2

Basic Microeconomics

The principal antitrust statutes in America were put into place over a hundred years ago. At that time economic theory offered little understanding of market outcomes beyond Adam Smith's original and intuitive insights. The formal modeling of those insights and of the benefits of competition versus monopoly were just beginning to appear in professional academic works, most notably, Alfred Marshall's *Principles of Economics*, vol. 1 (1890). A similarly rigorous understanding of what happens in that gray area between competition and monopoly would take some time to be developed and then some more time to be worked into the economics curriculum. Yet a sound understanding of the perfectly competitive and pure monopolized markets is, even by itself, quite insightful. Indeed, these models continue to provide useful starting points for interpreting much of what one reads about in the daily business press. They also reveal the primary intellectual force behind public policies designed to limit monopoly power. For all these reasons, we undertake in this chapter a review of the basic models of perfect competition and monopoly.

2.1 COMPETITION VERSUS MONOPOLY: THE POLES OF MARKET PERFORMANCE

Our review of the perfect competition and monopoly models is necessarily brief. We focus on firm profit-maximizing behavior and the resultant market outcome that such behavior implies. We take as given the derivation of an aggregate consumer demand for the product that defines the market of interest. This market demand curve describes the relationship between how much money consumers are willing to pay per unit of the good and the aggregate quantity of the good consumed. Figure 2.1 shows an example of a market demand curve—more specifically, a linear market demand curve, which can be described by the equation $P = A - BQ$. When we write the demand curve in this fashion with price on the left-hand side it is often called an inverse demand curve.[1] The vertical intercept A is the maximum willingness to

[1] The reason for this terminology is that traditionally in microeconomics, we think of quantity demanded as being the dependent variable, (left-hand side of the equation) and price, the independent variable, (right-hand side of the equation). However, when firms choose quantities and price adjusts to clear the market, it is preferable to put market price on the left-hand side, hence, the inverse demand function. Our discussion should make clear that the market demand curve can be thought of as the horizontal summation of the individual demand curve of each consumer. It is not, however, the horizontal summation of the demand curve facing each firm.

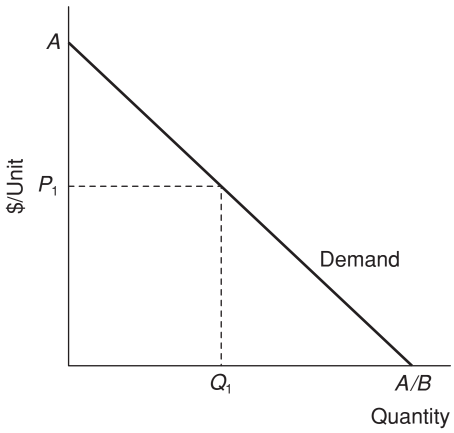

Figure 2.1 Market demand curve
The price P_1 is the marginal consumer valuation of an additional unit of output when current output is Q_1.

pay, or maximum reservation demand price that any consumer is willing to pay to have this good. At market prices greater than *A*, no one in this market wants to buy the product. As market price falls below *A*, demand for the product increases. For example if the market price of the good is P_1 then consumers will desire to purchase a quantity Q_1 of the good. The price P_1 is the most any consumer would pay to consume the last or the Q_1th unit of the good. The price P_1 describes consumer willingness to pay at the margin.

When we draw a demand curve we are implicitly thinking of some period of time over which the good is consumed. For example, we may want to look at consumer demand for the product per week, or per quarter, or per year. Similarly when we talk about firms producing the good, we want to consider their corresponding weekly, quarterly, or annual production of the good. The temporal period over which we define consumer demand and firm production typically affects what production technologies are available to the firm for producing the good. The shorter the time period, the fewer options any firm has for acquiring or hiring more inputs for use in production. Following the tradition in microeconomics, we distinguish between two time periods: the short-run and the long-run production periods. The short run is a sufficiently short time period for the industry so that no new production facilities—no new plant and equipment—can be brought on line. In the short run, neither the number of firms nor the fixed capital at each firm can be changed. By contrast, the long run is a production period sufficiently long so that firms can build new production facilities to meet market demand.

For either the short-run or the long-run scenario we are interested in determining when a market is in equilibrium. By this we mean finding an outcome at which the market is "at rest." A useful interpretation of a market equilibrium is a situation in which no consumer and no firm in the market has an incentive to change its decision on how much to buy or how much to sell. The precise meaning of this definition may vary depending on whether we consider the short run or the long run. In either case, the essential feature is the same. Equilibrium requires that no one has an incentive to change his or her trading decision.

2.1.1 Perfect Competition

A perfectly competitive firm is a "price taker." The price of its product is not something that the perfectly competitive firm chooses. Instead, that price is determined by the interaction of all the firms and consumers in the market for this good, and it is beyond the influence of any one of the perfectly competitive firms. This characterization only makes sense if each firm's potential supply of the product is "small" relative to market demand for the product. If a firm's supply of a good were large relative to the market, we would expect that that firm could influence the price at which the good was sold. An example of a "small" firm would be a wheat farmer in Kansas or, alternatively, a broker on the New York Stock Exchange trading IBM stock. Each is so small that any feasible change in behavior leaves the prices of wheat and IBM stock, respectively, unchanged.

