and Variable Costs

Just as we separated total cost into fixed and variable costs, we can also break average cost into fixed and variable components. **Average fixed cost** (AFC) is defined as FC/q. Since total fixed cost is a constant, dividing it by an increasing output gives a steadily

(1) Quantity q	(2) Fixed cost FC ($)	(3) Variable cost VC ($)	(4) Total cost $TC = FC + VC$ ($)	(5) Marginal cost per unit MC ($)	(6) Average cost per unit $AC = TC/q$ ($)	(7) Average fixed cost per unit $AFC = FC/q$ ($)	(8) Average variable cost per unit $AVC = VC/q$ ($)
0	55	0	55		Infinity	Infinity	Undefined
				30			
1	55	____	85		85	55	30
				25			
2	____	55	110		55	____	27½
				____			
3	55	75	130		43⅓	18⅓	25
				30			
4*	**55**	**105**	**160**	**40***	**40***	**13¾**	**26¼**
				50			
5	55	155	210		42	11	____
				70			
6	55	225	280		46⅔	9⅙	37½

*Minimum level of average cost.

TABLE 7-3. All Cost Concepts Derive from Total Cost Schedule

We can derive all the different cost concepts from the *TC* in column (4). Columns (5) and (6) are the important ones to concentrate on: marginal cost is calculated by subtraction of adjacent rows of *TC* and is shown in green. The starred *MC* of 40 at an output of 4 is the smoothed *MC* from Fig. 7-2(*b*). In column (6), note the point of minimum cost of $40 on the U-shaped *AC* curve in Fig. 7-2(*b*). (Can you see why the starred *MC* equals the starred *AC* at the minimum? Also, calculate and fill in all the missing numbers.)

falling average fixed cost curve [see column (7) of Table 7-3]. In other words, as a firm sells more output, it can spread its overhead cost over more and more units. For example, a software firm may have a large staff of programmers to develop a new game. The number of copies sold does not directly affect how many programmers are necessary, thus making them a fixed cost. So if the program is a best-seller, the *AFC* of the programmers is low; if the program is a failure, the *AFC* is high.

The dashed blue *AFC* curve in Figure 7-2(*b*) is a hyperbola, approaching both axes: it drops lower and lower, approaching the horizontal axis as the constant *FC* gets spread over more and more units. If we allow fractional units of *q*, *AFC* starts infinitely high as finite *FC* is spread over ever tinier *q*.

Average variable cost (*AVC*) equals variable cost divided by output, or $AVC = VC/q$. As you can see in both Table 7-3 and Figure 7-2(*b*), for this example *AVC* first falls and then rises.

The Relation between Average Cost and Marginal Cost

It is important to understand the link between average cost and marginal cost. We begin with three closely related rules:

1. When marginal cost is below average cost, it is pulling average cost down.
2. When *MC* is above *AC*, it is pulling up *AC*.
3. When *MC* just equals *AC*, *AC* is constant. At the bottom of a U-shaped *AC*, *MC* = *AC* = minimum *AC*.

To understand these rules, begin with the first one. If *MC* is below *AC*, this means that the last unit produced costs less than the average cost of all the previous units produced. This implies that the new *AC* (i.e., the *AC* including the last unit) must be less than the old *AC*, so *AC* must be falling.

We can illustrate this with an example. Looking at Table 7-3, we see that the *AC* of the first unit is 85.

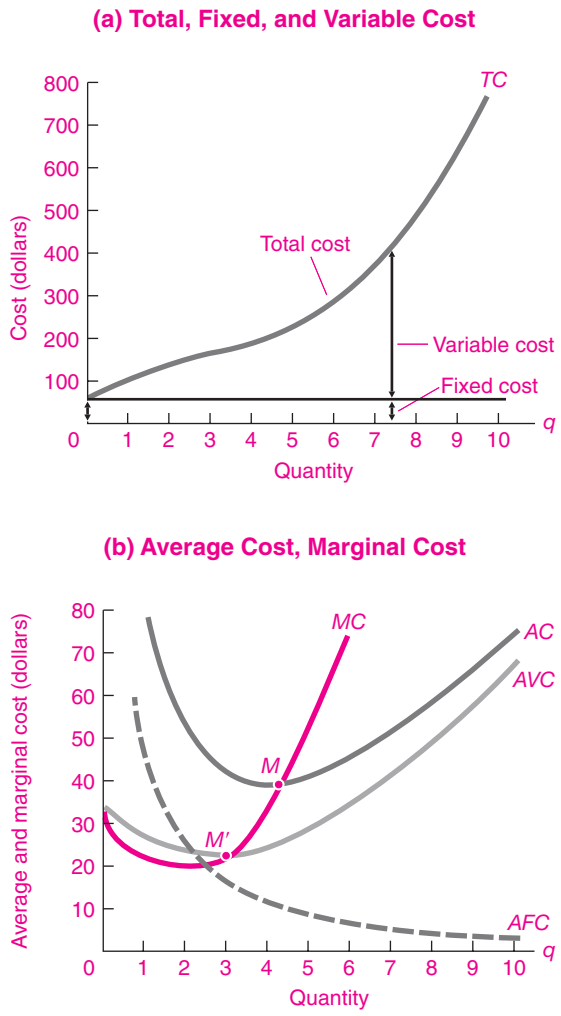


FIGURE 7-2. All Cost Curves Can Be Derived from the Total Cost Curve

(a) Total cost is made up of fixed cost and variable cost. **(b)** The green-colored curve of marginal cost falls and then rises, as indicated by the *MC* figures given in column (5) of Table 7-3. Note how *MC* intersects *AC* at its minimum.

The *MC* of the second unit is 25. This implies that the *AC* of the first 2 units is $(85 + 25)/2 = 55$. Because *MC* was below *AC*, this correctly implies that *AC* is falling.

The second rule is illustrated in Table 7-3 by the case of the sixth unit. The *AC* of 5 units is 42, and the *MC* between 5 and 6 units is 70. *MC* is pulling up *AC* as we see by the *AC* of the sixth unit, which is 46⅔.

q	*FC*	*VC*	*TC*	*MC*
3,998	55,000	104,920.03	159,920.03	
				39.98
3,999	55,000	104,960.01	159,960.01	
				39.99
4,000*	55,000	105,000.00	160,000.00	
				40.01
4,001	55,000	105,040.01	160,040.01	
				40.02
4,002	55,000	105,080.03	160,080.03	

*Production with minimum average cost.

TABLE 7-4. Take a Microscope to the *AC* and *MC* Calculations at the Minimum Point

This table magnifies the cost calculations around the minimum *AC* point. We assume for this calculation that the numbers in Table 7-3 are in thousands. Note how the marginal cost is a tiny bit below the minimum *AC* between 3999 and 4000 units and a tiny bit above it between 4000 and 4001 units.

The case of the fourth unit is a crucial one. At that level, note that *AC* is exactly equal to *MC* at a cost of 40. So the new *AC* is exactly equal to the old *AC* and is equal to *MC*. We illustrate the relationship in detail in Table 7-4, which focuses on the minimum *AC* level of production. For this table, we assume that the units in Table 7-3 are in thousands so that we can see tiny movements in output. See how *MC* is a tiny bit below *AC* when output is just below the minimum-*AC* point (and a tiny bit above *AC* when output is just above the minimum-*AC* point). If we were to increase the magnification further, we would come as close as we want to an exact equality of *MC* and *AC*.

You will improve your understanding of the relationship between *MC* and *AC* by studying Figure 7-2(*b*). Note that for the first 3 units, *MC* is below *AC*, and *AC* is therefore declining. At exactly 4 units, *AC* equals *MC*. Over 4 units, *MC* is above *AC* and pulling *AC* up. Graphically, that means the rising *MC* curve will intersect the *AC* curve precisely at its minimum point.

To summarize: In terms of our cost curves, if the *MC* curve is below the *AC* curve, the *AC* curve must be falling. By contrast, if *MC* is above *AC*, *AC* is rising. Finally, when *MC* is just equal to *AC*, the *AC* curve is flat. The *AC* curve is always pierced at its minimum point by a rising *MC* curve.

Batting Averages to Illustrate *MC* and *AC* Rules

We can illustrate the *MC*-*AC* relationship using batting averages. Let *AB* be your lifetime batting average up to this year (your average) and *MB* be your batting average for this year (your marginal). For simplicity, we also assume that there are 100 "at bats" each year.

When your *MB* is below *AB*, it will pull the new *AB* down. For example, suppose that your lifetime batting average for your first 3 years was .300 and your batting average for your fourth year was .100. Your new lifetime average or *AB* at the end of your fourth year is .250. Similarly, if your *MB* in your fourth year is higher than your lifetime average for your first 3 years, your lifetime average will be pulled up. If your batting average in the fourth year is the same as your lifetime average for the first 3 years, your lifetime average will not change (i.e., if *MB* = *AB*, then the new *AB* is equal to the old *AB*).

THE LINK BETWEEN PRODUCTION AND COSTS

What are the factors that determine the cost curves introduced above? The key elements are (1) factor prices and (2) the firm's production function.

Clearly the prices of inputs like labor and land are important ingredients of costs. Higher rents and higher wages mean higher costs, as any business manager will tell you. But costs also depend on the firm's technological opportunities. If technological improvements allow the firm to produce the same output with fewer inputs, the firm's costs will fall.

Indeed, if you know factor prices and the production function, you can calculate the cost curve. We can show the derivation of cost from production data and factor prices in the numerical example shown in Table 7-5. Suppose Farmer Smith rents 10 acres of land and can hire farm labor to produce wheat. Per period, land costs $5.5 per acre and labor costs $5 per worker. Using up-to-date farming methods, Smith can produce according to the production function shown in the first three columns of Table 7-5. In this example, land is a fixed cost (because Farmer Smith operates under a 10-year lease), while labor is a variable cost (because farmworkers can easily be hired and fired).

Using the production data and the input-cost data, for each level of output we calculate the total cost of production shown in column (6) of Table 7-5. As an example, consider the total cost of production for 3 tons of wheat. Using the given production function, Smith can produce this quantity with 10 acres of land and 15 farmhands. The total cost of producing 3 tons of wheat is (10 acres × $5.5 per acre) + (15 workers × $5 per worker) = $130. Similar calculations will give all the other total cost figures in column (6) of Table 7-5.

Note that these total costs are identical to the ones shown in Tables 7-1 through 7-3, so the other

(1) Output (tons of wheat)	(2) Land inputs (acres)	(3) Labor inputs (workers)	(4) Land rent ($ per acre)	(5) Labor wage ($ per worker)	(6) Total cost ($)
0	10	0	5.5	5	55
1	10	6	5.5	5	85
2	10	11	5.5	5	110
3	10	15	5.5	5	130
4	10	21	5.5	5	160
5	10	31	5.5	5	210
6	10	45	5.5	5	280

TABLE 7-5. Costs are Derived from Production Data and Input Costs

Farmer Smith rents 10 acres of wheatland and employs variable labor. According to the farming production function, careful use of labor and land allows the inputs and yields shown in columns (1) to (3) of the table. At input prices of $5.5 per acre and $5 per worker, we obtain Smith's cost of production shown in column (6). All other cost concepts (such as those shown in Table 7-3) can be calculated from the total cost data.

cost concepts shown in the tables (i.e., *MC, FC, VC, AC, AFC,* and *AVC*) are also applicable to the production-cost example of Farmer Smith.

Diminishing Returns and U-Shaped Cost Curves

Economists often draw cost curves like the letter "U" (the "U-shaped cost curves"). For a U-shaped cost curve, cost falls in the initial phase, reaches a minimum point, and finally begins to rise. Let's explore the reasons. Recall that Chapter 6's analysis of production distinguished two different time periods, the short run and the long run. The same concepts apply to costs as well:

- The *short run* is the period of time that is long enough to adjust variable inputs, such as materials and production labor, but too short to allow all inputs to be changed. In the short run, fixed or overhead factors such as plant and equipment cannot be fully modified or adjusted. Therefore, in the short run, labor and materials costs are typically variable costs, while capital costs are fixed.
- In the *long run,* all inputs can be adjusted—including labor, materials, and capital. Hence, in the long run, all costs are variable and none are fixed.[1]

Note that whether a particular cost is fixed or variable depends on the length of time we are considering. In the short run, for example, the number of planes that an airline owns is a fixed cost. But over the longer run, the airline can clearly control the size of its fleet by buying or selling planes. Indeed, there is an active market in used planes, making it relatively easy to dispose of unwanted planes. Typically, in the short run, we will consider capital to be the fixed cost and labor to be the variable cost. That is not always true (think of your college's tenured faculty), but generally labor inputs can be adjusted more easily than can capital.

Why is the cost curve U-shaped? Consider the short run in which capital is fixed but labor is variable. In such a situation, there are diminishing returns to the variable factor (labor) because each additional unit of labor has less capital to work with. As a result, the marginal cost of output will rise because the extra output produced by each extra labor unit is going down. In other words, diminishing returns to the variable factor will imply an increasing short-run marginal cost. This shows why diminishing returns lead to rising marginal costs.

Figure 7-3, which contains exactly the same data as Table 7-5, illustrates the point. It shows that the

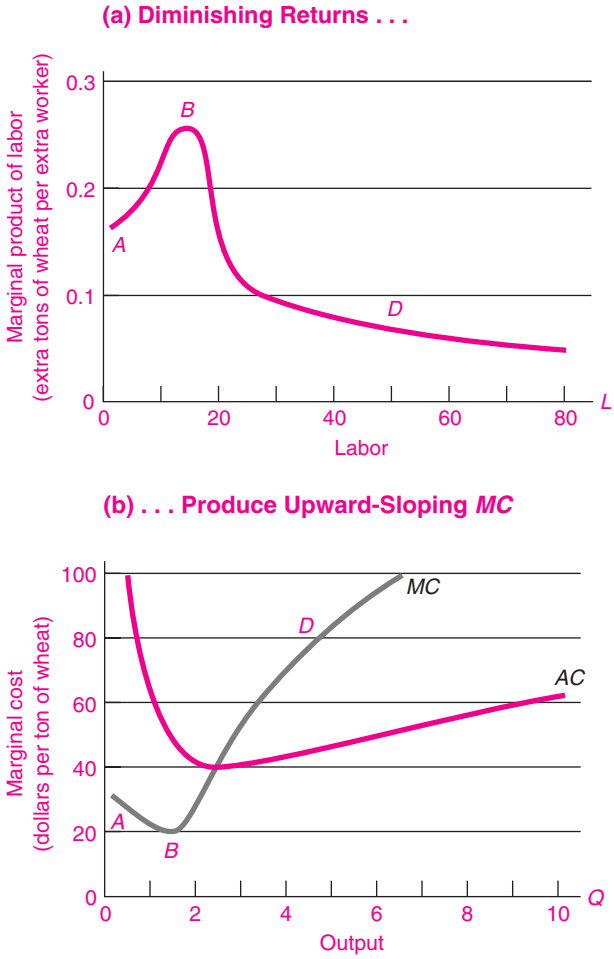


FIGURE 7-3. Diminishing Returns and U-Shaped Cost Curves

The U-shaped cost curves are based on diminishing returns in the short run. With fixed land and variable labor, the marginal product of labor in **(a)** first rises to the left of *B,* peaks at *B,* and then falls at *D* as diminishing returns to labor set in.

The cost curves in **(b)** are derived from the product curves and factor prices. Increasing and then diminishing marginal product of the variable factor gives U-shaped marginal and average cost curves.

[1] For a more complete discussion of the long and short runs, see Chapter 6.

region of increasing marginal product corresponds to falling marginal costs, while the region of diminishing returns implies rising marginal costs.

We can summarize the relationship between the productivity laws and the cost curves as follows:

In the short run, when factors such as capital are fixed, variable factors tend to show an initial phase of increasing marginal product followed by diminishing marginal product. The corresponding cost curves show an initial phase of declining marginal costs, followed by increasing *MC* after diminishing returns have set in.

CHOICE OF INPUTS BY THE FIRM

Marginal Products and the Least-Cost Rule

Every firm must decide *how* to produce its output. Should electricity be produced with oil or coal? Should cars be assembled in the United States or Mexico? Should classes be taught by faculty or graduate students? We now complete the link between production and cost by using the marginal product concept to illustrate how firms select the least-cost combinations of inputs.

In our analysis, we will rely on the fundamental assumption that *firms minimize their costs of production.* This cost-minimization assumption actually makes good sense not only for perfectly competitive firms but for monopolists or even nonprofit organizations like colleges or hospitals. It simply states that the firm should strive to produce its output at the lowest possible cost and thereby have the maximum amount of revenue left over for profits or for other objectives.

A simple example will illustrate how a firm might decide between different input combinations. Say a firm's engineers have calculated that the desired output level of 9 units could be produced with two possible options. In both cases, energy (E) costs \$2 per unit, while labor (L) costs \$5 per hour. Under option 1, the input mix is $E = 10$ and $L = 2$. Option 2 has $E = 4$ and $L = 5$. Which is the preferred option? At the market prices for inputs, total production costs for option 1 are $(\$2 \times 10) + (\$5 \times 2) = \$30$, while total costs for option 2 are $(\$2 \times 4) + (\$5 \times 5) = \$33$. Therefore, option 1 would be the preferred least-cost combination of inputs.

More generally, there are usually many possible input combinations, not just two. But we don't have to calculate the cost of every different combination of inputs in order to find the one which costs the least. Here's a simple way to find the least-cost combination: Start by calculating the marginal product of each input, as we did in Chapter 6. Then divide the marginal product of each input by its factor price. *This gives you the marginal product per dollar of input.* The cost-minimizing combination of inputs comes when the marginal product per dollar of input is equal for all inputs. That is, the marginal contribution to output of each dollar's worth of labor, of land, of oil, and so forth, must be just the same.

Following this reasoning, a firm will minimize its total cost of production when the marginal product per dollar of input is equalized for each factor of production. This is called the least-cost rule.

Least-cost rule: To produce a given level of output at least cost, a firm should buy inputs until it has equalized the marginal product per dollar spent on each input. This implies that

$$\frac{\text{Marginal product of } L}{\text{Price of } L} = \frac{\text{marginal product of } A}{\text{price of } A} = \cdots$$

This rule for firms is exactly analogous to what consumers do when they maximize utilities, as we saw in Chapter 5. In analyzing consumer choice, we saw that to maximize utility, consumers should buy goods so that the marginal utility per dollar spent on each consumer good is equalized for all commodities.

One way of understanding the least-cost rule is the following: Break each factor into packages worth \$1 each. (In our earlier energy-labor example, \$1 of labor would be ⅕ of an hour, while \$1 of energy would be ½ unit.) Then the least-cost rule states that the marginal product of each dollar-unit of input must be equalized. If the marginal products per \$1 of inputs were not equal, you could reduce the low-*MP*-per-dollar input and increase the high-*MP*-per-dollar input and produce the same output at lower cost.

A corollary of the least-cost rule is the substitution rule.

Substitution rule: If the price of one factor falls while all other factor prices remain the same, firms

will profit by substituting the now-cheaper factor for the other factors until the marginal products per dollar are equal for all inputs.

Let's take the case of labor (L). A fall in the price of labor will raise the ratio MP_L/P_L above the MP/P ratio for other inputs. Raising the employment of L lowers MP_L by the law of diminishing returns and therefore lowers MP_L/P_L. Lower price and MP of labor then bring the marginal product per dollar for labor back into equality with that ratio for other factors.

B. ECONOMIC COSTS AND BUSINESS ACCOUNTING

From General Motors down to the corner deli, businesses use more or less elaborate systems to keep track of their costs. Many of the cost categories in business accounting look very similar to the concepts of economic cost we learned above. But there are some important differences between how businesses measure costs and how economists would do it. In this section we will lay out the rudiments of business accounting and point out the differences and similarities with economic costs.

THE INCOME STATEMENT, OR STATEMENT OF PROFIT AND LOSS

Let us start with a small company, called Hot Dog Ventures, Inc. As the name suggests, this company sells gourmet frankfurters in a small store. The operation consists of buying the materials (hot dogs, top-flight buns, expensive mustard, espresso coffee beans) and hiring people to prepare and sell the food. In addition, the company has taken out a loan of $100,000 for its cooking equipment and other restaurant furnishings, and it must pay rent on its store. The founders of Hot Dog Ventures have big aspirations, so they incorporated the business and issued common stock (see Chapter 6 on forms of business organization).

To determine whether Hot Dog Ventures is earning a profit, we must turn to the **income statement,** or—as many companies prefer to call it—the *statement of profit and loss,* shown in Table 7-6. This statement reports the following: (1) Hot Dog Ventures' revenues from sales in 2009, (2) the expenses to be charged against those sales, and (3) the net income, or profits remaining after expenses have been deducted. This gives the fundamental identity of the income statement:

Net income (or profit) = total revenue − total expenses

This definition gives the famous "bottom line" of profits that firms want to maximize. And in many ways, business profits are close to an economist's definition of economic profits. Let's next examine the profit-and-loss statement in more detail, starting from the top. The first line gives the revenues, which were $250,000. Lines 2 through 9 represent the cost of different inputs into the production process. For example, the labor cost is the annual cost of employing labor, while rent is the annual cost of using the building. The selling and administrative costs include the costs of advertising the product and running the back office, while miscellaneous operating costs include the cost of electricity.

The first three cost categories—materials, labor cost, and miscellaneous operating costs—basically correspond to the variable costs of the firm, or its *cost of goods sold.* The next three categories, lines 6 through 8, correspond to the firm's fixed costs, since in the short run they cannot be changed.

Line 8 shows a term we haven't seen before, *depreciation,* which relates to the cost of capital goods. Firms can either rent capital or own their capital goods. In the case of the building, which Hot Dog Ventures rented, we deducted the rent in item (7) of the income statement.

When the firm owns the capital good, the treatment is more complicated. Suppose the cooking equipment has an estimated useful lifetime of 10 years, at the end of which it is useless and worthless. In effect, some portion of the cooking equipment is "used up" in the productive process each year. We call the amount used up "depreciation," and calculate that amount as the cost of the capital input for that year. **Depreciation** measures the annual cost of a capital input that a company actually owns itself.

The same reasoning would apply to any capital goods that a company owns. Trucks wear out, computers become obsolete, and buildings eventually begin to fall apart. For each of these, the company

Income Statement of Hot Dog Ventures, Inc.
(January 1, 2009 to December 31, 2009)

(1)	**Net sales (after all discounts and rebates)**		**$250,000**
	Less cost of goods sold:		
(2)	Materials	$ 50,000	
(3)	Labor cost	90,000	
(4)	Miscellaneous operating costs (utilities, etc.)	10,000	
(5)	Less overhead costs:		
(6)	Selling and administrative costs	15,000	
(7)	Rent for building	5,000	
(8)	Depreciation	15,000	
(9)	Operating expenses	$185,000	185,000
(10)	Net operating income		$ 65,000
	Less:		
(11)	Interest charges on equipment loan		6,000
(12)	State and local taxes		4,000
(13)	Net income (or profit) before income taxes		$ 55,000
(14)	Less: Corporation income taxes		18,000
(15)	**Net income (or profit) after taxes**		**$ 37,000**
(16)	Less: Dividends paid on common stock		15,000
(17)	Addition to retained earnings		$ 22,000

TABLE 7-6. The Income Statement Shows Total Sales and Expenses for a Period of Time

would take a depreciation charge. There are a number of different formulas for calculating each year's depreciation, but each follows two major principles: (*a*) The total amount of depreciation over the asset's lifetime must equal the capital good's historical cost or purchase price; (*b*) the depreciation is taken in annual accounting charges over the asset's accounting lifetime, which is usually related to the actual economic lifetime of the asset.

We can now understand how depreciation would be charged for Hot Dog Ventures. The equipment is depreciated according to a 10-year lifetime, so the $150,000 of equipment has a depreciation charge of $15,000 per year (using the simplest "straight-line" method of depreciation). If Hot Dog Ventures owned its store, it would have to take a depreciation charge for the building as well.

Adding up all the costs so far gives us the operating expenses (line 9). The net operating income is net revenues minus operating expenses (line 1 minus line 9). Have we accounted for all the costs of production yet? Not quite. Line 11 includes the annual cost of interest on the $100,000 loan. This should be thought of as the cost of borrowing the financial capital. While this is a fixed cost, it is typically kept separate from the other fixed costs. State and local taxes, such as property taxes, are treated as another expense. Deducting lines 11 and 12 gives a total of $55,000 in profits before income taxes. How are these profits divided? Approximately $18,000 goes to the federal government in the form of corporate income taxes. That leaves a profit of $37,000 after taxes. Dividends of $15,000 on the common stock are paid, leaving $22,000 to be plowed back as retained earnings in the business. Again, note that profits are a residual of sales minus costs.

THE BALANCE SHEET

Business accounting is concerned with more than the profits and losses that are the economic driving force. Business accounts also include the **balance sheet,** which is a picture of financial conditions on a given date. This statement records what a firm, person, or nation is worth at a given point in time. On one side of the balance sheet are the **assets** (valuable

properties or rights owned by the firm). On the other side are two items, the **liabilities** (money or obligations owed by the firm) and **net worth** (or net value, equal to total assets minus total liabilities).

One important distinction between the income statement and the balance sheet is that between stocks and flows. A **stock** represents the level of a variable, such as the amount of water in a lake or, in this case, the dollar value of a firm. A **flow** variable represents the change per unit of time, like the flow of water in a river or the flow of revenue and expenses into and out of a firm. *The income statement measures the flows into and out of the firm, while the balance sheet measures the stocks of assets and liabilities at the end of the accounting year.*

The fundamental identity or balancing relationship of the balance sheet is that total assets are balanced by total liabilities plus the net worth of the firm to its owners:

$$\text{Total assets} = \text{total liabilities} + \text{net worth}$$

We can rearrange this relationship to find:

$$\text{Net worth} = \text{assets} - \text{liabilities}$$

Let us illustrate this by considering Table 7-7, which shows a simple balance sheet for Hot Dog Ventures, Inc. On the left are assets, and on the right are liabilities and net worth. A blank space has been deliberately left next to the retained earnings entry because the only correct entry compatible with our fundamental balance sheet identity is $200,000. *A balance sheet must always balance because net worth is a residual defined as assets minus liabilities.* Suppose one item on a balance sheet changes (such as an increase in assets); then there must be a corresponding change on the balance sheet to maintain balance (a decrease in assets, an increase in liabilities, or an increase in net worth).

To illustrate how net worth always balances, suppose that hot dogs valued at $40,000 have spoiled. Your accountant reports to you: "Total assets are down $40,000; liabilities remain unchanged. This means total net worth has decreased by $40,000, and I have no choice but to write net worth down from the previous $210,000 to only $170,000." That's how accountants keep score.

We summarize our analysis of accounting concepts as follows:

1. The income statement shows the flow of sales, cost, and revenue over the year or accounting period. It measures the flow of dollars into and out of the firm over a specified period of time.
2. The balance sheet indicates an instantaneous financial picture or snapshot. It is like a measure of the stock of water in a lake. The major items are assets, liabilities, and net worth.

Balance Sheet of Hot Dog Ventures, Inc.
(December 31, 2009)

Assets		Liabilities and net worth	
		Liabilities	
Current assets:		Current liabilities:	
Cash	$ 20,000	Accounts payable	$ 20,000
Inventory	80,000	Notes payable	20,000
Fixed assets:		Long-term liabilities:	
Equipment	150,000	Bonds payable	100,000
Buildings	100,000		
		Net worth	
		Stockholders' equity:	
		Common stock	10,000
		Retained earnings	.....
Total	**$350,000**	**Total**	**$350,000**

TABLE 7-7. The Balance Sheet Records the Stock of Assets and Liabilities, plus Net Worth, of a Firm at a Given Point in Time

Accounting Conventions

In examining the balance sheet in Table 7-7, you might well ask, How are the values of the different items measured? How do the accountants know that the equipment is worth $150,000?

The answer is that accountants use a set of agreed-upon rules or accounting conventions to answer most questions. The most important assumption used in a balance sheet is that the value placed on almost every item reflects its *historical cost*. This differs from the economist's concept of "value," as we will see in the next section. For example, the inventory of hot-dog buns is valued at the price that was paid for them. A newly purchased fixed asset—a piece of equipment or a building—is valued at its purchase price (this being the historical-cost convention). Older capital is valued at its purchase price minus accumulated depreciation, thus accounting for the gradual decline in usefulness of capital goods. Accountants use historical cost because it reflects an objective evaluation and is easily verified.

In Table 7-7 current assets are convertible into cash within a year, while fixed assets represent capital goods and land. Most of the specific items listed are self-explanatory. Cash consists of coins, currency, and money on deposit in the bank. Cash is the only asset whose value is exact rather than an estimate.

On the liabilities side, accounts payable and notes payable are sums owed to others for goods bought or for borrowed funds. Bonds payable are long-term loans floated in the market. The last item on the balance sheet is net worth, which is also called "stockholders' equity." This has two components. The first is common stock, which represents what the stockholders originally contributed to the business. The second component is retained earnings. These are earnings reinvested in the business after the deduction of any distributions to shareholders, such as dividends. Recall from the income statement that Hot Dog Ventures had $22,000 of retained earnings for 2009. The net worth is the firm's assets less liabilities, when valued at historical cost. Confirm that net worth must equal $210,000 in Table 7-7.

Financial Finagling

Now that we have reviewed the principles of accounting, we see that there is considerable judgment involved in determining the exact treatment of certain items. In the late 1990s, under pressure to produce rapidly growing earnings, many companies manipulated their accounts to show glowing results or to paper over losses. Some of the most egregious examples included pretending that capital assets were current revenues (Enron, Global Crossing); capitalizing the outflow while recognizing the inflow as current revenues (Enron, Qwest); increasing the salvage value of trucks over time (Waste Management); increasing the value of the unused capacity of landfills even as they fill up (Waste Management); and reporting happy pro-forma numbers when the reality was unpleasant (Amazon.com, Yahoo, and Qualcomm, among a crowd of other dot-coms dead or alive).

To see how an accounting fraud works, let's take the example of Enron. Enron started off as a genuinely profitable business which owned the largest interstate network of natural-gas pipelines. To continue its rapid growth, it turned to trading natural-gas futures, and then it leveraged its business model into other markets.

Along the way, however, its profits began to decline and it hid the declines from investors. You might well ask, How could a large, publicly owned company like Enron have fooled virtually all of the people most of the time until 2001?

Its success in hiding its failures rested on four complementary factors. First, when troubles arose, Enron began to exploit ambiguities in accounting principles, such as the ones described above. One example was a deal called "Project Braveheart" with Blockbuster Video. This deal projected future revenues over the next 20 years with a present value of $111 million, and Enron accounted for them as current revenues even though the projections were based on highly dubious assumptions.

Second, the firm elected not to report the details of many financial transactions—for example, it hid hundreds of partnerships from its stockholders. Third, the board of directors and outside auditors were passive and did not challenge or in some cases even inquire into some details of Enron's accounts. Finally, the investment community, such as the large mutual funds, exercised little deep independent analysis of Enron's numbers even though at its peak Enron absorbed $70 billion of investors' funds.

The Enron case is a reminder that financial markets, accounting firms, and investment managers can be fooled into investing many billions of dollars

when firm insiders engage in aggressive accounting and fraudulent practices. A larger set of issues arose in 2007–2008 when a trillion dollars of poorly designed mortgage-backed securities got sound credit ratings from bond-rating agencies, but agencies and investors had little understanding of the income streams behind these securities. The history of such accounting and financial finagling is a reminder of the importance of sound accounting practices and the need for vigilant oversight by government and nongovernment bodies.

C. OPPORTUNITY COSTS

In this section we look at costs from yet another angle. Remember that one of the cardinal tenets of economics is that resources are scarce. That means every time we choose to use a resource one way, we've given up the opportunity to utilize it another way. That's easy to see in our own lives, where we must constantly decide what to do with our limited time and income. Should we go to a movie or study for next week's test? Should we travel to Mexico or buy a car? Should we get postgraduate or professional training or begin work right after college?

In each of these cases, making a choice in effect costs us the opportunity to do something else. The value of the best alternative forgone is called the opportunity cost, which we met briefly in Chapter 1 and develop more thoroughly here. The dollar cost of going to a movie instead of studying is the price of a ticket, but the opportunity cost also includes the possibility of getting a higher grade on the exam. The opportunity costs of a decision include all its consequences, whether they reflect monetary transactions or not.

Decisions have opportunity costs because choosing one thing in a world of scarcity means giving up something else. The **opportunity cost** is the value of the most valuable good or service forgone.

One important example of opportunity cost is the cost of going to college. If you went to a public university in your state in 2008, the total costs of tuition, books, and travel averaged about $7000. Does this mean that $7000 was your opportunity cost of going to school? Definitely not! You must include as well the *opportunity cost of the time* spent studying and going to classes. A full-time job for a college-age high school graduate averaged $26,000 in 2008. If we add up both the actual expenses and the earnings forgone, we would find that the opportunity cost of college was $33,000 (equal to $7000 + $26,000) rather than $7000 per year.

Business decisions have opportunity costs, too. Do all opportunity costs show up on the profit-and-loss statement? Not necessarily. In general, business accounts include only transactions in which money actually changes hands. By contrast, the economist always tries to "pierce the veil of money" to uncover the real consequences that lie behind the dollar flows and to measure the true *resource costs* of an activity. Economists therefore include all costs—whether they reflect monetary transactions or not.

There are several important opportunity costs that do not show up on income statements. For example, in many small businesses, the family may put in many unpaid hours, which are not included as accounting costs. Nor do business accounts include a capital charge for the owner's financial contributions. Nor do they include the cost of the environmental damage that occurs when a business dumps toxic wastes into a stream. But from an economic point of view, each of these is a genuine cost to the economy.

Let's illustrate the concept of opportunity cost by considering the owner of Hot Dog Ventures. The owner puts in 60 hours a week but earns no "wages." At the end of the year, as Table 7-6 showed, the firm earns a profit of $37,000—pretty good for a neophyte firm.

Or is it? The economist would insist that we should consider the value of a factor of production regardless of how the factor happens to be owned. We should count the owner's own labor as a cost even though the owner does not get paid directly but instead receives compensation in the form of profits. Because the owner has alternative opportunities for work, we must value the owner's labor in terms of the lost opportunities.

A careful examination might show that Hot Dog Ventures' owner could find a similar and equally interesting job working for someone else and earning $60,000. This represents the opportunity cost or earnings forgone because the owner decided to become the unpaid owner of a small business rather than the paid employee of another firm.

Therefore, the economist continues, let us calculate the true economic profits of the hot-dog firm. If we take the measured profits of $37,000 and subtract the $60,000 opportunity cost of the owner's labor, we find a net *loss* of $23,000. Hence, although the accountant might conclude that Hot Dog Ventures is economically viable, the economist would pronounce that the firm is an unprofitable loser.

What Was the Cost of the War in Iraq?

One of the most vexing questions facing Americans is to calculate how much the war in Iraq has cost. This issue involves questions of opportunity cost for the nation rather than for the firm, but the principles are similar. The Bush administration originally estimated that the war would be over quickly and that the costs would be around $50 billion. In reality, the war proved much longer and more expensive. According to a congressional report in 2008, the cumulative total spending on the campaigns in Iraq and Afghanistan was about $750 billion.

But economists Linda Bilmes and Joseph Stiglitz argue that even this large number underestimates the total because it does not take into account the entire opportunity cost of the war. One example of the understatement is that the pay of members of the military does not reflect the total costs to the nation because it underestimates costs in health care and other benefits. They write:

> When a young soldier is killed in Iraq or Afghanistan, his or her family will receive a U.S. government check for just $500,000 (combining life insurance with a "death gratuity")—far less than the typical amount paid by insurance companies for the death of a young person in a car accident. The "budgetary cost" of $500,000 is clearly only a fraction of the total cost society pays for the loss of life—and no one can ever really compensate the families. Moreover, disability pay seldom provides adequate compensation for wounded troops or their families. Indeed, in one out of five cases of seriously injured soldiers, someone in their family has to give up a job to take care of them.

Bilmes and Stiglitz also calculate that oil prices are higher because of the war, contributing to the increase in oil prices from $25 per barrel in 2003 to a peak of $155 a barrel in 2008.

When they add up all the opportunity costs through 2008, they conclude that the war in Iraq will cost the American people $3 trillion, or about $30,000 per household. While these numbers are subject to debate, they are a timely reminder of the difference between an accounting number and true economic or opportunity cost.

OPPORTUNITY COST AND MARKETS

At this point, however, you might well say: "Now I'm totally confused. First I learned that price is a good measure of true social cost in the marketplace. Now you tell me that opportunity cost is the right concept. Can't you economists make up your minds?"

Actually, there is a simple explanation: *In well-functioning markets, when all costs are included, price equals opportunity cost.* Assume that a commodity like wheat is bought and sold in a competitive market. If I bring my wheat to market, I will receive a number of bids from prospective buyers: $2.502, $2.498, and $2.501 per bushel. These represent the values of my wheat to, say, three different flour mills. I pick the highest—$2.502. The opportunity cost of this sale is the value of the best available alternative—that is, the second-highest bid, at $2.501—which is almost identical to the price that is accepted. As the market approaches perfect competition, the bids get closer and closer until, at the limit, the second-highest bid (which is our definition of opportunity cost) exactly equals the highest bid (which is the price). In competitive markets, numerous buyers compete for resources to the point where price is bid up to the best available alternative and is therefore equal to the opportunity cost.

Opportunity Cost outside Markets. The concept of opportunity cost is particularly crucial when you are analyzing transactions that take place outside markets. How do you measure the value of a road or a park? Of a health or safety regulation? Even the allocation of student time can be explained using opportunity cost.

- The notion of opportunity cost explains why students watch more TV the week after exams than the week before exams. Watching TV right before an exam has a high opportunity cost, for the alternative use of time (studying) has high value in improving grade performance and getting a good job. After exams, time has a lower opportunity cost.

- Or take the case of a proposal to drill for oil off the California coast. A storm of complaints is heard. A defender of the program states, "We need that oil to protect us from insecure foreign sources who are holding us hostage. We have plenty of seawater to go around. This is just good economics for the nation." In fact, it might be poor economics because of the opportunity cost. If drilling leads to oil spills that spoil the beaches, it might reduce the recreational value of the ocean. That opportunity cost might not be easily measured, but it is every bit as real as the value of oil under the waters.

The Road Not Traveled. Opportunity cost, then, is a measure of what has been given up when we make a decision. Consider what Robert Frost had in mind when he wrote,

> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

What other road did Frost have in mind? An urban life? A life where he would not be able to write of roads and walls and birches? Imagine the immeasurable opportunity cost to all of us if Robert Frost had taken the road more traveled by.

But let us return from the poetic to the practical. The crucial point to grasp is this:

Economic costs include, in addition to explicit money outlays, those opportunity costs incurred because resources can be used in alternative ways.