Because a perfectly competitive firm cannot influence the market price at which the good trades, the firm perceives that it can sell as much, or as little, as it wants to at that price. If the firm cannot sell as much as it wants to at the market price, then the implication is that selling more would require a fall in the price. But this would imply that the firm has some power over the market price—and so such a firm would not be a perfect competitor. If the firm can affect the price received by other producers, its actions have consequences that will affect other participants, leading the firm to engage in strategic behavior. Hence, to be a true perfectly competitive firm, the firm's output decision must not affect the going price. This feature may be illustrated in a graph by drawing the demand curve for a perfectly competitive firm as a horizontal line at the current market price. Note that a perfectly competitive firm faces a horizontal demand curve even though the market or industry demand curve describing demand faced by the entire industry is downward sloping.[2]

Like all firms, the perfectly competitive ones will each choose that output level which maximizes their individual profit. Profit is defined as the difference between the firm's revenue and its total costs. Revenue is just market price, P, times the firm's output, q. The firm's total cost is assumed to rise with the level of the firm's production according to some function, $C(q)$. It is important to understand that the firm's costs include the amount necessary to pay the owners of the firm's capital (that is, its stockholders) a normal or competitive return. This is a way of saying that input costs are properly measured as opportunity costs. That is, each input must be paid at least what that input could earn in its next best alternative employment. This is true for the capital employed by the firm as much as it is true for the labor and raw materials that the firm also uses. Generally speaking, the opportunity cost for the firm's capital is measured as the rate of return that the capital could earn if invested in other industries. This cost is then included in our measure of total cost, $C(q)$. In other words, the concept of profit we are using is that of economic profit and reflects net revenue above what is necessary to pay all of the firm's inputs at least what they could earn in alternative employment. The reason why this point is important is because it makes clear that when a firm earns no economic profit it does not mean that its stockholders go away empty-handed. It simply means that those stockholders do not earn more than a normal return on their investment.

A necessary condition for such profit maximization is that the firm chooses an output level such that the revenue received for the last unit produced, or the marginal revenue, just equals the cost incurred to produce that last unit, or the marginal cost. This condition for profit maximization holds for the output choice of any firm, be it a perfectly competitive one, or a monopoly. Since total revenue depends on the amount produced, marginal revenue is also

[2] This follows from the definition of a perfect competitor. One may wonder how each firm can face a horizontal demand curve while industry demand is downward sloped. The answer is that the demand curve facing the industry reflects the summation of the individual demand presented by each consumer—not the individual demand facing each firm.

dependent on q as described by the marginal revenue function, $MR(q)$. Because the perfectly competitive firm can sell as much as it likes at the going market price, each additional unit of output produced and sold generates additional revenue exactly equal to the current market price. That is, the marginal revenue function for a competitive firm is just $MR(q) = P$. Similarly, because total cost is a function of total output, q, so the marginal cost function also depends on q, according to the function $MC(q)$. This function describes the cost incurred by the firm for each successive unit of output produced.

Diagrams like those shown in Figures 2.2(a) and 2.2(b), respectively, are often used to illustrate the standard textbook model of the perfectly competitive firm and the perfectly competitive market in which the firm sells. For any market to be in equilibrium, the first order condition mentioned earlier must hold for each firm. For a competitive market, this means that for each firm the price received for a unit of output exactly equals the cost of producing that output at the margin. This condition is illustrated in Figures 2.2(a) and 2.2(b). The initial industry demand curve is D_1 and the market price is P_C. A firm producing output q_C incurs a marginal cost of production $MC(q_C)$ just equal to that price. Producing one more unit would incur an extra cost, as indicated by the marginal cost curve MC that exceeds the price at which that unit would sell. Conversely, producing less than q_C would save less in cost than it would sacrifice in revenue. When the firm produces q_C and sells it at market price, P_C, it is maximizing profit. It therefore has no incentive to change its choice of output. Hence, in a competitive equilibrium each firm must produce at a point where its marginal cost is just equal to the price.

Total market supply, Q_C, is the sum of each firm's output, q_C. Since each firm is maximizing profit, the condition $P = MC(q_C)$, will hold for each firm. If demand for the product increases and the market price rises to say P_1, each firm will revise its production decision and increase output to q_1, where $P_1 = MC(q_1)$. This will increase total production to Q_1. Indeed, because the firms' production decisions are governed by costs at the margin, the marginal cost curve of each firm provides the basis for determining the total supply at any given market price. As the price rises, we work out how each firm adjusts its profit-maximizing output by moving up its marginal cost function to a point where $P = MC(q)$ at this new price. Then we add up all the firms' revised decisions and compute the total output now supplied. Repeating this exercise for various prices reveals the industry supply function indicating the total output supplied at any given market price. It is illustrated by the curve S_1 in Figure 2.2(b). Since for each firm price is equal to its marginal cost, it must be the case that at each point on the supply function for every firm the incremental cost of the last unit produced is just equal to that price.

Consider a simple linear example where each firm's marginal cost curve is linear instead of curved as shown in Figure 2.2(a). Specifically, let the marginal cost of each firm be: $MC(q) = 4q + 8$. Given a market price P, the optimal output for any one competitive firm is then q such that $4q + 8 = P$, implying that the optimal output for each such firm satisfies $q = \frac{P}{4} - 2$.

If there are 80 such firms, total industry production Q at price P is 80 times q or $Q^S = 20P - 160$. Solving for P writes the resultant supply curve in the form implied by Figures 2.2(a) and 2.2(b) in which price appears on the vertical axis. This yields $P = 0.05Q^S + 8$. At a price of 8, each firm will produce zero output. Industry output will also be zero. A rise in P to 12 will induce each firm to raise its output to 1 unit, increasing industry output to 80. A further rise to $P = 16$ will lead every firm to raise its output to 2 units, implying a total supply of 160. We could repeat this exercise many times over, each time choosing a

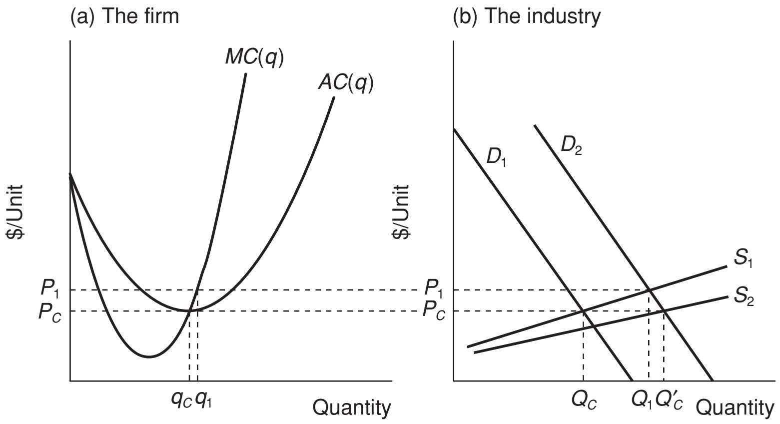

Figure 2.2 The long-run competitive equilibrium
Price P_1 is consistent with a short-run equilibrium in which each firm produces at a point where its marginal cost is equal to P_1. However, at P_1 price exceeds average cost and each firm earns a positive economic profit. This will encourage entry by new firms, shifting out the supply curve as shown in (b). The long-run competitive equilibrium occurs as price P_C in which each firm produces output level q_C and price equals both average and marginal cost.

different price. Plotting the industry output against each such price yields the industry supply curve. The important point to understand is that the derivation of that supply curve reflects the underlying first order condition for profit maximization—that is, each competitive firm choose a profit-maximizing level of output such that $P = MC(q)$.

In the example shown in Figures 2.2(a) and 2.2(b), the market initially clears at the price, P_C. Given the demand curve D_1, this equilibrium is consistent with the first order condition that each firm produce an output such that $P = MC(q)$. The requirement that each firm produces where marginal cost equals the market price is almost all that is required for a competitive equilibrium in the short run.[3] However, there is an additional condition that must be met in order for this to be a long-run competitive equilibrium. The condition is that in a long-run equilibrium each firm earns zero economic profit. This condition is also met in the initial equilibrium illustrated in Figure 2.2(a). At output q_C, each firm is just covering its cost of production, including the cost of hiring capital as well as labor and other inputs. In other words, a long-run competitive equilibrium requires that firms just "break even" and not earn any economic profit—revenue that exceeds the amount required to attract the productive inputs into the industry. This requirement can be stated differently. In the long run, the price of the good must just equal the average or per unit cost of producing the good. Again, both this zero profit condition and the further requirement that price equal marginal cost are satisfied in the initial equilibrium in which the industry demand curve is D_1 and the price is P_C.

If demand suddenly shifts to the level described by the demand curve, D_2, the existing industry firms will respond by increasing output. In so doing, these firms maximize profit by again satisfying the first requirement that they each produce where $P = MC(q)$. This leads

[3] We say almost because there may be a distinction between average variable cost and marginal cost. No production will occur at all in the short run if the firm cannot produce at a level that will cover its average variable cost.

each firm to expand its production from q_C to q_1, thereby raising the market output to Q_1. However, this short-run response does not satisfy the zero profit condition required for a long-run competitive equilibrium. At price P_1, the market price equals each firm's marginal cost but exceeds each firm's average cost. Hence, each firm earns a positive economic profit of $P_1 - AC(q_1)$ on each of the q_1 units it sells.

Such profit either induces new firms to enter the industry or existing firms to expand production. This expansion shifts the industry supply curve outward until the equilibrium price again just covered average cost. Figure 2.2(b) illustrates this by the shift in the industry supply curve to S_2. As drawn, this shift reestablishes the initial price, P_C. Each firm again produces output q_C at which the industry price equals both the firm's marginal cost and its average cost. Of course, total industry output is now higher at Q'_C. While each firm is producing the output q_C, there are now more firms. The point is that in a long run market equilibrium no firm has the incentive to change its production plan and in the long run, this means no firm wishes either to leave or to enter the market.

Practice Problem 2.1

Assume that the manufacturing of cellular phones is a perfectly competitive industry. The market demand for cellular phones is described by a linear demand function: $Q^D = \dfrac{6000 - 50P}{9}$. There are 50 manufacturers of cellular phones. Each manufacturer has the same production costs. These are described by long-run total and marginal cost functions of $TC(q) = 100 + q^2 + 10q$, and $MC(q) = 2q + 10$.

a. Show that a firm in this industry maximizes profit by producing $q = \dfrac{P - 10}{2}$
b. Derive the industry supply curve and show that it is $Q^S = 25P - 250$.
c. Find the market price and aggregate quantity traded in equilibrium.
d. How much output does each firm produce? Show that each firm earns zero profit in equilibrium.

Under perfect competition each firm's production of the good is small relative to the market. Now suppose that all these sellers become consolidated into one firm that is, by definition, a monopoly. Since the monopolist is the only supplier of the good, the monopoly is likely to be large relative to market demand. Specifically, the monopolist's demand curve is identical with the market demand curve. In complete contrast to the competitive firm, the monopoly firm is able to influence the price it receives for selling in this market. The monopolist's output decision will play a decisive role in determining the market-clearing price.