SUMMARY

A. Economic Analysis of Costs

1. Total cost (TC) can be broken down into fixed cost (FC) and variable cost (VC). Fixed costs are unaffected by any production decisions, while variable costs are incurred on items like labor or materials which increase as production levels rise.
2. Marginal cost (MC) is the extra total cost resulting from 1 extra unit of output. Average total cost (AC) is the sum of ever-declining average fixed cost (AFC) and average variable cost (AVC). Short-run average cost is generally represented by a U-shaped curve that is always intersected at its minimum point by the rising MC curve.
3. Useful rules to remember are

$$TC = FC + VC \quad AC = \frac{TC}{q} \quad AC = AFC + AVC$$

   At the bottom of U-shaped AC, $MC = AC$ = minimum AC.
4. Costs and productivity are like mirror images. When the law of diminishing returns holds, the marginal product falls and the MC curve rises. When there is an initial stage of increasing returns, MC initially falls.
5. We can apply cost and production concepts to a firm's choice of the best combination of factors of production. Firms that desire to maximize profits will want to minimize the cost of producing a given level of output. In this case, the firm will follow the least-cost rule: different factors will be chosen so that the marginal product per dollar of input is equalized for all inputs. This implies that $MP_L/P_L = MP_A/P_A = \cdots$.

B. Economic Costs and Business Accounting

6. To understand accounting, the most important relationships are:
   a. The character of the income statement (or profit-and-loss statement); the residual nature of profits; and depreciation of fixed assets.
   b. The fundamental balance sheet relationship between assets, liabilities, and net worth; the breakdown of each of these into financial and fixed assets; and the residual nature of net worth.

C. Opportunity Costs

7. The economist's definition of costs is broader than the accountant's. Economic cost includes not only the obvious out-of-pocket purchases or monetary transactions but also more subtle opportunity costs, such as the return to labor supplied by the owner of a firm. These opportunity costs are tightly constrained by the bids and offers in competitive markets, so price is close to opportunity cost for marketed goods and services.
8. The most important application of opportunity cost arises for nonmarket goods—those like clean air or health or recreation—which may be highly valuable even though they are not bought and sold in markets.

CONCEPTS FOR REVIEW

Analysis of Costs

total costs: fixed and variable
marginal cost
least-cost rule:

$$\frac{MP_L}{P_L} = \frac{MP_A}{P_A} = \frac{MP_{\text{any factor}}}{P_{\text{any factor}}}$$

$TC = FC + VC$
$AC = TC/q = AFC + AVC$

Accounting Concepts

income statement (profit-and-loss statement): sales, cost, profits
depreciation
fundamental balance sheet identity
assets, liabilities, and net worth
stocks vs. flows
opportunity cost
cost concepts in economics and accounting

FURTHER READING AND INTERNET WEBSITES

Further Reading

Advanced treatment of cost and production theory can be found in intermediate textbooks. See the list provided in Chapter 3.

You can find interesting articles on business cost, production, and decision problems in magazines such as *Business Week, Fortune, Forbes,* and *The Economist.* An excellent nontechnical analysis of the Enron fraud is contained in Paul M. Healy and Krishna G. Palepu, "The Fall of Enron," *Journal of Economic Perspectives,* Spring 2003, pp. 3–26.

The quotation on the cost of war is from Linda J. Bilmes and Joseph E. Stiglitz, "The Iraq War Will Cost Us $3 Trillion, and Much More," *Washington Post,* March 9, 2008, p. B1. Their full study is Joseph E. Stiglitz and Linda J. Bilmes, *The Three Trillion Dollar War: The True Cost of the Iraq Conflict* (Norton, New York, 2008).

Websites

Good case studies on costs and production can be found in the business press. See the websites of the business magazines listed above, *www.businessweek.com, www.fortune.com, www.forbes.com,* and *www.economist.com.* Some of these sites require a fee or subscription.

Information about individual firms is filed with the Securities and Exchange Commission and can be found at *www.sec.gov/edgarhp.htm.*

QUESTIONS FOR DISCUSSION

1. During his major-league career from 1936 to 1960, Ted Williams had 7706 at bats and 2654 hits.
   a. What was his lifetime batting average?
   b. In his last year, 1960, Williams had 310 at bats and 98 hits. What was his lifetime batting average at the end of 1959? What was his batting average for 1960?
   c. Explain the relationship between his average for 1959 and the change of his lifetime average from 1959 to 1960. State how this illustrates the relationship between *MC* and *AC*.
2. To the $55 of fixed cost in Table 7-3, add $90 of additional *FC*. Now calculate a whole new table, with the same *VC* as before but new *FC* = $145. What happens to *MC, AVC?* To *TC, AC, AFC?* Can you verify that minimum *AC* is now at $q^* = 5$ with *AC* = $60 = *MC?*
3. Explain why *MC* cuts *AC* and *AVC* at their minimum values (i.e., the bottom of their U-shaped cost curves).
4. "Compulsory military service allows the government to fool itself and the people about the true cost of a big army." Compare the budget cost and the opportunity cost of a voluntary army (where army pay is high) with those of compulsory service (where pay is low). What does the concept of opportunity cost contribute to analyzing the quotation?
5. Consider the data in Table 7-8, which contains a situation similar to that in Table 7-5.
   a. Calculate the *TC, VC, FC, AC, AVC,* and *MC.* On a piece of graph paper, plot the *AC* and *MC* curves.
   b. Assume that the price of labor doubles. Calculate a new *AC* and *MC.* Plot the new curves and compare them with those in **a.**

(1) Output (tons of wheat)	(2) Land inputs (acres)	(3) Labor inputs (workers)	(4) Land rent ($ per acre)	(5) Labor wage ($ per worker)
0	15	0	12	5
1	15	6	12	5
2	15	11	12	5
3	15	15	12	5
4	15	21	12	5
5	15	31	12	5
6	15	45	12	5
7	15	63	12	5

TABLE 7-8.

c. Now assume that total factor productivity doubles (i.e., that the level of output doubles for each input combination). Repeat the exercise in **b.** Can you see two major factors that tend to affect a firm's cost curves?

6. Explain the fallacies in each of the following:

a. Average costs are minimized when marginal costs are at their lowest point.

b. Because fixed costs never change, average fixed cost is a constant for each level of output.

c. Average cost is rising whenever marginal cost is rising.

d. The opportunity cost of drilling for oil in Yosemite Park is zero because no firm produces anything there.

e. A firm minimizes costs when it spends the same amount on each input.

7. In 2008, a fictitious software company named EconDisaster.com sold $7000 worth of a game called "Global Financial Meltdown." The company had salaries of $1000, rent of $500, and electricity use of $500, and it purchased a computer for $5000. The company uses straight-line depreciation with a lifetime of 5 years (this means that depreciation is calculated as the historical cost divided by the lifetime). It pays a corporation tax of 25 percent on profits and paid no dividends. Construct its income statement for 2008 based on Table 7-6.

8. Next, construct the balance sheet for EconDisaster.com for December 31, 2008. The company had no assets at the beginning of the year. The owners contributed $10,000 of start-up capital and obtained common stock. Net income and retained earnings can be calculated from question 7.

Appendix 7

PRODUCTION, COST THEORY, AND DECISIONS OF THE FIRM

The production theory described in Chapter 6 and the cost analysis of this chapter are among the fundamental building blocks of microeconomics. A thorough understanding of production and cost is necessary for an appreciation of how economic scarcity gets translated into prices in the marketplace. This appendix develops these concepts further and introduces the concept of an equal-product curve, or isoquant.

A NUMERICAL PRODUCTION FUNCTION

Production theory and cost analysis have their roots in the concept of a production function, which shows the maximum amount of output that can be produced with various combinations of inputs. Table 7A-1 starts with a numerical example of a constant-returns-to-scale production function, showing the amount of inputs along the axes and the amount of output at the grid points of the table.

Along the left-hand side are listed the varying amounts of land, going from 1 unit to 6 units. Along the bottom are listed amounts of labor, which also go from 1 to 6. Output corresponding to each land row and labor column is listed inside the table.

If we are interested in knowing exactly how much output there will be when 3 units of land and 2 units of labor are available, we count up 3 units of land and then go over 2 units of labor. The answer is seen to be 346 units of product. (Can you identify some other input combinations that will produce $q = 346$?) Similarly, we find that 3 units of land and 6 of labor produce 600 units of q. Remember that the production function shows the maximum output available given engineering skills and technical knowledge available at a particular time.

Land \ Labor	1	2	3	4	5	6
6	**346**	490	600	692	775	846
5	316	448	548	632	705	775
4	282	400	490	564	632	692
3	245	**346**	423	490	548	600
2	200	282	**346**	400	448	490
1	141	200	245	282	316	**346**

TABLE 7A-1. A Tabular Picture of a Production Function Relating Amount of Output to Varying Combinations of Labor and Land Inputs

When you have 3 land units and 2 labor units available, the engineer tells you the maximum obtainable output is 346 units. Note the different ways to produce 346. Do the same for 490. (The production function shown in the table is a special case of the Cobb-Douglas production function, one given by the formula $Q = 100\sqrt{2LA}$.)

THE LAW OF DIMINISHING MARGINAL PRODUCT

Table 7A-1 can nicely illustrate the law of diminishing returns. First, recall that the marginal product of labor is the extra production resulting from 1 additional unit of labor when land and other inputs are held constant. At any point in Table 7A-1, we can find the marginal product of labor by subtracting the output from the number on its right in the same row. Thus, when there are 2 units of land and 4 units of labor, the marginal product of an additional laborer would be 48, or 448 minus 400 in the second row.

By the "marginal product of land" we mean, of course, the extra product resulting from 1 additional unit of land when labor is held constant. It is calculated by comparing adjacent items in a given column. Thus, when there are 2 units of land and 4 units of labor, the marginal product of land is shown in the fourth column as 490 – 400, or 90.

We can easily find the marginal product of each of our two factors by comparing adjacent entries in the vertical columns or horizontal rows of Table 7A-1.

Having defined the concept of marginal product of an input, we now can easily define the law of diminishing returns: *The law of diminishing returns*

states that as we increase one input and hold other inputs constant, the marginal product of the varying input will, at least after some point, decline.

To illustrate this, hold land constant in Table 7A-1 by sticking to a given row—say, the row corresponding to land equal to 2 units. Now let labor increase from 1 to 2 units, from 2 to 3 units, and so forth. What happens to q at each step?

As labor goes from 1 to 2 units, the level of output increases from 200 to 282 units, or by 82 units. But the next dose of labor adds only 64 units, or 346 − 282. Diminishing returns have set in. Still further additions of a single unit of labor give us, respectively, only 54 extra units of output, 48 units, and finally 42 units. You can easily verify that the law holds for other rows and that the law holds when land is varied and labor held constant.

We can use this example to verify our intuitive justification of the law of diminishing returns—the assertion that the law holds because the fixed factor decreases relative to the variable factor. According to this explanation, each unit of the variable factor has less and less of the fixed factor to work with. So it is natural that extra product should drop off.

If this explanation is to hold water, output should increase proportionately when both factors are increased together. When labor increases from 1 to 2 and land simultaneously increases from 1 to 2, we should get the same increase in product as when both increase *simultaneously* from 2 to 3. This can be verified in Table 7A-1. In the first move we go from 141 to 282, and in the second move the product increases from 282 to 423, an equal jump of 141 units.

LEAST-COST FACTOR COMBINATION FOR A GIVEN OUTPUT

The numerical production function shows us the different ways to produce a given level of output. But which of the many possibilities should the firm use? If the desired level of output is $q = 346$, there are no less than four different combinations of land and labor, shown as A, B, C, and D in Table 7A-2.

As far as the engineer is concerned, each of these combinations is equally good at producing an output of 346 units. But the manager, interested in minimizing cost, wants to find the combination that costs the least.

	(1)	(2)	(3)	(4)
	Input Combinations		Total cost when P_L = \$2 P_A = \$3 (\$)	Total cost when P_L = \$2 P_A = \$1 (\$)
	Labor L	Land A		
A	1	6	20	_____
B	2	3	13	7
C	**3**	**2**	**12**	_____
D	6	1	15	_____

TABLE 7A-2. Inputs and Costs of Producing a Given Level of Output

Assume that the firm has chosen 346 units of output. Then it can use any of the four choices of input combinations shown as A, B, C, and D. As the firm moves down the list, production becomes more labor-intensive and less land-intensive. Fill in the missing numbers.

The firm's choice among the different techniques will depend on input prices. When P_L = \$2 and P_A = \$3, verify that the cost-minimizing combination is C. Show that lowering the price of land from \$3 to \$1 leads the firm to choose a more land-intensive combination at B.

Let us suppose that the price of labor is \$2 and the price of land \$3. The total costs when input prices are at this level are shown in the third column of Table 7A-2. For combination A, the total labor and land cost will be \$20, equal to $(1 \times \$2) + (6 \times \$3)$. Costs at B, C, and D will be, respectively, \$13, \$12, and \$15. At the assumed input prices, C is the least costly way to produce the given output.

If either of the input prices changes, the equilibrium proportion of the inputs will also change so as to use less of the input that has gone up most in price. (This is just like the substitution effect in Chapter 5's discussion of consumer demand.) As soon as input prices are known, the least-cost method of production can be found by calculating the costs of different input combinations.

Equal-Product Curves

The commonsense numerical analysis of the way in which a firm will combine inputs to minimize costs can be made more vivid by the use of diagrams. We will take the diagrammatic approach by putting together two new curves, the equal-product curve and the equal-cost line.

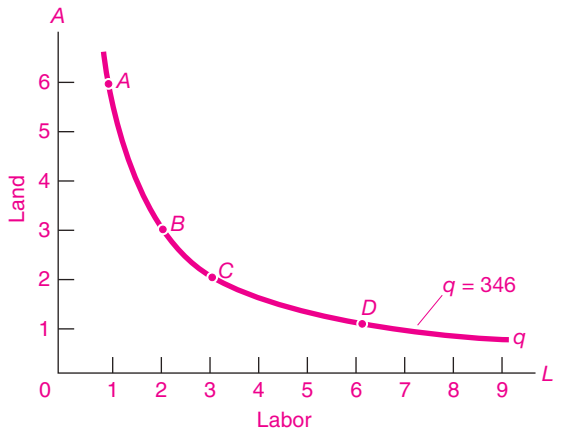


FIGURE 7A-1. Equal-Product Curve

All the points on the equal-product curve represent the different combinations of land and labor that can be used to produce the same 346 units of output.

Let's turn Table 7A-1 into a continuous curve by drawing a smooth curve through all the points that yield $q = 346$. This smooth curve, shown in Figure 7A-1, indicates all the different combinations of labor and land that yield an output of 346 units. This is called an **equal-product curve** or **isoquant** and is analogous to the consumer's indifference curve discussed in the appendix to Chapter 5. You should be able to draw on Figure 7A-1 the corresponding equal-product curve for output equal to 490 by getting the data from Table 7A-1. Indeed, an infinite number of such equal-product contour lines could be drawn in.

Equal-Cost Lines

Given the price of labor and land, the firm can evaluate the total cost for points *A, B, C,* and *D* or for any other point on the equal-product curve. The firm will minimize its costs when it selects that point on its equal-product curve that has the lowest total cost.

An easy technique for finding the least-cost method of production is to construct **equal-cost lines.** This is done in Figure 7A-2, where the family of parallel straight lines represents a number of equal-cost curves when the price of labor is \$2 and the price of land \$3.

To find the total cost for any point, we simply read off the number appended to the equal-cost line going through that point. The lines are all straight and parallel because the firm is assumed to be able to buy all it wishes of either input at constant prices. The lines are somewhat flatter than 45° because the price of labor P_L is somewhat less than the price of land P_A. More precisely, we can always say that the arithmetic value of the slope of each equal-cost line must equal the ratio of the price of labor to that of land—in this case $P_L/P_A = \frac{2}{3}$.

Equal-Product and Equal-Cost Contours: Least-Cost Tangency

Combining the equal-product and equal-cost lines, we can determine the optimal, or cost-minimizing, position of the firm. Recall that the optimal input combination comes at that point where the given output of $q = 346$ can be produced at least cost. To find such a point, simply superimpose the single green equal-product curve upon the family of blue equal-cost lines, as shown in Figure 7A-3. The firm will always keep moving along the green convex curve of Figure 7A-3 as long as it is able to cross over to lower cost lines. Its equilibrium will therefore be at *C,* where the equal-product curve touches (but does not cross) the lowest equal-cost line. This is a point of tangency, where the slope of the equal-product curve just matches the slope of an equal-cost line and the curves are just kissing.

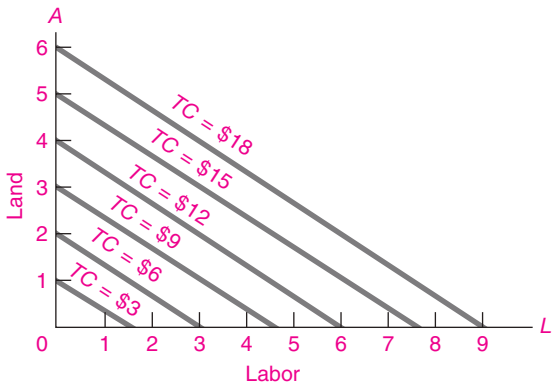


FIGURE 7A-2. Equal-Cost Lines

Every point on a given equal-cost line represents the same total cost. The lines are straight because factor prices are constant, and they all have a negative slope equal to the ratio of labor price to land price, \$2/\$3, and hence are parallel.

Substituting Inputs to Minimize Cost of Production

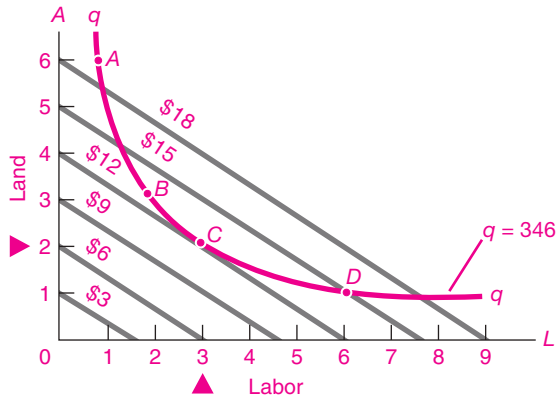


FIGURE 7A-3. Least-Cost Input Combination Comes at *C*

The firm desires to minimize its costs of producing a given output of 346. It thus seeks out the least expensive input combination along its green equal-product curve. It looks for the input combination that is on the lowest of the equal-cost lines. Where the equal-product curve touches (but does not cross) the lowest equal-cost line is the least-cost position. This tangency means that factor prices and marginal products are proportional, with equalized marginal products per dollar.

We already know that the slope of the equal-cost curves is P_L/P_A. But what is the slope of the equal-product curve? Recall from Chapter 1's appendix that the slope at a point of a curved line is the slope of the straight line tangent to the curve at the point in question. For the equal-product curve, this slope is a "substitution ratio" between the two factors. It depends upon the relative marginal products of the two factors of production, namely, MP_L/MP_A—just as the rate of substitution between two goods along a consumer's indifference curve was earlier shown to equal the ratio of the marginal utilities of the two goods (see the appendix to Chapter 5).

Least-Cost Conditions

Using our graphical apparatus, we have therefore derived the conditions under which a firm will minimize its costs of production:

1. The ratio of marginal products of any two inputs must equal the ratio of their factor prices:

$$\text{Substitution ratio} = \frac{\text{marginal product of labor}}{\text{marginal product of land}}$$

$$= \begin{matrix}\text{slope of}\\ \text{equal-product}\\ \text{curve}\end{matrix} = \frac{\text{price of labor}}{\text{price of land}}$$

2. We can also rewrite condition 1 in a different and illuminating way. From the last equation it follows that the marginal product per dollar received from the (last) dollar of expenditure must be the same for every productive input:

$$\frac{\text{Marginal product of } L}{\text{Price of } L} =$$

$$\frac{\text{marginal product of } A}{\text{price of } A} = \ldots$$

But you should not be satisfied with abstract explanations. Always remember the commonsense economic explanation which shows how a firm will distribute its expenditure among inputs to equalize the marginal product per dollar of spending.

SUMMARY TO APPENDIX

1. A production-function table lists the output that can be produced for each labor column and each land row. Diminishing returns to one variable factor, when other factors are held fixed or constant, can be shown by calculating the decline of marginal products in any row or column.
2. An equal-product curve or isoquant depicts the alternative input combinations that produce the same level of output. The slope, or substitution ratio, along such an equal-product curve equals relative marginal products (e.g., MP_L/MP_A). Curves of equal total cost are parallel lines with slopes equal to factor-price ratios (P_L/P_A). Least-cost equilibrium comes at the tangency point, where an equal-product curve touches but does not cross the lowest *TC* curve. In least-cost equilibrium, marginal products are proportional to factor prices, with equalized marginal product per dollar spent on all factors (i.e., equalized MP_i/P_i).

CONCEPTS FOR REVIEW

equal-product curves, isoquants
parallel lines of equal *TC*
substitution ratio = MP_L/MP_A
P_L/P_A as the slope of parallel equal-*TC* lines
least-cost tangency condition: $MP_L/MP_A = P_L/P_A$ or $MP_L/P_L = MP_A/P_A$

QUESTIONS FOR DISCUSSION

1. Show that raising labor's wage while holding land's rent constant will steepen the blue equal-cost lines and move tangency point *C* in Figure 7A-3 northwest toward *B,* with the now-cheaper input substituted for the input which is now more expensive. If we substitute capital for labor, restate the result. Should union leaders recognize this relationship?
2. What is the least-cost combination of inputs if the production function is given by Table 7A-1 and input prices are as shown in Figure 7A-3, where $q = 346$? What would be the least-cost ratio for the same input prices if output doubled to $q = 692$? What has happened to the "factor intensity," or land-labor ratio? Can you see why this result would hold for any output change under constant returns to scale?

CHAPTER

8

Analysis of Perfectly Competitive Markets

Cost of production would have no effect on competitive price if it could have none on supply.

John Stuart Mill

We have described how the market mechanism performs a kind of miracle every day, providing our daily necessities like bread and a vast array of high-quality goods and services without central control or direction. Exactly how does this market mechanism work?

The answer begins with the two sides to every market—supply and demand. These two components must be put together to understand how the market as a whole behaves. This first chapter on industrial organization analyzes the behavior of perfectly competitive markets; these are idealized markets in which firms and consumers are too small to affect the price. The first section shows how competitive firms behave, after which some special cases are examined. The chapter concludes by showing that a perfectly competitive industry will be efficient. After having surveyed the central case of perfect competition, we move on in the following chapters to other forms of market behavior, such as monopolies.

A. SUPPLY BEHAVIOR OF THE COMPETITIVE FIRM

BEHAVIOR OF A COMPETITIVE FIRM

We begin with an analysis of perfectly competitive firms. If you own such a firm, how much should you produce? How much wheat should Farmer Smith produce if wheat sells at $6 per bushel?

Our analysis of perfectly competitive firms relies on two key assumptions. First, we will assume that our competitive firm *maximizes profits*. Second, we reiterate that perfect competition is a world of *atomistic firms who are price-takers*.

Profit Maximization

Profits are like the net earnings or take-home pay of a business. They represent the amount a firm can pay in dividends to the owners, reinvest in new plant

and equipment, or employ to make financial investments. All these activities increase the value of the firm to its owners.

Firms maximize profits because that maximizes the economic benefit to the owners of the firm. Allowing lower-than-maximum profits is like asking for a pay cut, which few business owners will voluntarily undertake.

Profit maximization requires the firm to manage its internal operations efficiently (prevent waste, encourage worker morale, choose efficient production processes, and so forth) and to make sound decisions in the marketplace (buy the correct quantity of inputs at least cost and choose the optimal level of output).

Because profits involve both costs and revenues, the firm must have a good grasp of its cost structure. Turn back to Table 7-3 in the previous chapter to make sure you are clear on the important concepts of total cost, average cost, and marginal cost.

Perfect Competition

Perfect competition is the world of *price-takers*. A perfectly competitive firm sells a *homogeneous product* (one identical to the product sold by others in the industry). The firm is so small relative to its market that it cannot affect the market price; it simply takes the price as given. When Farmer Smith sells a homogeneous product like wheat, she sells to a large pool of buyers at the market price of $6 per bushel. Just as consumers must generally accept the prices that are charged by Internet access providers or movie theaters, so must competitive firms accept the market prices of the wheat or oil that they produce.

We can depict a price-taking perfect competitor by examining the way demand looks to a perfectly competitive firm. Figure 8-1 shows the contrast between the industry demand curve (the *DD* curve) and the demand curve facing a single competitive firm (the *dd* curve). Because a competitive industry is populated by firms that are small relative to the market, the firm's segment of the demand curve is only a tiny segment of the industry's curve. Graphically, the competitive firm's portion of the demand curve is so small that, to the lilliputian eye of the perfect competitor, the firm's *dd* demand curve looks completely horizontal or infinitely elastic. Figure 8-1 illustrates how the elasticity of demand for a single competitor appears very much greater than that for the entire market.

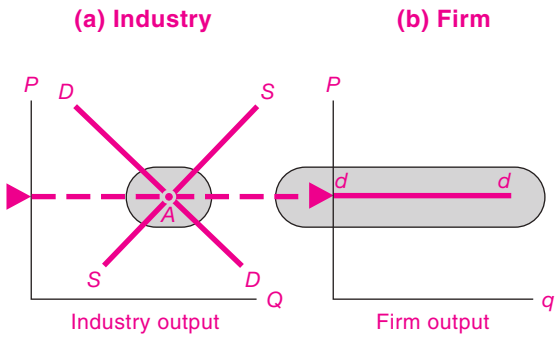


FIGURE 8-1. Demand Curve Is Completely Elastic for a Perfectly Competitive Firm

The industry demand curve on the left has inelastic demand at the market equilibrium at *A*. However, the demand curve for the perfectly competitive firm on the right is horizontal (i.e., completely elastic). The demand curve on the right is horizontal because a perfect competitor has such a small fraction of the market that it can sell all it wants at the market price.

Because competitive firms cannot affect the price, the price for each unit sold is the extra revenue that the firm will earn. For example, at a market price of $40 per unit, the competitive firm can sell all it wants at $40. If it decides to sell 101 units rather than 100 units, its revenue goes up by exactly $40.

Here are the major points to remember:

1. Under **perfect competition,** there are many small firms, each producing an identical product and each too small to affect the market price.
2. The perfect competitor faces a completely horizontal demand (or *dd*) curve.
3. The extra revenue gained from each extra unit sold is therefore the market price.

Competitive Supply Where Marginal Cost Equals Price

Suppose *you* are managing Bob's oil operations and are responsible for setting the profit-maximizing output. How would you go about this task? Examine Table 8-1, which contains the same cost data as Tables 7-3 and 7-4 in the previous chapter. This table adds a further assumption that the market price of oil is $40 per unit.

Supply Decision of Competitive Firm

(1) Quantity q	(2) Total cost TC ($)	(3) Marginal cost per unit MC ($)	(4) Average cost AC ($)	(5) Price P ($)	(6) Total revenue $TR = q \times P$ ($)	(7) Profit $\pi = TR - TC$ ($)
0	55,000					
1,000	85,000	**27**	85	40	40,000	−45,000
2,000	110,000	**22**	55	40	80,000	−30,000
3,000	130,000	**21**	43.33	40	120,000	−10,000
3,999	159,960.01	**38.98**	40.000+	40	159,960	−0.01
		39.99				
4,000	**160,000**	**40**	**40**	**40**	**160,000**	**0**
		40.01				
4,001	160,040.01	**40.02**	40.000+	40	160,040	−0.01
5,000	210,000	**60**	42	40	200,000	−10,000

TABLE 8-1. Profit Is Maximized at Production Level Where Marginal Cost Equals Price

The first four columns use the same cost data as that analyzed in Tables 7-3 and 7-4 of the previous chapter. Column (5) shows the price of $40 that is received by the price-taking perfect competitor. Total revenue is price times quantity, while profit is total revenue less total cost.

This table shows that the maximum profit comes at that output where price equals *MC*. If output is raised above $q = 4000$, the additional revenue of $40 per unit is less than the marginal cost, so profit is lowered. What happens to profit if output is raised when $q < 4000$?

You might take a guess and sell 3000 units. This yields total revenue of $40 × 3000 = $120,000, with total cost of $130,000, so the firm incurs a loss of $10,000. From economics, you have learned to think about *marginal* or incremental decisions. So you analyze the effect of selling an additional unit. The revenue from each unit is $40, while the marginal cost at that volume is only $21. This implies that the additional revenue outweighs the marginal cost of 1 more unit. So you analyze a production level of 4000 units. At this output, the firm has revenues of $40 × 4000 = $160,000 and costs of $160,000, so profits are zero.

What would happen if you increase output to 5000 units? At this output, the firm has revenues of $40 × 5000 = $200,000 and costs of $210,000. Now you're losing $10,000 again. What went wrong? When you look at your accounts, you see that at the output level of 5000, the marginal cost is $60. This is more than the market price of $40, so you are losing $20 (equal to price minus *MC*) on the last unit produced.

Now you see the light: *The maximum profit comes at that output where marginal cost equals price.*

The reason underlying this proposition is that the competitive firm can always make additional profit as long as the price is greater than the marginal cost of the last unit. Total profit reaches its peak—is maximized—when there is no longer any extra profit to be earned by selling extra output. At the maximum-profit point, the last unit produced brings in an amount of revenue exactly equal to that unit's cost. What is that extra revenue? It is the price per unit. What is that extra cost? It is the marginal cost.

Let's test this rule by looking at Table 8-1. Starting at the profit-maximizing output of 4000 units, if

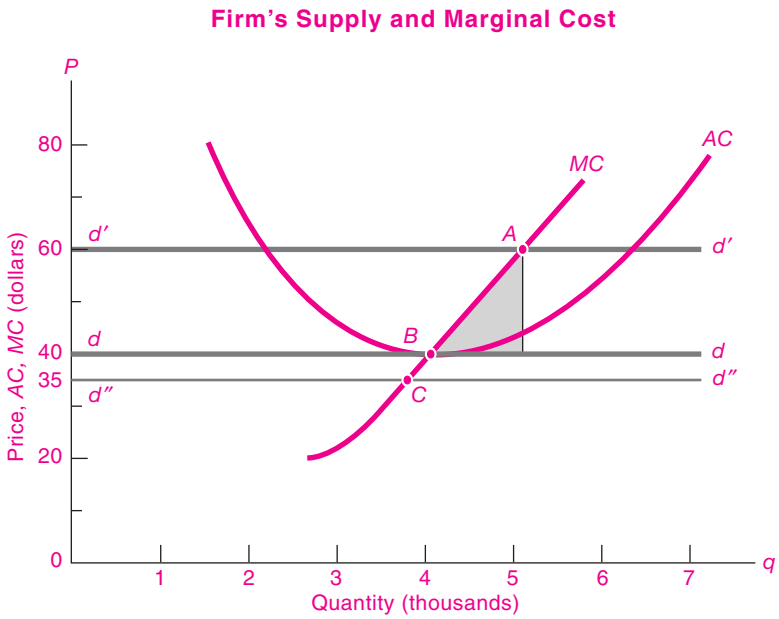


FIGURE 8-2. Firm's Supply Curve Is Its Rising Marginal Cost Curve

For a profit-maximizing competitive firm, the upward-sloping marginal cost (*MC*) curve is the firm's supply curve. For market price at $d'd'$, the firm will supply output at the intersection point at *A*. Explain why intersection points at *B* and *C* represent equilibria for prices at *d* and d'' respectively. The shaded blue region represents the loss from producing at *A* when price is $40.

Bob sells 1 more unit, that unit would bring a price of $40 while the marginal cost of that unit is $40.01. So the firm would lose money on the 4001st unit. Similarly, the firm would lose $0.01 if it produced 1 less unit. This shows that the firm's maximum-profit output comes at exactly $q = 4000$, where price equals marginal cost.

Rule for a firm's supply under perfect competition: A firm will maximize profits when it produces at that level where marginal cost equals price:

$$\text{Marginal cost} = \text{price} \quad \text{or} \quad MC = P$$

Figure 8-2 shown above illustrates a firm's supply decision diagrammatically. When the market price of output is $40, the firm consults its cost data in Table 8-1 and finds that the production level corresponding to a marginal cost of $40 is 4000 units. Hence, at a market price of $40, the firm will wish to produce and sell 4000 units. We can find that profit-maximizing amount in Figure 8-2 at the intersection of the price line at $40 and the *MC* curve at point *B*.

We designed this example so that at the profit-maximizing output the firm has zero profits, with total revenues equal to total costs. Point *B* is the **zero-profit point,** the production level at which the firm makes zero economic profits; at the zero-profit point, price equals average cost, so revenues just cover costs.

What if the firm chooses the wrong output? Suppose the firm chooses output level *A* in Figure 8-2 when the market price is $40. It would be losing money because the last units have marginal cost above price. We can calculate the loss of profit if the firm mistakenly produces at *A* by the shaded blue triangle in Figure 8-2. This depicts the surplus of *MC* over price for production between *B* and *A*.

The general rule then is:

A profit-maximizing firm will set its output at that level where marginal cost equals price. Diagrammatically, this means that a firm's marginal cost curve is also its supply curve.

Total Cost and the Shutdown Condition

Our general rule for firm supply leaves open one possibility—that the price will be so low that the firm will want to shut down. Isn't it possible that at the $P = MC$ equilibrium, Bob may be losing a truckful of money and would want to shut down? In general, a firm will want to shut down in the short run when it can no longer cover its variable costs.

For example, suppose the firm were faced with a market price of $35, shown by the horizontal *d″d″* line in Figure 8-2. At that price, *MC* equals price at point *C,* a point at which the price is actually less than the average cost of production. Would the firm want to keep producing even though it was incurring a loss?

The surprising answer is that the firm should *not* necessarily shut down if it is losing money. The firm should *minimize its losses,* which is the same thing as maximizing profits. Producing at point *C* would result in a loss of only $20,000, whereas shutting down would involve losing $55,000 (which is the fixed cost). The firm should therefore continue to produce.

To understand this point, remember that a firm must still cover its contractual commitments even when it produces nothing. In the short run, the firm must pay fixed costs such as interest to the bank, rentals on the oil rigs, and directors' salaries. The balance of the firm's costs are variable costs, such as those for materials, production workers, and fuel, which would have zero cost at zero production. It will be advantageous to continue operations, with *P* at least as high as *MC,* as long as revenue covers variable costs.

The critically low market price at which revenues just equal variable costs (or, equivalently, at which losses exactly equal fixed costs) is called the **shutdown point.** For prices above the shutdown point, the firm will produce along its marginal cost curve because, even though the firm might be losing money, it would lose more money by shutting down. For prices below the shutdown point, the firm will produce nothing at all because by shutting down the firm will lose only its fixed costs. This gives the shutdown rule:

Shutdown rule: The shutdown point comes where revenues just cover variable costs or where losses are equal to fixed costs. When the price falls below average variable costs, the firm will maximize profits (minimize its losses) by shutting down.

Figure 8-3 shows the shutdown and zero-profit points for a firm. The zero-profit point comes where price is equal to *AC,* while the shutdown point comes where price is equal to *AVC.* Therefore, the firm's supply curve is the solid green line in Figure 8-3. It first goes up the vertical axis to the price corresponding to the shutdown point; next jumps to the shutdown point at *M′,* where *P* equals the level of *AVC;* and then continues up the *MC* curve for prices above the shutdown price.

Zero-Profit and Shutdown Prices

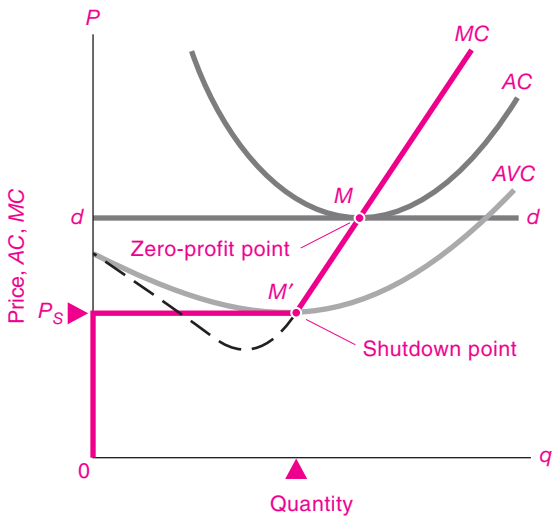


FIGURE 8-3. Firm's Supply Curve Travels Down the *MC* Curve to the Shutdown Point

The firm's supply curve corresponds to its *MC* curve as long as revenues exceed variable costs. Once price falls below P_s, the shutdown point, losses are greater than fixed costs, and the firm shuts down. Hence the solid green curve is the firm's supply curve.

The analysis of shutdown conditions leads to the surprising conclusion that profit-maximizing firms may in the short run continue to operate even though they are losing money. This condition will hold particularly for firms that are heavily indebted and therefore have high fixed costs (the airlines being a good example). For these firms, as long as losses are less than fixed costs, profits are maximized and losses are minimized when they pay the fixed costs and still continue to operate.

Unemployed Rigs in the Drilling Industry

A striking example of the shutdown rule at work was seen in the oil industry. New oil wells are drilled by "oil rigs." Each oil rig is like a little business, which can operate or shut down depending upon profitability. When a price war broke out among oil

producers in 1999, many shut down, and the number of rigs in operation in the United States declined to under 500. Had the oil fields run dry? Not at all. Rather, production was discouraged because the price of oil was so low. It was the profits, not the wells, that dried up.

What happened to drilling activity during the oil-price surge of the 2000s? From 2002 to 2008, when oil prices quadrupled, the number of rigs in operation went up by a factor of almost 4. In effect, as prices rose, these firms moved up along an upward-sloping *MC* supply curve similar to the one shown in Figure 8-3.

B. SUPPLY BEHAVIOR IN COMPETITIVE INDUSTRIES

Our discussion up to now has concerned only the individual firm. But a competitive market comprises many firms, and we are interested in the behavior of all firms together, not just a single firm. How can we move from the one to the many? From Bob's operation to the entire oil industry?

SUMMING ALL FIRMS' SUPPLY CURVES TO GET MARKET SUPPLY

Suppose we are dealing with a competitive market for oil. At a given price, firm A will bring a given quantity of oil to market, firm B will bring another quantity, as will firms C, D, and so on. In each case, the quantity supplied will be determined by each firm's marginal costs. The *total* quantity brought to market at a given price will be the *sum* of the individual quantities that all firms supply at that price.[1]

This reasoning leads to the following relationship between individual and market supplies for a perfectly competitive industry:

The market supply curve for a good in a perfectly competitive market is obtained by adding horizontally the supply curves of all the individual producers of that good.

Figure 8-4 illustrates this rule for two firms. We obtain the industry's *SS* supply curve by horizontal addition at each price of the firms' individual supply

[1] Recall that the *DD* market demand curve is similarly obtained by horizontal summation of individual *dd* demand curves.