2.1.2 Monopoly

Since the monopolist's demand curve slopes downward, any increased production by the monopolist will lead to a price reduction. For instance, a monopolist who was selling Q_1 units at price P_1 will find that increasing production to Q_2 units will cause the market price to fall from P_1 to P_2. The good news is that, by selling the additional output, the monopolist earns additional revenue. However, the bad news is that the original Q_1 units no longer sell at a price of P_1. Now, these units sell for only P_2 each. Often it is the case and we will assume

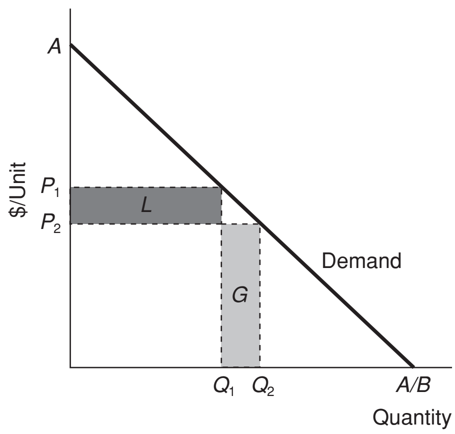

Figure 2.3 The marginal revenue from increased production for a monopolist
An increase in production from Q_1 to Q_2 causes a gain in revenues approximated by area G and a loss in revenues approximated by area L. The net change or marginal revenue is therefore $G - L$. Note, because the firm is a monopolist, this is also the net revenue gain generated by cutting price from P_1 to P_2.

it here that the monopolist cannot charge the first Q_1 customers a high price and the next $Q_2 - Q_1$ customers a lower price for the same commodity. The fact that such price discrimination is ruled out means that the monopolist must sell at the market-clearing price to all consumers and, therefore, that increases in the monopolist's total output will reduce the equilibrium market price.

Accordingly, the monopolist is very different from the competitive firm that reckons that every additional unit sold will bring in revenue equal to the current market price. Instead, the monopolist knows that every unit sold will bring in marginal revenue less than the existing price. Because the additional output can be sold only if the price declines, the marginal revenue from an additional unit sold is not market price but something less.

Marginal revenue for a monopolist is illustrated by the shaded areas G and L in Figure 2.3. These areas reflect the two forces affecting the monopolist's revenue when the monopolist increases output from Q_1 to Q_2, and thereby causes the price to fall from P_1 to P_2. Area G is equal to the new price P_2 times the rise in output, $Q_2 - Q_1$. It is the revenue gain that comes from selling more units. Area L equals the amount by which the price falls, $P_1 - P_2$, times the original output level, Q_1. This reflects the revenue lost on the initial Q_1 units as a result of cutting the price to P_2. The net change in the monopolist's revenue is the difference between the gain and the loss or $G - L$.

We can be more precise about this. Let $\Delta Q = Q_2 - Q_1$, and $\Delta P = P_1 - P_2$. The slope of the monopolist's (inverse) demand curve may then be expressed $\frac{\Delta P}{\Delta Q}$. If we describe this demand curve (which of course is also the market demand curve) as a linear relation, $P = A - BQ$, that slope is also equal to the term, $-B$, i.e., $\frac{\Delta P}{\Delta Q} = -B$. In other words, an increase in output ΔQ[4] leads to a decline in price ΔP equal to $-B\Delta Q$. Since total revenue is defined

[4] Under perfect competition, firm output is different from industry output. So we use a lower case q to refer to firm output and an upper case Q for industry output. Under monopoly, firm output is the market output and so we use Q to describe both.

as price per unit times the number of units sold, we can write total revenue as a function of the firm's output decision, or $R(Q) = P(Q)Q = (A - BQ)Q$. As just shown in Figure 2.3, the change in revenue, $\Delta R(Q)$, due to the increase in output ΔQ, is the sum of two effects. The first is the revenue gain, $P_2\Delta Q$. The second is the revenue loss, $Q_1\Delta P$. Hence,

$$\Delta R(Q) = P_2\Delta Q - Q_1\Delta P = (A - BQ_2)\Delta Q - Q_1(B\Delta Q) \tag{2.1}$$

where we have used the demand curve to substitute $A - BQ_2$ for P_2 in the first term on the right-hand side. $MR(Q)$, is measured on a per unit basis. Hence, we must divide the change in revenue shown in equation (2.1) by the change in output, ΔQ, to obtain marginal revenue. This yields

$$MR(Q) = \frac{\Delta R(Q)}{\Delta Q} = A - BQ_2 - BQ_1 \approx A - 2BQ \tag{2.2}$$

Here we have used the approximation, $B(Q_1 + Q_2) \approx 2BQ$. This will be legitimate so long as we are talking about small changes in output, i.e., so long as Q_2 is fairly close to Q_1.

Equation (2.2)—sometimes referred to as the "twice as steep rule"—is quite important, and we will make frequent reference to it throughout the text. It not only illustrates that the monopolist's marginal revenue is less than the current price but, for the case of linear demand, also demonstrates the precise relationship between price and marginal revenue. The equation for the monopolist's marginal revenue function, $MR(Q) = A - 2BQ$, has the same price intercept A as the monopolist's demand curve but twice the slope, $-2B$ versus just $-B$. In other words, when the market demand curve is linear, the monopolist's marginal revenue curve starts from the same vertical intercept as that demand curve, but is everywhere twice as steeply sloped. The monopolist's marginal revenue curve must then lie everywhere below the inverse demand curve.