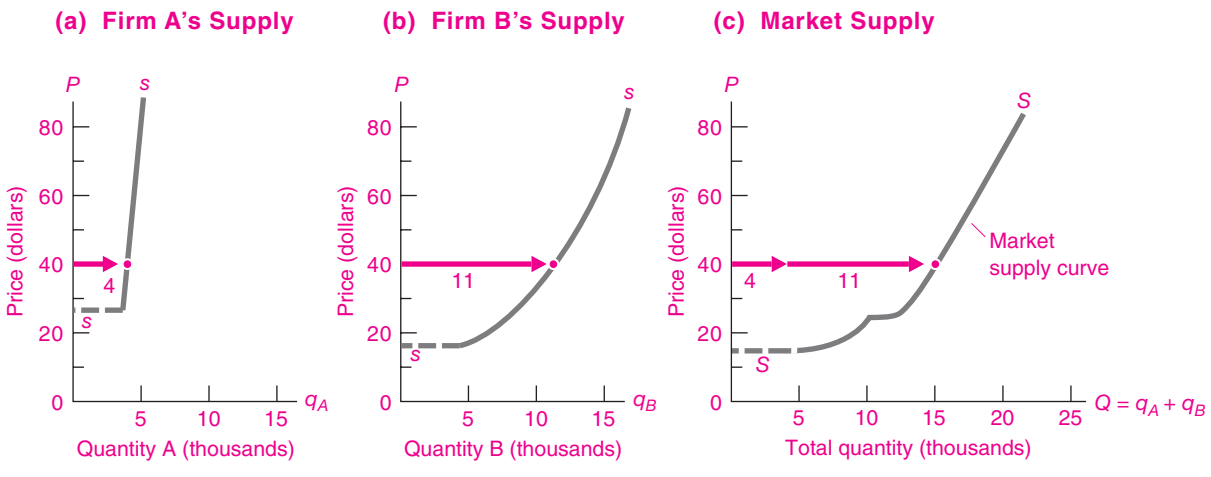


FIGURE 8-4. Add All Firms' Supply Curves to Derive Market Supply

The diagrams show how the market supply curve (*SS*) is derived from two individual supply curves (*ss*). We horizontally add quantities supplied by each firm at $40 to get total market supply at $40. This applies at each price and to any number of firms. If there are 1000 firms identical to firm A, the market supply curve would look like firm A's supply curve with a thousandfold change of horizontal scale.

curves. At a price of $40, firm A will supply 4000 units while firm B will supply 11,000 units. Therefore, the industry will supply a total of 15,000 units at a price of $40. If there are 2 million rather than 2 firms, we would still derive industry output by adding all the 2 million individual-firm quantities at the going price. Horizontal addition of output at each price gives us the industry supply curve.

SHORT-RUN AND LONG-RUN EQUILIBRIUM

Economists have observed that demand shifts produce greater price adjustments and smaller quantity adjustments in the short run than they do in the long run. We can understand this observation by distinguishing two time periods for market equilibrium that correspond to different cost categories: (1) *short-run equilibrium,* when output changes must use the same fixed amount of capital, and (2) *long-run equilibrium,* when capital and all other factors are variable and there is free entry and exit of firms into and from the industry.

Entry and Exit of Firms

The birth (entry) and death (exit) of firms are important factors that affect the evolution of a market economy. Firms *enter* an industry either when they are newly formed or when an existing firm decides to start production in a new sector. Firms *exit* when they stop producing; they might leave voluntarily because a line of production is unprofitable, or they might go bankrupt if the entire firm cannot pay its bills. We say that there is *free entry and exit* when there are no barriers to entry or exit. Barriers to entry include such factors as government regulations or intellectual property rights (e.g., patents or software).

Many people are surprised by the large number of births and deaths of firms in a dynamic economy like the United States. For example, there were 6.5 million registered businesses at the beginning of 2003. In that year, 748,000 new businesses were born and 658,000 went out of business. The riskiest industry was Internet providers, where 30 percent of jobs were lost because of firm deaths in that year. The safest industry was colleges, where only 4 percent of jobs were lost by college closings.

Most firms exit quietly, but sometimes large firms have a noisy exit, as occurred when the telecommunications giant WorldCom, with $104 billion of assets, went under because of a massive accounting fraud. Although the smooth cost curves do not always capture the drama of entry and exit, the underlying logic of *P, MC,* and *AC* is a powerful force driving the growth and decline of major industries.

Let's illustrate the distinction between short-run and long-run equilibriums with an example. Consider the market for fresh fish supplied by a local fishing fleet. Suppose the demand for fish increases; this case is shown in Figure 8-5(*a*) as a shift from *DD* to *D′D′*. With higher prices, fishing captains will want to increase their catch. In the short run, they cannot build new boats, but they can hire extra crews and work longer hours. Increased inputs of variable factors will produce a greater quantity of fish along the *short-run supply curve* S_SS_S shown in Figure 8-5(*a*). The short-run supply curve intersects the new demand curve at *E′*, the point of short-run equilibrium.

The high prices lead to high profits, which in the long run coax out more shipbuilding and attract more sailors into the industry. Additionally, new firms may start up or enter the industry. This gives us the *long-run supply curve* S_LS_L in Figure 8-5(*b*) and the long-run equilibrium at *E″*. The intersection of the long-run supply curve with the new demand curve yields the long-run equilibrium attained when all economic conditions (including the number of ships, shipyards, and firms) have adjusted to the new level of demand.

Long-Run Industry Supply. What is the shape of the long-run supply curve for an industry? Suppose that an industry has free entry of identical firms. If the identical firms use general inputs, such as unskilled labor, that can be attracted from the vast ocean of other uses without affecting the prices of those general inputs, we get the case of constant costs shown by the horizontal S_LS_L supply curve in Figure 8-6.

By contrast, suppose some of the inputs used in the industry are in relatively short supply—for example, fertile vineyard land for the wine industry or scarce beachfront properties for summer vacations. Then the supply curve for the wine or vacation industry must be upward-sloping, as shown by S_LS_L' in Figure 8-6.

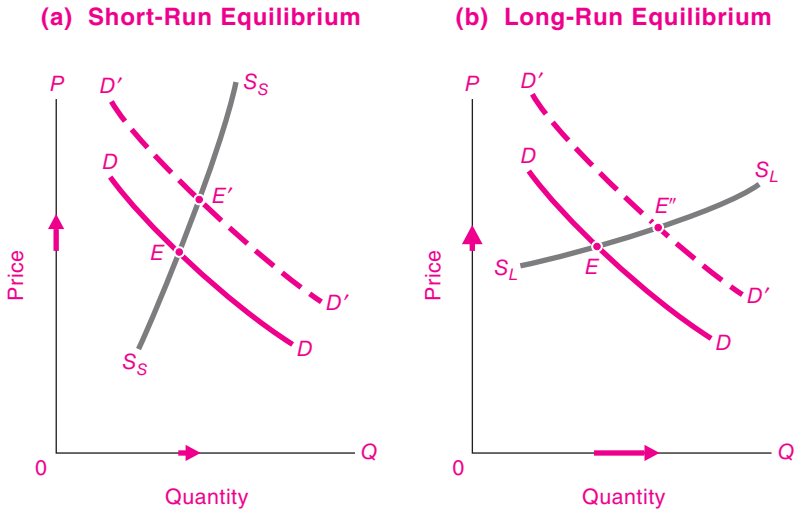


FIGURE 8-5. Effect of Increase in Demand on Price Varies in Different Time Periods

We distinguish between periods in which firms have time to make **(a)** adjustments in variable factors such as labor (short-run equilibrium) and **(b)** full adjustment of all factors, fixed as well as varying (long-run equilibrium). The longer the time for adjustments, the greater the elasticity of supply response and the smaller the rise in price.

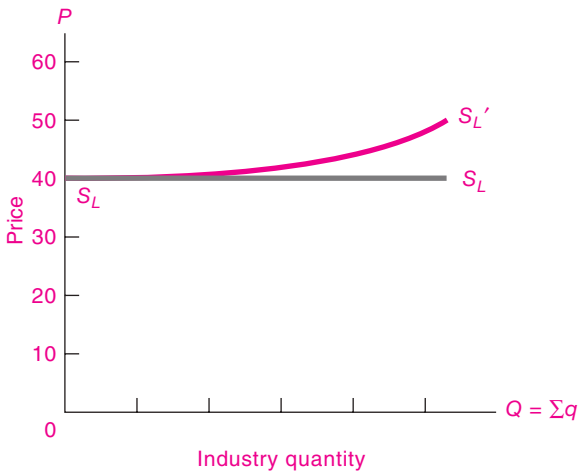


FIGURE 8-6. Long-Run Industry Supply Depends on Cost Conditions

With entry and exit free and any number of firms able to produce on identical, unchanged cost curves, the long-run S_LS_L curve will be horizontal at each firm's minimum average cost or zero-profit price. If the industry uses a specific factor, such as scarce beachfront property, the long-run supply curve must slope upward like S_LS_L', as higher production employs less well-suited inputs.

The long-run supply curve of industries using scarce factors rises because of diminishing returns. For example, take the case of the rare vineyard land. As firms apply increasing inputs of labor to fixed land, they receive smaller and smaller increments of wine-grape output. But each dose of labor costs the same in wages, so the *MC* of wine rises. This long-run rising *MC* means that the long-run supply curve must be rising.

The Long Run for a Competitive Industry

Our analysis of zero-profit conditions showed that firms might stay in business for a time even though they are unprofitable. This situation is possible particularly for firms with high fixed capital costs. With this analysis we can understand why in business downturns many of America's largest companies, such as General Motors, stayed in business even though they were losing billions of dollars.

Such losses raise a troubling question: Is it possible that capitalism is heading toward "euthanasia of the capitalists," a situation where increased competition produces chronic losses? For this question, we

need to analyze the *long-run shutdown conditions*. We showed that firms shut down when they can no longer cover their variable costs. But in the long run, *all* costs are variable. A firm that is losing money can pay off its bonds, release its managers, and let its leases expire. In the long run, all commitments are once again options. Hence, in the long run firms will produce only when price is at or above the zero-profit condition where price equals average cost.

There is, then, a critical zero-profit point below which long-run price cannot remain if firms are to stay in business. In other words, long-run price must cover out-of-pocket costs such as labor, materials, equipment, taxes, and other expenses, along with opportunity costs such as competitive return on the owner's invested capital. That means long-run price must be equal to or above total long-run average cost.

Take the case where price falls below this critical zero-profit level. Unprofitable firms will start leaving the industry. Since fewer firms are producing, the short-run market supply curve will shift to the left, and the price will therefore rise. Eventually, the price will rise enough so that the industry is no longer unprofitable. So, even though we produce very few horseshoes today compared to a century ago, horseshoe manufacturing will earn a zero long-run profit.

Consider the opposite case of a profitable industry such as developing computer games. At the beginning, the price starts above total long-run average cost, so firms are making positive economic profits. Now suppose entry into the industry is absolutely free in the long run, so any number of identical firms can come into the industry and produce at exactly the same costs as those firms already in the industry. In this situation, new firms are attracted by prospective profits, the short-run supply curve shifts to the right, and price falls. Eventually price falls to the zero-profit level, so it is no longer profitable for other firms to enter the industry. Thus, even though computer games might be a thriving industry, it would earn a zero long-run profit.

The conclusion is that in the long run, the price in a competitive industry will tend toward the critical point where revenues just cover full competitive costs. Below this critical long-run price, firms would leave the industry until price returns to long-run average cost. Above this long-run price, new firms would enter the industry, thereby forcing market price back down to the long-run equilibrium price where all competitive costs are just covered.

Zero-profit long-run equilibrium: In a competitive industry populated by identical firms with free entry and exit, the long-run equilibrium condition is that price equals marginal cost equals the minimum long-run average cost for each identical firm:

$P = MC =$ minimum long-run $AC =$ zero-profit price

This is the long-run **zero-economic-profit** condition.

We have reached a surprising conclusion about the long-run profitability of competitive capitalism. The forces of competition tend to push firms and industries toward a zero-profit long-run state. In the long run, competitive firms will earn the normal return on their investment, but no more. Profitable industries tend to attract entry of new firms, thereby driving down prices and reducing profits toward zero. By contrast, firms in unprofitable industries leave to seek better profit opportunities; prices and profits then tend to rise. *The long-run equilibrium in a perfectly competitive industry is therefore one with no economic profits.*

C. SPECIAL CASES OF COMPETITIVE MARKETS

This section probes more deeply into supply-and-demand analysis. We first consider certain general propositions about competitive markets and then continue with some special cases.

GENERAL RULES

We analyzed above the impact of demand and supply shifts in competitive markets. These findings apply to virtually any competitive market, whether it is for codfish, brown coal, Douglas fir, Japanese yen, IBM stock, or petroleum. Are there any general rules? The propositions that follow investigate the impact of shifts in supply or demand upon the price and quantity bought and sold. Remember always that by a shift in demand or supply we mean a shift of the demand or supply curve or schedule, not a movement along the curve.

Demand rule: (*a*) Generally, an increase in demand for a commodity (the supply curve being unchanged) will raise the price of the commodity. (*b*) For most commodities, an increase in demand will also increase the quantity demanded. A decrease in demand will have the opposite effects.

Supply rule: (*c*) An increase in supply of a commodity (the demand curve being constant) will generally lower the price and increase the quantity bought and sold. (*d*) A decrease in supply has the opposite effects.

These two rules of supply and demand summarize the qualitative effects of shifts in supply and demand. But the quantitative effects on price and quantity depend upon the exact shapes of the supply and demand curves. In the cases that follow, we will see the response for a number of important cost and supply situations.

Constant Cost

Production of many manufacturing items, such as textiles, can be expanded by merely duplicating factories, machinery, and labor. Producing 200,000 shirts per day simply requires that we do the same thing as we did when we were manufacturing 100,000 per day but on a doubled scale. In addition, assume that the textile industry uses land, labor, and other inputs in the same proportions as the rest of the economy.

In this case the long-run supply curve *SS* in Figure 8-7 is a horizontal line at the constant level of unit costs. A rise in demand from *DD* to *D′D′* will shift the new intersection point to *E′*, raising *Q* but leaving *P* the same.

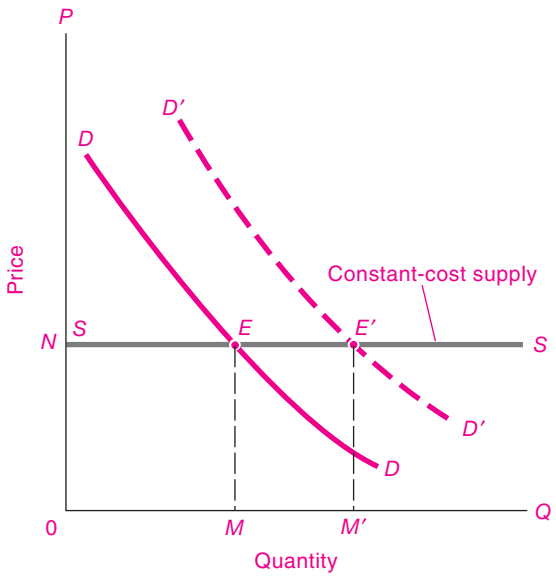


FIGURE 8-7. Constant-Cost Case

Increasing Costs and Diminishing Returns

The last section discussed industries, such as for wine or beach properties, where a product uses an input in limited supply. In the case of wine vineyards, good sites are limited in number. The annual output of wine can be increased to some extent by adding more labor to each acre of land. But the law of diminishing returns will eventually operate if variable factors of production, such as labor, are added to fixed amounts of a factor such as land.

As a result of diminishing returns, the marginal cost of producing wine increases as wine production rises. Figure 8-8 shows the rising supply curve *SS*. How will price be affected by an increase in demand? The figure shows that higher demand will increase the price of this good even in the long run with identical firms and free entry and exit.

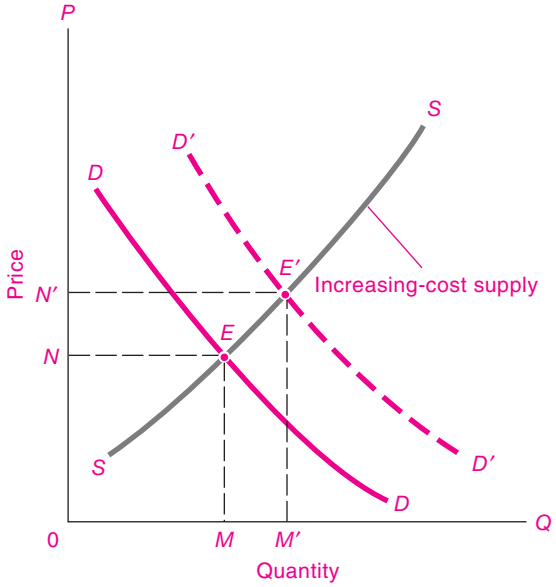


FIGURE 8-8. Increasing-Cost Case

Fixed Supply and Economic Rent

Some goods or productive factors are completely fixed in amount, regardless of price. There is only one *Mona Lisa* by da Vinci. Nature's original endowment of land can be taken as fixed in amount. Raising the price offered for land cannot create an additional corner at 57th Street and Fifth Avenue in New York City. Raising the pay of top managers is unlikely to change their effort. When the quantity supplied is constant at every price, the payment for the use of such a factor of production is called **rent** or **pure economic rent.**

When supply is independent of price, the supply curve is vertical in the relevant region. Land will continue to contribute to production no matter what its price. Figure 8-9 shows the case of land, for which a higher price cannot coax out any increase in output.

An increase in the demand for a fixed factor will affect only the price. Quantity supplied is unchanged.

When a tax is placed upon the fixed commodity, the tax is completely paid by (or "shifted" back to) the supplier (say, the landowner). The supplier absorbs the entire tax out of economic rent. The consumer buys exactly as much of the good or service as before and at no higher price.

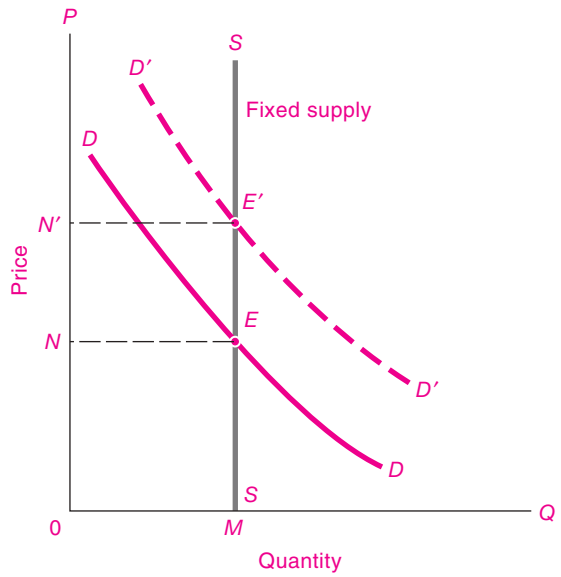


FIGURE 8-9. Factors with Fixed Supply Earn Rent

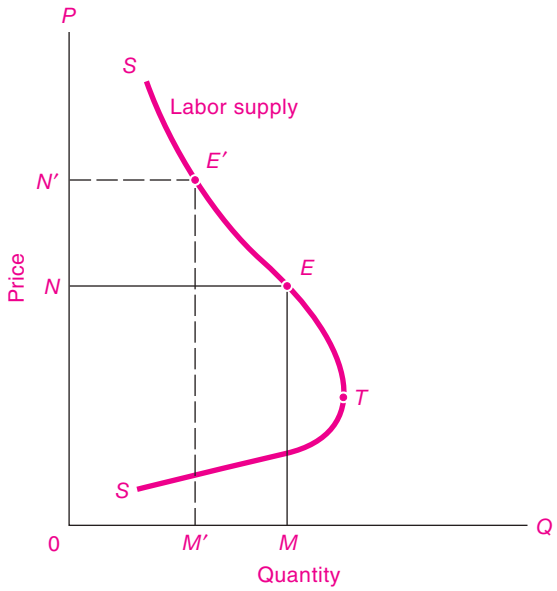


FIGURE 8-10. Backward-Bending Supply Curve

Backward-Bending Supply Curve

Firms in poor countries sometimes found that when they raised wages, the local workers worked fewer hours. When the wage was doubled, instead of continuing to work 6 days a week, the workers might work 3 days and go fishing for the other 3 days. The same has been observed in high-income countries. As improved technology raises real wages, people feel that they want to take part of their higher earnings in the form of more leisure and early retirement. Chapter 5 described income and substitution effects, which explain why a supply curve might *bend backward.*

Figure 8-10 shows what a supply curve for labor might look like. At first the labor supplied rises as higher wages coax out more labor. But beyond point *T*, higher wages lead people to work fewer hours and to take more leisure. An increase in demand raises the price of labor, as was stated in the demand rule at the beginning of this section. But note why we were cautious to add "for most commodities" to demand rule (*b*), for now the increase in demand decreases the quantity of labor supplied.

Shifts in Supply

All the above discussions dealt with a shift in demand and no shift in supply. To analyze the supply rule,

we must now shift supply, keeping demand constant. If the law of downward-sloping demand is valid, increased supply must decrease price and increase quantity demanded. You should draw your own supply and demand curves and verify the following quantitative corollaries of the supply rule:

(c') An increased supply will decrease P most when demand is inelastic.
(d') An increased supply will increase Q least when demand is inelastic.

What are commonsense reasons for these rules? Illustrate with cases of elastic demand for autos and of inelastic demand for electricity.

D. EFFICIENCY AND EQUITY OF COMPETITIVE MARKETS

EVALUATING THE MARKET MECHANISM

One of the remarkable features of the last decade has been the "rediscovery of the market." Many countries have abandoned the heavy-handed interventionism of government command and regulation for the decentralized coordination of the invisible hand. Having reviewed the basic operation of competitive markets, let's ask how well they perform. Do they deserve high grades for satisfying people's economic needs? Is society getting many guns and much butter for a given amount of inputs? Or does the butter melt on the way to the store, while the guns have crooked barrels? We will provide an overview of the efficiency of competitive markets in this chapter.

The Concept of Efficiency

Efficiency is one of the central concepts in all of economics. In a general sense, an economy is efficient when it provides its consumers with the most desired set of goods and services, given the resources and technology of the economy.[2] A more precise definition uses the concept of *Pareto efficiency* (alternatively called *allocative efficiency, Pareto optimality,* or sometimes simply *efficiency*).

Pareto efficiency (or sometimes just **efficiency**) occurs when no possible reorganization of production or distribution can make anyone better off without making someone else worse off. Under conditions of allocative efficiency, one person's satisfaction or utility can be increased only by lowering someone else's utility.

We can think of the concept of efficiency intuitively in terms of the production-possibility frontier. An economy is clearly inefficient if it is inside the *PPF*. If we move out to the *PPF*, no one need suffer a decline in utility. At a minimum, an efficient economy is on its *PPF*. But efficiency goes further and requires not only that the right mix of goods be produced but also that these goods be allocated among consumers to maximize consumer satisfactions.

Efficiency of Competitive Equilibrium

One of the most important results in all economics is that the allocation of resources by perfectly competitive markets is efficient. This important result assumes that all markets are perfectly competitive and that there are no externalities like pollution or imperfect information. In this section, we use a simplified example to illustrate the general principles underlying the efficiency of competitive markets.

Consider an idealized situation where all individuals are identical. Further assume: (*a*) Each person works at growing food. As people increase their work and cut back on their leisure hours, each additional hour of sweaty labor becomes increasingly tiresome. (*b*) Each extra unit of food consumed brings diminished marginal utility (MU).[3] (*c*) Because food production takes place on fixed plots of land, by the law of diminishing returns each extra minute of work brings less and less extra food.

Figure 8-11 shows supply and demand for our simplified competitive economy. When we sum

[2] Economic efficiency is different from engineering efficiency, and sometimes it will be economical to use a production method that is *less* efficient from an engineering point of view. For example, physics shows that more energy can be converted to electricity if combustion occurs at 2500°C than at 1000°C. Yet the higher temperature might require exotic metals and designs and cost more. So the lower temperature would be economically efficient even though the higher temperature would have higher thermodynamic efficiency.

[3] To keep matters at their simplest, we measure welfare in fixed "utils" of leisure time (or "disutils" of sweaty labor time). We further assume that each hour of forgone leisure has a constant marginal utility, so all utilities and costs are reckoned in these leisure-labor units.

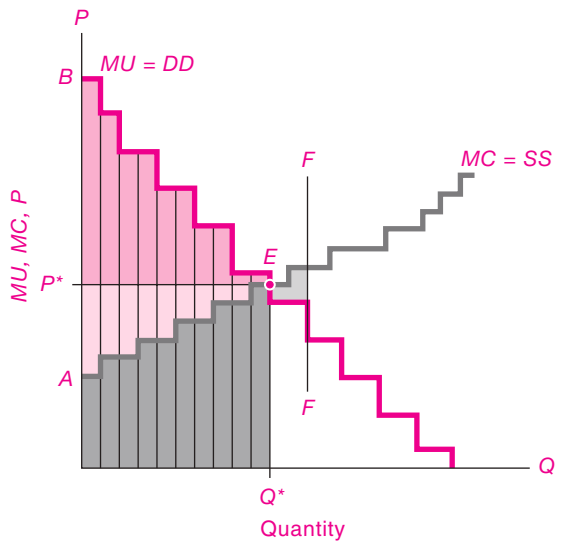


FIGURE 8-11. At Competitive Equilibrium Point *E*, the Marginal Costs and Utilities of Food Are Exactly Balanced

Many identical farmer-consumers bring their food to market. The *MC* = *SS* curve adds together the individual marginal cost curves, while the *MU* = *DD* curve is the horizontal sum of consumer valuations of food. At competitive market equilibrium *E*, the utility gain from the last unit of food equals the utility cost (in terms of forgone leisure).

The figure also illustrates economic surplus. The cost of producing food is shown by the dark blue slices. The light-colored green slices above the *SS* curve and below the price line add up to the "producer surplus." The dark-colored green slices under *DD* and above the price line are the "consumer surplus." The sum of the consumer and producer surpluses is "economic surplus." At the competitive equilibrium at *E*, economic surplus is maximized. Verify that production at *FF* lowers total surplus.

horizontally the identical supply curves of our identical farmers, we get the upward-stepping *MC* curve. As we saw earlier in this chapter, the *MC* curve is also the industry's supply curve, so the figure shows *MC* = *SS*. Also, the demand curve is the horizontal summation of the identical individuals' marginal utility (or demand-for-food) curves; it is represented by the downward-stepping *MU* = *DD* curve for food in Figure 8-11.

The intersection of the *SS* and *DD* curves shows the competitive equilibrium for food. At point *E*, farmers supply exactly what consumers want to purchase at the equilibrium market price. Each person will be working up to the critical point where the declining marginal-utility-of-consuming-food curve intersects the rising marginal-cost-of-growing-food curve.

Figure 8-11 shows a new concept, **economic surplus,** which is the green area between the supply and demand curves at the equilibrium. The economic surplus is the sum of the consumer surplus that we met in Chapter 5, which is the area between the demand curve and the price line, and the **producer surplus,** which is the area between the price line and the *SS* curve. The producer surplus includes the rent and profits to firms and owners of specialized inputs in the industry and indicates the excess of revenues over cost of production. The economic surplus is the welfare or net utility gain from production and consumption of a good; it is equal to the consumer surplus plus the producer surplus.

A careful analysis of the competitive equilibrium will show that it maximizes the economic surplus available in that industry. For this reason, it is economically efficient. At the competitive equilibrium at point *E* in Figure 8-11, the representative consumer will have higher utility or economic surplus than would be possible with any other feasible allocation of resources.

Another way of seeing the efficiency of the competitive equilibrium is by comparing the economic effect of a small change from the equilibrium at *E*. As the following three-step process shows, if *MU* = *P* = *MC*, then the allocation is efficient.

1. *P* = *MU*. Consumers choose food purchases up to the amount where *P* = *MU*. As a result, every person is gaining *P* utils of satisfaction from the last unit of food consumed. (Utils of satisfaction are measured in terms of the constant marginal utility of leisure, as discussed in footnote 3.)
2. *P* = *MC*. As producers, each person is supplying food up to the point where the price of food exactly equals the *MC* of the last unit of food supplied (the *MC* here being the cost in terms of the forgone leisure needed to produce the last unit of food). The price then is the utils of leisure-time satisfaction lost because of working to grow that last unit of food.
3. Putting these two equations together, we see that *MU* = *MC*. This means that the utils gained from the last unit of food consumed exactly equal the

leisure utils lost from the time needed to produce that last unit of food. *It is exactly this condition—that the marginal gain to society from the last unit consumed equals the marginal cost to society of that last unit produced—which guarantees that a competitive equilibrium is efficient.*

Equilibrium with Many Consumers and Markets

Let us now turn from our simple parable about identical farmer-consumers to an economy populated by millions of different firms, hundreds of millions of people, and countless commodities. Can a perfectly competitive economy still be efficient in this more complex world?

The answer is "yes," or better yet, "yes, if . . ." Efficiency requires some stringent conditions that are addressed in later chapters. These include having reasonably well-informed consumers, perfectly competitive producers, and no externalities like pollution or incomplete knowledge. For such economies, a system of perfectly competitive markets will earn the economist's gold star of Pareto efficiency.

Figure 8-12 illustrates how a competitive system brings about a balance between utility and cost for a single commodity with nonidentical firms and consumers. On the left, we add horizontally the demand curves for all consumers to get the market curve *DD* in the middle. On the right, we add all the different *MC* curves to get the industry *SS* curve in the middle.

At the competitive equilibrium at point *E*, consumers on the left get the quantity they are willing to purchase of the good at the price reflecting efficient social *MC*. On the right, the market price also allocates production efficiently among firms. The blue area under *SS* in the middle represents the minimized sum of the blue cost areas on the right. Each firm is setting its output so that $MC = P$. Production efficiency is achieved because there is no reorganization of production that would allow the same level of industry output to be produced at lower cost.

Many Goods. Our economy produces not only food but also clothing, movies, and many other

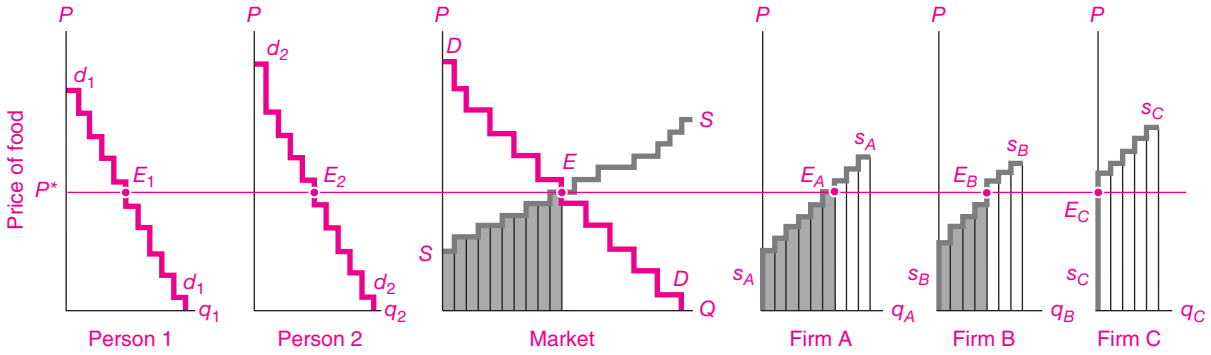


FIGURE 8-12. Competitive Market Integrates Consumers' Demands and Producers' Costs

(a) Individual demands are shown on the left. We add the consumers' *dd* curves horizontally to obtain the market demand *DD* curve in the middle.

(b) The market brings together all consumer demands and firm supplies to reach market equilibrium at *E*. The horizontal price-of-food line shows where each consumer on the left and each producer on the right reach equilibrium. At *P**, see how each consumer's *MU* is equated to each firm's *MC*, leading to allocative efficiency.

(c) For each competitive firm, profits are maximized when the supply curve is given by the rising *MC* curve. The blue area depicts each firm's cost of producing the amount at *E*. At prices equal to marginal cost, the industry produces output at the least total cost.

commodities. How does our analysis apply when consumers must choose among many products?

The principles are exactly the same, but now we recall one further condition: Utility-maximizing consumers spread their dollars among different goods until the marginal utility of the last dollar is equalized for each good consumed. In this case, as long as the ideal conditions are met, a competitive economy is efficient with a multitude of goods and factors of production.

In other words, a perfectly competitive economy is efficient when marginal private cost equals marginal social cost and when both equal marginal utility. Each industry must balance *MC* and *MU*. For example, if movies have 2 times the *MC* of hamburgers, the *P* and the *MU* of movies must also be twice those of hamburgers. Only then will the *MU*s, which are equal to the *P*s, be equal to the *MC*s. By equating price and marginal cost, competition guarantees that an economy can attain allocative efficiency.

The perfectly competitive market is a device for synthesizing (*a*) the willingness of consumers possessing dollar votes to pay for goods with (*b*) the marginal costs of those goods as represented by firms' supply. Under certain conditions, competition guarantees efficiency, in which no consumer's utility can be raised without lowering another consumer's utility. This is true even in a world of many factors and products.

Marginal Cost as a Benchmark for Efficiency

This chapter has shown the importance of marginal cost in attaining an efficient allocation of resources. But the importance of marginal cost extends far beyond perfect competition. Using marginal cost to achieve productive efficiency holds for any society or organization trying to make the most effective use of its resources—whether that entity is a capitalist or socialist economy, a profit-maximizing or nonprofit organization, a university or a church, or even a family.

The essential role of marginal cost is this: Suppose you have an objective that can be reached using several approaches, each of which is costly. In deciding how much of each approach to use, always do so by equating the marginal cost among the different approaches. Only when marginal costs are equalized can we squeeze the maximum from our scarce resources.

The use of marginal cost as a benchmark for efficient resource allocation is applicable not just to profit-maximizing firms but to all economic problems, indeed to all problems involving scarcity. Suppose that you have been charged with solving a critical environmental problem, such as global warming. You will soon find that marginal cost will be crucial to attaining your environmental objectives most efficiently. By ensuring that the marginal costs of reducing greenhouse-gas emissions are equalized in every industry and in every corner of the world, you can guarantee that your environmental objectives are being reached at the lowest possible costs.

Marginal cost is a fundamental concept for efficiency. For any goal-oriented organization, efficiency requires that the marginal cost of attaining the goal should be equal in every activity. In a market, an industry will produce its output at minimum total cost only when each firm's *MC* is equal to a common price.

QUALIFICATIONS

We have now seen the essence of the invisible hand—the remarkable efficiency properties of competitive markets. But we must quickly qualify the analysis by pointing to shortcomings of the market.

There are two important areas where markets fail to achieve a social optimum. First, markets may be inefficient in situations where pollution or other externalities are present or when there is imperfect competition or information. Second, the distribution of incomes under competitive markets, even when it is efficient, may not be socially desirable or acceptable. We will review both of these points in later chapters, but it will be useful to describe each of these shortcomings briefly here.

Market Failures

What are the market failures which spoil the idyllic picture assumed in our discussion of efficient markets? The important ones are imperfect competition, externalities, and imperfect information.

Imperfect Competition. When a firm has market power in a particular market (say it has a monopoly

because of a patented drug or a local electricity franchise), the firm can raise the price of its product above its marginal cost. Consumers buy less of such goods than they would under perfect competition, and consumer satisfaction is reduced. This kind of reduction of consumer satisfaction is typical of the inefficiencies created by imperfect competition.

Externalities. Externalities are another important market failure. Recall that externalities arise when some of the side effects of production or consumption are not included in market prices. For example, a power company might pump sulfurous fumes into the air, causing damage to neighboring homes and to people's health. If the power company does not pay for the harmful impacts, pollution will be inefficiently high and consumer welfare will suffer.

Not all externalities are harmful. Some are beneficial, such as the externalities that come from knowledge-generating activities. For example, when Chester Carlson invented xerography, he became a millionaire; but he still received only a tiny fraction of the benefits when the world's secretaries and students were relieved of billions of hours of drudgery. Another positive externality arises from public-health programs, such as inoculation against smallpox, cholera, or typhoid; an inoculation protects not only the inoculated person but also others whom that person might otherwise have infected.

Imperfect Information. A third important market failure is imperfect information. The invisible-hand theory assumes that buyers and sellers have complete information about the goods and services they buy and sell. Firms are assumed to know about all the production functions for operating in their industry. Consumers are presumed to know about the quality and prices of goods—such as whether the financial statements of firms are accurate and whether the drugs they use are safe and efficacious.

Clearly, reality is far from this idealized world. The critical question is, How damaging are departures from perfect information? In some cases, the loss of efficiency is slight. I will hardly be greatly disadvantaged if I buy a chocolate ice cream that is slightly too sweet or if I don't know the exact temperature of the beer that flows from the tap. In other cases, the loss is severe. Take the case of steel mogul Eben Byers, who a century ago took Radithor, sold as a cure-all, to relieve his ailments. Later analysis showed that Radithor was actually distilled water laced with radium. Byers died a hideous death when his jaw and other bones disintegrated. This kind of invisible hand we don't need.

One of the important tasks of the government is to identify those areas where informational deficiencies are economically significant—such as in finance—and then to find appropriate remedies.

Two Cheers for the Market, but Not Three

We have seen that markets have remarkable efficiency properties. But can we therefore conclude that laissez-faire capitalism produces the greatest happiness of the greatest numbers? Does the market necessarily result in the fairest possible use of resources? The answers are no and no.

People are not equally endowed with purchasing power. A system of prices and markets may be one in which a few people have most of the income and wealth. They may have inherited scarce land or oil properties or manage a big corporation or a profitable hedge fund. Some are very poor through no fault of their own, while others are very rich through no virtue of their own. So the weighting of dollar votes, which lie behind the individual demand curves, may be unfair.

An economy with great inequality is not necessarily inefficient. The economy might be squeezing a large quantity of guns and butter from its resources. But the rich few may be eating the butter and feeding it to their cats, while the guns are mainly protecting the butter of the rich.

A society does not live on efficiency alone. A society may choose to alter market outcomes to improve the equity or fairness of the distribution of income and wealth. Nations may levy progressive taxes on those with high incomes and wealth and use the proceeds to finance food, schools, and health care for the poor. But there are vexing questions here. How much should the rich be taxed? What programs will best benefit the poor? Should immigrants be included in the benefit programs? Should capital be taxed at the same rate as labor? Should the nonworking poor get government help?

There are no scientifically correct answers to these questions. Positive economics cannot say how much governments should intervene to correct the inequalities and inefficiencies of the marketplace. These normative questions are appropriately answered through political debate and fair elections. But economics can offer valuable insights into the merit of alternative interventions so that the goals of a modern society can be achieved in the most effective manner.

SUMMARY

A. Supply Behavior of the Competitive Firm

1. A perfectly competitive firm sells a homogeneous product and is too small to affect the market price. Competitive firms are assumed to maximize their profits. To maximize profits, the competitive firm will choose that output level at which price equals the marginal cost of production, that is, $P = MC$. Diagrammatically, the competitive firm's equilibrium will come where the rising MC supply curve intersects its horizontal demand curve.
2. Variable costs must be taken into consideration in determining a firm's short-run shutdown point. Below the shutdown point, the firm loses more than its fixed costs. It will therefore produce nothing when price falls below the shutdown price.
3. A competitive industry's long-run supply curve, $S_L S_L$, must take into account the entry of new firms and the exodus of old ones. In the long run, all of a firm's commitments expire. It will stay in business only if price is at least as high as long-run average costs. These costs include out-of-pocket payments to labor, lenders, material suppliers, or landlords and opportunity costs, such as returns on the property assets owned by the firm.

B. Supply Behavior in Competitive Industries

4. Each firm's rising MC curve is its supply curve. To obtain the supply curve of a group of competitive firms, we add horizontally their separate supply curves. The supply curve of the industry hence represents the marginal cost curve for the competitive industry as a whole.
5. Because firms can adjust production over time, we distinguish two different time periods: (*a*) short-run equilibrium, when variable factors like labor can change but fixed factors like capital and the number of firms cannot, and (*b*) long-run equilibrium, when the numbers of firms and plants, and all other conditions, adjust completely to the new demand conditions.
6. In the long run, when firms are free to enter and leave the industry and no one firm has any particular advantage of skill or location, competition will eliminate any excess profits earned by existing firms in the industry. So, just as free exit implies that price cannot fall below the zero-profit point, free entry implies that price cannot exceed long-run average cost in long-run equilibrium.
7. When an industry can expand its production without pushing up the prices of its factors of production, the resulting long-run supply curve will be horizontal. When an industry uses factors specific to it, such as scarce beachfront property, its long-run supply curve will slope upward.