In Figure 2.4, we show both the market demand curve and the corresponding marginal revenue curve facing the monopolist. Again profit maximization requires that a firm produce

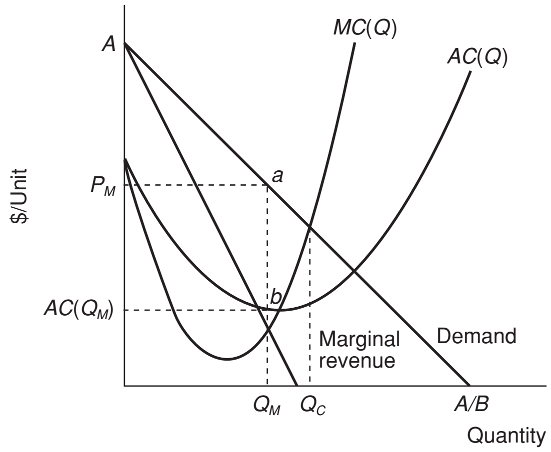

Figure 2.4 The textbook monopoly case
The monopolist maximizes profit by choosing the output Q_M at which marginal revenue equals marginal cost. The price at which this output can be sold is identified by the demand curve as P_M, which exceeds marginal cost. Profit is *abcd*. The competitive industry would have instead produced Q_C, at which point price equals marginal cost.

Derivation Checkpoint

The Calculus of Competition

For those familiar with calculus, the competitive firm's problem may be solved by first writing the firm's profit π as a function of its output q, or as $\pi(q)$ which, in turn, is defined as the difference between revenue $R(q)$ and cost $C(q)$. If we then recognize that revenue is just price times quantity or $R(q) = Pq$, we obtain:

$$\pi(q) = R(q) - C(q) = Pq - C(q)$$

Maximization of the firm's profit requires taking the derivative of the profit function with respect to q and setting it equal to zero. Recall however that the competitive firm takes P as given. Hence, the standard maximization procedure yields:

$$\frac{d\pi}{dq} = P - C'(q) = 0$$

Since $C'(q)$ is the change in cost as one more unit is produced it is precisely what we call marginal cost. Hence the profit-maximizing condition for the competitive firm is to choose the output q for which marginal cost $C'(q)$ equals price P.

For the monopoly firm, its output is the same as industry output Q and so its price is not given but instead declines with output as the firm moves down its demand curve. That is, the monopolist does not face a single price but instead a price function $P(Q)$, which is really the inverse demand curve. Hence, the monopolist's profit maximization problem is to choose output Q so as to maximize:

$$\pi(Q) = R(Q) - C(Q) = P(Q)Q - C(Q)$$

Again, standard maximization techniques yield:

$$\frac{d\pi}{dQ} = P(Q) + QP'(Q) - C'(Q) = 0$$

The sum, $P(Q) + QP'(Q)$, is the firm's marginal revenue. The monopolist will maximize profit by producing where marginal cost equals marginal revenue. For a linear demand curve of the form of $P(Q) = A - BQ$ we have $P'(Q) = -B$. Hence, in this case, the firm's marginal revenue is $A - BQ - BQ$, or $A - 2BQ$. The monopolist's marginal revenue curve has the same intercept as its demand curve but is twice as steeply sloped.

Note that the profit-maximizing condition above can also be written as

$$P(Q) - C'(Q) = -QP'(Q)$$

Dividing both sides by $P(Q)$ we then have

$$\frac{P(Q) - C'(Q)}{P(Q)} = -\frac{QP'(Q)}{P(Q)} = \frac{1}{\eta}$$

Where η is what economists call the elasticity of demand—a measure of how responsive the quantity demanded is to price movements. It is formally defined as:

$$\eta = \frac{P(Q)}{Q}\frac{1}{P'(Q)}$$

up to the point where the marginal revenue associated with the last unit of output just covers the marginal cost of producing that unit. This is true for the monopoly firm as well as for the perfectly competitive firm. The key and important difference here is that for the monopoly firm, marginal revenue is less than price. For the monopoly firm, the profit-maximizing rule of marginal revenue equal to marginal cost, or $MR(Q) = MC(Q)$, holds at the output Q_M. The profit-maximizing monopolist produces at this level and sells each unit at the price P_M. Observe that, at this output level, the revenue received from selling the last unit of output MR is less than the price at which that output is sold, $MR(Q_M) < P_M$. It is this fact that leads the monopolist to produce an output below the (short-run) equilibrium output of a competitive industry, Q_C.

We have also drawn the average cost function for the monopoly firm in Figure 2.4. The per unit or average cost of producing the output level Q_M, described on the average cost curve by $AC(Q_M)$, is less than the price P_M at which the monopolist sells the good. This means, of course, that total revenue exceeds total cost, and so the monopolist earns a positive economic profit. The monopoly profit is shown as the rectangle $P_M abAC(Q_M)$. Furthermore, because the monopolist is the only firm in this market, and because we assume that no other firm can enter and supply this good, this market outcome is a long-run equilibrium. Each consumer buys as much as he wants to at price P_M and, given these cost conditions, the monopolist has no incentive to sell more or to sell less. Even in the long run, there is no tendency under monopoly for the market price to equal the unit cost of production.