C. Special Cases of Competitive Markets

8. Recall the general rules that apply to competitive supply and demand: Under the demand rule, an increase in the demand for a commodity (the supply curve being unchanged) will generally raise the price of the commodity and also increase the quantity demanded. A decrease in demand will have the opposite effects.

   Under the supply rule, an increase in the supply of a commodity (the demand curve being constant) will generally lower the price and increase the quantity sold. A decrease in supply has the opposite effects.
9. Important special cases include constant and increasing costs, completely inelastic supply (which produces economic rents), and backward-bending supply. These special cases will explain many important phenomena found in markets.

D. Efficiency and Equity of Competitive Markets

10. The analysis of competitive markets sheds light on the efficient organization of a society. Allocative or Pareto efficiency occurs when there is no way of reorganizing production and distribution such that everyone's satisfaction can be improved.
11. Under ideal conditions, a competitive economy attains allocative efficiency. Efficiency requires that all firms are perfect competitors and that there are no externalities like pollution or imperfect information. Efficiency implies that economic surplus is maximized, where economic surplus equals consumer surplus plus producer surplus.
12. Efficiency comes because (*a*) when consumers maximize satisfaction, the marginal utility (in terms of leisure) just equals the price; (*b*) when competitive

producers supply goods, they choose output so that marginal cost just equals price; (*c*) since $MU = P$ and $MC = P$, it follows that $MU = MC$.

13. There are exacting limits on the social optimality of competitive markets.
    a. Pareto efficiency requires perfect competition, complete information, and no externalities. When all three conditions are met, this will lead to the important efficiency condition:

Price ratio = marginal cost ratio = marginal utility ratio

    b. The most perfectly competitive markets may not produce a fair distribution of income and consumption. Societies may therefore decide to modify the laissez-faire market outcomes. Economics has the important role of analyzing the relative costs and benefits of alternative kinds of interventions.
14. Marginal cost is a fundamental concept for attaining any goal, not just for profits. Efficiency requires that the marginal cost of attaining the goal be equal in every activity.

CONCEPTS FOR REVIEW

Competitive Supply

$P = MC$ as maximum-profit condition
firm's *ss* supply curve and its *MC* curve
zero-profit condition, where $P = MC = AC$
shutdown point, where $P = MC = AVC$
summing individual *ss* curves to get industry *SS*
short-run and long-run equilibrium
long-run zero-profit condition
producer surplus + consumer surplus = economic surplus
efficiency = maximizing economic surplus

Efficiency and Equity

allocative efficiency, Pareto efficiency
conditions for allocative efficiency: $MU = P = MC$
efficiency of competitive markets
efficiency vs. equity

FURTHER READING AND INTERNET WEBSITES

Further Reading

The efficiency of perfect competition is one of the major findings of microeconomics. Advanced books in microeconomics, such as those listed in Chapter 4, can give insights into the basic findings.

Nobel Prizes in economics were awarded to Kenneth Arrow, John Hicks, and Gerard Debreu for their contributions to developing the theory of perfect competition and its relationship to economic efficiency. Their essays surveying the field are highly useful and are contained in Assar Lindbeck, *Nobel Lectures in Economics* (University of Stockholm, 1992). See also the Nobel website listed below for the Nobel citations for these economists.

Websites

For the citations of Arrow, Hicks, and Debreu, look at the website *www.nobel.se/economics/index.html* to read about the importance of their contributions and how they relate to economics.

QUESTIONS FOR DISCUSSION

1. Explain why each of the following statements about profit-maximizing competitive firms is incorrect. Restate each one correctly.
    a. A competitive firm will produce output up to the point where price equals average variable cost.
    b. A firm's shutdown point comes where price is less than minimum average cost.
    c. A firm's supply curve depends only on its marginal cost. Any other cost concept is irrelevant for supply decisions.

d. The $P = MC$ rule for competitive industries holds for upward-sloping, horizontal, and downward-sloping MC curves.

e. The competitive firm sets price equal to marginal cost.

2. Suppose you are a perfectly competitive firm producing computer memory chips. Your production capacity is 1000 units per year. Your marginal cost is \$10 per chip up to capacity. You have a fixed cost of \$10,000 if production is positive and \$0 if you shut down. What are your profit-maximizing levels of production and profit if the market price is (*a*) \$5 per chip, (*b*) \$15 per chip, and (*c*) \$25 per chip? For case (*b*), explain why production is positive even though profits are negative.

3. One of the most important rules of economics, business, and life is the *sunk-cost principle,* "Let bygones be bygones." This means that sunk costs (which are bygone in the sense that they are unrecoverably lost) should be ignored when decisions are being made. Only future costs, involving marginal and variable costs, should count in making rational decisions.

To see this, consider the following: We can calculate fixed costs in Table 8-1 as the cost level when output is 0. What are fixed costs? What is the profit-maximizing level of output for the firm in Table 8-1 if price is \$40 while fixed costs are \$0? \$55,000? \$100,000? \$1,000,000,000? Minus \$30,000? Explain the implication for a firm trying to decide whether to shut down.

4. Examine the cost data shown in Table 8-1. Calculate the supply decision of a profit-maximizing competitive firm when price is \$21, \$40, and \$60. What would the level of total profit be for each of the three prices? What would happen to the exit or entry of identical firms in the long run at each of the three prices?

5. Using the cost data shown in Table 8-1, calculate the price elasticity of supply between $P = 40$ and $P = 40.02$ for the individual firm. Assume that there are 2000 identical firms, and construct a table showing the industry supply schedule. What is the industry price elasticity of supply between $P = 40$ and $P = 40.02$?

6. Examine Figure 8-12 to see that competitive firm C is not producing at all. Explain the reason why the profit-maximizing output level for firm C is at $q_c = 0$. What would happen to total industry cost of production if firm C produced 1 unit while firm B produced 1 less unit than the competitive output level?

Say that firm C is a mom-and-pop grocery store. Why would chain grocery stores A and B drive C out of business? How do you feel about keeping C in business? What would be the economic impact of legislation that divided the market into three equal parts between the mom-and-pop store and chain stores A and B?

7. Often, consumer demand for a commodity will depend upon the use of durable goods, such as housing or transportation. In such a case, demand will show a time-varying pattern of response similar to that of supply. A good example is gasoline. In the short run the stock of automobiles is fixed, while in the long run consumers can buy new automobiles or bicycles.

What is the relationship between the time period and the price elasticity of demand for gasoline? Sketch the short-run and long-run demand curves for gasoline. Show the impact of a decline in the supply of gasoline in both periods. Describe the impact of an oil shortage on the price of gasoline and the quantity demanded in both the long run and the short run. State two new rules of demand, (*c*) and (*d*), parallel to the rules of supply (*c*) and (*d*) discussed in the General Rules portion of Section C above, that relate the impact of a shift in supply on price and quantity in the long run and the short run.

8. Interpret this dialogue:

A: "How can competitive profits be zero in the long run? Who will work for nothing?"

B: "It is only *excess* profits that are wiped out by competition. Managers get paid for their work; owners get a normal return on capital in competitive long-run equilibrium—no more, no less."

9. Consider three firms which are emitting sulfur into the California air. We will call supply the units of pollution control or reduction. Each firm has a cost-of-reduction schedule, and we will say that these schedules are given by the MC curves of firms A, B, and C in Figure 8-12.

a. Interpret the "market" supply or MC schedule for reducing sulfur emissions, shown in the middle of Figure 8-12.

b. Say that the pollution-control authority decides to seek 10 units of pollution control. What is the efficient allocation of pollution control across the three firms?

c. Say that the pollution-control authority decides to have the first two firms produce 5 units each of pollution control. What is the additional cost?

d. Say that the pollution-control authority decides upon a "pollution charge" to reduce pollution to 10 units. Can you identify what the appropriate charge would be using Figure 8-12? Can you say how each firm would respond? Would the pollution reduction be efficient?

e. Explain the importance of marginal cost in the efficient reduction of pollution in this case.

10. In any competitive market, such as illustrated in Figure 8-11, the area above the market price line and below the DD curve is consumer surplus (see the discussion in Chapter 5). The area above the SS curve

and below the price line is producer surplus and equals profits plus rent to the firms in the industry or owners of specialized inputs to the industry. The sum of the producer and consumer surpluses is economic surplus and measures the net contribution of that good to utility above the cost of production.

Can you find any reorganization of production that would increase the economic surplus in Figure 8-11 as compared to the competitive equilibrium at point *E*? If the answer is no, then the equilibrium is allocationally efficient (or Pareto efficient). Define allocational efficiency; then answer the question and explain your answer.

CHAPTER

Imperfect Competition and Monopoly

9

The best of all monopoly profits is a quiet life.

J. R. Hicks

Perfect competition is an idealized market of atomistic firms who are price-takers. In fact, while they are easily analyzed, such firms are hard to find. When you buy your car from Ford or Toyota, your hamburgers from McDonald's or Wendy's, or your computer from Dell or Apple, you are dealing with firms large enough to affect the market price. Indeed, most markets in the economy are dominated by a handful of large firms, often only two or three. Welcome to the world you live in, the world of imperfect competition.

A. PATTERNS OF IMPERFECT COMPETITION

We shall see that for a given technology, prices are higher and outputs are lower under imperfect competition than under perfect competition. But imperfect competitors have virtues along with these vices. Large firms exploit economies of large-scale production and are responsible for much of the innovation that propels long-term economic growth. If you understand how imperfectly competitive markets work, you will have a much deeper understanding of modern industrial economies.

Recall that a perfectly competitive market is one in which no firm is large enough to affect the market price. By this strict definition, few markets in the U.S. economy are perfectly competitive. Think of the following: aircraft, aluminum, automobiles, computer software, breakfast cereals, chewing gum, cigarettes, electricity distribution, refrigerators, and wheat. How many of these are sold in perfectly competitive markets? Certainly not aircraft, aluminum, or automobiles. Until World War II there was only one aluminum company, Alcoa. Even today, the four largest U.S. firms produce three-quarters of U.S. aluminum output. The world commercial-aircraft market is dominated by only two firms, Boeing and Airbus. In the automotive industry, too, the top five automakers (including Toyota and Honda) have almost 80 percent of the U.S. car and light-truck market. The software industry shows rapid innovation, yet for most individual software applications, from tax accounting to gaming, a few firms have most of the sales.

What about breakfast cereals, chewing gum, cigarettes, and refrigerators? These markets are dominated even more completely by a relatively small number of companies. Nor does the retail market in electricity meet the definition of perfect competition. In most localities, a single company distributes all the electricity used by the population. Very few of us will find it economical to build a windmill to generate our own power!

Looking at the list above, you will find that only wheat falls within our strict definition of perfect

competition. All the other goods, from autos to cigarettes, fail the competitive test for a simple reason: Some of the firms in the industry can affect the market price by changing the quantity they sell. To put it another way, they have *some* control over the price of their output.

Definition of Imperfect Competition

If a firm can affect the market price of its output, the firm is classified as an imperfect competitor.

Imperfect competition prevails in an industry whenever individual sellers can affect the price of their output. The major kinds of imperfect competition are monopoly, oligopoly, and monopolistic competition.

Imperfect competition does not imply that a firm has absolute control over the price of its product. Take the cola market, where Coca-Cola and Pepsi together have the major share of the market, and imperfect competition clearly prevails. If the average price of other producers' sodas in the market is 75 cents, Pepsi may be able to set the price of a can at 70 or 80 cents and still remain a viable firm. The firm could hardly set the price at $40 or 5 cents a can because at those prices it would go out of business. We see, then, that an imperfect competitor has some but not complete discretion over its prices.

Moreover, the amount of discretion over price will differ from industry to industry. In some imperfectly competitive industries, the degree of monopoly power is very small. In the retail computer business, for example, more than a few percent difference in price will usually have a significant effect upon a firm's sales. In the market for computer operating systems, by contrast, Microsoft has a virtual monopoly and has great discretion about the price of its Windows software.

Graphical Depiction. Figure 9-1 shows graphically the difference between the demand curves faced by perfectly and imperfectly competitive firms. Figure 9-1 (*a*) reminds us that a perfect competitor faces a horizontal demand curve, indicating that it can sell all it wants at the going market price. An imperfect competitor, in contrast, faces a downward-sloping demand curve. Figure 9-1 (*b*) shows that if an imperfectly competitive firm increases its sales, it will definitely depress the market price of its output as it moves down its *dd* demand curve.

Another way of seeing the difference between perfect and imperfect competition is by considering

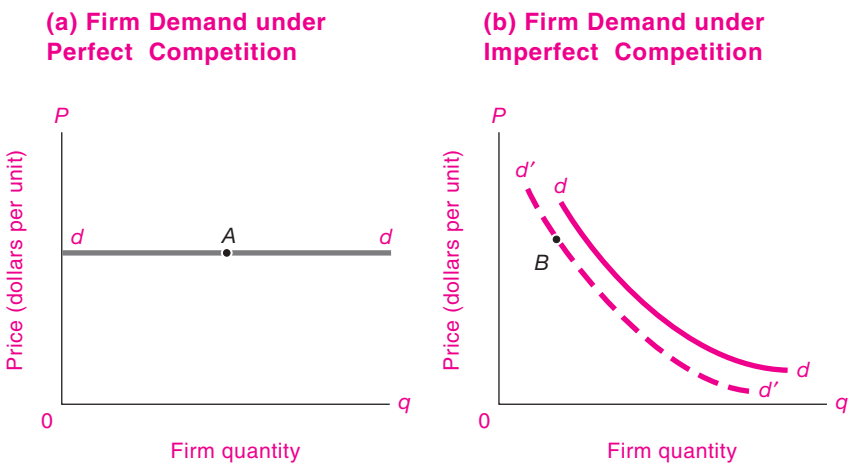


FIGURE 9-1. Acid Test for Imperfect Competition Is Downward Tilt of Firm's Demand Curve

(a) The perfectly competitive firm can sell all it wants along its horizontal *dd* curve without depressing the market price. **(b)** But the imperfect competitor will find that its demand curve slopes downward as higher price drives sales down. And unless it is a sheltered monopolist, a cut in its rivals' prices will appreciably shift its own demand curve leftward to *d′d′*.

the price elasticity of demand. For a perfect competitor, demand is perfectly elastic; for an imperfect competitor, demand has a finite elasticity. As an exercise in use of the elasticity formulas, calculate the elasticities for the perfect competitor in Figure 9-1(*a*) and the imperfect competitor at point *B* in 9-1(*b*).

The fact that the demand curves of imperfect competitors slope down has an important implication: Imperfect competitors are *price-makers* not *price-takers*. They must decide on the price of their product, while perfect competitors take the price as given.

VARIETIES OF IMPERFECT COMPETITORS

A modern industrial economy like the United States is a jungle populated with many species of imperfect competition. The dynamics of the personal computer industry, driven by rapid improvements in technology, are different from the patterns of competition in the not-so-lively funeral industry. Nevertheless, much can be learned about an industry by paying careful attention to its market structure, particularly the number and size of sellers and how much of the market the largest sellers control. Economists classify imperfectly competitive markets into three different market structures.

Monopoly

At one pole of the competitive spectrum is the perfect competitor, which is one firm among a vast multitude of firms. At the other pole is the **monopoly,** which is a single seller with complete control over an industry. (The word comes from the Greek words *mono* for "one" and *polist* for "seller.") A monopolist is the only firm producing in its industry, and there is no industry producing a close substitute. Moreover, for now we assume that the monopolist must sell everything at the same price—there is no price discrimination.

True monopolies are rare today. Most monopolies persist because of some form of government regulation or protection. For example, a pharmaceutical company that discovers a new wonder drug may be granted a patent, which gives it monopoly control over that drug for a number of years. Another important example of monopoly is a franchised local utility, such as the firm that provides your household water. In such cases there is truly a single seller of a service with no close substitutes. One of the few examples of a monopoly without government license is Microsoft Windows, which has succeeded in maintaining its monopoly through large investments in research and development, rapid innovation, network economies, and tough (and sometimes illegal) tactics against its competitors.

But even monopolists must always be looking over their shoulders for potential competitors. The pharmaceutical company will find that a rival will produce a similar drug; telephone companies that were monopolists a decade ago now must reckon with cellular telephones; Bill Gates worries that some small firm is waiting in the wings to unseat Microsoft's monopolistic position. *In the long run, no monopoly is completely secure from attack by competitors.*

Oligopoly

The term **oligopoly** means "few sellers." Few, in this context, can be a number as small as 2 or as large as 10 or 15 firms. The important feature of oligopoly is that each individual firm can affect the market price. In the airline industry, the decision of a single airline to lower fares can set off a price war which brings down the fares charged by all its competitors.

Oligopolistic industries are common in the U.S. economy, especially in the manufacturing, transportation, and communications sectors. For example, there are only a few car makers, even though the automobile industry sells many different models. The same is true in the market for household appliances: stores are filled with many different models of refrigerators and dishwashers, all made by a handful of companies. You might be surprised to know that the breakfast cereal industry is an oligopoly dominated by a few firms even though there seem to be endless varieties of cereals.

Monopolistic Competition

The final category we examine is **monopolistic competition.** In this situation, a large number of sellers produce differentiated products. This market structure resembles perfect competition in that there are many sellers, none of whom has a large share of the market. It differs from perfect competition in that the products sold by different firms are not identical. **Differentiated products** are ones whose important characteristics vary. Personal computers, for example, have differing characteristics such as speed, memory, hard disk, modem, size, and weight. Because computers are differentiated, they can sell at slightly different prices.

The classic case of monopolistic competition is the retail gasoline market. You may go to the local Shell station, even though it charges slightly more, because it is on your way to work. But if the price at Shell rises more than a few pennies above the competition, you might switch to the Merit station a short distance away.

This example illustrates the importance of location in product differentiation. It takes time to go to the bank or the grocery store, and the amount of time needed to reach different stores will affect our shopping choices. The *whole price* of a good includes not just its dollar price but also the opportunity cost of search, travel time, and other non-dollar costs. Because the whole prices of local goods are lower than those in faraway places, people generally tend to shop close to home or to work. This consideration also explains why large shopping complexes are so popular: they allow people to buy a wide variety of goods while economizing on shopping time. Today, shopping on the Internet is increasingly important because, even when shipping costs are added, the time required to buy the good online can be very low compared to getting in your car or walking to a shop.

Product quality is an increasingly important part of product differentiation today. Goods differ in their characteristics as well as their prices. Most personal computers can run the same software, and there are many manufacturers. Yet the personal computer industry is a monopolistically competitive industry, because computers differ in speed, size, memory, repair services, and ancillaries like CDs, DVDs, Internet connections, and sound systems. Indeed, a whole batch of monopolistically competitive computer magazines is devoted to explaining the differences among the computers produced by the monopolistically competitive computer manufacturers!

Competition vs. Rivalry

When studying oligopolies, it is important to recognize that imperfect competition is not the same as no competition. Indeed, some of the most vigorous rivalries in the

Types of Market Structures

Structure	Number of producers and degree of product differentiation	Part of economy where prevalent	Firm's degree of control over price	Methods of marketing
Perfect competition	Many producers; identical products	Financial markets and agricultural products	None	Market exchange or auction
Imperfect competition				
Monopolistic competition	Many producers; many real or perceived differences in product	Retail trade (pizzas, beer, . . .), personal computers	Some	Advertising and quality rivalry; administered prices
Oligopoly	Few producers; little or no difference in product	Steel, chemicals, . . .		
	Few producers; products are differentiated	Cars, word-processing software, . . .		
Monopoly	Single producer; product without close substitutes	Franchise monopolies (electricity, water); Microsoft Windows; patented drugs	Considerable	Advertising

TABLE 9-1. Alternative Market Structures

Most industries are imperfectly competitive. Here are the major features of different market structures.

economy occur in markets where there are but a few firms. Just look at the cutthroat competition in the airline industry, where two or three airlines may fly a particular route but still engage in periodic fare wars.

How can we distinguish the rivalry of oligopolists from perfect competition? Rivalry encompasses a wide variety of behavior to increase profits and market share. It includes advertising to shift out the demand curve, price cuts to attract business, and research to improve product quality or develop new products. Perfect competition says nothing about rivalry but simply means that no single firm in the industry can affect the market price.

Table 9-1 on page 172 gives a picture of the various possible categories of imperfect and perfect competition. This table is an important summary of the different kinds of market structures and warrants careful study.

SOURCES OF MARKET IMPERFECTIONS

Why do some industries display near-perfect competition while others are dominated by a handful of large firms? Most cases of imperfect competition can be traced to two principal causes. First, industries tend to have fewer sellers when there are significant economies of large-scale production and decreasing costs. Under these conditions, large firms can simply produce more cheaply and then undersell small firms, which cannot survive.

Second, markets tend toward imperfect competition when there are "barriers to entry" that make it difficult for new competitors to enter an industry. In some cases, the barriers may arise from government laws or regulations which limit the number of competitors. In other cases, there may be economic factors that make it expensive for a new competitor to break into a market. We will examine both sources of imperfect competition.

Costs and Market Imperfection

The technology and cost structure of an industry help determine how many firms that industry can support and how big they will be. The key is whether there are economies of scale in an industry. If there are economies of scale, a firm can decrease its average costs by expanding its output, at least up to a point. That means bigger firms will have a cost advantage over smaller firms.

When economies of scale prevail, one or a few firms will expand their outputs to the point where they produce most of the industry's total output. The industry then becomes imperfectly competitive. Perhaps a single monopolist will dominate the industry; a more likely outcome is that a few large sellers will control most of the industry's output; or there might be a large number of firms, each with slightly different products. Whatever the outcome, we must inevitably find some kind of imperfect competition instead of the atomistic perfect competition of price-taking firms.

We can see how the relationship between the size of the market and the scale economies helps determine the market structure. There are three interesting cases, illustrated in Figure 9-2.

1. To understand further how costs may determine market structure, let's first look at a case which is favorable for perfect competition. Figure 9-2(*a*) shows an industry where the point of minimum average cost is reached at a level of output that is tiny relative to the market. As a result, this industry can support the large number of efficiently operating firms that are needed for perfect competition. Figure 9-2(*a*) illustrates the cost curves in the perfectly competitive farm industry.
2. An intermediate case is an industry with economies of scale that are large relative to the size of the industry. Numerous detailed econometric and engineering studies confirm that many nonagricultural industries show declining average long-run costs. For example, Table 9-2 shows the results of a study of six U.S. industries. For these cases, the point of minimum average cost occurs at a large fraction of industry output.

   Now consider Figure 9-2(*b*), which shows an industry where firms have minimum average costs at a sizable fraction of the market. The industry demand curve allows only a small number of firms to coexist at the point of minimum average cost. Such a cost structure will lead to oligopoly. Most manufacturing industries in the United States—including steel, automobiles, cement, and oil—have a demand and cost structure similar to the one in Figure 9-2(*b*). These industries will tend to be oligopolistic, since they can support only a few large producers.
3. A final important case is natural monopoly. A **natural monopoly** is a market in which the

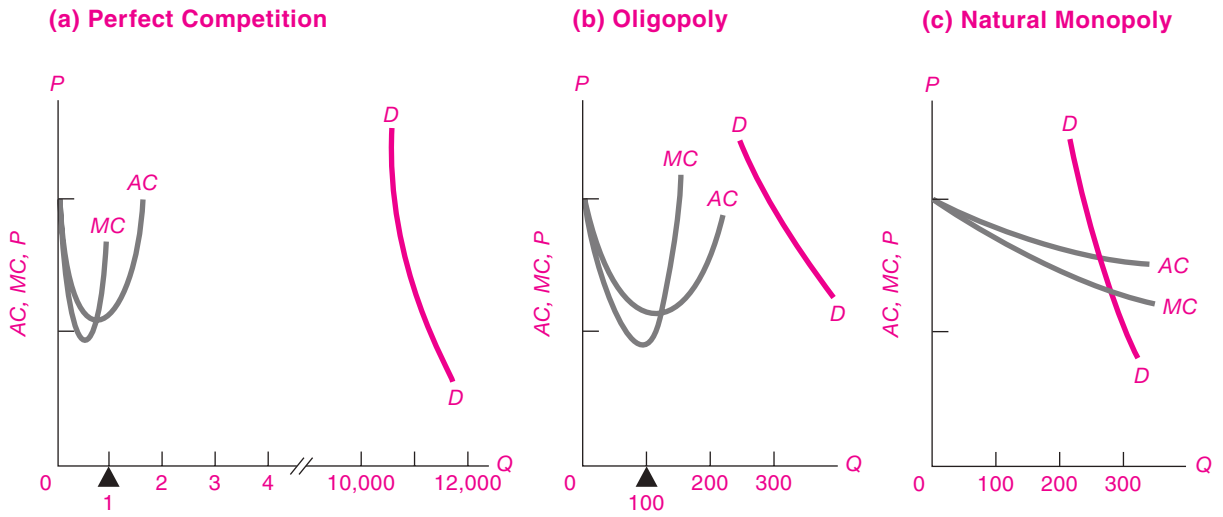


FIGURE 9-2. Market Structure Depends on Relative Cost and Demand Factors

Cost and demand conditions affect market structures. In perfectly competitive (**a**), total industry demand *DD* is so vast relative to the efficient scale of a single seller that the market allows viable coexistence of numerous perfect competitors. In (**b**), costs turn up at a higher level of output relative to total industry demand *DD*. Coexistence of numerous perfect competitors is impossible, and oligopoly will emerge. When costs fall rapidly and indefinitely, as in the case of natural monopoly in (**c**), one firm can expand to monopolize the industry.

Industry	(1) Share of U.S. output needed by a single firm to exploit economies of scale (%)	(2) Actual average market share of each of the top three firms (%)	(3) Reasons for economies of large-scale operations
Beer brewing	10–14	13	Need to create a national brand image and to coordinate investment
Cigarettes	6–12	23	Advertising and image differentiation
Glass bottles	4–6	22	Need for central engineering and design staff
Cement	2	7	Need to spread risk and raise capital
Refrigerators	14–20	21	Marketing requirements and length of production runs
Petroleum	4–6	8	Need to spread risk on crude-oil ventures and coordinate investment

TABLE 9-2. Industrial Competition Is Based on Cost Conditions

This study examined the impact of cost conditions on concentration patterns. Column (1) shows the estimate of the point where the long-run average cost curve begins to turn up, as a share of industry output. Compare this with the average market share of each of the top three firms in column (2).

Source: F. M. Scherer and David Ross, *Industrial Market Structure and Economic Performance,* 3d ed. (Houghton Mifflin, Boston, 1990).

industry's output can be efficiently produced only by a single firm. This occurs when the technology exhibits significant economies of scale over the entire range of demand. Figure 9-2(*c*) shows the cost curves of a natural monopolist. With perpetual increasing returns to scale, average and marginal costs fall forever. As output grows, the firm can charge lower and lower prices and still make a profit, since its average cost is falling. Peaceful competitive coexistence of thousands of perfect competitors will be impossible because one large firm is so much more efficient than a collection of small firms.

Some important examples of natural monopolies are the local distribution in telephone, electricity, gas, and water as well as long-distance links in railroads, highways, and electrical transmission. Many of the most important natural monopolies are "network industries" (see the discussion in Chapter 6).

Technological advances, however, can undermine natural monopolies. Most of the U.S. population is now served by at least two cellular telephone networks, which use radio waves instead of wires and are undermining the old natural monopoly of the telephone companies. We see a similar trend today in cable TV as competitors invade these natural monopolies and are turning them into hotly contested oligopolies.

Barriers to Entry

Although cost differences are the most important factor behind market structures, barriers to entry can also prevent effective competition. **Barriers to entry** are factors that make it hard for new firms to enter an industry. When barriers are high, an industry may have few firms and limited pressure to compete. Economies of scale act as one common type of barrier to entry, but there are others, including legal restrictions, high cost of entry, advertising, and product differentiation.

Legal Restrictions. Governments sometimes restrict competition in certain industries. Important legal restrictions include patents, entry restrictions, and foreign-trade tariffs and quotas. A *patent* is granted to an inventor to allow temporary exclusive use (or monopoly) of the product or process that is patented. For example, pharmaceutical companies are often granted valuable patents on new drugs in which they have invested hundreds of millions of research-and-development dollars. Patents are one of the few forms of government-granted monopolies that are generally approved of by economists. Governments grant patent monopolies to encourage inventive activity. Without the prospect of monopoly patent protection, a company or a sole inventor might be unwilling to devote time and resources to research and development. The temporarily high monopoly price and the resulting inefficiency is the price society pays for the invention.

Governments also impose *entry restrictions* on many industries. Typically, utilities, such as telephone, electricity distribution, and water, are given *franchise monopolies* to serve an area. In these cases, the firm gets an exclusive right to provide a service, and in return the firm agrees to limit its prices and provide universal service in its region even when some customers might be unprofitable.

Free trade is often controversial, as we will see in the chapter on that subject. But one factor that will surprise most people is how important international trade is to promoting vigorous competition.

Historians who study the tariff have written, "The tariff is the mother of trusts." (See question 10 at the end of this chapter for an analysis of this subject.) This is because government-imposed *import restrictions* have the effect of keeping out foreign competitors. It could very well be that a single country's market for a product is only big enough to support two or three firms in an industry, while the world market is big enough to support a large number of firms.

We can see the effect of restricting foreign competition in terms of Figure 9-2. Suppose a small country like Belgium or Benin decides that only *its* national airlines should provide airline service in the country. It is unlikely that such tiny airlines could have an efficient fleet of airplanes, reservation and repair systems, and Internet support. Service to Belgium and Benin would be poor, and prices would be high. What has happened is that the protectionist policy has changed the industry structure from Figure 9-2(*b*) to 9-2(*c*).

When markets are broadened by abolishing tariffs in a large free-trade area, vigorous and effective competition is encouraged and monopolies tend to lose their power. One of the most dramatic examples

of increased competition has come in the European Union, which has lowered tariffs among member countries steadily over the last three decades and has benefited from larger markets for firms and lower concentration of industry.

High Cost of Entry. In addition to legally imposed barriers to entry, there are economic barriers as well. In some industries the price of entry simply may be very high. Take the commercial-aircraft industry, for example. The high cost of designing and testing new airplanes serves to discourage potential entrants into the market. It is likely that only two companies—Boeing and Airbus—can afford the $10 to $20 billion that the next generation of aircraft will cost to develop.

In addition, companies build up intangible forms of investment, and such investments might be very expensive for any potential new entrant to match. Consider the software industry. Once a spreadsheet program (like Excel) or a word-processing program (like Microsoft Word) has achieved wide acceptability, potential competitors find it difficult to make inroads into the market. Users, having learned one program, are reluctant to switch to another. Consequently, in order to get people to try a new program, any potential entrant will need to run a big promotional campaign, which would be expensive and may still result in failure to produce a profitable product. (Recall our discussion of network effects in Chapter 6.)

Advertising and Product Differentiation. Sometimes it is possible for companies to create barriers to entry for potential rivals by using advertising and product differentiation. Advertising can create product awareness and loyalty to well-known brands. For example, Pepsi and Coca-Cola spend hundreds of millions of dollars per year advertising their brands, which makes it very expensive for any potential rivals to enter the cola market.

In addition, product differentiation can impose a barrier to entry and increase the market power of producers. In many industries—such as breakfast cereals, automobiles, household appliances, and cigarettes—it is common for a small number of manufacturers to produce a vast array of different brands, models, and products. In part, the variety appeals to the widest range of consumers. But the enormous number of differentiated products also serves to discourage potential competitors. The demands for each of the individual differentiated products will be so small that they will not be able to support a large number of firms operating at the bottom of their U-shaped cost curves. The result is that perfect competition's *DD* curve in Figure 9-2(*a*) contracts so far to the left that it becomes like the demand curves of oligopoly or monopoly shown in Figure 9-2(*b*) and (*c*). Hence, differentiation, like tariffs, produces greater concentration and more imperfect competition.

Branding and Differentiated Products

One important part of modern business strategy is to establish a brand. Suppose, for example, that all the Coca-Cola factories were to collapse in an earthquake. What would happen to the value of Coca-Cola's stock price? Would it go to zero?

The answer, according to finance specialists, is that, even with no tangible assets, Coca-Cola would still be worth about $67 billion. This is the company's *brand value*. A product's brand involves the perception of taste and quality in the minds of consumers. Brand value is established when a firm has a product that is seen as better, more reliable, or tastier than other products, branded or nonbranded.

In a world of differentiated products, some firms earn fancy profits because of the value of their brands. The following table shows recent estimates of the top 10 brands:

Rank	Brand	Brand value, 2006 ($, billion)
1	Coca-Cola	67
2	Microsoft	60
3	IBM	56
4	GE	49
5	Intel	32
6	Nokia	30
7	Toyota	28
8	Disney	28
9	McDonald's	27
10	Mercedes-Benz	22

Source: *BusinessWeek*, available on the Internet at *http://www.businessweek.com/*.

Thus, for Coca-Cola, the market value of the firm was $67 billion more than would be justified by its plant,

equipment, and other assets. How do firms establish and maintain brand value? First, they usually have an innovative product, such as a new drink, a cute cartoon mouse, or a high-quality automobile. Second, they maintain their brand value by heavy advertising, even associating a deadly product like Marlboro cigarettes (brand rank 14) with a good-looking cowboy in a romantic sunset with beautiful horses. Third, they protect their brands using intellectual property rights such as patents and copyrights. In one sense, brand value is the residue of past innovative activity.

B. MONOPOLY BEHAVIOR

We begin our survey of the behavior of imperfect competitors with an analysis of the polar case of monopoly. We need a new concept, marginal revenue, which will have wide applications for other market structures as well. The major conclusion will be that monopolistic practices lead to inefficiently high prices and low outputs and therefore reduce consumer welfare.

THE CONCEPT OF MARGINAL REVENUE

Price, Quantity, and Total Revenue

Suppose that you have a monopoly on a new kind of computer game called *Monopolia*. You wish to maximize your profits. What price should you charge, and what output level should you produce?

To answer these questions, we need a new concept, *marginal revenue* (or *MR*). From the firm's demand curve, we know the relationship between price (P) and quantity sold (q). These are shown in columns (1) and (2) of Table 9-3 and as the blue demand curve (dd) for the monopolist in Figure 9-3(a).

We next calculate the total revenue at each sales level by multiplying price times quantity. Column (3) of Table 9-3 shows how to calculate the **total revenue** (TR), which is simply $P \times q$. Thus 0 units bring in TR of 0; 1 unit brings in $TR = \$180 \times 1 = \$180$; 2 units bring in $\$160 \times 2 = \$320$; and so forth.

In this example of a straight-line or linear demand curve, total revenue at first rises with output, since the reduction in P needed to sell the extra q is moderate in this upper, elastic range of the demand curve. But when we reach the midpoint of the straight-line demand curve, TR reaches its maximum. This comes at $q = 5$, $P = \$100$, with $TR = \$500$. Increasing q beyond this point brings the firm into the inelastic demand region. For inelastic demand, reducing price increases sales less than proportionally, so total revenue falls. Figure 9-3(b) shows TR to be dome-shaped, rising from zero at a very high price to a maximum of \$500 and then falling to zero as price approaches zero.

How could you find the price at which revenues are maximized? You would see in Table 9-3 that TR is maximized when $q = 5$ and $P = 100$. This is the point where the demand elasticity is exactly 1.

Note that the price per unit can be called *average revenue* (AR) to distinguish it from total revenue. Hence, we get $P = AR$ by dividing TR by q (just as we earlier got AC by dividing TC by q). Verify that if column (3) had been written down before column (2), we could have filled in column (2) by division.

Marginal Revenue and Price

The final new concept is marginal revenue. **Marginal revenue** (MR) is the change in revenue that is generated by an additional unit of sales. MR can be either positive or negative.

Table 9-3 shows marginal revenue in column (4). MR is calculated by subtracting the total revenues of adjacent outputs. When we subtract the TR we get by selling q units from the TR we get by selling $q + 1$ units, the difference is extra revenue or MR. Thus, from $q = 0$ to $q = 1$, we get $MR = \$180 - \$0$. From $q = 1$ to $q = 2$, MR is $\$320 - \$180 = \$140$.

MR is positive until we arrive at $q = 5$ and negative from then on. What does the strange notion of negative marginal revenue mean? That the firm is paying people to take its goods? Not at all. Negative MR means that in order to sell additional units, the firm must decrease its price on earlier units so much that its total revenues decline.

For example, when the firm sells 5 units, it gets

$$TR\ (5\ \text{units}) = 5 \times \$100 = \$500$$

Now say the firm wishes to sell an additional unit of output. Because it is an imperfect competitor, it can increase sales only by lowering price. So to sell 6 units, it lowers the price from \$100 to \$80. It gets

Total and Marginal Revenue

(1) Quantity q	(2) Price $P = AR = TR/q$ ($)	(3) Total revenue $TR = P \times q$ ($)	(4) Marginal revenue MR ($)
0	200	0	
			+180
1	180	180	
			+140
2	160	320	
			+100
3	140	420	
			+60
4	120	480	+40
			+20
5	100	500	
			___
6	80	480	
			−60
7	60	___	
			−100
8	40	320	
			−140
9	___	180	
			−180
10	0	0	

TABLE 9-3. Marginal Revenue Is Derived from Demand Schedule

Total revenue (TR) in column (3) comes from multiplying P by q. To get marginal revenue (MR), we increase q by a unit and calculate the change in total revenue. MR is less than P because of the lost revenue from lowering the price on previous units to sell another unit of q. Note that MR is positive when demand is elastic. But after demand turns inelastic, MR becomes negative even though price is still positive.

\$80 of revenue from the sixth unit, but it gets only $5 \times \$80$ on the first 5 units, yielding

$$TR(6 \text{ units}) = (5 \times \$80) + (1 \times \$80)$$
$$= \$400 + \$80 = \$480$$

Marginal revenue between 5 and 6 units is \$480 − \$500 = −\$20. The necessary price reduction on the first 5 units was so large that, even after adding in the sale of the sixth unit, total revenue fell. This is what happens when MR is negative. To test your understanding, fill in the blanks in columns (2) to (4) of Table 9-3.

Note that even though MR is negative, AR, or price, is still positive. Do not confuse marginal revenue with average revenue or price. Table 9-3 shows

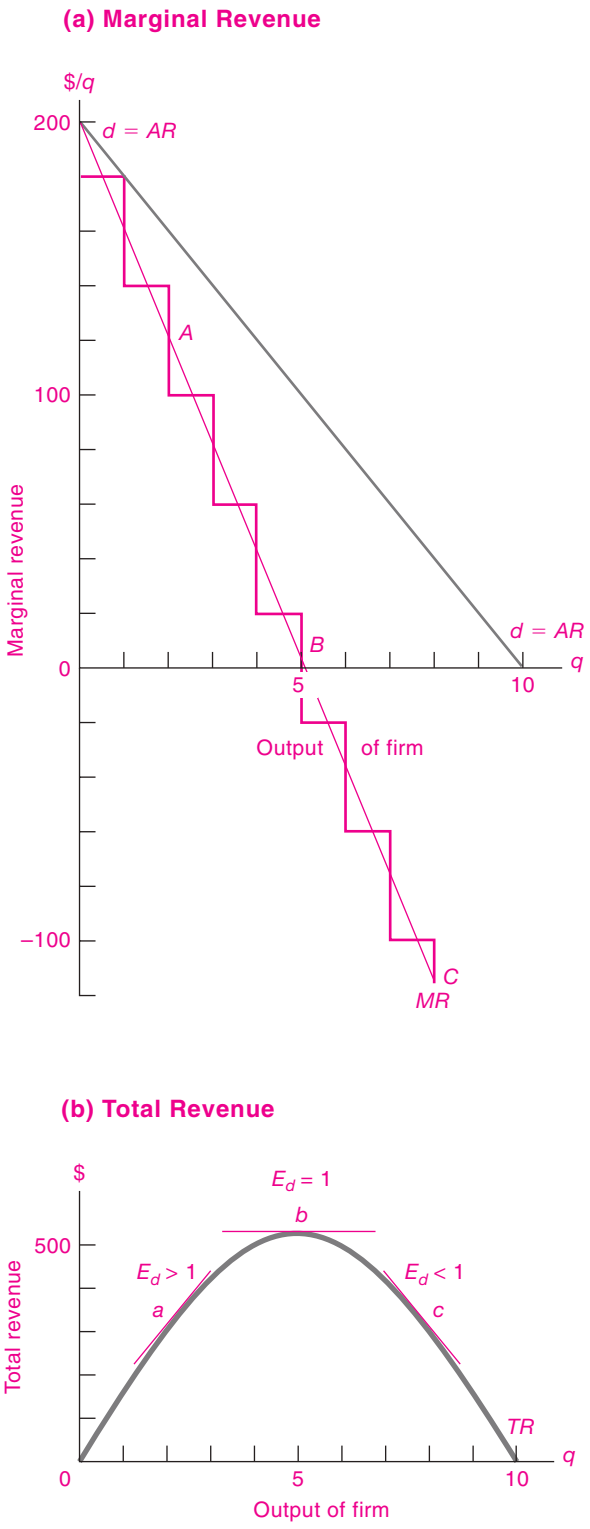


FIGURE 9-3. Marginal Revenue Curve Comes from Demand Curve

(a) The steps show the increments of total revenue from each extra unit of output. *MR* falls below *P* from the beginning. *MR* becomes negative when *dd* turns inelastic. Smoothing the incremental steps of *MR* gives the smooth, thin green *MR* curve, which in the case of straight line *dd* will always have twice as steep a slope as *dd*.

(b) Total revenue is dome-shaped—rising from zero where $q = 0$ to a maximum (where *dd* has unitary elasticity) and then falling back to zero where $P = 0$. If we graph *TR* as a smooth blue line in (*b*), this gives smoothed green *MR* in (*a*).

Source: Table 9-3.

that they are different. In addition, Figure 9-3(*a*) plots the demand (*AR*) curve and the marginal revenue (*MR*) curve. Scrutinize Figure 9-3(*a*) to see that the plotted green steps of *MR* definitely lie below the blue *dd* curve of *AR*. In fact, *MR* turns negative when *AR* is halfway down toward zero.

Elasticity and Marginal Revenue

What is the relationship between the price elasticity of demand and marginal revenue? Marginal revenue is positive when demand is elastic, zero when demand is unit-elastic, and negative when demand is inelastic.

This result is an important implication of the definition of elasticity that we used in Chapter 4. Recall that demand is elastic when a price decrease leads to a revenue increase. In such a situation, a price decrease raises output demanded so much that revenues rise, so marginal revenue is positive. For example, in Table 9-3, as price falls in the elastic region from $P = \$180$ to $P = \$100$, output demanded rises sufficiently to raise total revenue, and marginal revenue is positive.

What happens when demand is unit-elastic? A percentage price cut then just matches the percentage output increase, and marginal revenue is therefore zero. Can you see why marginal revenue is always negative in the inelastic range? Why is the marginal revenue for the perfect competitor's infinitely elastic demand curve always positive?

Table 9-4 shows the important elasticity relationships. Make sure you understand them and can apply them.

If demand is	Relation of q and P	Effect of q on TR	Value of marginal revenue (MR)
Elastic ($E_D > 1$)	% change q > % change P	Higher q raises TR	$MR > 0$
Unit-elastic ($E_D = 1$)	% change q = % change P	Higher q leaves TR unchanged	$MR = 0$
Inelastic ($E_D < 1$)	% change q < % change P	Higher q lowers TR	$MR < 0$

TABLE 9-4. Relationships of Demand Elasticity, Output, Price, Revenue, and Marginal Revenue

Here are the key points to remember:

1. *Marginal revenue* (MR) is the change in revenue that is generated by an additional unit of sales.
2. Price = average revenue ($P = AR$).
3. With downward-sloping demand,
   $P > MR$
   = P − reduced revenue on all previous units.
4. Marginal revenue is positive when demand is elastic, zero when demand is unit-elastic, and negative when demand is inelastic.
5. For perfect competitors, $P = MR = AR$.

PROFIT-MAXIMIZING CONDITIONS

Now return to the question of how a monopolist should set its quantity and price if it wants to maximize profits. By definition, total profit equals total revenue minus total costs; in symbols, $TP = TR - TC = (P \times q) - TC$. We will show that *maximum profit will occur when output is at that level where the firm's marginal revenue is equal to its marginal cost.*

One way to determine this maximum-profit condition is by using a table of costs and revenues, such as Table 9-5. To find the profit-maximizing quantity and price, compute total profit in column (5). This column tells us that the monopolist's best quantity, which is 4 units, requires a price of $120 per unit. This produces a total revenue of $480, and, after subtracting total costs of $250, we calculate total profit to be $230. A glance shows that no other price-output combination has as high a level of total profit.

We get more insight using a second approach, which is to compare marginal revenue in column (6) with marginal cost in column (7). As long as each additional unit of output provides more revenue than it costs, the firm's profit will increase as output increases. So the firm should continue to increase its output as long as MR is greater than MC.

On the other hand, suppose that MR is less than MC at a given output. This means that increasing output lowers profits, so the firm should cut back on output. Clearly, the best-profit point comes where marginal revenue exactly equals marginal cost. The rule for finding maximum profit is therefore:

The maximum-profit price (P^*) and quantity (q^*) of a monopolist come where the firm's marginal revenue equals its marginal cost:

$MR = MC$, at the maximum-profit P^* and q^*

These examples show the logic of the $MC = MR$ rule for maximizing profits, but we always want to understand the intuition behind the rules. Look for a moment at Table 9-5 and suppose that the monopolist is producing $q = 2$. At that point, its MR for producing 1 full additional unit is +$100, while its MC is $20. Thus, if it produced 1 additional unit, the firm would make additional profits of $MR - MC = \$100 - \$20 = \$80$. Indeed, column (5) of Table 9-5 shows that the extra profit gained by moving from 2 to 3 units is exactly $80.

Thus, when MR exceeds MC, additional profits can be made by increasing output; when MC exceeds MR, additional profits can be made by decreasing q. Only when $MR = MC$ can the firm maximize profits, because there are no additional profits to be made by changing its output level.

Monopoly Equilibrium in Graphs

Figure 9-4 shows the monopoly equilibrium. Part (*a*) combines the firm's cost and revenue curves. The maximum-profit point comes at that output where MC equals MR, which is given at their intersection at E. The monopoly equilibrium, or maximum-profit point, is at an output of $q^* = 4$. To find the profit-maximizing price, we run vertically up from E to the

Summary of Firm's Maximum Profit

(1) Quantity q	(2) Price P ($)	(3) Total revenue TR ($)	(4) Total cost TC ($)	(5) Total profit TP ($)	(6) Marginal revenue MR ($)	(7) Marginal cost MC ($)	
0	200	0	145	−145			
					+180	30	$MR > MC$
1	180	180	175	+5			
					+140	25	
2	160	320	200	+120			
					+100	20	
3	140	420	220	+200			
					+60	30	
4*	**120***	**480**	**250**	**+230**	+40	40	$MR = MC$
					+20	50	
5	100	500	300	+200			
					−20	70	
6	80	480	370	+110			
					−60	90	
7	60	420	460	−40			
					−100	110	$MR < MC$
8	40	320	570	−250			

*Maximum-profit equilibrium.

TABLE 9-5. Equating Marginal Cost to Marginal Revenue Gives Firm's Maximum-Profit q and P

Total and marginal costs of production are now brought together with total and marginal revenues. The maximum-profit condition is where $MR = MC$, with $q^* = 4$, $P^* = \$120$, and maximum $TP = \$230 = (\$120 \times 4) - \$250$.

dd curve at *G*, where $P^* = \$120$. The fact that average revenue at *G* lies above average cost at *F* guarantees a positive profit. The actual amount of profit is given by the green area in Figure 9-4(*a*).

The same story is told in part (*b*) with curves of total revenue, cost, and profit. Total revenue is dome-shaped. Total cost is ever rising. The vertical difference between them is total profit, which begins negative and ends negative. In between, *TP* is positive, reaching its maximum of \$230 at $q^* = 4$.

We add one further important geometric point. *The slope of a total value is a marginal value.* (You can refresh your memory on this by looking at page 22 in Chapter 1's appendix.) So look at point *G* in Figure 9-4(*b*). If you carefully calculate the slope at that point, you will see that it is \$40 per unit. This means that every unit of additional output produces \$40 of additional revenue, which is just the definition of *MR*. So the slope of the *TR* curve is *MR*. Similarly, the slope of the *TC* curve is *MC*. Note that at $q = 4$, *MC* is also \$40 per unit. At $q = 4$, marginal cost and marginal revenue are equal. At that point total profit (*TP*) reaches its maximum, and an additional unit adds exactly equal amounts to costs and revenues.

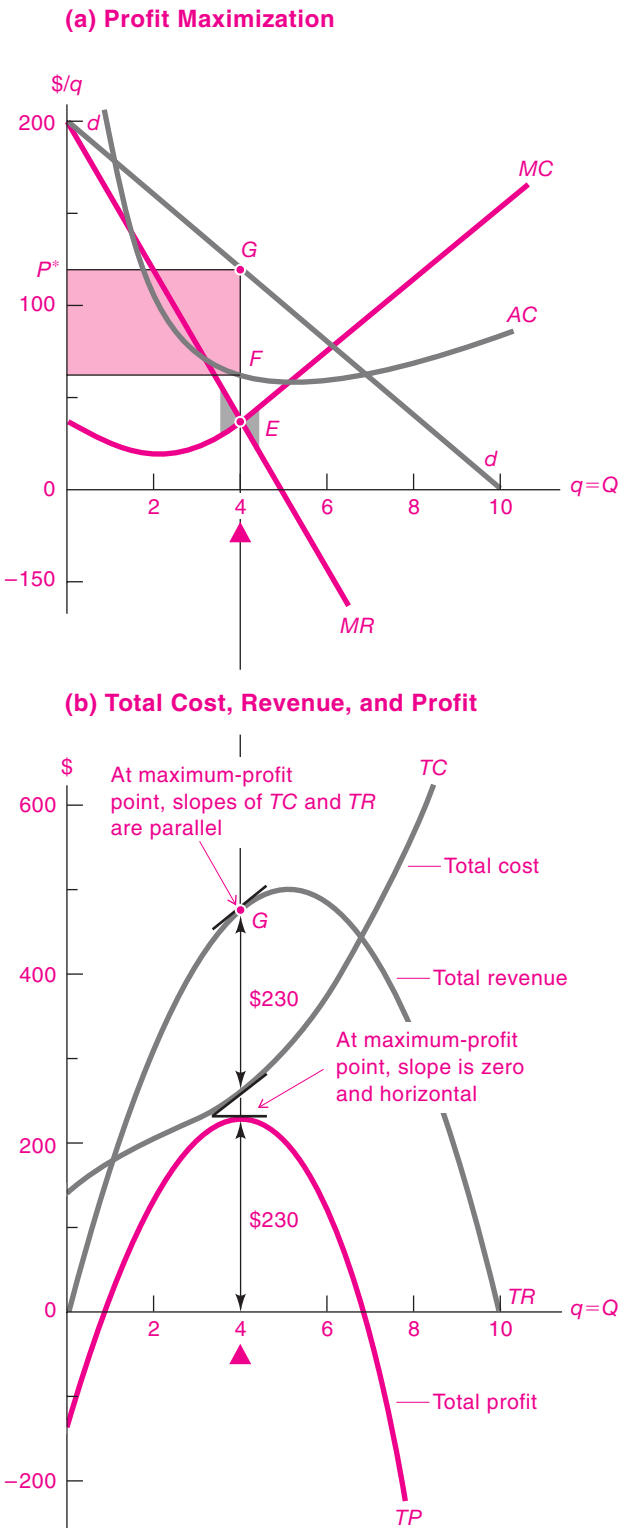


FIGURE 9-4. Profit-Maximizing Equilibrium Can Be Shown Using Either Total or Marginal Curves

(a) At *E*, where *MC* intersects *MR*, the monopolist gets maximum profits. Price is on the demand curve at *G*, above *E*. Since *P* is above *AC*, the maximized profit is a positive profit. (Can you explain why the blue triangles of shading on either side of *E* show the reduction in total profit that would come from a departure from *MR* = *MC*?)

Panel **(b)** tells the same story of maximizing profit as does **(a),** but it uses total concepts rather than marginal concepts. The *TR* curve shows the total revenue, while the *TC* curve shows total cost. Total profit is equal to *TR* minus *TC*, shown geometrically as the vertical distance from *TR* to *TC*. The slope of each curve is that curve's marginal value (e.g., *MR* is the slope of *TR*). At the maximum profit, *TR* and *TC* are parallel and therefore have equal slopes, *MR* = *MC*.

At the maximum-profit output, the blue slopes of *TR* and *TC* (which are *MR* and *MC*) are parallel and therefore equal.

A monopolist will maximize its profits by setting output at the level where *MC* = *MR*. Because the monopolist has a downward-sloping demand curve, this means that *P* > *MR*. Because price is above marginal cost for a profit-maximizing monopolist, the monopolist reduces output below the level that would be found in a perfectly competitive industry.

Perfect Competition as a Polar Case of Imperfect Competition

Although we have applied the *MC* = *MR* rule to monopolists that desire to maximize profits, this rule is actually applicable far beyond the present analysis. A little thought shows that the *MC* = *MR* rule applies with equal validity to a profit-maximizing perfect competitor. We can see this in two steps:

1. MR *for a perfect competitor.* What is *MR* for a perfect competitor? For a perfect competitor, the sale of extra units will never depress price, and the "lost revenue on all previous *q*" is therefore equal to zero. Price and marginal revenue are identical for perfect competitors.

   Under perfect competition, price equals average revenue equals marginal revenue (*P* = *AR* = *MR*). A perfect competitor's *dd* curve and its *MR* curve coincide as horizontal lines.

2. MR = P = MC *for a perfect competitor.* In addition, we can see that the logic of profit maximization for monopolists applies equally well to perfect competitors, but the result is a little different. Economic logic shows that profits are maximized at that output level where *MC* equals *MR*. But by step 1 above, for a perfect competitor, *MR* equals *P*. Therefore, the *MR* = *MC* profit-maximization condition becomes the special case of *P* = *MC* that we derived in the last chapter for a perfect competitor:

   Because a perfect competitor can sell all it wants at the market price, *MR* = *P* = *MC* at the maximum-profit level of output.

You can see this result visually by redrawing Figure 9-4(*a*). If the graph applied to a perfect competitor, the *dd* curve would be horizontal at the market price, and it would coincide with the *MR* curve. The profit-maximizing *MR* = *MC* intersection would also come at *P* = *MC*. We see then how the general rule for profit maximization applies to perfect as well as imperfect competitors.

THE MARGINAL PRINCIPLE: LET BYGONES BE BYGONES

We close this chapter with a more general point about the use of marginal analysis in economics. While economic theory will not necessarily make you fabulously wealthy, it does introduce you to some new ways of thinking about costs and benefits. *One of the most important lessons of economics is that you should look at the marginal costs and marginal benefits of decisions and ignore past or sunk costs.* We might put this as follows:

> Let bygones be bygones. Don't look backward. Don't cry over spilt milk or moan about yesterday's losses. Make a hard-headed calculation of the extra costs you'll incur by any decision, and weigh these against its extra advantages. Make a decision based on marginal costs and marginal benefits.

This is the **marginal principle,** which means that people will maximize their incomes or profits or satisfactions by counting only the marginal costs and marginal benefits of a decision. There are countless situations in which the marginal principle applies. We have just seen that the marginal principle of equating marginal cost and marginal revenue is the rule for profit maximization by firms.

Loss Aversion and the Marginal Principle

An interesting application is the behavior of people who are selling their houses. Behavioral economists have observed that people often resist selling their house for less than the dollar purchase price even in the face of steep declines in local housing prices.

For example, suppose you bought your house in San Jose for $250,000 in 2005 and wanted to sell it in 2008. Because of the decline in housing prices, comparable houses sold for $200,000 in 2008. As was the case for millions of people in the last few years, you are faced with a nominal dollar loss.

Studies show that you might well set the price at your purchase price of $250,000 and wait for several months without a single serious offer. This is what behavioral economists call "loss aversion," meaning that people resist taking a loss even though it is costly to hold on to an asset. This behavior has been verified in housing markets, where people subject to a loss set higher asking prices and wait longer for sales.

Economists counsel against this kind of behavior. It would be better to observe the marginal principle. Forget about what you paid for your house. Just get the best price you can.

Monopolists of the Gilded Age

Economic abstractions sometimes hide the human drama of monopoly, so we close this section by recounting one of the most colorful periods of American business history. Because of changing laws and customs, monopolists in today's America bear little resemblance to the brilliant, unscrupulous, and often dishonest robber barons of the Gilded Age (1870–1914). Legendary figures like Rockefeller, Gould, Vanderbilt, Frick, Carnegie, Rothschild, and Morgan were driven to create entire industries like railroads or oil, provide their finance, develop the western frontier, destroy their competitors, and pass on fabulous fortunes to their heirs.

The last three decades of nineteenth-century America experienced robust economic growth lubricated by tremendous graft and corruption. Daniel Drew was a cattle rustler, horse trader, and railroader who mastered the trick of "watering the stock." This practice involved depriving his cattle of water until they reached the slaughterhouse; he then induced a great thirst with salt and allowed the beasts to engorge themselves on water just before being weighed. Later, tycoons would "water their stock" by inflating the value of their securities.

The railroaders of the American frontier west were among the most unscrupulous entrepreneurs on record. The transcontinental railroads were funded with vast federal land grants, aided by bribes and stock gifts to numerous members of Congress and the cabinet. Shortly after the Civil War, the wily railroader Jay Gould attempted to corner the entire gold supply of the United States, and with it the nation's money supply. Gould later promoted his railroad by describing the route of his northern line—snowbound much of the year—as a tropical paradise, filled with orange groves, banana plantations, and monkeys. By century's end, all the bribes, land grants, watered stock, and fantastic promises had led to the greatest rail system in the world.

The story of John D. Rockefeller epitomizes the nineteenth-century monopolist. Rockefeller saw visions of riches in the fledgling oil industry and began to organize oil refineries. He was a meticulous manager and sought to bring "order" to the quarrelsome wildcatters. He bought up competitors and consolidated his hold on the industry by persuading the railroads to give him deep and secret rebates and supply information about his competitors. When competitors stepped out of line, Rockefeller's railroads refused to ship their oil and even dumped it on the ground. By 1878, John D. controlled 95 percent of the pipelines and oil refineries in the United States. Prices were raised and stabilized, ruinous competition was ended, and monopoly was achieved.

Rockefeller devised an ingenious new device to ensure control over his alliance. This was the "trust," in which the stockholders turned their shares over to "trustees" who would then manage the industry to maximize its profits. Other industries imitated the Standard Oil Trust, and soon trusts were set up in kerosene, sugar, whiskey, lead, salt, and steel.

These practices so upset agrarians and populists that the nation soon passed antitrust laws (see Chapter 10). In 1910, the Standard Oil Corporation was dissolved in the first great victory by the Progressives against "Big Business." Ironically, Rockefeller actually profited from the breakup because the price of Standard Oil shares soared when they were offered to the public.

Great monopolies produced great wealth. Whereas the United States had three millionaires in 1861, there were 4000 of them by 1900 ($1 million at the turn of the century is equivalent to about $100 million in today's dollars).

Great wealth in turn begot conspicuous consumption (a term introduced into economics by Thorstein Veblen in *The Theory of the Leisure Class,* 1899). Like European popes and aristocrats of an earlier era, American tycoons wanted to transform their fortunes into lasting monuments. The wealth was spent in constructing princely palaces such as the "Marble House," which can still be seen in Newport, Rhode Island; in buying vast art collections, which form the core of the great American museums like New York's Metropolitan Museum of Art; and in launching foundations and universities such as those named after Stanford, Carnegie, Mellon, and Rockefeller. Long after their private monopolies were broken up by the government or overtaken by competitors, and long after their wealth was largely dissipated by heirs and overtaken by later generations of entrepreneurs, the philanthropic legacy of the robber barons continues to shape American arts, science, and education.[1]

[1] See the Further Reading section for books on this topic.

SUMMARY

A. Patterns of Imperfect Competition

1. Most market structures today fall somewhere on a spectrum between perfect competition and pure monopoly. Under imperfect competition, a firm has some control over its price, a fact seen as a downward-sloping demand curve for the firm's output.
2. Important kinds of market structures are (*a*) monopoly, where a single firm produces all the output in a given industry; (*b*) oligopoly, where a few sellers of a similar or differentiated product supply the industry; (*c*) monopolistic competition, where a large number of small firms supply related but somewhat differentiated products; and (*d*) perfect competition, where a large number of small firms supply an identical product. In the first three cases, firms in the industry face downward-sloping demand curves.
3. Economies of scale, or decreasing average costs, are the major source of imperfect competition. When

firms can lower costs by expanding their output, perfect competition is destroyed because a few companies can produce the industry's output most efficiently. When the minimum efficient size of plants is large relative to the national or regional market, cost conditions produce imperfect competition.

4. In addition to declining costs, other forces leading to imperfect competition are barriers to entry in the form of legal restrictions (such as patents or government regulation), high entry costs, advertising, and product differentiation.

B. Monopoly Behavior

5. We can easily derive a firm's total revenue curve from its demand curve. From the schedule or curve of total revenue, we can then derive marginal revenue, which denotes the change in revenue resulting from an additional unit of sales. For the imperfect competitor, marginal revenue is less than price because of the lost revenue on all previous units of output that will result when the firm is forced to drop its price in order to sell an extra unit of output. That is, with demand sloping downward,

$$P = AR > MR = P - \text{lost revenue on all previous } q$$

6. Recall Table 9-4's rules relating demand elasticity, price and quantity, total revenue, and marginal revenue.
7. A monopolist will find its maximum-profit position where $MR = MC$, that is, where the last unit it sells brings in extra revenue just equal to its extra cost. This same $MR = MC$ result can be shown graphically by the intersection of the MR and MC curves or by the equality of the slopes of the total revenue and total cost curves. In any case, *marginal revenue = marginal cost* must always hold at the equilibrium position of maximum profit.
8. For perfect competitors, marginal revenue equals price. Therefore, the profit-maximizing output for a perfect competitor comes where $MC = P$.
9. Economic reasoning leads to the important *marginal principle*. In making decisions, count marginal future advantages and disadvantages, and disregard sunk costs that have already been paid. Be wary of loss aversion.

CONCEPTS FOR REVIEW

Patterns of Imperfect Competition

perfect vs. imperfect competition
monopoly, oligopoly, monopolistic competition
product differentiation
barriers to entry (government and economic)

Marginal Revenue and Monopoly

marginal (or extra) revenue, *MR*
$MR = MC$ as the condition for maximizing profits
$MR = P$, $P = MC$, for perfect competitors
natural monopoly
the marginal principle

FURTHER READING AND INTERNET WEBSITES

Further Reading

The theory of monopoly was developed by Alfred Marshall around 1890; see his *Principles of Economics,* 9th ed. (Macmillan, New York, 1961).

An excellent review of monopoly and industrial organization is F. M. Scherer and David Ross, *Industrial Market Structure and Economic Performance,* 3rd ed. (Houghton Mifflin, Boston, 1990).

The Gilded Age period gave birth to "yellow journalism" in the United States and fostered many muckraking histories, such as Matthew Josephson, *The Robber Barons* (Harcourt Brace, New York, 1934). A more balanced recent account is Ron Chernow, *Titan: The Life of John D. Rockefeller, Sr.* (Random House, New York, 1998).

For a study of loss aversion in the housing market, see David Genesove and Christopher Mayer, "Loss Aversion and Seller Behavior: Evidence from the Housing Market," *Quarterly Journal of Economics,* 2001. The foundation of this theory is in Amos Tversky and Daniel Kahneman, "Loss Aversion in Riskless Choice: A Reference-Dependent Model," *Quarterly Journal of Economics,* 1991.

Websites

An important legal case over the last decade has concerned whether Microsoft had a monopoly on PC operating systems. This is thoroughly discussed in the "Findings of Fact" of the Microsoft antitrust case by Judge Thomas Penfield Jackson (November 5, 1999). His opinion and further developments can be found at *www.microsoft.com/presspass/legalnews.asp.*

QUESTIONS FOR DISCUSSION

1. Suppose a monopolist owns a mineral spring. Answer and demonstrate each of the following:
   a. Assume that the cost of production is zero. What is the elasticity of demand at the profit-maximizing quantity?
   b. Assume that the *MC* of production is always \$1 per unit. What is the elasticity of demand at the profit-maximizing quantity?
2. Explain why each of the following statements is false. For each, write the correct statement.
   a. A monopolist maximizes profits when $MC = P$.
   b. The higher the price elasticity, the higher is a monopolist's price above its *MC*.
   c. Monopolists ignore the marginal principle.
   d. Monopolists will maximize sales. They will therefore produce more than perfect competitors and their price will be lower.
3. What is *MR*'s numerical value when *dd* has unitary elasticity? Explain.
4. In his opinion on the Microsoft antitrust case, Judge Jackson wrote: "[T]hree main facts indicate that Microsoft enjoys monopoly power. First, Microsoft's share of the market for Intel-compatible PC operating systems is extremely large and stable. Second, Microsoft's dominant market share is protected by a high barrier to entry. Third, and largely as a result of that barrier, Microsoft's customers lack a commercially viable alternative to Windows." (See the website reference, section 34, in this chapter's Further Readings.) Why are these elements related to monopoly? Are all three necessary? If not, which ones are crucial? Explain your reasoning.
5. Estimate the numerical price elasticities of demand at points *A* and *B* in Figure 9-1. (*Hint:* You may want to review the rule for calculating elasticities in Figure 4-5.)
6. Redraw Figure 9-4(*a*) for a perfect competitor. Why is *dd* horizontal? Explain why the horizontal *dd* curve coincides with *MR*. Then proceed to find the profit-maximizing *MR* and *MC* intersection. Why does this yield the competitive condition $MC = P$? Now redraw Figure 9-4(*b*) for a perfect competitor. Show that the slopes of *TR* and *TC* must still match at the maximum-profit equilibrium point for a perfect competitor.
7. Banana Computer Company has fixed costs of production of \$100,000, while each unit costs \$600 of labor and \$400 of materials and fuel. At a price of \$3000, consumers would buy no Banana computers, but for each \$10 reduction in price, sales of Banana computers increase by 1000 units. Calculate marginal cost and marginal revenue for Banana Computer, and determine its monopoly price and quantity.
8. Show that a profit-maximizing monopolist will never operate in the price-inelastic region of its demand curve.
9. Explain the error in the following statement: "A firm out to maximize its profits will always charge the highest price that the traffic will bear." State the correct result, and use the concept of marginal revenue to explain the difference between the correct and the erroneous statements.
10. Recall from pp. 183–184 how trusts were organized to monopolize industries like oil and steel. Explain the saying, "The tariff is the mother of trusts." Use Figure 9-2 to illustrate your analysis. Use the same diagram to explain why lowering tariffs and other trade barriers reduces monopoly power.
11. *For students who like calculus:* You can show the condition for profit maximization easily using calculus. Define $TP(q)$ = total profits, $TC(q)$ = total costs, and $TR(q)$ = total revenues. Marginal this-or-that is the derivative of this-or-that with respect to output, so $dTR/dq = TR'(q) = MR$ = marginal revenue.
    a. Explain why $TP = TR - TC$.
    b. Show that a maximum of the profit function comes where $TC'(q) = TR'(q)$. Interpret this finding.

CHAPTER

10

Competition among the Few

Look at the airline price wars of 1992. When American Airlines, Northwest Airlines, and other U.S. carriers went toe-to-toe in matching and exceeding one another's reduced fares, the result was record volumes of air travel—and record losses. Some estimates suggest that the overall losses suffered by the industry that year exceed the combined profits for the entire industry from its inception.

Akshay R. Rao, Mark E. Bergen, and Scott Davis
"How to Fight a Price War"

Earlier chapters analyzed the market structures of perfect competition and complete monopoly. If you look out the window at the American economy, however, you will see that such polar cases are rare. Most industries lie between these two extremes and are populated by a small number of firms competing with each other.

What are the key features of these intermediate types of imperfect competitors? How do they set their prices and outputs? To answer these questions, we look closely at what happens under oligopoly and monopolistic competition, paying special attention to the role of concentration and strategic interaction. We then introduce the elements of game theory, which is an important tool for understanding how people and businesses interact in strategic situations. The final section reviews the different public policies used to combat monopolistic abuses, focusing on regulation and antitrust laws.

A. BEHAVIOR OF IMPERFECT COMPETITORS

Look back at Table 9-1, which shows the following kinds of market structures: (1) *Perfect competition* is found when a large number of firms produce an identical product. (2) *Monopolistic competition* occurs when a large number of firms produce slightly differentiated products. (3) *Oligopoly* is an intermediate form of imperfect competition in which an industry is dominated by a few firms. (4) *Monopoly* is the most concentrated market structure, in which a single firm produces the entire output of an industry.

How do we measure the power of firms in an industry to control price and output? How do the different species behave? We begin with these issues.

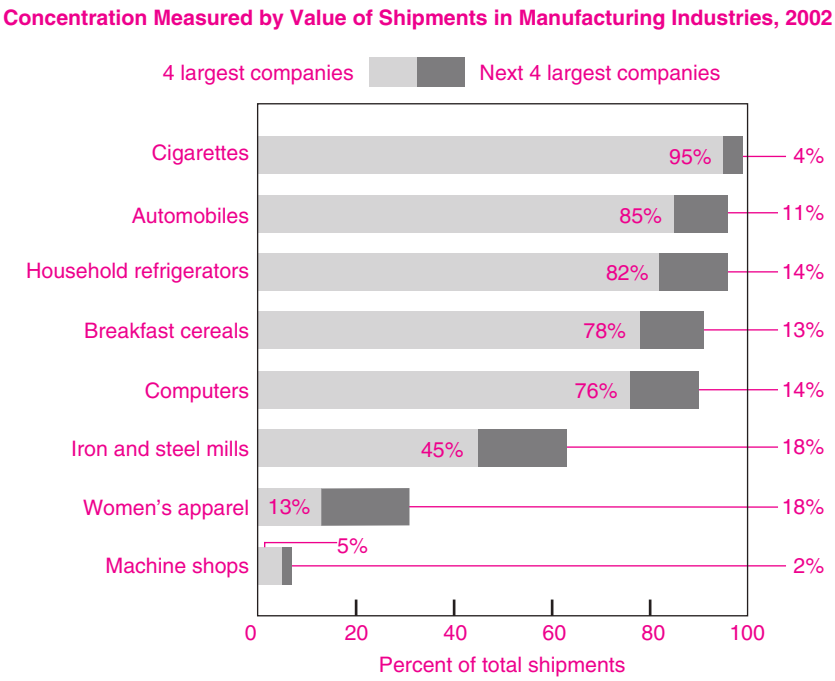


FIGURE 10-1. Concentration Ratios Are Quantitative Measures of Market Power

For refrigerators, automobiles, and many other industries, a few firms produce most of the domestic output. Compare this with the ideal of perfect competition, in which each firm is too small to affect the market price.

Source: U.S. Bureau of the Census, 2002 data.

Measures of Market Power

In many situations—such as deciding whether the government should intervene in a market or whether a firm has abused its monopoly position—economists need a quantitative measure of the extent of a firm's market power. **Market power** signifies the degree of control that a single firm or a small number of firms have over the price and production decisions in an industry.

The most common measure of market power is the *concentration ratio* for an industry, illustrated in Figure 10-1. The **four-firm concentration ratio** measures the fraction of the market or industry accounted for by the four largest firms. Similarly, the eight-firm concentration ratio is the percent of the market taken by the top eight firms. The market is customarily measured by domestic sales, shipments, or output. In a pure monopoly, the four-firm and eight-firm concentration ratios would be 100 percent because one firm produces 100 percent of the output; under perfect competition, both ratios would be close to zero because even the largest firms produce only a tiny fraction of industry output.

Many economists believe that traditional concentration ratios do not adequately measure market power. An alternative, which better captures the role of dominant firms, is the **Herfindahl-Hirschman Index (HHI).** This is calculated by summing the squares of each participant's market share. Perfect competition would have an HHI of near zero because each firm produces only a small percentage of the total output, while complete monopoly would have an HHI of 10,000 because one firm produces 100 percent of the output ($100^2 = 10{,}000$). (For the formula and an example, see question 2 at the end of this chapter.)

Warning on Concentration Measures

Although concentration measures are widely used, they are often misleading because of international competition and competition from closely related industries. Conventional concentration measures such as those shown in Figure 10-1 exclude imports and include only domestic production. Because foreign

competition is very intense in the manufacturing sector, the actual market power of domestic firms is much smaller than is indicated by measures of market power based solely on domestic production. For example, the conventional concentration measures shown in Figure 10-1 indicate that the top four U.S. automotive firms had 85 percent of the U.S. market. If we include imports as well, however, these top four U.S. firms had only 43 percent of the U.S. market.

In addition to ignoring international competition, traditional concentration measures ignore the impact of competition from other, related industries. For example, concentration ratios have historically been calculated for a narrow industry definition, such as "wired telecommunications carriers." Sometimes, however, strong competition comes from other quarters. For example, cellular telephones are a major threat to traditional wired telephone service even though the two are produced by different industries. Even though the four-firm concentration ratio for wired carriers alone is 60 percent, the four-firm ratio for all telecommunications carriers is only 46 percent, so the definition of a market can strongly influence the calculation of the concentration ratios.

In the end, some measure of market power is essential for many legal purposes, such as aspects of antitrust law, examined later in this chapter. A careful delineation of the market to include all the relevant competitors can be helpful in determining whether monopolistic abuses are in fact a real threat.

THE NATURE OF IMPERFECT COMPETITION

In analyzing the determinants of concentration, economists have found that three major factors are at work in imperfectly competitive markets. These factors are economies of scale, barriers to entry, and strategic interaction (the first two were analyzed in the previous chapter, and the third is the subject of detailed examination in the next section):

- *Costs.* When the minimum efficient size of operation for a firm occurs at a sizable fraction of industry output, only a few firms can profitably survive and oligopoly is likely to result.
- *Barriers to competition.* When there are large economies of scale or government restrictions to entry, these will limit the number of competitors in an industry.
- *Strategic interaction.* When only a few firms operate in a market, they will soon recognize their interdependence. **Strategic interaction,** which is a genuinely new feature of oligopoly that has inspired the field of game theory, occurs when each firm's business depends upon the behavior of its rivals.

Why are economists particularly concerned about industries characterized by imperfect competition? The answer is that such industries behave in certain ways that are inimical to the public interest. For example, imperfect competition generally leads to prices that are above marginal costs. Sometimes, without the spur of competition, the quality of service deteriorates. Both high prices and poor quality are undesirable outcomes.

As a result of high prices, oligopolistic industries often (but not always) have supernormal profits. The profitability of the highly concentrated tobacco and pharmaceutical industries has been the target of political attacks on numerous occasions. Careful studies show, however, that concentrated industries tend to have only slightly higher rates of profit than unconcentrated ones.

Historically, one of the major defenses of imperfect competition has been that large firms are responsible for most of the research and development (R&D) and innovation in a modern economy. There is certainly some truth in this idea, for highly concentrated industries sometimes have high levels of R&D spending per dollar of sales as they try to achieve a technological edge over their rivals. At the same time, individuals and small firms have produced many of the greatest technological breakthroughs. We review the economics of innovation in Chapter 11.

THEORIES OF IMPERFECT COMPETITION

While the concentration of an industry is important, it does not tell the whole story. Indeed, to explain the behavior of imperfect competitors, economists have developed a field called *industrial organization*. We cannot cover this vast area here. Instead, we will focus on three of the most important cases of imperfect competition—collusive oligopoly, monopolistic competition, and small-number oligopoly.

Collusive Oligopoly

The degree of imperfect competition in a market is influenced not just by the number and size of firms but by their behavior. When only a few firms operate in a market, they see what their rivals are doing and react. For example, if there are two airlines operating along the same route and one raises its fare, the other must decide whether to match the increase or to stay with the lower fare, undercutting its rival. *Strategic interaction* is a term that describes how each firm's business strategy depends upon its rivals' business behavior.

When there are only a small number of firms in a market, they have a choice between *cooperative* and *noncooperative* behavior. Firms act noncooperatively when they act on their own without any explicit or implicit agreements with other firms. That's what produces price wars. Firms operate in a cooperative mode when they try to minimize competition. When firms in an oligopoly actively cooperate with each other, they engage in **collusion.** This term denotes a situation in which two or more firms jointly set their prices or outputs, divide the market among themselves, or make other business decisions jointly.

During the early years of American capitalism, before the passage of effective antitrust laws, oligopolists often merged or formed a trust or cartel (recall Chapter 9's discussion of trusts, page 184). A **cartel** is an organization of independent firms, producing similar products, that work together to raise prices and restrict output. Today, with only a few exceptions, it is strictly illegal in the United States and most other market economies for companies to collude by jointly setting prices or dividing markets.

Nonetheless, firms are often tempted to engage in tacit collusion, which occurs when they refrain from competition without explicit agreements. When firms tacitly collude, they often quote identical high prices, pushing up profits and decreasing the risk of doing business. In recent years, sellers of online music, diamonds, and kosher Passover products have been investigated for price fixing, while private universities, art dealers, airlines, and the telephone industry have been accused of collusive behavior.

The rewards for successful collusion can be great. Consider an industry where four firms have tired of ruinous price wars. They agree to charge the same price and share the market. They form a **collusive oligopoly** and set a price which maximizes their joint profits. By joining together, the four firms in effect become a monopolist.

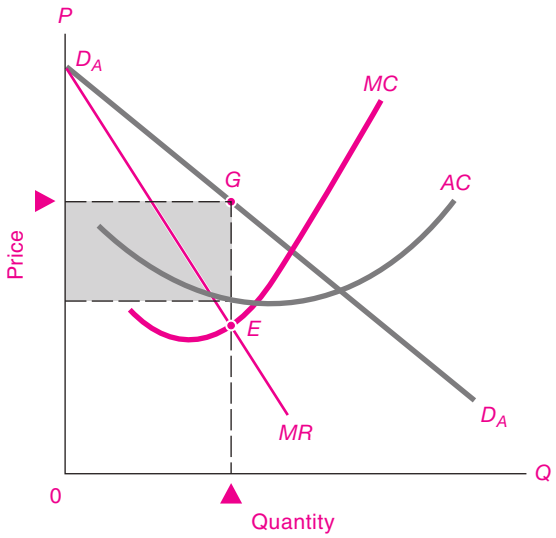


FIGURE 10-2. Collusive Oligopoly Looks Much Like Monopoly

After experience with disastrous price wars, firms will surely recognize that each price cut is canceled by competitors' price cuts. So oligopolist A may estimate its demand curve D_AD_A by assuming that others will be charging similar prices. When firms collude to set a jointly profit-maximizing price, the price will be very close to that of a single monopolist. Can you see why profits are equal to the blue rectangle?