Reality Checkpoint

Hung Up on Monopoly

It is not always easy to find examples of the classic monopoly behavior described in economics textbooks. However, Tyco International's control of the plastic hanger market in the late 1990s may have come pretty close. Retail firms such as J. C. Penney and K-Mart use only plastic hangers to display their clothing goods. Starting in about 1994, Tyco used mergers and acquisitions of rival firms to gain control of 70 to 80 percent of the market for plastic hangers. In a number of geographic regions, Tyco became the only plastic hanger firm available. In 1996, Tyco acquired a Michigan-based hanger firm, Batts, that was one of the largest suppliers to the Midwest region. Immediately thereafter, Tyco raised prices by 10 percent to all its customers. Some clients grumbled but most accepted the higher prices. Others though, such as K-Mart and VF (makers of Lee and Wrangler jeans) informed Tyco that they had an alternative hanger supplier, namely, a company called WAF. For a brief moment, Tyco appears to have backed off raising the price. Yet the firm's underlying strategy soon became clear. In the fall of 1999, Tyco bought the WAF Corporation. Within a few months, it not only raised prices to all its customers again but, this time, it also added in a new delivery charge. Tyco also pursued an aggressive repurchase program so as to corner the market on used hangers. If it did not control the supply of this alternative to new hangers, Tyco would have faced increasing difficulty in charging a high price.

Source: M. Maremont, "Lion's Share: For Plastic Hangers You Almost Need to Go to Tyco International," *Wall Street Journal*, February 15, 2000, p. A1.

9781405176323_4_002.qxd 10/12/07 7:58 PM Page 30

2.2 **Practice Problem**

Now suppose that the manufacturing of cellular phones, as described in Practice Problem 2.1, is monopolized. The monopolist has 50 identical plants to run. Each plant has the same cost function as described in Practice Problem 2.1. The overall marginal cost function for the multiplant monopolist[5] is described by $MC(Q) = 10 + Q/25$. The market demand is also assumed to be the same as in Practice Problem 2.1.

Recall $Q^D = \dfrac{600 - 50P}{9}$

a. Show that the monopolist's marginal revenue function is $MR(Q) = 120 - 18Q/50$.
b. Show that the monopolist's profit-maximizing output level is $Q_M = 275$. What price does the monopolist set to sell this level of output?
c. What is the profit earned at each one of the monopolist's plants?

2.2 PROFIT TODAY VERSUS PROFIT TOMORROW: FIRM DECISION-MAKING OVER TIME

Both the competition and the monopoly models described in the previous section are somewhat vague with respect to time. While some distinction is made between the short run and the long run, neither concept explicitly confronts the notion of a unit of time such as a day, a week, a month, or a year, or of how many such units constitute say, the long run. To maximize profit in the long run requires, for example, only that the firm make all necessary adjustments to its inputs in order to produce at the optimum level, and then repeatedly choose this input–output combination in every individual period. From the standpoint of decision-making then, the long run is envisioned as a single market period and the assumption that the firm will seek to maximize profit is unambiguous in its meaning.

However, the recognition that the long run is a series of individual, finite time periods extending far into the future also raises the possibility that each such period will not be the same. Here, the choice may well be between taking an action that yields profit immediately versus taking an action that yields perhaps greater profit but not until many periods later. In such a setting, the meaning of maximizing profit is less clear. Is it better or worse to have more profit later and less profit now? How does one compare profit in one period with profit in another? Such questions must be answered if we are to provide a useful analysis of the strategic interaction among firms over time.

Sacrificing profit today means incurring a cost. Hence, the problem just described arises anytime that a cost is incurred in the present in return for benefits to be realized much later. Firms often face such a trade-off. A classic example is the decision to build a new manufacturing plant. If the plant is constructed now, the firm will immediately incur the expense

[5] Strictly speaking, the monopolist is a multiplant one because he now has 50 plants to run. The profit-maximizing monopolist will want to allocate total production across the 50 plants in such a way that marginal cost of producing the last unit of output is the same in each plant. Therefore, the monopolist derives his overall marginal cost function in a manner similar to how we constructed the supply function for the competitive industry. This point is further explained in Chapter 3 in the section on multiplant monopolies.

of hiring architects and construction workers and the buying of building materials, machinery, and equipment. It will only be sometime later—after the plant is built and running smoothly—that the firm will actually begin to earn some profit or return on this investment.

In order to understand how firms make decisions in which the costs and benefits are experienced not just in one period but instead over time, we borrow some insights from financial markets. After all, the comparison of income received (or foregone) at different points in time is really what financial markets are all about. Think for a moment. If one buys some stock in say, Microsoft, one has to give up some funds today—namely, the price of a share in Microsoft times the number of shares bought. Of course, investors do this every day. Thousands of Microsoft shares are bought each day of the week. These investors are thus sacrificing some of their current income—which could alternatively be used to purchase a Caribbean vacation, or wardrobe, or other consumer goods—to buy these shares. Why do investors do this? The answer is that they do so in the expectation that those shares will pay dividends and will also appreciate in value over time. That is, stockholders buy shares of stock and incur the associated investment expense now, in the hope that the ownership of those shares will generate income as dividends and capital gains later.

In short, the financial markets are explicitly involved in trading current for future income. Accordingly, we can use the techniques of those markets to evaluate similar trades of current versus future profit that a firm might make. The key insight that we borrow from financial markets is the notion of present value or discounting. To understand the concept of discounting, imagine that a friend (a trustworthy friend) has asked to borrow \$1,000 for twelve months. Suppose further that for you to lend her money requires that you withdraw \$1,000 from your checking account, an account that pays 3 percent interest per year. In other words, you will have to lose about \$30 of interest income by making this withdrawal. Although you like your friend very much, you may not see just precisely why you should make her a gift of \$30. Therefore, you agree to lend her the \$1,000 today if, a year from now, she pays you not only the \$1,000 of principal but also an additional \$30 in interest. Your friend will likely agree. After all, if she borrowed from the bank directly she would have to pay at least as much. The bank cannot afford to pay you 3 percent per year if it does not charge an interest rate at least as high when it loans those funds out. In fact, the bank will probably charge an interest rate a bit higher to cover its expenses. So, it makes sense for your friend to sign a contract (or perhaps just shake hands on the deal) requiring that you give her \$1,000 today and that she give you \$1,030 in 12 months.