Figure 10-2 illustrates oligopolist A's situation, where there are four firms with identical cost and demand curves. We have drawn A's demand curve, D_AD_A, assuming that the other three firms always charge the same price as firm A.

The maximum-profit equilibrium for the collusive oligopolist is shown in Figure 10-2 at point *E*, the intersection of the firm's *MC* and *MR* curves. Here, the appropriate demand curve is D_AD_A. The optimal price for the collusive oligopolist is shown at point *G* on D_AD_A, above point *E*. This price is identical to the monopoly price: it is above marginal cost and earns each of the colluding oligopolists a handsome monopoly profit.

When oligopolists collude to maximize their joint profits, taking into account their mutual

interdependence, they will produce the monopoly output and price and earn the monopoly profit.

Although many oligopolists would be delighted to earn such high profits, in reality many obstacles hinder effective collusion. First, collusion is illegal. Second, firms may "cheat" on the agreement by cutting their price to selected customers, thereby increasing their market share. Clandestine price cutting is particularly likely in markets where prices are secret, where goods are differentiated, where there is more than a handful of firms, or where the technology is changing rapidly. Third, the growth of international trade means that many companies face intensive competition from foreign firms as well as from domestic companies.

Indeed, experience shows that running a successful cartel is a difficult business, whether the collusion is explicit or tacit.

A long-running thriller in this area is the story of the international oil cartel known as the Organization of Petroleum Exporting Countries, or OPEC. OPEC is an international organization which sets production quotas for its members, which include Saudi Arabia, Iran, and Algeria. Its stated goal is "to secure fair and stable prices for petroleum producers; an efficient, economic and regular supply of petroleum to consuming nations; and a fair return on capital to those investing in the industry." Its critics claim it is really a collusive monopolist attempting to maximize the profits of producing countries.

OPEC became a household name in 1973, when it reduced production sharply and oil prices skyrocketed. But a successful cartel requires that members set a low production quota and maintain discipline. Every few years, price competition breaks out when some OPEC countries ignore their quotas. This happened in a spectacular way in 1986, when Saudi Arabia drove oil prices from $28 per barrel down to below $10.

Another problem faced by OPEC is that it must negotiate production quotas rather than prices. This leads to high levels of price volatility because demand is unpredictable and highly price-inelastic in the short run. Oil producers became rich in the 2000s as prices soared, but the cartel had little control over actual events.

The airline industry is another example of a market with a history of repeated—and failed—attempts at collusion. It would seem a natural candidate for collusion. There are only a few major airlines, and on many routes there are only one or two rivals. But just look back to the quote at the beginning of the chapter, which describes one of the recent price wars in the United States. Airline bankruptcy is so frequent that some airlines spend more time bankrupt than solvent. Indeed, the evidence shows that the only time an airline can charge supernormal fares is when it has a near-monopoly on all flights to a city.

Monopolistic Competition

At the other end of the spectrum from collusive oligopolies is **monopolistic competition.** Monopolistic competition resembles perfect competition in three ways: there are many buyers and sellers, entry and exit are easy, and firms take other firms' prices as given. The distinction is that products are identical under perfect competition, while under monopolistic competition they are differentiated.

Monopolistic competition is very common—just scan the shelves at any supermarket and you'll see a dizzying array of different brands of breakfast cereals, shampoos, and frozen foods. Within each product group, products or services are different, but close enough to compete with each other. Here are some other examples of monopolistic competition: There may be several grocery stores in a neighborhood, each carrying the same goods but at different locations. Gas stations, too, all sell the same product, but they compete on the basis of location and brand name. The several hundred magazines on a newsstand rack are monopolistic competitors, as are the 50 or so competing brands of personal computers. The list is endless.

The important point to recognize is that each seller has some freedom to raise or lower prices because of product differentiation (in contrast to perfect competition, where sellers are price-takers). Product differentiation leads to a downward slope in each seller's demand curve.

Figure 10-3 might represent a monopolistically competitive computer magazine which is in short-run equilibrium with a price at *G*. The firm's *dd* demand curve shows the relationship between sales and its price when other magazine prices are unchanged; its demand curve slopes downward since this magazine is a little different from everyone else's because

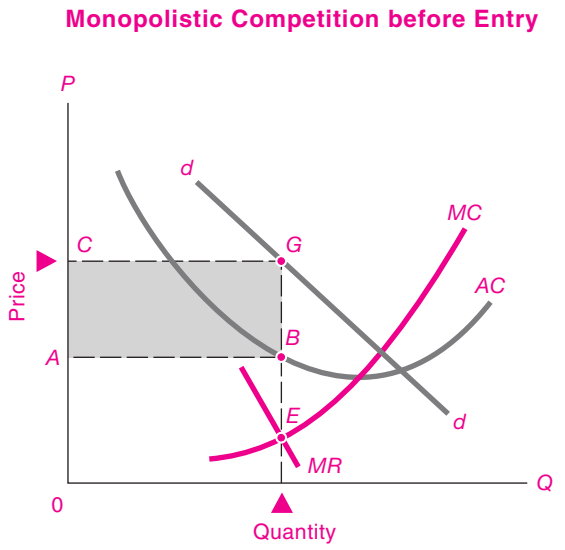


FIGURE 10-3. Monopolistic Competitors Produce Many Similar Goods

Under monopolistic competition, numerous small firms sell differentiated products and therefore have downward-sloping demand. Each firm takes its competitors' prices as given. Equilibrium has $MR = MC$ at E, and price is at G. Because price is above AC, the firm is earning a profit, area $ABGC$.

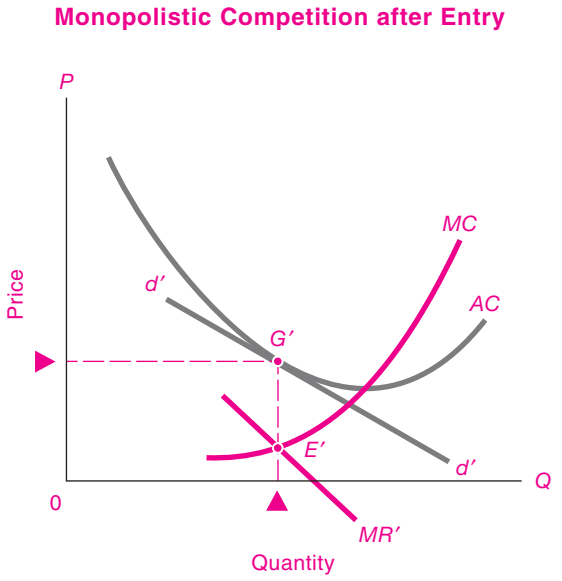


FIGURE 10-4. Free Entry of Numerous Monopolistic Competitors Wipes Out Profit

The typical seller's original profitable dd curve in Figure 10-3 will be shifted downward and leftward to $d'd'$ by the entry of new rivals. Entry ceases only when each seller has been forced into a long-run, no-profit tangency such as at G'. At long-run equilibrium, price remains above MC, and each producer is on the left-hand declining branch of its long-run AC curve.

of its special focus. The profit-maximizing price is at G. Because price at G is above average cost, the firm is making a handsome profit represented by area $ABGC$.

But our magazine has no monopoly on writers or newsprint or insights on computers. Firms can enter the industry by hiring an editor, having a bright new idea and logo, locating a printer, and hiring workers. Since the computer magazine industry is profitable, entrepreneurs bring new computer magazines into the market. With their introduction, the demand curve for the products of existing monopolistically competitive computer magazines shifts leftward as the new magazines nibble away at our magazine's market.

The ultimate outcome is that computer magazines will continue to enter the market until all economic profits (including the appropriate opportunity costs for owners' time, talent, and contributed capital) have been beaten down to zero. Figure 10-4 shows the final long-run equilibrium for the typical seller. In equilibrium, the demand is reduced or shifted to the left until the new $d'd'$ demand curve just touches (but never goes above) the firm's AC curve. Point G' is a long-run equilibrium for the industry because profits are zero and no one is tempted to enter or forced to exit the industry.

This analysis is well illustrated by the personal computer industry. Originally, such computer manufacturers as Apple and Compaq made big profits. But the personal computer industry turned out to have low barriers to entry, and numerous small firms entered the market. Today, there are dozens of firms, each with a small share of the computer market but no economic profits to show for its efforts.

The monopolistic competition model provides an important insight into American capitalism: The rate of profit will in the long run be zero in this kind of imperfectly competitive industry as firms enter with new differentiated products.

In the long-run equilibrium for monopolistic competition, prices are above marginal costs but economic profits have been driven down to zero.

Critics of capitalism argue that monopolistic competition is inherently inefficient. They point to an excessive number of trivially different products that lead to wasteful duplication and expense. To understand the reasoning, look back at the long-run equilibrium price at G' in Figure 10-4. At that point, price is above marginal cost and output is reduced below the ideal competitive level.

This economic critique of monopolistic competition has considerable appeal: It takes real ingenuity to demonstrate the gains to human welfare from adding Apple Cinnamon Cheerios to Honey Nut Cheerios and Whole Grain Cheerios. It is hard to see the reason for gasoline stations on every corner of an intersection.

But there is logic to the differentiated goods and services produced by a modern market economy. The great variety of products fills many niches in consumer tastes and needs. Reducing the number of monopolistic competitors might lower consumer welfare because it would reduce the diversity of available products. People will pay a premium to be free to choose among various options.

Rivalry among the Few

For our third example of imperfect competition, we turn back to markets in which only a few firms compete. This time, instead of focusing on collusion, we consider the fascinating case where firms have a strategic interaction with each other. Strategic interaction is found in any market which has relatively few competitors. Like a tennis player trying to outguess her opponent, each business must ask how its rivals will react to changes in key business decisions. If GE introduces a new model of refrigerator, what will Whirlpool, its principal rival, do? If American Airlines lowers its transcontinental fares, how will United react?

Consider as an example the market for air shuttle services between New York and Washington, currently served by Delta and US Airways. This market is called a **duopoly** because industry output is produced by only two firms. Suppose that Delta has determined that if it cuts fares 10 percent, its profits will rise as long as US Airways does not match its cut but its profits will fall if US Airways does match its price cut. If they cannot collude, Delta must make an educated guess as to how US Airways will respond to its price moves. Its best approach would be to estimate how US Airways would react to each of its actions and then to maximize profits *with strategic interaction recognized*. This analysis is the province of game theory, discussed in Section B of this chapter.

Similar strategic interactions are found in many large industries: in television, in automobiles, even in economics textbooks. Unlike the simple approaches of monopoly and perfect competition, it turns out that there is no simple theory to explain how oligopolists behave. Different cost and demand structures, different industries, even different managerial temperaments will lead to different strategic interactions and to different pricing strategies. Sometimes, the best behavior is to introduce some randomness into the response simply to keep the opposition off balance.

Competition among the few introduces a completely new feature into economic life: It forces firms to take into account competitors' reactions to price and output decisions and brings strategic considerations into their markets.

PRICE DISCRIMINATION

When firms have market power, they can sometimes increase their profits through price discrimination. **Price discrimination** occurs when the same product is sold to different consumers for different prices.

Consider the following example: You run a company selling a successful personal-finance program called MyMoney. Your marketing manager comes in and says:

> Look, boss. Our market research shows that our buyers fall into two categories: (1) our current customers, who are locked into MyMoney because they keep their financial records using our program, and (2) potential new buyers who have been using other programs. Why don't we raise our price, but give a rebate to new customers who are willing to switch from our competitors? I've run the numbers. If we raise our price from $20 to $30 but give a $15 rebate for people who have been using other financial programs, we will make a bundle.

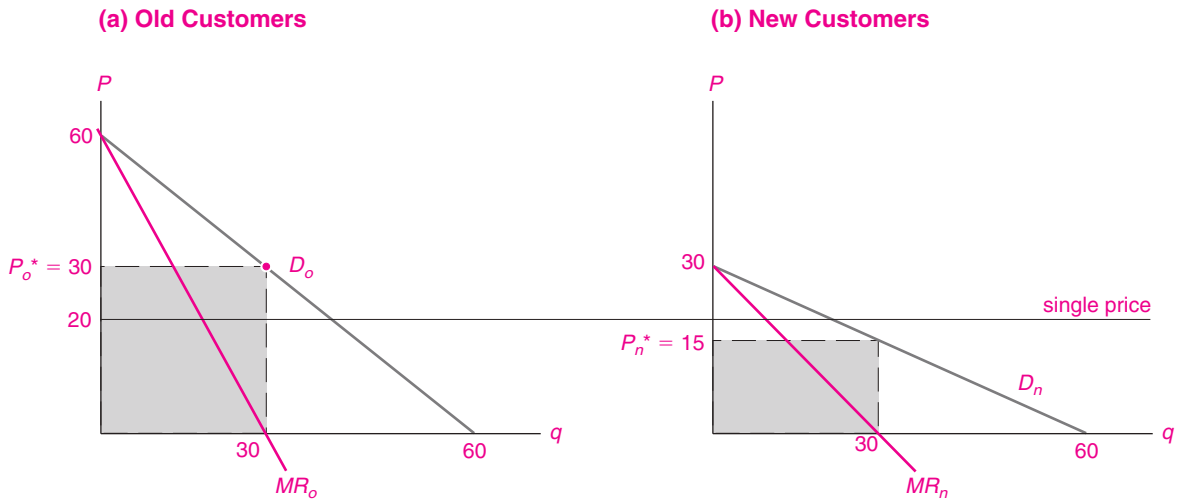


FIGURE 10-5. Firms Can Increase Their Profits through Price Discrimination

You are a profit-maximizing monopoly seller of computer software with zero marginal cost. Your market contains established customers in **(a)** and new customers in **(b)**. Old customers have more inelastic demand because of the high costs of switching to other programs.

If you must set a single price, you will maximize profits at a price of $20 and earn profits of $1200. But suppose you can segment your market between locked-in current users and reluctant new buyers. This would increase your profits to ($30 × 30) + ($15 × 30) = $1350.

You are intrigued by the suggestion. Your house economist constructs the demand curves in Figure 10-5. Her research indicates that your old customers have more price-inelastic demand than your potential new customers because new customers must pay substantial switching costs. If your rebate program works and you succeed in segmenting the market, the numbers show that your profits will rise from $1200 to $1350. (To make sure you understand the analysis, use the data shown in Figure 10-5 to estimate the monopoly price and profits if you set a single monopoly price and if you price-discriminate between the two markets.)

Price discrimination is widely used today, particularly with goods that are not easily transferred from the low-priced market to the high-priced market. Here are some examples:

- Identical textbooks are sold at lower prices in Europe than in the United States. What prevents wholesalers from purchasing large quantities abroad and undercutting the domestic market? A protectionist import quota prohibits the practice. However, as an individual, you might well reduce the costs of your books by buying them abroad through online bookstores.
- Airlines are the masters of price discrimination (review our discussion of "Elasticity Air" in Chapter 4). They segment the market by pricing tickets differently for those who travel in peak or off-peak times, for those who are business or pleasure travelers, and for those who are willing to stand by. This allows them to fill their planes without eroding revenues.
- Local utilities often use "two-part prices" (sometimes called nonlinear prices) to recover some of their overhead costs. If you look at your telephone or electricity bill, it will generally have a "connection" price and a "per-unit" price of service. Because connection is much more price-inelastic than the per-unit price, such two-part pricing allows sellers to lower their per-unit prices and increase the total quantity sold.
- Firms engaged in international trade often find that foreign demand is more elastic than domestic demand. They will therefore sell at

lower prices abroad than at home. This practice is called "dumping" and is sometimes banned under international-trade agreements.

- Sometimes a company will actually *degrade* its top-of-the-line product to make a less capable product, which it will then sell at a discounted price to capture a low-price market. For example, IBM inserted special commands to slow down its laser printer from 10 pages per minute to 5 pages per minute so that it could sell the slow model at a lower price without cutting into sales of its top model.

What are the economic effects of price discrimination? Surprisingly, they often improve economic welfare. To understand this point, recall that monopolists raise their price and lower their sales to increase profits. In doing so, they may capture the market for eager buyers but lose the market for reluctant buyers. By charging different prices for those willing to pay high prices (who get charged high prices) and those willing to pay only lower prices (who may sit in the middle seats or get a degraded product, but at a lower price), the monopolist can increase both its profits and consumer satisfactions.[1]

B. GAME THEORY

Strategic thinking is the art of outdoing an adversary, knowing that the adversary is trying to do the same to you.

Avinash Dixit and Barry Nalebuff,
Thinking Strategically (1991)

Economic life is full of situations in which people or firms or countries compete for profits or dominance. The oligopolies that we analyzed in the previous section sometimes break out into economic warfare. Such rivalry was seen in the last century when Vanderbilt and Drew repeatedly cut shipping rates on their parallel railroads. In recent years, airlines would occasionally launch price wars to attract customers and sometimes end up ruining everyone (see this chapter's introductory quote). But airlines learned that they needed to think and act strategically. Before an airline cuts its fares, it needs to consider how its rivals will react, and how it should then react to that reaction, and so on.

Once decisions reach the stage of thinking about what your opponent is thinking, and how you would then react, you are in the world of *game theory*. This is the analysis of situations involving two or more interacting decision makers who have conflicting objectives. Consider the following findings of game theorists in the area of imperfect competition:

- As the number of noncooperative oligopolists becomes large, the industry price and quantity tend toward the perfectly competitive outcome.
- If firms succeed in colluding, the market price and quantity will be close to those generated by a monopoly.
- Experiments suggest that as the number of firms increases, collusive agreements become more difficult to police and the frequency of cheating and noncooperative behavior increases.
- In many situations, there is no stable equilibrium for an oligopolistic market. Strategic interplay may lead to unstable outcomes as firms threaten, bluff, start price wars, punish weak opponents, signal their intentions, or simply exit from the market.

Game theory analyzes the ways in which two or more players choose strategies that jointly affect each other. This theory, which sounds frivolous, is in fact fraught with significance and was largely developed by John von Neumann (1903–1957), a Hungarian-born mathematical genius. Game theory has been used by economists to study the interaction of oligopolists, union-management disputes, countries' trade policies, international environmental agreements, reputations, and a host of other topics.

Game theory offers insights for politics and warfare, as well as for everyday life. For example, game theory suggests that in some circumstances a carefully chosen random pattern of behavior may be the best strategy. Inspections to catch illegal drugs or weapons should sometimes search randomly rather than predictably. Likewise, you should occasionally bluff at poker, not simply to win a pot with a weak hand but also to ensure that other players do not drop out

[1] For an example of how perfect price discrimination improves efficiency, see question 3 at the end of this chapter.

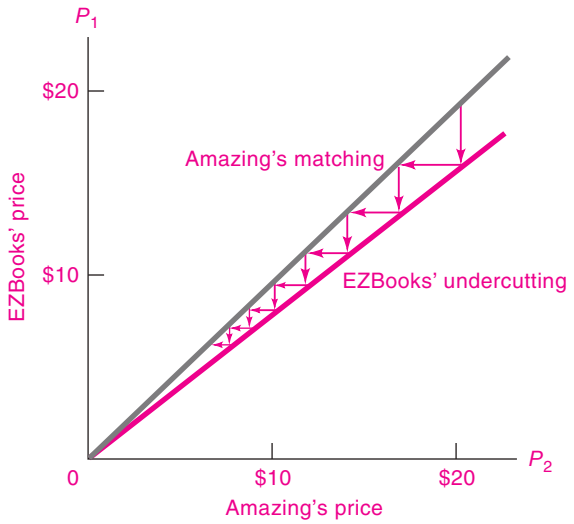


FIGURE 10-6. What Happens When Two Firms Insist on Undercutting Each Other?

Trace through the steps by which dynamic price cutting leads to ever-lower prices for two rivals.

when you bet high on a good hand. We will sketch out some of the major concepts of game theory in this section.

Thinking about Price Setting

Let's begin by analyzing the dynamics of price cutting. You are the head of an established firm, Amazing.com, whose motto is "We will not be undersold." You open your browser and discover that EZBooks.com, an upstart Internet bookseller, has an advertisement that says, "We sell for 10 percent less." Figure 10-6 shows the dynamics. The vertical arrows show EZBooks' price cuts; the horizontal arrows show Amazing's responding strategy of matching each price cut.

By tracing through the pattern of reaction and counterreaction, you can see that this kind of rivalry will end in mutual ruin at a zero price. Why? Because the only price compatible with both strategies is a price of zero: 90 percent of zero is zero.

Finally, it dawns on the two firms: When one firm cuts its price, the other firm will match the price cut. Only if the firms are shortsighted will they think that they can undercut each other for long. Soon each begins to ask, What will my rival do if I cut my price, or raise my price, or leave it alone? *Once you begin to consider how others will react to your actions, you have entered the realm of game theory.*

BASIC CONCEPTS

We will illustrate the basic concepts of game theory by analyzing a **duopoly price game.** A duopoly is a market which is served by only two firms. For simplicity, we assume that each firm has the same cost and demand structure. Further, each firm can choose whether to charge its normal price or lower its price below marginal costs and try to drive its rival into bankruptcy and then capture the entire market. The novel element in the duopoly game is that the firm's profits will depend on its rival's strategy as well as on its own.

A useful tool for representing the interaction between two firms or people is a two-way **payoff table.** A payoff table is a means of showing the strategies and the payoffs of a game between two players. Figure 10-7 shows the payoffs in the duopoly price game for our two companies. In the payoff table, a firm can choose between the strategies listed in its rows or columns. For example, EZBooks can choose between its two columns and Amazing can choose between its two rows. In this example, each firm decides whether to charge its normal price or to start a price war by choosing a lower price.

Combining the two decisions of each duopolist gives four possible outcomes, which are shown in the four cells of the table. Cell A, at the upper left, shows the outcome when both firms choose the normal price; D is the outcome when both choose to conduct a price war; and B and C result when one firm has a normal price and one a war price.

The numbers inside the cells show the **payoffs** of the two firms, that is, the profits earned by each firm for each of the four outcomes. The number in the lower left shows the payoff to the player on the left (Amazing); the entry in the upper right shows the payoff to the player at the top (EZBooks). Because the firms are identical, the payoffs are mirror images.

Alternative Strategies

Now that we have described the basic structure of a game, we next consider the behavior of the players. The new element in game theory is analyzing not only your own actions but also the interaction

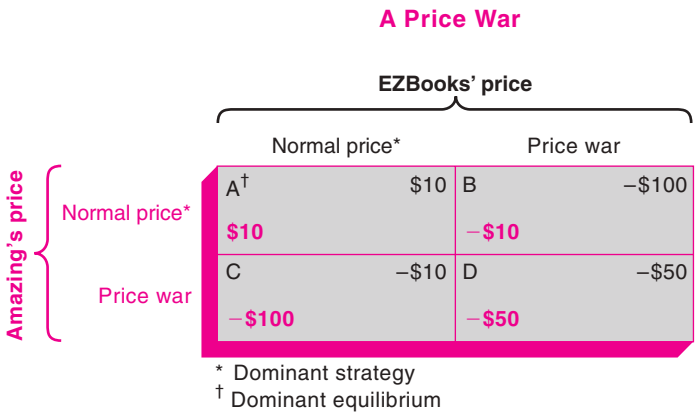


FIGURE 10-7. A Payoff Table for a Price War

The payoff table shows the payoffs associated with different strategies. Amazing has a choice between two strategies, shown as its two rows; EZBooks can choose between its two strategies, shown as two columns. The entries in the cells show the profits for the two players. For example, in cell C, Amazing plays "price war" and EZBooks plays "normal price." The result is that Amazing has green profit of −$100 while EZBooks has blue profit of −$10. Thinking through the best strategies for each player leads to the dominant equilibrium in cell A.

between your goals and moves and those of your opponent. But in trying to outwit your opponent, you must always remember that your opponent is trying to outwit you.

The guiding philosophy in game theory is the following: Pick your strategy by asking what makes most sense for you assuming that your opponents are analyzing your strategy and doing what is best for them.

Let's apply this maxim to the duopoly example. First, note that our two firms have the highest joint profits in outcome A. Each firm earns $10 when both follow a normal-price strategy. At the other extreme is the price war, where each cuts its price and runs a big loss.

In between are two interesting strategies where only one firm engages in the price war. In outcome C, for example, EZBooks follows a normal-price strategy while Amazing engages in a price war. Amazing takes most of the market but loses a great deal of money because it is selling below cost; EZBooks is actually better off selling at a normal price rather than responding.

Dominant Strategy. In considering possible strategies, the simplest case is that of a **dominant strategy.** This situation arises when one player has a single best strategy *no matter what strategy the other player follows.*

In our price-war game, for example, consider the options open to Amazing. If EZBooks conducts business as usual with a normal price, Amazing will get $10 of profit if it plays the normal price and will lose $100 if it declares economic war. On the other hand, if EZBooks starts a war, Amazing will lose $10 if it follows the normal price but will lose even more if it also engages in economic warfare. You can see that the same reasoning holds for EZBooks. Therefore, no matter what strategy the other firm follows, each firm's best strategy is to have the normal price. *Charging the normal price is a dominant strategy for both firms in this particular price-war game.*

When both (or all) players have a dominant strategy, we say that the outcome is a **dominant equilibrium.** We can see that in Figure 10-7, outcome A is a dominant equilibrium because it arises from a situation where both firms are playing their dominant strategies.

Nash Equilibrium. Most interesting situations do not have a dominant equilibrium, and we must therefore look further. We can use our duopoly example to

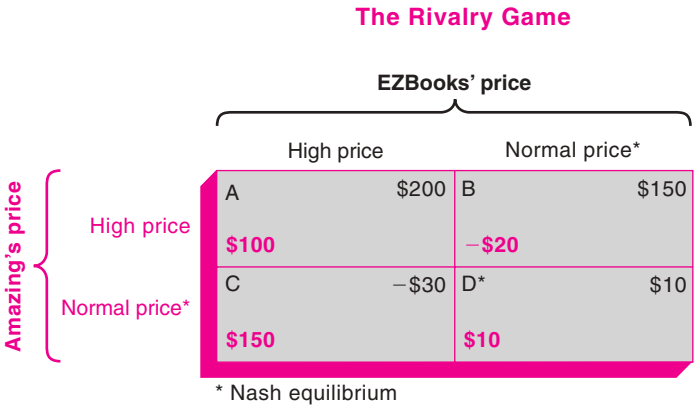


FIGURE 10-8. Should a Duopolist Try the Monopoly Price?

In the rivalry game, each firm can earn $10 by staying at its normal price. If both raise price to the high monopoly level, their joint profits will be maximized. However, each firm's temptation to "cheat" and raise its profits by lowering price ensures that the normal-price Nash equilibrium will prevail in the absence of collusion.

explore this case. In this example, which we call the *rivalry game,* each firm considers whether to charge its normal price or to raise its price toward the monopoly price and try to earn monopoly profits.

The rivalry game is shown in Figure 10-8. The firms can stay at their normal-price equilibrium, which we found in the price-war game. Or they can raise their price in the hopes of earning monopoly profits. Our two firms have the highest *joint* profits in cell A; here they earn a total of $300 when each follows a high-price strategy. Situation A would surely come about if the firms could collude and set the monopoly price. At the other extreme is the competitive-style strategy of the normal price, where each rival has profits of $10.

In between are two interesting strategies where one firm chooses a normal-price and one a high-price strategy. In cell C, for example, EZBooks follows a high-price strategy but Amazing undercuts. Amazing takes most of the market and has the highest profit of any situation, while EZBooks actually loses money. In cell B, Amazing gambles on high price, but EZBooks' normal price means a loss for Amazing.

Amazing has a dominant strategy in this new game. It will always have a higher profit by choosing a normal price. On the other hand, the best strategy for EZBooks depends upon what Amazing does. EZBooks would want to play normal if Amazing plays normal and would want to play high if Amazing plays high.

This leaves EZBooks with a dilemma: Should it play high and hope that Amazing will follow suit? Or play safe? Here is where game theory becomes useful. EZBooks should choose its strategy by first putting itself in Amazing's shoes. By doing so, EZBooks will find that Amazing should play normal regardless of what EZBooks does because playing normal is Amazing's dominant strategy. EZBooks should assume that Amazing will follow its best strategy and play normal, which therefore means that EZBooks should play normal. *This illustrates the basic rule of game theory: You should choose your strategy based on the assumption that your opponents will act in their own best interest.*

The approach we have described is a deep concept known as the **Nash equilibrium,** named after mathematician John Nash, who won a Nobel Prize for his discovery. In a Nash equilibrium, no player can gain anything by changing his own strategy, given the other player's strategy. The Nash equilibrium is also sometimes called the **noncooperative equilibrium** because each party chooses the strategy which is best for himself—without collusion or cooperation and without regard for the welfare of society or any other party.

Let us take a simple example: Assume that other people drive on the right-hand side of the road. What is your best strategy? Clearly, unless you are suicidal, you should also drive on the right-hand side. Moreover, a situation where everyone drives on the right-hand side is a Nash equilibrium: as long as everybody else is driving on the right-hand side, it will not be in anybody's interest to start driving on the left-hand side.

[Here is a technical definition of the Nash equilibrium for the advanced student: Suppose

that player A picks strategy S_A* while player B picks strategy S_B*. The pair of strategies (S_A*, S_B*) is a Nash equilibrium if neither player can find a better strategy to play assuming that the other player sticks to his original strategy. This discussion focuses on two-person games, but the analysis, and particularly the important Nash equilibrium, can be usefully extended to many-person or "*n*-person" games.]

You should verify that the starred strategies in Figure 10-8 constitute a Nash equilibrium. That is, neither player can improve its payoffs from the (normal, normal) equilibrium as long as the other doesn't change its strategy. Verify that the dominant equilibrium shown in Figure 10-7 is also a Nash equilibrium.

The Nash equilibrium (also called the noncooperative equilibrium) is one of the most important concepts of game theory and is widely used in economics and the other social sciences. Suppose that each player in a game has chosen a best strategy (the one with the highest payoff) *assuming* that all the other players keep their strategies unchanged. An outcome where all players follow this strategy is called a Nash equilibrium. Game theorists have shown that a competitive equilibrium is a Nash equilibrium.

Games, Games, Everywhere . . .

The insights of game theory pervade economics, the social sciences, business, and everyday life. In economics, for example, game theory can help explain trade wars as well as price wars.

Game theory can also suggest why foreign competition may lead to greater price competition. What happens when Chinese or Japanese firms enter a U.S. market where domestic firms had tacitly colluded on a strategy that led to high oligopolistic prices? The foreign firms may "refuse to play the game." They did not agree to the rules, so they may cut prices to increase their share of the market. Collusion among the domestic firms may break down because they must lower prices to compete effectively with the foreign firms.

A key feature in many games is the attempt on behalf of players to build *credibility*. You are credible if you can be expected to keep your promises and carry out your threats. But you cannot gain credibility simply by making promises. Credibility must be consistent with the incentives of the game.

How can you gain credibility? Here are some examples: Central banks earn reputations for being tough on inflation by adopting politically unpopular policies. Even greater credibility comes when the central bank is independent of the elected branches. Businesses make credible promises by writing contracts that inflict penalties if they do not perform as promised. A more perilous strategy is for an army to burn its bridges behind it. Because there is no retreat, the threat that they will fight to the death is a credible one.

The short discussion here provides a tiny peek at the vast terrain of game theory. This area has been enormously useful in helping economists and other social scientists think about situations where small numbers of people are well informed and try to outwit each other. Students who go on in economics, business, management, and even national security will find that using game theory can help them think strategically.

C. PUBLIC POLICIES TO COMBAT MARKET POWER

Economic analysis shows that monopolies produce economic waste. How significant are these inefficiencies? What can public policy do to reduce monopolistic harms? We address these two questions in this final section.

ECONOMIC COSTS OF IMPERFECT COMPETITION

The Cost of Inflated Prices and Reduced Output

Our analysis has shown how imperfect competitors reduce output and raise prices, thereby producing less (and charging more) than would be forthcoming in a perfectly competitive industry. This can be seen most clearly for monopoly, which is the most extreme version of imperfect competition. To see how and why monopoly keeps output too low, imagine that all other industries are efficiently organized. In such a world, price is the correct economic standard or measure of scarcity; price measures both the marginal utility of consumption to households and the marginal cost of production to firms.

Now Monopoly Inc. enters the picture. A monopolist is not a wicked firm—it doesn't rob people or force its goods down consumers' throats. Rather, Monopoly Inc. exploits the fact that it is the sole seller and raises its price above marginal cost (i.e., $P > MC$). Since $P = MC$ is necessary for economic efficiency, the marginal value of the good to consumers is therefore above its marginal cost. The same is true for oligopoly and monopolistic competition, as long as companies hold prices above marginal cost.

The Static Costs of Imperfect Competition

We can depict the efficiency losses from imperfect competition by using a simplified version of our monopoly diagram, here shown in Figure 10-9.

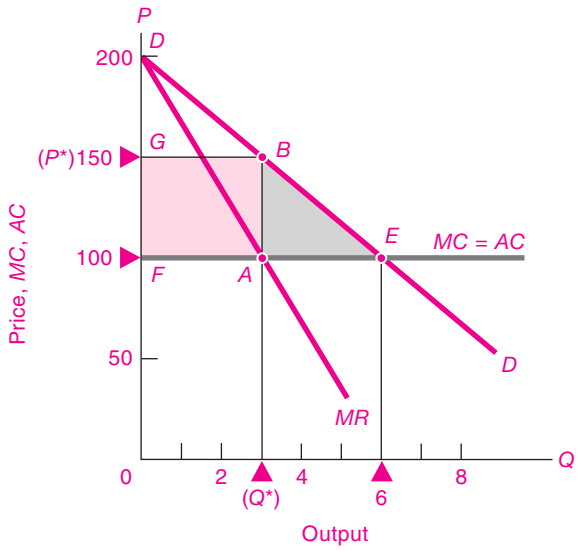


FIGURE 10-9. Monopolists Cause Economic Waste by Restricting Output

Monopolists make their output scarce and thereby drive up price and increase profits. If the industry were competitive, equilibrium would be at point *E*, where economic surplus is maximized.

At the monopolist's output at point *B* (with $Q = 3$ and $P = 150$), price is above *MC*, and consumer surplus is lost. Adding together all the consumer-surplus losses between $Q = 3$ and $Q = 6$ leads to economic waste from monopoly equal to the blue shaded area *ABE*. In addition, the monopolist has monopoly profits (that would have been consumer surplus) given by the green shaded region *GBAF*.

If the industry were perfectly competitive, the equilibrium would be reached at point *E*, where $P = MC$. Under universal perfect competition, this industry's quantity would be 6 with a price of 100.

Now consider the impact of monopoly. The monopoly might be created by a foreign-trade tariff or quota, by a labor union which monopolizes the supply of labor, or by a patent on a new product. The monopolist would set its *MC* equal to *MR* (not to industry *P*), displacing the equilibrium to the lower $Q = 3$ and the higher $P = 150$ in Figure 10-9. The area *GBAF* is the monopolist's profit, which compares with a zero-profit competitive equilibrium.

The inefficiency loss from monopoly is sometimes called **deadweight loss.** This term refers to the loss of economic welfare arising from distortions in prices and output such as those due to monopoly, as well as those due to taxation, tariffs, or quotas. Consumers might enjoy a great deal of consumer surplus if a new anti-pain drug is sold at marginal cost; however, if a firm monopolizes the product, consumers will lose more surplus than the monopolist will gain. That net loss in economic welfare is called deadweight loss.

We can picture the deadweight loss from a monopoly diagrammatically in Figure 10-9. Point *E* is the efficient level of production at which $P = MC$. For each unit that the monopolist reduces output below *E*, the efficiency loss is the vertical distance between the demand curve and the *MC* curve. The total deadweight loss from the monopolist's output restriction is the sum of all such losses, represented by the blue triangle *ABE*.

The technique of measuring the costs of market imperfections by "little triangles" of deadweight loss, such as the one in Figure 10-9, can be applied to most situations where output and price deviate from the competitive levels.

This cost calculation is sometimes called the "static cost" of monopoly. It is static because it assumes that the technology for producing output is unchanging. Some economists believe that imperfect competitors may have "dynamic benefits" if they generate more rapid technological change than do perfectly competitive markets. We will return to this question in the next chapter's discussion of innovation.

Public Policies on Imperfect Competition

How can nations reduce the harmful effects of monopolistic practices? Three approaches are often recommended by economists and legal scholars:

1. Historically, the first tool used by governments to control monopolistic practices was economic regulation. As this practice has evolved over the last century, economic regulation now allows specialized regulatory agencies to oversee the prices, outputs, entry, and exit of firms in regulated industries such as public utilities and transportation. It is, in effect, limited government control without government ownership.
2. The major method now used for combating excessive market power is the use of antitrust policy. Antitrust policies are laws that prohibit certain kinds of behavior (such as firms' joining together to fix prices) or curb certain market structures (such as pure monopolies and highly concentrated oligopolies).
3. More generally, anticompetitive abuses can be avoided by encouraging competition wherever possible. There are many government policies that can promote vigorous rivalry even among large firms. It is particularly crucial to reduce barriers to entry in all industries. That means encouraging small businesses and not walling off domestic markets from foreign competition.

We will review the first two approaches in the balance of this chapter.

REGULATING ECONOMIC ACTIVITY

Economic regulation of American industry goes back more than a century. The first federal regulation applied to transportation, with the Interstate Commerce Commission (ICC) in 1887. The ICC was designed as much to prevent price wars and to guarantee service to small towns as it was to control monopoly. Later, federal regulation spread to banks in 1913, to electric power in 1920, and to communications, securities markets, labor, trucking, and air travel during the 1930s.

Economic regulation involves the control of prices, entry and exit conditions, and standards of service. Such regulation is most important in industries that are natural monopolies. (Recall that a natural monopoly occurs when the industry's output can be efficiently produced only by a single firm.) Prominent examples of industries regulated today include public utilities (electricity, natural gas, and water) and telecommunications (telephone, radio, cable TV, and more generally the electromagnetic spectrum). The financial industry has been regulated since the 1930s, with strict rules specifying what banks, brokerage firms, and insurance companies can and cannot do. Since 1977, many economic regulations have been loosened or lifted, such as those on the airlines, trucking, and securities firms.

Why Regulate Industry?

Regulation restrains the unfettered market power of firms. What are the reasons why governments might choose to override the decisions made in the marketplace? The first reason is to *prevent abuses of market power* by monopolies or oligopolies. A second major reason is to *remedy informational failures*, such as those which occur when consumers have inadequate information. A third reason is to *correct externalities* like pollution. The third of these reasons pertains to social regulation and is examined in the chapter on environmental economics; we review the first two reasons in this section.

Containing Market Power

The traditional view is that regulatory measures should be taken to reduce excessive market power. More specifically, governments should regulate industries where there are too few firms to ensure vigorous rivalry, particularly in the extreme case of natural monopoly.

We know from our discussion of declining costs in earlier chapters that pervasive economies of scale are inconsistent with perfect competition; we will find oligopoly or monopoly in such cases. But the point here is even more extreme: *When there are such powerful economies of scale or scope that only one firm can survive, we have a natural monopoly.*

Why do governments sometimes regulate natural monopolies? They do so because a natural monopolist, enjoying a large cost advantage over its potential competitors and facing price-inelastic demand, can jack up its price sharply, obtain enormous monopoly profits, and create major economic inefficiencies. Hence, regulation allows society to enjoy the benefits

of a natural monopoly while preventing the super-high prices that might result if it were unregulated. A typical example is local water distribution. The cost of gathering water, building a water-distribution system, and piping water into every home is sufficiently large that it would not pay to have more than one firm provide local water service. This is a natural monopoly. Under economic regulation, a government agency would provide a franchise to a company in a particular region. That company would agree to provide water to all households in that region. The government would review and approve the prices and other terms that the company would then present to its customers.

Another kind of natural monopoly, particularly prevalent in network industries, arises from the requirement for standardization and coordination through the system for efficient operation. Railroads need standard track gauges, electrical transmission requires load balancing, and communications systems require standard codes so that different parts can "talk" to each other.

In earlier times, regulation was justified on the dubious grounds that it was needed to prevent cut-throat or destructive competition. This was one argument for continued control over railroads, trucks, airlines, and buses, as well as for regulation of the level of agricultural production. Economists today have little sympathy for this argument. After all, competition will increase efficiency, and ruinously low prices are exactly what an efficient market system *should* produce.

Remedying Information Failures

Consumers often have inadequate information about products in the absence of regulation. For example, testing pharmaceutical drugs is expensive and scientifically complex. The government regulates drugs by allowing only the sale of those drugs which are proved "safe and efficacious." Government also prohibits false and misleading advertising. In both cases, the government is attempting to correct for the market's failure to provide information efficiently on its own.

One area where regulating the provision of information is particularly critical is financial markets. When people buy stocks or bonds of private companies, they are placing their fortunes in the hands of people about whom they know next to nothing. Before buying shares of ZYX.com, I will examine their financial statements to determine what their sales, earnings, and dividends have been. But how can I know exactly how they measure earnings? How can I be sure that they are reporting this information honestly?

This is where government regulation of financial markets steps in. Most regulations of the financial industry serve the purpose of improving the quantity and quality of information so that markets can work better. When a company sells stocks or bonds in the United States, it is required to issue copious documentation of its current financial condition and future prospects. Companies' books must be certified by independent auditors.

Occasionally, particularly in times of speculative frenzies, companies will bend or even break the rules. This happened on a large scale in the late 1990s and early 2000s, particularly in communications and many "high-tech" firms. When these illegal practices were made public, Congress passed a new law in 2002; this law made it illegal to lie to an auditor, established an independent board to oversee accountants, and provided new oversight powers to the Securities and Exchange Commission (SEC). Some argue that this kind of law should be welcomed by honest businesses; tough reporting standards are beneficial to financial markets because they reduce informational asymmetries between buyers and sellers, promote trust, and encourage financial investment.

Stanford's John McMillan uses an interesting analogy to describe the role of government regulation. Sports are contests in which individuals and teams strive to defeat opponents with all their strength. But the participants must adhere to a set of extremely detailed rules; moreover, referees keep an eagle eye on players to make sure that they obey the rules, with appropriately scaled penalties for infractions. Without carefully crafted rules, a game would turn into a bloody brawl. Similarly, government regulations, along with a strong legal system, are necessary in a modern economy to ensure that overzealous competitors do not monopolize, pollute, defraud, mislead, maim, or otherwise mistreat workers and consumers. This sports analogy reminds us that the government still has an important role to play in monitoring the economy and setting the rules of the road.

ANTITRUST LAW AND ECONOMICS

A second important government tool for promoting competition is antitrust law. The purpose of antitrust policies is to provide consumers with the economic benefits of vigorous competition. Antitrust laws attack anticompetitive abuses in two different ways: First, they prohibit certain kinds of *business conduct,* such as price fixing, that restrain competitive forces. Second, they restrict some *market structures,* such as monopolies, that are considered most likely to restrain trade and abuse their economic power in other ways. The framework for antitrust policy was set by a few key legislative statutes and by more than a century of court decisions.

The Framework Statutes

Antitrust law is like a huge forest that has grown from a handful of seeds. The statutes on which the law is based are so concise and straightforward that they can be quoted in Table 10-1; it is astounding how much law has grown from so few words.

Sherman Act (1890). Monopolies had long been illegal under the common law, based on custom and past judicial decisions. But the body of laws proved ineffective against the mergers, cartels, and trusts that swept through the American economy in the 1880s. (Reread the section on the monopolists of the Gilded Age in Chapter 9 to get a flavor of the cutthroat tactics of that era.)

In 1890, populist sentiments led to the passage of the Sherman Act, which is the cornerstone of American antitrust law. Section 1 of the Sherman Act prohibits contracts, combinations, and conspiracies "in restraint of trade." Section 2 prohibits "monopolizing" and conspiracies to monopolize. Neither the statute nor the accompanying discussion contained any clear notion about the exact meaning

The Antitrust Laws

Sherman Antitrust Act (1890, as amended)

§1. Every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations, is declared to be illegal.

§2. Every person who shall monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several States, or with foreign nations, shall be deemed guilty of a felony. . . .

Clayton Antitrust Act (1914, as amended)

§2. It shall be unlawful . . . to discriminate in price between different purchasers of commodities of like grade and quality . . . where the effect of such discrimination may be substantially to lessen competition or tend to create a monopoly in any line of commerce. . . . *Provided,* That nothing herein contained shall prevent differentials which make only due allowance for differences in the cost. . . .

§3. That it shall be unlawful for any person . . . to lease or make a sale or contract . . . on the condition, agreement, or understanding that the lessee or purchaser thereof shall not use or deal in the . . . commodities of a competitor . . . where the effect . . . may be to substantially lessen competition or tend to create a monopoly in any line of commerce.

§7. No [corporation] . . . shall acquire . . . the whole or any part . . . of another [corporation] . . . where . . . the effect of such an acquisition may be substantially to lessen competition, or to tend to create a monopoly.

Federal Trade Commission Act (1914, as amended)

§5. Unfair methods of competition . . . and unfair or deceptive acts or practices . . . are declared unlawful.

TABLE 10-1. American Antitrust Law Is Based on a Handful of Statutes

The Sherman, Clayton, and Federal Trade Commission Acts laid the foundation for American antitrust law. Interpretation of these acts has fleshed out modern antitrust doctrines.

of monopoly or which actions were prohibited. The meaning was fleshed out in later case law.

Clayton Act (1914). The Clayton Act was passed to clarify and strengthen the Sherman Act. It outlawed *tying contracts* (in which a customer is forced to buy product B if she wants product A); it ruled *price discrimination* and exclusive dealings illegal; it banned *interlocking directorates* (in which some people would be directors of more than one firm in the same industry) and *mergers* formed by acquiring common stock of competitors. These practices were not illegal *per se* (meaning "in itself") but only when they might substantially lessen competition. The Clayton Act emphasized prevention as well as punishment.

Another important element of the Clayton Act was that it specifically provided antitrust immunity to labor unions.

Federal Trade Commission Acts. In 1914 the Federal Trade Commission (FTC) was established to prohibit "unfair methods of competition" and to warn against anticompetitive mergers. In 1938, the FTC was also empowered to ban false and deceptive advertising. To enforce its powers, the FTC can investigate firms, hold hearings, and issue cease-and-desist orders.

BASIC ISSUES IN ANTITRUST LAW: CONDUCT AND STRUCTURE

While the basic antitrust statutes are straightforward, it is not easy in practice to decide how to apply them to specific situations of industry conduct or market structure. Actual law has evolved through an interaction of economic theory and case law.

One key issue that arises in many cases is, What is the relevant market? For example, what is the "telephone" industry in Albuquerque, New Mexico? Is it all information industries, or only telecommunications, or only wired telecommunications, or wired phones in all of New Mexico, or just in some specific zip code? In recent U.S. cases, the market has been defined to include products which are reasonably close substitutes. If the price of land-line telephone service goes up and people switch to cell-phone service in significant numbers, then these two products would be considered to be in the same industry. If by contrast few people buy more newspapers when the price of phone service increases, then newspapers are not in the telephone market.

Illegal Conduct

Some of the earliest antitrust decisions concerned illegal behavior. The courts have ruled that certain kinds of collusive behavior are illegal per se; there is simply no defense that will justify these actions. The offenders cannot defend themselves by pointing to some worthy objective (such as product quality) or mitigating circumstance (such as low profits).

The most important class of per se illegal conduct is agreements among competing firms to fix prices. Even the severest critic of antitrust policy can find no redeeming virtue in price fixing. Two other practices are illegal in all cases:

- *Bid rigging,* in which different firms agree to set their bids so that one firm will win the auction, usually at an inflated price, is always illegal.
- *Market allocation schemes,* in which competitors divide up markets by territory or by customers, are anticompetitive and hence illegal per se.

Many other practices are less clear-cut and require some consideration of the particular circumstances:

- *Price discrimination,* in which a firm sells the same product to different customers at different prices, is unpopular but generally not illegal. (Recall the discussion of price discrimination earlier in this chapter.) To be illegal, the discrimination must not be based on differing production costs, and it must injure competition.
- *Tying contracts,* in which a firm will sell product A only if the purchaser buys product B, are generally illegal only if the seller has high levels of market power.
- What about *ruinously low prices?* Suppose that because of Wal-Mart's efficient operations and low prices, Pop's grocery store goes out of business. Is this illegal? The answer is no. Unless Wal-Mart did something else illegal, simply driving its competitors bankrupt because of its superior efficiency is not illegal.

Note that the practices on this list relate to a firm's *conduct.* It is the acts themselves that are illegal, not the structure of the industry in which the acts take place. Perhaps the most celebrated example is the great electric-equipment conspiracy. In 1961, the electric-equipment industry was found guilty of collusive price agreements. Executives of the largest companies—such as GE and Westinghouse—conspired to raise

prices and covered their tracks like characters in a spy novel by meeting in hunting lodges, using code names, and making telephone calls from phone booths. The companies agreed to pay extensive damages to their customers for overcharges, and some executives were jailed for their antitrust violations.

Structure: Is Bigness Badness?

The most visible antitrust cases concern the structure of industries rather than the conduct of companies. These cases consist of attempts to break up or limit the conduct of dominant firms.

The first surge of antitrust activity under the Sherman Act focused on dismantling existing monopolies. In 1911, the Supreme Court ordered that the American Tobacco Company and Standard Oil be broken up into many separate companies. In condemning these flagrant monopolies, the Supreme Court enunciated the important "rule of reason." Only *unreasonable* restraints of trade (mergers, agreements, and the like) came within the scope of the Sherman Act and were considered illegal.

The rule-of-reason doctrine virtually nullified the antitrust laws' attack on monopolistic mergers, as shown by the *U.S. Steel case* (1920). J. P. Morgan had put this giant together by merger, and at its peak it controlled 60 percent of the market. But the Supreme Court held that pure size or monopoly by itself was no offense. In that period, as they do today, the cases that shaped the economic landscape focused on illegal monopoly structures more than anticompetitive conduct.

In recent years, two important cases have set the ground rules for monopolistic structure and behavior. In the *AT&T case,* the Department of Justice filed a far-reaching suit. For most of the twentieth century, the American Telephone and Telegraph (AT&T, sometimes called the Bell System) was a vertically and horizontally integrated regulated monopoly supplier of telecommunications services. In 1974, the Department of Justice filed an antitrust suit, contending that AT&T had monopolized the regulated long-distance market by anticompetitive means, such as preventing MCI and other carriers from connecting to the local markets, and had monopolized the telecommunications-equipment market by refusing to purchase equipment from non-Bell suppliers.

Faced with the prospect of losing the antitrust suit, the company settled in a consent decree in 1982. The local Bell operating companies were divested (legally separated) from AT&T and were regrouped into seven large regional telephone holding companies. AT&T retained its long-distance operations as well as Bell Labs (the research organization) and Western Electric (the equipment manufacturer). The net effect was an 80 percent reduction in the size and sales of the Bell System.

The dismantling of the Bell System set off a breathtaking revolution in the telecommunications industry. New technologies are changing the telecommunications landscape: cellular phone systems are eating away at the natural monopoly of Alexander Graham Bell's wire-based system; telephone companies are joining forces to bring television signals into homes; fiber-optic lines are beginning to function as data superhighways, carrying vast amounts of data around the country and the world; the Internet is linking people and places together in ways that were unimagined a decade ago. One clear lesson of the breakup of the Bell System is that monopoly is not necessary for rapid technological change.

The most recent major antitrust case involved the giant software company *Microsoft.* In 1998, the federal government and 19 states lodged a far-reaching suit alleging that Microsoft had illegally maintained its dominant position in the market for operating systems and had used that dominance to leverage itself into other markets, such as the Internet browser market. The government claimed that "Microsoft has engaged in a broad pattern of unlawful conduct with the purpose and effect of thwarting emerging threats to its powerful and well-entrenched operating system monopoly." Although a monopoly acquired by fair means is legal, acting to stifle competition is illegal.

In his "Findings of Fact," Judge Jackson declared that Microsoft was a monopoly that had controlled more than 90 percent of the market share for PC operating systems since 1990 and that Microsoft had abused its market power and caused "consumer harm by distorting competition." Judge Jackson found that Microsoft had violated Sections 1 and 2 of the Sherman Act. He found that "Microsoft maintained its monopoly power by anticompetitive means, attempted to monopolize the Web browser market, and violated the Sherman Act by unlawfully tying its Web browser to its operating system."

The Department of Justice proposed the radical step of separating Microsoft along functional lines. This "divestiture" would require a separation of Microsoft into two separate, independent companies. One company ("WinCo") would own Microsoft's Windows and other operating-system businesses, and the other ("AppCo") would own the applications and other businesses. Judge Jackson accepted the Department of Justice's remedy recommendation with no modifications.

But then the case took a bizarre twist when it turned out that Judge Jackson had been holding private heart-to-heart discussions with journalists even as he was trying the case. He was chastised for his unethical conduct and removed from the case. Shortly thereafter, the Bush administration decided it would not seek to separate Microsoft but would settle for "conduct" remedies. These measures would restrict Microsoft's conduct through steps such as prohibiting contractual tying and discriminatory pricing as well as ensuring the interoperability of Windows with non-Windows software. After extensive further hearings, the case was settled in November 2002 with Microsoft intact but under the watchful eye of the government and the courts.

Antitrust Laws and Efficiency

Economic and legal views toward regulation and antitrust have changed dramatically over the last three decades. Increasingly, economic regulation and antitrust laws are aimed toward the goal of improving economic efficiency rather than combating businesses simply because they are big or profitable.

What has prompted the changing attitude toward antitrust policy? First, economists found that concentrated industries sometimes had outstanding performance. That is, while concentrated industries might have static inefficiencies, these were more than outweighed by their dynamic efficiencies. Consider Intel, Microsoft, and Boeing. They have had substantial market shares, but they have also been highly innovative and commercially successful.

A second thrust of the new approach to regulation and antitrust arose from new findings on the nature of competition. Considering both experimental evidence and observation, many economists believe that intense rivalry will spring up even in oligopolistic markets as long as collusion is strictly prohibited. Indeed, in the words of Richard Posner, formerly a law professor and currently a federal judge,

> The only truly unilateral acts by which firms can get or keep monopoly power are practices like committing fraud on the Patent Office or blowing up a competitor's plant, and fraud and force are in general adequately punished under other statutes.

In this view, the only valid purpose of the antitrust laws should be to replace existing statutes with a simple prohibition against *agreements*—explicit or tacit—that unreasonably restrict competition.

A final reason for the reduced activism in antitrust has been growing globalization in many concentrated industries. As more foreign firms gain a foothold in the American economy, they tend to compete vigorously for a share of the market and often upset established sales patterns and pricing practices as they do so. For example, when the U.S. sales of Japanese automakers increased, the cozy coexistence of the Big Three American auto firms dissolved. Many economists believe that the threat of foreign competition is a much more powerful tool for enforcing market discipline than are antitrust laws.

SUMMARY

A. Behavior of Imperfect Competitors

1. Recall the four major market structures: (*a*) *Perfect competition* is found when no firm is large enough to affect the market price. (*b*) *Monopolistic competition* occurs when a large number of firms produce slightly differentiated products. (*c*) *Oligopoly* is an intermediate form of imperfect competition in which an industry is dominated by a few firms. (*d*) *Monopoly* comes when a single firm produces the entire output of an industry.
2. Measures of concentration are designed to indicate the degree of market power in an imperfectly competitive industry. Industries which are more concentrated tend to have higher levels of R&D expenditures, but on average their profitability is not higher.
3. High barriers to entry and complete collusion can lead to collusive oligopoly. This market structure produces a price and quantity relation similar to that under monopoly.
4. Another common structure is the monopolistic competition that characterizes many retail industries. Here we see many small firms, with only slight differences in the characteristics of their products (such as different locations of gasoline stations or different types of breakfast cereals). Product differentiation leads each firm to face a downward-sloping demand curve as each firm is free to set its own prices. In the long run, free entry extinguishes profits as these industries show an equilibrium in which their *AC* curves are tangent to their demand curves. In this tangency equilibrium, prices are above marginal costs, but the industry exhibits greater diversity of quality and service than would occur under perfect competition.
5. A final situation recognizes the strategic interplay that is present when an industry has but a handful of firms. When a small number of firms compete in a market, they must recognize their strategic interactions. Competition among the few introduces a completely new feature into economic life: It forces firms to take into account competitors' reactions to price and output decisions and brings strategic considerations into these markets.
6. Price discrimination occurs when the same product is sold to different consumers at different prices. This practice often occurs when sellers can segment their market into different groups.

B. Game Theory

7. Economic life contains many situations with strategic interaction among firms, households, governments, or others. Game theory analyzes the way that two or more parties, who interact in an arena such as a market, choose actions or strategies that jointly affect all participants.
8. The basic structure of a game includes the players, who have different possible actions or strategies, and the payoffs, which describe the various possible profits or other benefits that the players might obtain under each outcome. The key new concept is the payoff table of a game, which displays information about the strategies and the payoffs or profits of the different players for all possible outcomes.
9. The key to choosing strategies in game theory is for players to think about their opponent's goals as well as their own, never forgetting that the other side is doing the same. When playing a game in economics or any other field, assume that your opponent will choose his or her best option. Then pick your strategy to maximize your benefit, always assuming that your opponent is similarly analyzing your options.
10. Sometimes a dominant strategy is available—one that is best no matter what the opposition does. More often, we find a Nash equilibrium (or noncooperative equilibrium), in which no player can improve his or her payoff as long as the other player's strategy remains unchanged.

C. Public Policies to Combat Market Power

11. Monopoly power often leads to economic inefficiency when prices rise above marginal cost, costs are bloated by lack of competitive pressure, and product quality deteriorates.
12. Economic regulation involves the control of prices, production, entry and exit conditions, and standards of service in a particular industry. The normative view of economic regulation is that government intervention is appropriate when there are major market failures. These include excess market power in an industry, an inadequate supply of information for consumers and workers, and externalities such as pollution. The strongest case for economic regulation comes in regard to natural monopolies. Natural monopoly occurs when average costs are falling for every level of output, so the most efficient

organization of the industry requires production by a single firm.

13. Antitrust policy, prohibiting anticompetitive conduct and preventing monopolistic structures, is the primary way that public policy limits abuses of market power by large firms. This policy grew out of legislation like the Sherman Act (1890) and the Clayton Act (1914). The primary purposes of antitrust policy are (*a*) to prohibit anticompetitive activities (which include agreements to fix prices or divide up territories, price discrimination, and tie-in agreements) and (*b*) to break up illegal monopoly structures. In today's legal theory, such structures are those that have excessive market power (a large share of the market) and also engage in anticompetitive acts.
14. Legal antitrust policy has been significantly influenced by economic thinking during the last three decades. As a result, antitrust policy now focuses almost exclusively on improving efficiency and ignores earlier populist concerns with bigness itself.

CONCEPTS FOR REVIEW

Models of Imperfect Competition

concentration: concentration ratios, HHI
market power
strategic interaction
tacit and explicit collusion
imperfect competition:
  collusive oligopoly
  monopolistic competition
  small-number oligopoly
no-profit equilibrium in monopolistic competition
inefficiency of $P > MC$

Game Theory

players, strategies, payoffs
payoff table
dominant strategy and equilibrium
Nash or noncooperative equilibrium

Policies for Imperfect Competition

deadweight losses
reasons for regulation:
  market power
  externalities
  information failures

Antitrust Policy

Sherman, Clayton, and FTC Acts
natural monopoly
per se prohibitions vs. the "rule of reason"
efficiency-oriented antitrust policy

FURTHER READING AND INTERNET WEBSITES

Further Reading

An excellent review of industrial organization is Dennis W. Carlton and Jeffrey M. Perloff, *Modern Industrial Organization* (Addison-Wesley, New York, 2005).

Game theory was developed in 1944 by John von Neumann and Oscar Morgenstern and published in *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, N.J., 1980). An entertaining review of game theory by two leading microeconomists is Avinash K. Dixit and Barry J. Nalebuff, *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday* (Norton, New York, 1993). A nontechnical biography of John Nash by journalist Silvia Nasar, *A Beautiful Mind: A Biography of John Forbes Nash Jr.* (Touchstone Books, New York, 1999), is a vivid history of game theory and of one of its most brilliant theorists.

Law and economics advanced greatly under the influence of scholars like Richard Posner, now a circuit court judge. His book, *Antitrust Law: An Economic Perspective* (University of Chicago Press, 1976), is a classic.

Websites

Game theorists have set up a number of sites. See particularly those by David Levine of UCLA at *levine.sscnet.ucla.edu* and Al Roth of Harvard at *www.economics.harvard.edu/~aroth/alroth.html*.

OPEC has its site at *www.opec.org*. This site makes interesting reading from the point of view of oil producers, many of which are Arab countries.

Data and methods pertaining to concentration ratios can be found in a Bureau of the Census publication at *www.census.gov/epcd/www/concentration.html*.

An excellent website with links to many issues on antitrust is *www.antitrust.org*. The homepage for the Antitrust Division of the Department of Justice, at *www.usdoj.gov/atr/public/div_stats/211491.htm*, contains an overview of antitrust issues.

QUESTIONS FOR DISCUSSION

1. Review collusive oligopoly and monopolistic competition, which are two theories of imperfect competition discussed in this chapter. Draw up a table that compares perfect competition, monopoly, and the two theories with respect to the following characteristics: (*a*) number of firms; (*b*) extent of collusion; (*c*) price vs. marginal cost; (*d*) price vs. long-run average cost; (*e*) efficiency.
2. Consider an industry whose firms have the following sales:

Firm	Sales
Appel Computer	1000
Banana Computer	800
Cumquat Computer	600
Dellta Computer	400
Endive Computer	300
Fettucini Computer	200
Grapefruit Computer	150
Hamburger Computer	100
InstantCoffee Computer	50
Jasmine Computer	1

The Herfindahl-Hirschman Index (HHI) is defined as

$$\text{HHI} = (\text{market share of firm 1 in \%})^2 + (\text{market share of firm 2 in \%})^2 + \cdots + (\text{market share of last firm in \%})^2$$

   a. Calculate the four-firm and six-firm concentration ratios for the computer industry.
   b. Calculate the HHI for the industry.
   c. Suppose that Appel Computer and Banana Computer were to merge with no change in the sales of any of the different computers. Calculate the new HHI.
3. "Perfect price discrimination" occurs when each consumer is charged his or her maximum price for the product. When this happens, the monopolist is able to capture the entire consumer surplus. Draw a demand curve for each of six consumers and compare (*a*) the situation in which all consumers face a single price with (*b*) a market under perfect price discrimination. Explain the paradoxical result that perfect price discrimination removes the inefficiency of monopoly.
4. The government decides to tax a monopolist at a constant rate of $\$x$ per unit. Show the impact upon output and price. Is the post-tax equilibrium closer to or further from the ideal equilibrium of $P = MC$?
5. Show that a profit-maximizing, unregulated monopolist will never operate in the price-inelastic region of its demand curve. Show how regulation can force the monopolist into the inelastic portion of its demand curve. What will be the impact of an increase in the regulated price of a monopolist upon revenues and profits when it is operating on (*a*) the elastic portion of the demand curve, (*b*) the inelastic portion of the demand curve, and (*c*) the unit-elastic portion of the demand curve?
6. Make a list of the industries that you feel are candidates for the title "natural monopoly." Then review the different strategies for intervention to prevent exercise of monopoly power. What would you do about each industry on your list?
7. Firms often lobby for tariffs or quotas to provide relief from import competition.
   a. Suppose that the monopolist shown in Figure 10-9 has a foreign competitor that will supply output perfectly elastically at a price slightly above the monopolist's $AC = MC$ but below P. Show the impact of the foreign competitor's entry into the market.
   b. What would be the effect on the price and quantity if a prohibitive tariff were levied on the foreign good? (A prohibitive tariff is one that is so high as to effectively wall out all imports.) What would be the effect of a small tariff? Use your analysis to explain the statement, "The tariff is the mother of monopoly."
8. Explain in words and with the use of diagrams why a monopolistic equilibrium leads to economic inefficiency relative to a perfectly competitive equilibrium. Why is the condition $MC = P = MU$ of Chapter 8 critical for this analysis?
9. Consider the *prisoner's dilemma*, one of the most famous of all games. Molly and Knuckles are partners in crime. The district attorney interviews each separately, saying, "I have enough on both of you to send you to jail for a year. But I'll make a deal with you: If you *alone* confess, you'll get off with a 3-month sentence, while your partner will serve 10 years. If you *both* confess, you'll both get 5 years." What should Molly do? Should she confess and hope to get a short sentence? Three months are preferable to the year she would get if she remains silent. But wait. There is an even better reason for confessing. Suppose Molly doesn't confess and, unbeknownst to her, Knuckles does confess. Molly stands to get 10 years! It's clearly better in this situation for Molly to confess and get

5 years rather than 10 years. Construct a payoff table like that in Figure 10-8. Show that each player has a dominant strategy, which is to confess, and both therefore end up with long prison terms. Then show what would happen if they could make binding commitments not to confess.

10. In his Findings of Fact in the Microsoft case, Judge Jackson wrote: "It is indicative of monopoly power that Microsoft felt that it had substantial discretion in setting the price of its Windows 98 upgrade product (the operating system product it sells to existing users of Windows 95). A Microsoft study from November 1997 reveals that the company could have charged $49 for an upgrade to Windows 98—there is no reason to believe that the $49 price would have been unprofitable—but the study identifies $89 as the revenue-maximizing price. Microsoft thus opted for the higher price." Explain why these facts would indicate that Microsoft is not a perfect competitor. What further information would be needed to prove Microsoft was a monopoly?

11. In long-run equilibrium, both perfectly competitive and monopolistically competitive markets achieve a tangency between the firm's *dd* demand curve and its *AC* average cost curve. Figure 10-4 shows the tangency for a monopolistic competitor, while Figure 10-10 displays the tangency for a perfect competitor. Discuss the similarities and differences in the two situations with respect to:

- **a.** The elasticity of the demand curve for the firm's product
- **b.** The extent of divergence between price and marginal cost
- **c.** Profits
- **d.** Economic efficiency

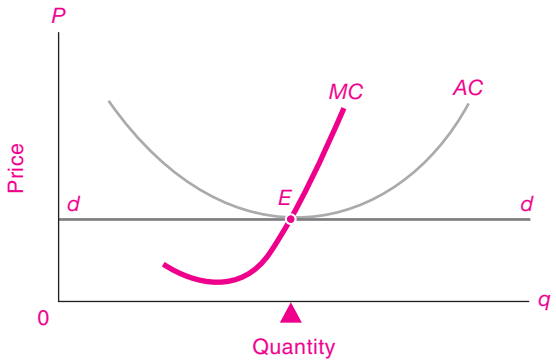


FIGURE 10-10. Perfect Competition

12. Reread the history of OPEC. Draw a set of supply and demand curves in which supply is completely price-inelastic. Show that a cartel that sets a quantity target (the inelastic supply curve) will experience more volatile prices if demand is price-inelastic than if demand is price-elastic when (*a*) the demand curve shifts horizontally by a certain quantity (such as would occur with an unanticipated demand shock) or (*b*) there is a shift in the supply curve (say, due to cheating by a cartel member).

CHAPTER

Economics of Uncertainty

11

Pearls lie not on the seashore. If thou desirest one, thou must dive for it.

Chinese proverb

Life is full of uncertainties. Suppose that you are in the oil business. You might be in charge of a joint venture in Siberia. What obstacles would you face? You would face major risks that plague oil producers everywhere—the risks of a price plunge, of embargoes, or of an attack on your tankers by some hostile regime. Added to these are the uncertainties of operating in uncharted terrain: you are unfamiliar with the geological formations, with the routes for getting the oil to the market, with the success rate on drilling wells, and with the skills of the local workforce.

In addition to these uncertainties are the political risks involved in dealing with an increasingly autocratic and nationalistic government in Moscow, along with the problems that arise from occasional wars and from corrupt elements in a country where bribes are common and the rule of law is insecure. And your partners may turn out to be unscrupulous fellows who take advantage of their local knowledge to get more than their fair share.

The economic issues in your joint venture present complexities that are not captured in our elementary theories. Many of these issues involve *risk, uncertainty,* and *information.* Our oil company must deal with the uncertainties of drilling, of volatile prices, and of shifting markets. Likewise, households must contend with uncertainty about future wages or employment and about the return on their investments in education or in financial assets. Additionally, some people suffer from misfortunes such as devastating hurricanes, earthquakes, or illnesses. The first section of this chapter discusses the fundamental economics of uncertainty.

How do individuals and societies cope with uncertainties? One important approach is through insurance. The second section deals with the fundamentals of insurance, including the important concept of social insurance. The third section applies the concept of social insurance to health care, which is a growing political and social dilemma in the United States. We conclude with an examination of the economics of information and apply this to the rise of the Internet.

No study of the realities of economic life is complete without a thorough study of the fascinating questions involved in decision making under uncertainty and the economics of information.

A. ECONOMICS OF RISK AND UNCERTAINTY

Our analysis of markets presumed that costs and demands were known for certain. In reality, business life is teeming with risk and uncertainty. We described

the uncertainties involved in a joint venture for oil in Siberia, but these problems are not confined to the oil business. Virtually all firms face uncertainties about their output and input prices. They may find that their markets are shrinking because of a recession or that credit is hard to find in a financial crisis. Furthermore, the behavior of their competitors cannot be forecast in advance. The essence of business is to invest now in order to make profits in the future, in effect putting fortunes up as hostage to future uncertainties. Economic life is a risky business.

Modern economics has developed useful tools to incorporate uncertainty into the analysis of business and household behavior. This section examines the role of markets in spreading risks over space and time and analyzes the theory of individual behavior under uncertainty. These topics are but a brief glimpse into the fascinating world of risk and uncertainty in economic life.

SPECULATION: SHIPPING ASSETS OR GOODS ACROSS SPACE AND TIME

We begin by considering the role of speculative markets. **Speculation** involves buying and selling in order to make profits from fluctuations in prices. A speculator wants to buy low and sell high. The item might be grain, oil, eggs, stocks, or foreign currencies. Speculators do not buy these items for their own sake. The last thing they want is to see the egg truck show up at their door. Rather, they make a profit from price changes.

Many people think of speculation as a slightly sinister activity, particularly when it arises from accounting frauds and inside information. But speculation can be beneficial to society. The economic function of speculators is to "move" goods from periods of abundance to periods of scarcity. Even though speculators may never see a barrel of oil or a Brazilian bond, they can help even out the price and yield differences of these items among regions or over time. They do this by buying when goods are abundant and prices are low and selling when goods are scarce and prices are high, and this indeed can improve a market's efficiency.

Arbitrage and Geographic Price Patterns

The simplest case is one in which speculative activity reduces or eliminates regional price differences by buying and selling the same commodity. This activity is called **arbitrage,** which is the purchase of a good or asset in one market for immediate resale in another market in order to profit from a price discrepancy.

Let's say that the price of wheat is 50 cents per bushel higher in Chicago than in Kansas City. Further, suppose that the costs of insurance and transportation are 10 cents per bushel. An *arbitrager* (someone engaged in arbitrage) can purchase wheat in Kansas City, ship it to Chicago, and make a profit of 40 cents per bushel. As a result of market arbitrage, the differential will be reduced so that the price difference between Chicago and Kansas City can never exceed 10 cents per bushel. *As a result of arbitrage, the price difference between markets will generally be less than the cost of moving the good from one market to the other.*

The frenzied activities of arbitragers—talking on the phone simultaneously to several brokers in several markets, searching out price differentials, trying to eke out a tiny profit every time they can buy low and sell high—tend to align the prices of identical products in different markets. Once again, we see the invisible hand at work—the lure of profit acts to smooth out price differentials across markets and make markets function more efficiently.

Speculation and Price Behavior over Time

Forces of speculation will tend to establish definite patterns of prices over time as well as over space. But the difficulties of predicting the future make this pattern less than perfect: we have an equilibrium that is constantly being disturbed but is always in the process of reforming itself—rather like a lake's surface under the play of the winds.

Consider the simplest case of a crop like corn that is harvested once a year and can be stored for future use. To avoid shortages, the crop must last for the entire year. Since no one passes a law regulating the storage of corn, how does the market bring about an efficient pattern of pricing and use over the year? The equilibrium is set by the activities of speculators trying to make a profit.

A well-informed corn speculator realizes that if all the corn is thrown on the market immediately after the autumn harvest, it will fetch a very low price because there will be a glut on the market. Several

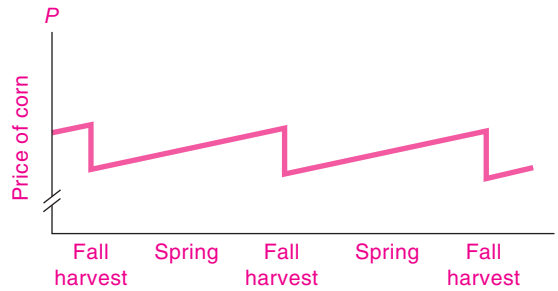


FIGURE 11-1. Speculators Even Out the Price of a Commodity over Time

When a good is stored, the expected price rise must match holding costs. In equilibrium, price is lowest at harvest time, rising gently with accumulated storage, insurance, and interest costs until the next harvest. This flexible pattern tends to even out consumption over the seasons. Otherwise, a harvest glut would cause very low autumn price and sky-high spring price.

months later, when corn is running short, the price will tend to skyrocket. In this case, speculators can make a profit by (1) purchasing some of the autumn crop while it is cheap, (2) putting it into storage, and (3) selling it later when the price has risen.

As a result of the speculative activities, the autumn price increases, the spring supply of corn increases, and the spring price declines. The process of speculative buying and selling tends to even out the supply, and therefore the price, over the year. Figure 11-1 shows the behavior of prices over an idealized yearly cycle.

Interestingly, if there is brisk competition among speculators, none of them will make excess profits. The returns to speculators will include the interest on invested capital, the appropriate earnings for their time, and a risk premium to compensate them for the noninsurable risks that they bear.

Speculation reveals the invisible-hand principle at work. By evening out supplies and prices, speculation actually increases economic efficiency. By moving goods over time from periods of abundance to periods of scarcity, the speculator is buying where the price and marginal utility of the good are low and selling where the price and marginal utility are high. By pursuing their private interests (profits), speculators are at the same time increasing the public interest (total utility).

Shedding Risks through Hedging

One important function of speculative markets is to allow people to shed risks through hedging. **Hedging** consists of reducing the risk involved in owning an asset or commodity by making an offsetting sale of that asset. Let's see how it works. Consider someone who owns a corn warehouse. She buys 2 million bushels of Kansas corn in the fall, stores it for 6 months, and sells it in the spring at a 10-cents-per-bushel profit, just covering her costs.

The problem is that corn prices tend to fluctuate. If the price of corn rises, she makes a large windfall gain. But if the price falls sharply, the decrease could completely wipe out her profits. How can the warehouse owner make a living storing only corn while avoiding the risks of corn-price fluctuations?

She can avoid the corn-price risk by *hedging her investments*. The owner hedges by selling the corn the moment it is bought rather than waiting until it is shipped 6 months later. Upon buying 2 million bushels of corn in September, she sells the corn immediately for delivery in the future at an agreed-upon price that will just yield a 10-cents-per-bushel storage cost. She thereby protects herself against all corn-price risk. *Hedging allows businesses to insulate themselves from the risk of price changes.*

The Economic Impacts of Speculation

But who buys the corn, and why? Someone agrees to buy the warehouse owner's corn now for future delivery. This buyer might be a baker who has a contract to sell bread in 6 months and wants to lock in the price. Or perhaps an ethanol plant needs corn for next year's production. Or the buyer might be a group of investors who believe that corn prices will rise and that they will therefore make a supernormal return on their investment. Someone, somewhere, and at the right price, has an economic incentive to take on the risk of corn-price fluctuations.

Speculative markets serve to improve the price and allocation patterns across space and time as well as to help transfer risks. If we look behind the veil of money, we see that ideal speculation reallocates goods from times of feast (when prices are low) to times of famine (when prices are high).

Our discussion has suggested that ideal speculative markets can increase economic efficiency. Let's see how. Say that identical consumers have utility

schedules in which satisfaction in one year is independent of that in every other year. Now suppose that in the first of 2 years there is a big crop—say, 3 units per person—while the second year has a small crop of only 1 unit per person. If this crop deficiency could be foreseen perfectly, how should the consumption of the 2-year, 4-unit total be spread over the 2 years? Neglecting storage, interest, and insurance costs, *total utility and economic efficiency for the 2 years together will be maximized only when consumption is equal in each year.*

Why is uniform consumption better than any other division of the available total? Because of the law of diminishing marginal utility. This is how we might reason: "Suppose I consume more in the first year than in the second. My marginal utility (*MU*) in the first year will be low, while it will be high in the second year. So if I carry some crop over from the first to the second year, I will be moving consumption from low-*MU* times to high-*MU* times. When consumption levels are equalized, *MU*s will be equal and I will be maximizing my total utility."

A graph can illuminate this argument. If we measure utility in dollars, with each dollar always denoting the same marginal utility, the demand curves for the risky commodity would look just like the marginal utility schedule of Figure 5-1 on page 85. The two curves of Figure 11-2(*a*) show what would happen with no carryover and with unequal consumption. Here, price is determined first at A_1, where higher S_1S_1 intersects *DD*, and second at A_2, where the lower supply S_2S_2 intersects *DD*. Total utility of the blue shaded areas would add up to only (4 + 3 + 2) + 4, or \$13.

But with optimal carryover of 1 unit to the second year, as shown in Figure 11-2(*b*), *P*s and *Q*s will be equalized at E_1 and E_2, and the total utility of the shaded areas will add up to (4 + 3) + (4 + 3), or \$14 per person. A little analysis can show that the gain in utility of \$1 is measured by Figure 11-2(*b*)'s dark green block, which represents the excess of the second unit's marginal utility over that of the third. This shows why the equality of marginal utilities, which is achieved by ideal speculation, is optimal.