Quite explicitly, you and your friend have just negotiated a trade of present funds for future funds. In fact, you have established the exact terms at which such a trade can take place. One thousand dollars today may be exchanged for \$1,030 one year from now. Of course, matters would have been a bit different if the interest rate that your bank paid on deposits had been 5 percent. In that case, you would have asked your friend for \$50 (5 percent of \$1,000) in repayment beyond the \$1,000 originally borrowed. That would have been the only repayment that would truly compensate you for your loss of the interest on your bank deposit. In general, if we denote the interest rate as r, then we have that \$1,000 today exchanges for $(1 + r)$ times \$1,000 in one year. If we now become even more general and consider an initial loan amount different from \$1,000, say of Y, we will quickly see that the same logic implies that Y today trades for $(1 + r)Y$ paid in twelve months.

There is, however, an alternative way to view the transactions just described. Instead of asking how much money one will receive in a year for giving up \$1,000 or Y now, we can reverse the question. That is, we can ask instead how much we have to pay today in order to get a particular payment one year from the present. For example, we could ask how much does it cost right now to buy a contract requiring that the other party to the deal pay us \$1,030 in a year. If the interest rate is 3 percent, the answer is easy. It is simply \$1,000. In fact, this is the contract with your friend that we just considered. You essentially paid \$1,000 to purchase a promise from your friend to pay you \$1,030 in one year. The intuition is that at an interest rate of 3 percent, the banks and the financial markets are saying that in return for a deposit of \$1,000 they promise to pay \$1,030 in one year. In other words, we can buy the contract we are thinking about for exactly \$1,000 from the banks. There's no sense in paying more for it from anyone else, and no one else is going to accept less. Therefore, when the interest rate is 3 percent, the market is saying that the current price of a contract promising to pay \$1,030 in one year is exactly \$1,030/(1.03) or \$1,000. Since price is just the economist's term for value, we call this the present value or, more completely, the present discounted value of \$1,030 due in twelve months.

More generally, the present value of a piece of paper (e.g., a loan contract or share of stock) promising its owner a payment of Z in one period is just $Z/(1 + r)$. The term $1/(1 + r)$ is typically referred to as the discount factor and is often presented just as R. In other words, $R = 1/(1 + r)$. Hence, the present value of Z dollars one year from now is often written as RZ. The source of the adjective discount should be clear. Income that does not arrive until a year from now is not as valuable as income received today. Instead, the value of such future income is discounted. This has nothing to do with inflation and any possible cheapening of the currency over time. It simply reflects the fact that individuals prefer to have their consumption now and have to be paid a premium—an interest rate return—in order to be persuaded to wait.

What if the term of the loan had been for two years? Let us return again to our original example of a \$1,000 loan at 3 percent interest. If your friend had initially asked to borrow the funds for two years, your reasoning might have gone as follows. Making a two-year loan to my friend requires that I take \$1,000 out of my checking account today. Not making the loan means that the \$1,000 stays in the bank. In this case, I will earn 3 percent over the next 12 months and, accordingly, start the next year with \$1,030 in the bank. I will then earn 3 percent on this amount over the next or second year. Accordingly, by refusing my friend and keeping the funds in the bank, I will have on deposit \$1,030(1.03) = \$1,060.90 in two years. Therefore, I will only lend my friend the funds for two years if she in turn promises to pay me \$1,060.90—the same as I could have earned at the bank—when the loan expires 24 months from now. Note that the amount \$1,060.90 can be alternatively expressed as $\$1,000(1.03)(1.03) = \$1,000(1.03)^2$. In general, a loan today of amount Y will yield $Y(1 + r)^2$ or YR^{-2} in two years. By extension, a loan of Y dollars for t years will yield an amount of $Y(1 + r)^t$ or YR^{-t} when it matures t years from now.

As before, we can turn the question around and ask how much we need to pay currently in order to receive an amount of Z dollars at some date t periods into the future. The answer follows immediately from our work above. It is R^tZ. How do we know this? If we put the amount R^tZ dollars in an interest-bearing account today, then the amount that can be withdrawn in t periods is, by our previous logic, $(R^tZ)R^{-t} = Z$. So, clearly, the present discounted value of an amount Z to be received t periods in the future is just R^tZ.

The only remaining question is how to value a claim that provides different amounts at different dates in the future. For example, consider the construction of a plant that will, after completion in one year, generate Z_1 in net revenue; and then a net revenue of Z_2 two years from now; Z_3 three years from now, and so on. What is the present value of this stream of future net revenues? Well, the present value of Z_1 in one period is, as we know, RZ_1. Similarly, the present value of the Z_2 to be received in two periods is R^2Z_2. If we continue in this manner we will work out the present value of the income received at each particular date. The present value of this entire stream will then simply be the sum of all these individual present values. In general, the present value PV of a stream of income receipts to be received at different dates extending T periods into the future is:

$$PV = RZ_1 + R^2Z_2 + R^3Z_3 + \ldots + R^TZ^T = \sum_{t=1}^{T} R^tZ_t \tag{2.3}$$