While this discussion has focused on commodities, most speculation today involves financial assets such as stocks, bonds, mortgages, and foreign exchange. Every day, literally trillions of dollars of assets change hands as people speculate, hedge, and invest their

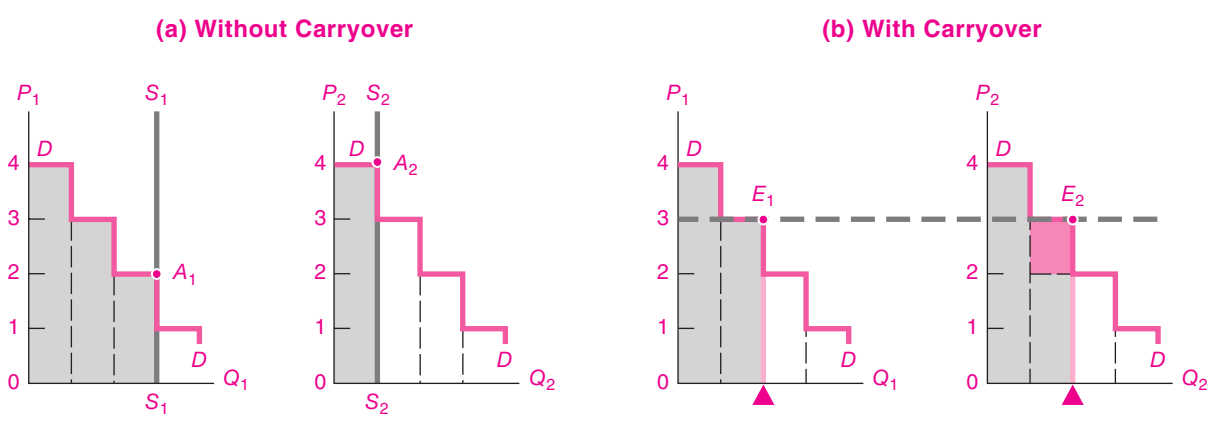


FIGURE 11-2. Speculative Storage Can Improve Efficiency

The blue areas measure total utility enjoyed each year. Carrying 1 unit to the second year equalizes *Q* and also *P* and *MU* and increases total utility by the amount of the dark green block.

This diagram will apply equally well to a number of situations. It could be labeled "**(a)** Without Arbitrage across Regional Markets" and "**(b)** With Arbitrage across Markets." We can also use this diagram to illustrate risk aversion if we label it "**(a)** With a Risky Gamble" and "**(b)** Without a Risky Gamble." Insurance then serves to move people from **(a)** to **(b)** by spreading the risks across many independent potential gambles.

funds. The general principles underlying financial speculation, hedging, and arbitrage are exactly the same as those outlined here, although the stakes are even higher.

Ideal speculation serves the important function of reducing undesired variations in consumption. In a world where individuals are averse to risk, speculation can increase total utility and allocational efficiency.

RISK AND UNCERTAINTY

What are people's attitudes toward risk? Why do people try to insulate themselves from many important risks? How can market institutions like insurance help individuals avoid major risks? Why do markets fail to provide insurance in some circumstances? We turn now to these issues.

Whenever you drive a car, own a house, join the army, or invest in the stock market, you are risking life, limb, or fortune. People generally want to avoid major risks to their income, consumption, and health. When people avoid risks, they are *risk-averse*.

A person is **risk-averse** when the pain from losing a given amount of income is greater in magnitude than the pleasure from gaining the same amount of income.

For example, suppose that we are offered a risky coin flip in which we will win \$1000 if the coin comes up heads and lose \$1000 if the coin comes up tails. This bet has an *expected value* of 0 (equal to a probability of ½ times \$1000 plus a probability of ½ times − \$1000). A bet which has a zero expected value is called a fair bet. If we turn down all fair bets, we are risk-averse.

In terms of the utility concept that we analyzed in Chapter 5, risk aversion is the same as *diminishing marginal utility of income*. Being risk-averse implies that the gain in utility achieved by getting an extra amount of income is less than the loss in utility from losing the same amount of income. For a fair bet (such as flipping a coin for \$1000), the expected dollar value is zero. But in terms of utility, the expected utility value is negative because the utility you stand to win is less than the utility you stand to lose.

Figure 11-2 illustrates the concept of risk aversion. Say that situation (*b*) is the initial position, in which you have equal amounts of consumption in states 1 and 2, consuming 2 units in both states. Someone comes to you and says, "Let's flip a coin for 1 unit." This person is in effect offering you the chance to move to situation (*a*), where you would have 3 units of consumption if the coin came up heads and 1 unit if tails. By careful calculation, you see that if you refuse the bet and stay in situation (*b*), the expected value of utility is 7 utils (= ½ × 7 utils + ½ × 7 utils), whereas if you accept the bet, the expected value of utility is 6½ utils (= ½ × 9 utils + ½ × 4 utils). This example shows that if you are risk-averse, with diminishing marginal utility, you will avoid actions that increase uncertainty without some expectation of gain.

Say that I am a corn farmer. While I clearly must contend with the weather, I prefer to avoid corn-price risks. Suppose that there are two equally likely outcomes with prices of \$3 and \$5 per bushel, so the expected value of the corn price is \$4 per bushel. Unless I can shed the price risk, I am forced into a lottery where I must sell my 10,000-bushel crop for either \$30,000 or \$50,000 depending upon the flip of the corn-price coin.

Because I am risk-averse, I would prefer a sure thing to such a lottery. The prospect of losing \$10,000 is more painful than the prospect of gaining \$10,000 is pleasant. If my income is cut to \$30,000, I will have to cut back on important spending, such as replacing an aging tractor. On the other hand, the extra \$10,000 might be less critical, going toward luxuries like a winter vacation. I therefore decide to hedge my price risk by selling my corn for the expected-value price of \$4 per bushel.

People are generally risk-averse, preferring a sure thing to uncertain levels of consumption: people prefer outcomes with less uncertainty and the same average values. For this reason, activities that reduce the uncertainties of consumption lead to improvements in economic welfare.

The Troubling Rise in Gambling

Gambling has historically been a "vice" that was—along with illegal drugs, commercial sex, alcohol, and tobacco—discouraged by the state. Attitudes about such activities ebb and flow. Over the last two decades, attitudes toward gambling

became permissive as those toward drugs and tobacco hardened. Overall, gambling has been one of the fastest-growing sectors of the (legal) economy.

Gambling is a different animal from speculation. While ideal speculative activity increases economic welfare, gambling raises serious economic issues. To begin with, aside from recreational value, gambling does not create goods and services. In the language of game theory, described in the previous chapter, gambling is a "negative-sum game" for the players—the customers are (almost) sure to lose in the long run because the house takes a cut of all bets. In addition, by its very nature, gambling increases income inequality. People who sit down to the gambling table with the same amount of money go away with widely different amounts. A gambler's family must expect to be on top of the world one week only to be living on crumbs and remorse when luck changes. Some observers also believe that gambling has adverse social impacts. These include addiction to gambling, neighborhood crime, political corruption, and infiltration of gambling by organized crime.

Given the substantial economic case against gambling, how can we understand the recent trend to legalize gambling and operate government lotteries? One reason is that when states are starved for tax revenues, they look under every tree for new sources; they rationalize lotteries and casinos as a way to channel private vices to the public interest by skimming off some of the revenues to finance public projects. In addition, legal gambling may drive out illegal numbers rackets and take some of the profitability out of organized crime. Notwithstanding these rationales, many observers raise questions about an activity in which the state profits by promoting irrational behavior among those who can least afford it.

B. THE ECONOMICS OF INSURANCE

Most people would like to avoid the risks of losing life, limb, and house. But risks cannot simply be buried. When a house burns down, when someone is hurt in an automobile accident, or when a hurricane destroys New Orleans—someone, somewhere, must bear the cost.

Markets handle risks by **risk spreading.** This process takes risks that would be large for one person and spreads them around so that they are but small risks for a large number of people. The major form of risk spreading is **insurance,** which is a kind of gambling in reverse.

For example, in buying fire insurance on a house, homeowners seem to be betting with the insurance company that the house will burn down. If it does not, the owners forfeit the small premium charge. If it does burn down, the company must reimburse the owners for the loss at an agreed-upon rate. What is true of fire insurance is equally true of life, accident, automobile, or any other kind of insurance.

The insurance company is spreading risks by pooling many different risks: it may insure millions of houses or lives or cars. The advantage for the insurance company is that what is unpredictable for one individual is highly predictable for a population. Say that the Inland Fire Insurance Company insures 1 million homes, each worth $100,000. The chance that a house will burn down is 1 in 1000 per year. The expected value of losses to Inland is then .001 × $100,000 = $100 per house per year. Inland charges each homeowner $100 plus another $100 for administration and for reserves.

Each homeowner is faced with the choice between the *certain* loss of $200 for each year or the *possible* 1-in-1000 catastrophic loss of $100,000. Because of risk aversion, the household will choose to buy insurance that costs more than the expected value of the household's loss in order to avoid the small chance of a catastrophic loss. Insurance companies can set a premium that will earn the company a profit and at the same time produce a gain in expected utility of individuals. Where does the economic gain come from? It arises from the law of diminishing marginal utility.

Insurance breaks large risks into small pieces and then sells these smaller pieces in return for a small risk premium. Although insurance appears to be just another form of gambling, it actually has the opposite effect. Whereas nature deals us risks, insurance helps reduce individual risks by spreading them out.

Capital Markets and Risk Sharing

Another form of risk sharing takes place in the capital markets because the financial ownership of *physical* capital can be spread among many owners through the vehicle of corporate *financial* ownership.

Take the example of investment to develop a new commercial aircraft. A completely new design, including research and development, might require $5 billion of investment spread over 10 years. Yet there is no guarantee that the plane will find a large-enough commercial market to repay the invested funds. Few people have the wealth or inclination to undertake such a risky venture.

Market economies accomplish this task through publicly owned corporations. A company like Boeing is owned by millions of people, none of whom owns a major portion of the shares. In a hypothetical case, divide Boeing's ownership equally among 10 million individuals. Then the $5 billion investment becomes $500 per person, which is a risk that many would be willing to bear if the returns on Boeing stock appear attractive.

By spreading the ownership of risky investments among a multitude of owners, capital markets can spread risks and encourage much larger investments and risks than would be tolerable for individual owners.

MARKET FAILURES IN INFORMATION

Our analysis up to now has assumed that investors and consumers are well informed about the risks they face and that speculative and insurance markets function efficiently. In reality, markets involving risk and uncertainty are plagued by market failures. Two of the major failures are adverse selection and moral hazard. When these are present, markets may give the wrong signals, incentives may get distorted, and sometimes markets may simply not exist. Because of market failures, governments may decide to step in and offer social insurance.

Moral Hazard and Adverse Selection

While insurance is a useful device for reducing risks, sometimes insurance is not available. The reason is that efficient insurance markets can thrive only under limited conditions.

What are the conditions for efficient insurance markets? First, there must be a large number of insurable events. Only then will companies be able to spread the risks so that what is a large risk to an individual will become a small risk to many people.

Moreover, the events must be statistically independent. No prudent insurance company would sell all its fire-insurance policies in the same building or sell only hurricane insurance in Miami. Insurance companies try to diversify their coverage among many independent risks.

Additionally, there must be sufficient experience regarding such events so that insurance companies can reliably estimate the losses. For example, after the September 11 terrorist attacks, private terrorism insurance was canceled because insurance companies could not get reliable estimates of the chances of future attacks (see question 3 at the end of this chapter).

Finally, the insurance must be relatively free of moral hazard. **Moral hazard** is at work when insurance increases risky behavior and thereby changes the probability of loss. In many situations moral hazard is unimportant. Few people will risk death because they have a generous life-insurance policy. In some areas, moral hazard is severe. Studies indicate that the presence of insurance increases the amount of cosmetic surgery, and most medical-insurance policies consequently exclude this procedure.

When these ideal conditions are met—when there are many outcomes, all more or less independent, and when the probabilities can be accurately gauged and are not contaminated by moral hazard—private insurance markets can function efficiently.

Sometimes, private insurance is limited or expensive because of adverse selection. **Adverse selection** arises when the people with the highest risk are also those who are most likely to buy the insurance. Adverse selection can lead to a market where only the people with the highest risks are insured, or even to a situation where there is no market at all.

A good example occurs when a company is offering life insurance to a population made up of smokers and nonsmokers. Suppose the company cannot determine whether a person is a smoker, or perhaps there is a government policy which says that companies cannot differentiate among people on the basis of their personal behavior. However, people know their smoking habits. We see here the phenomenon of asymmetric information between buyer and seller. **Asymmetric information** occurs when buyers and sellers have different information on important facts, such as a person's health status or the quality of a good being sold.

Suppose that the company starts by setting a price based on the average mortality rate of the population. At this price, many smokers buy the insurance, but most nonsmokers do not. This means that people have sorted themselves unfavorably for the company—there is adverse selection. Soon the data begin to come in, and the company learns that its experience is much worse than it had forecast.

What happens next might be that the company raises the premiums on its insurance. As the price rises, more of the nonsmokers drop out, and the experience becomes even worse. Perhaps the price rises so high that even the smokers stop buying insurance. In the worst case, the market just dries up completely.

We see that the policy of uniform market pricing has led to adverse selection—raising the cost, limiting the coverage, and producing an incomplete market. Another example is the market for "lemons" such as used cars, where only the worst cars are sold, and the prices of used cars in equilibrium are reduced. Such market failures are particularly severe when there is asymmetric information between buyers and sellers.

Would You Invest in a Company for Grade Insurance?

A friend of yours proposes the following scheme: He wants you to invest in a start-up company called G-Insure.com, which offers grade insurance for students. In return for a modest premium, the company promises to compensate students for 100 percent of the income loss from poor grades. This seems like a good idea because income risks are very large for most people.

On reflection, can you see why G-Insure.com is almost sure to be a bad idea? The reason is that grades depend too much on individual effort and the market would therefore be infected with moral hazard and adverse selection. Students would be tempted to study less (moral hazard), and students who expected to have lower grades would be more prone to buy grade insurance (adverse selection). These problems might even produce a "missing market"—one in which supply and demand intersect at a zero level of grade insurance. So the company will either have no business or lose piles of money.

SOCIAL INSURANCE

When market failures are so severe that the private market cannot provide coverage in an effective manner, governments turn to **social insurance.** This consists of mandatory programs, with broad or universal coverage, funded by taxes or fees. These programs are insurance because they cover risky situations such as unemployment, illness, or low incomes during retirement. The taxing and regulatory powers of the government, plus the ability to prevent adverse selection through universal coverage, can make government insurance a welfare-improving measure. The rationale for social insurance was explained as follows by the distinguished public-policy economist Martin Feldstein:[1]

> There are two distinct reasons for providing social insurance. Both reflect the asymmetry of information. The first is that asymmetric information weakens the functioning of private insurance markets. The second is the inability of the government to distinguish between those who are poor in old age or when unemployed because of bad luck or an irrational lack of foresight from those who are intentionally "gaming" the system by not saving in order to receive transfer payments.

The key point is that social insurance is provided when the requirements of private insurance are not met. Perhaps the risks are not independent, as when many people simultaneously become unemployed in a recession. Perhaps adverse selection is serious, as when people choose to buy catastrophic health insurance soon after they learn they have a terrible disease. Perhaps the risks cannot be easily evaluated, as in the case of insurance against terrorist attacks. In each of these cases, the private market functioned poorly or not at all, so the government stepped in with social insurance.

Let's spend a moment on the example of unemployment insurance. This is an example of a private market that cannot function because so many of the requirements for private insurance are violated: moral hazard is high (people may decide to become unemployed if benefits are generous); there is severe adverse selection (those who often lose jobs are more likely to participate); spells of unemployment are not independent (they tend to occur together

[1] See the reference in this chapter's Further Reading section.

during business-cycle recessions); and business cycles are unpredictable, so the risks cannot be accurately measured. At the same time, some countries feel that people should have a safety net under them should they lose their job. As a result, governments generally step in to provide unemployment insurance.

The next section discusses the important case of government-provided health care, which for many countries is the largest program of social insurance.

Social insurance is provided by governments when private insurance markets cannot function effectively and society believes that individuals should have a social safety net for the most severe risks such as unemployment, illness, and low incomes.

C. HEALTH CARE: THE PROBLEM THAT WON'T GO AWAY

Health care is the single largest government program of the U.S. federal government. For 2008, expenditures on health care totaled close to $700 billion—larger even than the military budget. Most of this spending was on the social insurance program called Medicare, which provides subsidized health care for the elderly. The balance was health care for the poor, the disabled, and veterans.

The U.S. health-care system is controversial both because it is expensive and because a large number of people are not covered by insurance or other programs. Health-care spending rose from 4 percent of national output (GDP) in 1940 to 7 percent in 1970 and reached 16 percent in 2008. Yet almost 16 percent of the nonelderly population has no coverage. This has been called the problem that can't be solved and won't go away.

THE ECONOMICS OF MEDICAL CARE

Why has health care been so controversial? In the United States, the health-care system is a partnership between the market system and the government. In recent years, this system has produced some remarkable accomplishments. Many terrible diseases, such as smallpox and polio, have been eradicated. Life expectancy—one of the key indexes of health—has improved more in developing countries since 1900 than it did during the entire prior span of recorded history. Advances in medical technology—from arthroscopic knee surgery to sophisticated anticancer drugs—have enabled more people to live pain-free and productive lives.

Even with these great achievements, major health problems remain unsolved in the United States: Infant mortality is higher than in many countries with lower incomes; many Americans are without health insurance coverage; great disparities in care exist between the rich and the poor; and communicable diseases like AIDS and tuberculosis are spreading.

The issue that most concerns the public, the business community, and political leaders is the exploding cost of health care. Virtually everyone agrees that the U.S. health system has contributed greatly to the nation's health, but many worry that it is becoming unaffordable.

Special Economic Features of Health Care

The health-care system in the United States has three characteristics that have contributed to the rapid growth of the health-care sector in recent years: a high income elasticity, rapid technological advance, and the increasing insulation of consumers from prices.

Health care has a high income elasticity, indicating that ensuring a long and fit life becomes increasingly important as people are able to pay for other essential needs. Goods with high income elasticities, other things held constant, tend to take a growing share of consumer income as income rises.

Health care has enjoyed rapid improvements in medical technology over the last century. Advances in fundamental biomedical knowledge, discovery and use of a wide variety of vaccines and pharmaceuticals, progress in understanding the spread of communicable diseases, and increasing public awareness of the role of individual behavior in areas such as smoking, drinking, and driving—all these have contributed to the remarkable improvement in the health of Americans. The new and improved technologies have created new markets and stimulated spending in the health-care sector.

Additionally, spending on health care has risen rapidly because of the increased subsidization of

medical care over recent decades. Health-care coverage in the United States is largely provided by employers as a tax-free fringe benefit. Tax-free status is, in effect, a government subsidy. In 1960, 60 percent of medical expenses were paid directly by consumers; by 2007, only 15 percent of medical expenses were paid out of pocket. This phenomenon is sometimes called the "third-party payment effect" to indicate that when a third party pays the bill, the consumer often ignores the cost.

All these forces (high income elasticity, the development of new technologies, and the increasing scope of third-party payments) contribute to the rapid growth of expenditures on medical care.

Health Care as a Social Insurance Program

Why is health care a social insurance program? Three reasons are cited by experts on health-care economics:

1. Many parts of the health-care system, such as the prevention of communicable diseases and the development of basic science, are *public goods* that the market will not provide efficiently. Eradication of smallpox benefited billions of potential victims, yet no firm could collect even a small fraction of the benefits of the eradication program. When one person stops smoking because of knowledge of its dangers, or when another person uses condoms after learning how AIDS is transmitted, these activities are no less valuable to others. This syndrome will lead to underinvestment in public health improvements by the market.
2. A second set of market failures arises because of the failure of private insurance markets. One significant reason for this failure is the presence of *asymmetric information* among patients, doctors, and insurance companies. Medical conditions are often isolated occurrences for patients, so such asymmetric information between doctors and patients means that patients may be completely dependent upon doctors' recommendations regarding the appropriate level of health care. Sometimes, as when patients are wheeled into the emergency room, they may be incapacitated and unable to choose treatment strategies for themselves, so demand depends even more upon the recommendations of the suppliers. Special protection must be given to ensure that consumers do not unwittingly purchase unnecessary, poor-quality, or high-cost services.

   There are also informational asymmetries between the patient and the insurance provider. People may know more about their medical condition than do insurance companies. Low-risk individuals may choose not to buy health insurance. This leads to *adverse selection*, which increases the average riskiness of the group and subsequently increases the cost for those who do participate. It is not surprising that healthy people in their twenties are those most likely to be uninsured.
3. A third concern of government policy is *equity*—to provide a minimum standard of medical care for all. In part, good health care is increasingly viewed as a basic right in wealthy countries. But good health care is also a good social investment. Inadequate health care is particularly harmful for poor people not only because they tend to be sicker than wealthier individuals but also because their incomes are almost entirely derived from their labor. A healthier population is a more productive population because healthier people have higher earnings and require less medical care.

   Inadequate health care is most costly for children. The medical condition of poor and minority children in the United States has in some dimensions actually worsened in recent years. Sick children are handicapped from the start: they are less likely to attend school, perform more poorly when they do attend, are more likely to drop out, and are less likely to get good jobs with high pay when they grow up. No country can prosper when a significant fraction of its children have inadequate medical care.

Rationing Health Care

Whether or not a country provides equal health care for all its residents, health care must be rationed because supply is limited. Until we get to the point where every symptom of every hypochondriac can be extensively examined, probed, and treated, it will be necessary to leave some perceived medical need unsatisfied. There is no choice but to ration health care.

However, it is not obvious *how* we are to ration such a good. Most goods and services are rationed

by the purse. Prices ration out the limited supply of fancy cars and mansions, as well as the not-so-fancy food and shoes, to those who most want and can afford them. In many areas of health care, by contrast, we do not allow prices to ration out services to the highest bidders. For example, we do not auction off liver transplants or blood or emergency-room access to the highest bidder. Rather, we desire that these goods be allocated equitably.

The subsidization of health care leads to shortages, and demand for the good must therefore be limited in some other way. This phenomenon is known as *nonprice rationing*. Many of us have experienced this kind of rationing when we wait in line for a good or service. When price is not allowed to rise to balance supply and demand, some other mechanism must be found to "clear the market."

Figure 11-3 illustrates nonprice rationing in the medical market. Suppose that there are only Q_0 units of medical care available with a consumer demand function of *DD*. The market-clearing price would come at *C*, where quantities supplied and demanded are equal. However, because the consumer pays only 20 percent of the costs out of pocket, the quantity demanded is Q_1. The *AB* segment is unsatisfied demand, which is subject to nonprice rationing; the greater the subsidy, the more nonprice rationing must be used.

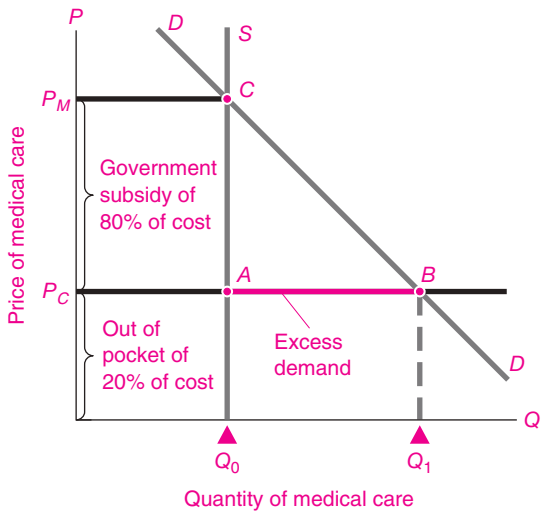


FIGURE 11-3. Free Health Care Leads to Nonprice Rationing

When governments provide free or subsidized access to medical care, some way must be found to ration out the limited services. In the example of a government subsidy, when the quantity demanded exceeds the quantity supplied, the excess demand *AB* must be choked off by some mechanism other than price. Most often, people must wait for nonemergency services, sometimes for hours, sometimes for months.

Health care is an economic commodity like shoes and gasoline. Physicians' services, nursing care, hospitalization, and other services are limited in supply. The demands of consumers—summing up the critical, the reasonable, the marginal, and the nonsensical—outstrip the available resources. But the resources must somehow be rationed out. Rationing of health care according to dollar votes is unacceptable because it does too much damage to the public health, leaves crucial demands unmet, and impoverishes many. What should be the scope of the market, and what nonmarket mechanism should be used where the market is supplanted? These questions are the crux of the great debate about medical care.

D. INNOVATION AND INFORMATION

One of the most important topics in economics is the economics of information. Information includes things as varied as e-mails, songs, new vaccines, and even the textbook you are reading. Information is a very different kind of commodity from things like pizza and shoes because information is expensive to produce but cheap to reproduce. Because of the unusual nature of information, it is subject to market failures, so we need to develop different kinds of public policies to regulate it—the law of "intellectual property."

Schumpeter's Radical Innovation

We set the stage for our discussion by returning to the economics of imperfect competition that we discussed in the previous two chapters. We learned that imperfect competitors set prices too high, earn supernormal profits, and neglect product quality.

This dismal view of monopoly was challenged by one of the great economists of the last century,

Joseph Schumpeter. He argued that the essence of economic development is innovation and that monopolists are in fact the wellsprings of innovation in a capitalist economy.

Joseph Schumpeter: Economist as Romantic

Born in the Austrian Empire, Joseph Schumpeter (1883–1950), a legendary scholar whose research ranged widely in the social sciences, led a flamboyant private life.

He began studying law, economics, and politics at the University of Vienna—then one of the world centers of economics and the home of the "Austrian School" that today reveres laissez-faire capitalism. As a professor, he was often the champion of his students. Six months into his teaching career, he charged into the library and scolded the librarian for not allowing his students to have free use of the books. After trading insults, the librarian challenged Schumpeter to a duel. Schumpeter won by nicking the librarian on the shoulder, and his students thereupon had unlimited access to the library.

In between dueling, insulting the stodgy faculty by showing up at faculty meetings in riding pants, and carousing, Schumpeter devoted himself to introducing economic theory to the European continent, founding the Econometric Society, and traveling to England and America. He later moved to Harvard University, where he eventually became embittered as the theories of his great rival, John Maynard Keynes, swept the profession.

Schumpeter's writings covered much of economics, sociology, and history, but his first love was economic theory. Schumpeter's early classic, *The Theory of Economic Development* (1911), broke with the traditional static analysis of its time by emphasizing the importance of the entrepreneur or innovator, the person who introduces "new combinations" in the form of new products or methods of organization. Innovations result in temporary supernormal profits, which are eventually eroded away by imitations. Ever the romantic, Schumpeter saw in the entrepreneur the hero of capitalism, the person of "superior qualities of intellect and will," motivated by the will to conquer and the joy of creation.

His magisterial *History of Economic Analysis* (published posthumously in 1954) is a superb survey of the emergence of modern economics. His "popular" book, *Capitalism, Socialism, and Democracy* (1942), laid out his startling hypothesis on the technological superiority of monopoly and developed the theory of competitive democracy, which later grew into public-choice theory. (See question 7 at the end of this chapter.) He ominously predicted that capitalism would wither away because of disenchantment among the elites. Were he alive today, he might well join in the conservative complaint that the welfare state drains the economic vitality from the market economy.

The Economics of Information

Modern economics emphasizes the special problems involved in the **economics of information.** Information is a fundamentally different commodity from normal goods. Because information is costly to produce but cheap to reproduce, markets in information are subject to severe market failures.

Consider the production of a software program, such as Windows Vista. Developing this program took several years and cost Microsoft many billions of dollars. Yet you can purchase a legal copy for about $220 or buy an illegal pirated copy for $5. The same phenomenon is at work in pharmaceuticals, entertainment, and other areas where much of the value of a good comes from the information it contains. In each of these areas, the research and development on the product may be an expensive process that takes years. But once the information is recorded on paper, in a computer, or on a compact disc, it can be reproduced and used by a second person essentially for free.

The inability of firms to capture the full monetary value of their inventions is called **inappropriability.** Inventions are not fully appropriable because other firms may imitate or pirate an invention, in which case the other firms may derive some of the benefits of the inventive investments; sometimes, imitators may drive down the price of the new product, in which case consumers would get some of the rewards. Case studies have found that the *social return* to invention (the value of an invention to all consumers and producers) is many times the appropriable *private return* to the inventor (the monetary value of the invention to the inventor).

Information is expensive to produce but cheap to reproduce. To the extent that the rewards to invention are inappropriable, we would expect private research and development to be underfunded, with the most significant underinvestment in basic

research because that is the least appropriable kind of information. The inappropriability and high social return on research lead most governments to subsidize basic research in the fields of health and science and to provide special incentives for other creative activities.

Intellectual Property Rights

Governments have long recognized that creative activities need special support because the rewards for producing valuable information are reduced by imitation. The U.S. Constitution authorizes Congress "to promote the Progress of Science and useful Arts, by securing, for limited Times, to Authors and Inventors, the exclusive Right to their respective Writings and Discoveries." Thus special laws governing patents, copyrights, business and trade secrets, and electronic media create **intellectual property rights.** The purpose is to give the owner special protection against the material's being copied and used by others without compensation to the owner or original creator.

The earliest intellectual property right was the **patent,** under which the U.S. government creates an exclusive use (in effect, a limited monopoly) over a "novel, nonobvious, and useful" invention for a limited period, currently set at 20 years. Similarly, copyright laws provide legal protection against unauthorized copying of original works in different media such as text, music, video, art, software, and other information goods.

Why would governments actually encourage monopolies? In effect, patents and copyrights grant property rights to inventors over books, music, and ideas. By allowing inventors to have exclusive use of their intellectual property, the government increases the degree of appropriability and thereby increases the incentives for people to invent useful new products, write books, compose songs, and write computer software. A patent also requires disclosure of the technological details of the invention, which encourages further invention and lawful imitation. Examples of successful patents include those on the cotton gin, the telephone, the Xerox machine, and many profitable drugs.

The Dilemma of the Internet

Inventions that improve communications are hardly limited to the modern age. But the rapid growth of electronic storage, access, and transmission of information highlights the dilemma of providing incentives for creating new information. Many new information technologies have large up-front or sunk costs but virtually zero marginal costs. With the low cost of electronic information systems like the Internet, it is technologically possible to make large amounts of information available to everyone, everywhere, at close to zero marginal cost. Perfect competition cannot survive here because a price equal to a zero marginal cost will yield zero revenues and therefore no viable firms.

The economics of the information economy highlights the conflict between efficiency and incentives. On the one hand, all information might be provided free of charge—free economics textbooks, free movies, free songs. Free provision of information looks economically efficient because the price would thereby be equal to the marginal cost, which is zero. But a zero price on intellectual property would destroy the profits and therefore reduce the incentives to produce new books, movies, and songs because creators would reap little return from their creative activity. Society has struggled with this dilemma in the past. But with the costs of reproduction and transmission so much lower for electronic information than for traditional information, finding sensible public policies and enforcing intellectual property rights is becoming ever more difficult.

Experts emphasize that intellectual property laws are often hard to enforce, especially when they apply across national borders. The United States has a long-running trade dispute charging that China condones the illegal copying of American movies, musical recordings, and software. A DVD movie that sells for $25 in the United States can be purchased for 50 cents in China. The U.S. copyright industries estimate that 85 to 95 percent of all their members' copyrighted works sold in 2007 in China were pirated.

In a world increasingly devoted to developing new knowledge—much of it intangible, like music, movies, patents, new drugs, and software—governments must find a middle ground in intellectual property rights. If intellectual property rights are too strong, this will lead to high prices and monopoly losses, while too weak intellectual property laws will discourage invention and innovation.

SUMMARY

A. Economics of Risk and Uncertainty

1. Economic life is full of uncertainty. Consumers face uncertain incomes and employment patterns as well as the threat of catastrophic losses; businesses have uncertain costs, and their revenues contain uncertainties about price and production.
2. In well-functioning markets, arbitrage, speculation, and insurance help smooth out the unavoidable risks. Speculators are people who buy and sell assets or commodities with an eye to making profits on price differentials across markets. They move goods across regions from low-price to high-price markets, across time from periods of abundance to periods of scarcity, and even across uncertain states of nature to periods when chance makes goods scarce.
3. The profit-seeking action of speculators and arbitragers tends to create certain equilibrium patterns of price over space, time, and risks. These market equilibria are zero-profit outcomes where the marginal costs and marginal utilities in different regions, times, or uncertain states of nature are in balance. To the extent that speculators moderate price and consumption instability, they are part of the invisible-hand mechanism that performs the socially useful function of reallocating goods from feast times (when prices are low) to famine times (when prices are high).
4. Speculative markets allow individuals to hedge against unwelcome risks. The economic principle of risk aversion, which derives from diminishing marginal utility, implies that individuals will not accept risky situations with zero expected value. Risk aversion implies that people will buy insurance to reduce the potentially disastrous declines in utility from fire, death, or other calamities.

B. The Economics of Insurance

5. Insurance and risk spreading tend to stabilize consumption in different states of nature. Insurance takes large individual risks and spreads them so broadly that they become acceptable to a large number of individuals. Insurance is beneficial because, by helping to equalize consumption across different uncertain states, it raises the expected level of utility.
6. The conditions necessary for the operation of efficient insurance markets are stringent: there must be large numbers of independent events and little chance of moral hazard or adverse selection. When market failures such as adverse selection arise, prices may be distorted or markets may simply not exist.
7. If private insurance markets fail, the government may step in to provide social insurance. Social insurance is provided by governments when private insurance markets cannot function effectively and society believes that individuals should have a social safety net for major risks such as unemployment, illness, and low incomes. Even in the most laissez-faire of advanced market economies today, governments insure against unemployment and health risks in old age.

C. Health Care: The Problem That Won't Go Away

8. Health care is the largest social insurance program. The health-care market is characterized by multiple market failures that lead governments to intervene. Health-care systems have major externalities. Additionally, the asymmetric information between doctors and patients leads to uncertainties about the appropriate treatment and level of care, and the asymmetry between patients and insurance companies leads to adverse selection in the purchase of insurance. Finally, because health care is so important to human welfare and to labor productivity, most governments strive to provide a minimum standard of health care to the population.
9. When the government subsidizes health care and attempts to provide universal coverage, there will be excess demand for medical services. One of the challenges is to develop efficient and equitable mechanisms of nonprice rationing.

D. Innovation and Information

10. Schumpeter emphasized the importance of the innovator, who introduces "new combinations" in the form of new products and new methods of organization and is rewarded by temporary entrepreneurial profits.
11. Today, the economics of information emphasizes the difficulties involved in the efficient production and distribution of new and improved knowledge. Information is different from ordinary goods because it is expensive to produce but cheap to reproduce. The inability of firms to capture the full monetary value of their inventions is called inappropriability. To increase appropriability, governments create intellectual property rights governing patents, copyrights, trade secrets, and electronic media. The growth of electronic information systems like the Internet has increased the dilemma of how to efficiently price information services.

CONCEPTS FOR REVIEW

Risk, Uncertainty, and Insurance

arbitrage leading to regional equalization of prices
ideal seasonal price pattern
speculation, arbitrage, hedging
risk aversion and diminishing marginal utility
consumption stability vs. instability
insurance and risk spreading
moral hazard, adverse selection
social insurance
nonprice rationing

Economics of Information

information economics
inappropriability, protection of intellectual property rights, dilemma of efficient production of knowledge
market failure in information

FURTHER READING AND INTERNET WEBSITES

Further Reading

The concept of social insurance was described by Martin Feldstein in "Rethinking Social Insurance," *American Economic Review,* March 2005 and available at *www.nber.org/feldstein/aeajan8.pdf.*

For an analysis of gambling, see William R. Eadington, "The Economics of Casino Gambling," *Journal of Economic Perspectives,* Summer 1999.

The Schumpeterian hypothesis was developed in Joseph Schumpeter, *Capitalism, Socialism, and Democracy* (Harper & Row, New York, 1942).

Many of the economic, business, and policy issues involved in the new information economy are covered in a nontechnical book by two eminent economists, Carl Shapiro and Hal R. Varian, *Information Rules* (Harvard Business School Press, Cambridge, Mass., 1998). A discussion of the economics of the Internet is contained in Jeffrey K. MacKie-Mason and Hal Varian, "Economic FAQs about the Internet," *Journal of Economic Perspectives*, Summer 1994, p. 92.

A discussion by the U.S. government of Chinese infringement of intellectual property rights can be found at *www.ustr.gov/Document_Library/Reports_Publications/Section_Index.html.*

Websites

One of the most interesting websites about the Internet and intellectual property rights is compiled by Hal R. Varian, chief economist of Google and former dean of the School of Information Management and Systems at the University of California at Berkeley. This site, called "The Economics of the Internet, Information Goods, Intellectual Property and Related Issues," is at *www.sims.berkeley.edu/resources/infoecon*.

Information on the American health-care system is usefully compiled by the National Center on Health Statistics at *www.cdc.gov/nchs/*.

QUESTIONS FOR DISCUSSION

1. Suppose a friend offers to flip a fair coin, with you paying your friend $100 if it comes up heads and your friend paying you $100 if it comes up tails. Explain why the expected dollar value is $0. Then explain why the expected utility value is negative if you are risk-averse.
2. Consider the example of grade insurance (see page 218). Suppose that with a grade-insurance policy, students would be compensated $5000 a year for each point that their grade point average fell below the top grade (the resulting number might be an estimate of the impact of grades on future earnings). Explain why the presence of grade insurance would produce moral hazard and adverse selection. Why would moral hazard and adverse selection make insurance companies reluctant to sell grade insurance? Are you surprised that you cannot buy grade insurance?
3. After the terrorist attacks of September 11, 2001, most insurance companies canceled their insurance coverage for terrorism. According to President Bush, "More than $15 billion in real estate transactions have been canceled or put on hold because owners and investors could not obtain the insurance protection they need."

As a result, the federal government stepped in to provide coverage for up to $90 billion in claims. Using the principles of insurance, explain why insurance companies might decline to insure property against terrorist attacks. Explain whether or not you think the federal program is an appropriate form of social insurance.

4. In the early nineteenth century, little of the nation's agricultural output was sold in markets, and transportation costs were very high. What would you expect to have been the degree of price variation across regions as compared with that of today?
5. Assume that a firm is making a risky investment (say, spending $2 billion developing a competitor to Windows). Can you see how the diversified ownership of this firm could allow near-perfect risk spreading on the software investment?
6. Health insurance companies sometimes do not allow new participants to be covered on "existing conditions," or preexisting illnesses. Explain why this policy might alleviate problems of adverse selection.
7. Joseph Schumpeter wrote as follows:

> The modern standard of life of the masses evolved during the period of relatively unfettered "big business." If we list the items that enter the modern workman's budget and, from 1899 on, observe the course of their prices, we cannot fail to be struck by the rate of the advance which, considering the spectacular improvement in qualities, seems to have been greater and not smaller than it ever was before. Nor is this all. As soon as we inquire into the individual items in which progress was most conspicuous, the trail leads not to the doors of those firms that work under conditions of comparatively free competition but precisely to the doors of the large concerns—which, as in the case of agricultural machinery, also account for much of the progress in the competitive sector—and a shocking suspicion dawns upon us that big business may have had more to do with creating that standard of life than keeping it down. (*Capitalism, Socialism, and Democracy*)

Use this passage to describe the tradeoff between "static" monopoly inefficiencies and "dynamic" efficiencies of technological change.

8. Long-term care for the elderly involves helping individuals with activities (such as bathing, dressing, and toileting) that they cannot perform for themselves. How were these needs taken care of a century ago? Explain why moral hazard and adverse selection make long-term-care insurance so expensive today that few people choose to buy it.
9. Economic studies have found that the private rate of return on inventions is typically as low as one-third of the social return. Explain this finding in terms of the economics of innovation.