A special case of equation (2.3) occurs when the income received in each period Z_t is the same, that is, when $Z_1 = Z_2 = \ldots = Z_T = \bar{Z}$. In that case, the present value of the total stream is:

$$PV = \frac{\bar{Z}}{(1 - R)}(R - R^{T+1}) \tag{2.4}$$

An even more special case occurs when not only is the income receipt constant at $Z = \bar{Z}$, but the stream persists into the indefinite future so that the terminal period T approaches infinity. In that case, since the discount factor R is less than one, the term R^{T+1} in equation (2.4) goes to zero. Hence, when the stream is both constant and perpetual, the present value formula becomes:

$$PV = \bar{Z}\left(\frac{R}{1 - R}\right) = \frac{\bar{Z}}{r} \tag{2.5}$$

Thus, if the interest rate r were 3 percent, a promise to pay a constant \$30 forever would have a present value of PV = \$30/0.03 = \$1,000. Note that for all our present value formulas, an increase in the real interest rate r implies a decrease in the discount factor R. In turn, this means that a rise in the interest rate implies a decrease in the present value of any given future income stream.

Again, it is important to remember the context in which these equations have been developed. Often firm decision-making has a temporal dimension. Indeed, our focus on long-run equilibria implies that we are considering just such decisions. Hence, we need to consider trade-offs that are made over time. An expense may need to be incurred now in order to reap additional profit at some future date or dates. The simple dictum maximize profit does not have a clear meaning in such cases. The only way of evaluating the desirability of such a trade-off over time is to discount, that is, translate the future dollar inflows into a current or present value that may then be compared with the current expense necessary to secure those future receipts. If the present value of the future income is not at least as great as the value of the necessary expense, then the trade-off is not favorable. If, for instance, a plant costs \$3 million to build, and will generate future profit with a discounted present value of only \$2 million, it is not a desirable investment, and we would not expect a rational firm to undertake it.[6] In short, our assumption that firms maximize profit must now be qualified to mean that firms maximize the present value of all current and future profit. Of course, for one-period problems, this is identical with the assumption that firms simply maximize profit.

[6] We have treated the problem as one of current expenses versus future receipts. Of course, future costs should be discounted as well.

Reality Checkpoint

Piracy on the High (Air)waves: Discounting, Monopoly Power, and Public Policy

In the United States, direct satellite television is largely provided by two services, Direct TV and the Dish Network. Subscribers to these services pay between $30 and $80 per month, depending on the package of TV programs they wish to view. In return, they get a satellite dish, receiver and decoder that permits them to view the programming in as many as four rooms in their house. In recent years, the number of satellite TV subscribers has grown rapidly to something in the order of 20 million. Yet even with this growth, satellite TV still has only about a third of the subscribers that its competitor, cable TV, has.

However, there is one category of subscriber in which satellite TV does outperform cable. This is the category of illegal, nonpaying subscribers. The satellite TV firms have some idea as to how many dishes have been sold and installed over the years. Their estimate leaves them with about one to three million more dishes in place than actually subscribe to satellite services. The firms reckon that at least half of these represent users who obtain the service illegally by tapping into the satellite transmission.

To engage in such theft, the would-be airwaves pirate needs the basic hardware equipment—the dish and receiver—and also a smart card that tells the receiver what programs to decode. Anyone can buy such a smart card—new or used—on eBay.com and other sites. This is where the pirates enter the picture. A number of firms buy the smart card, then use hackers to break the code and write a script that tells the receiver to unscramble everything. Some illicit firms then simply sell the cards. Others sell the script that reprograms the card.

Of course, the satellite firms are aware of all this. They therefore periodically send out an Electronic Counter Measure (ECM) signal that puts out a new code and/or corrupts unauthorized cards. However, the best satellite pirates have become quite good at detecting when an ECM is coming and quickly write new scripts that restore the cards' operating ability.

How much are these illegal services worth? Consumers who subscribe to Direct TV or the Dish Network would expect to pay something like $75 per month for the complete package that includes all the channels of these networks. This typically includes the hardware, which is "rented" for free. Assume that the typical consumer has a current residence and therefore satellite TV horizon of five years. Then with an annual interest rate of 4 percent (compounded monthly) this implies a present value of about $4,114 for complete and legitimate satellite service over this time span. Illegal users pay about $225 to acquire their own hardware. They also pay about $25 to subscribe to the hacker services that provide them with updated scripts to keep their cards working. This works to a present value of just under $1,600, implying a saving of $2,500. However, illegal users do take some risks. Recently, the satellite firms have cracked down on the pirate companies and, in the process, obtained the lists of their customers. Those customers face very large potential fines. Typically, the satellite firms offer those caught the option of paying $5,000 to avoid further legal charges. Yet even with all their best efforts, the satellite firms reckon that the typical illicit consumer has at most a one-in-three chance of being caught and paying that fine. Thus, the expected value of the fine is $5,000/3 = $1,667. If it takes two years on average to catch the thief, then the present value of the fine—again assuming a four percent annual interest rate—is $1,540. Hence, an educated guess of the total expected cost to the illicit satellite user in present value terms is $1,600 + $1,667 = $3,267. Including the risk of getting caught has reduced the savings to $847. Even in a black market, consumers earn some surplus.

Source: D. Lieberman "Millions of Pirates Are Plundering Satellite TV," *USA Today*, December 2, 2004, p. C1.

However, we will need to be familiar with the idea of discounting and the present value of future profits in the second half of the book when we take up such issues as collusion and research and development, which often have a multiperiod dimension.

2.3

Practice Problem

Suite Enterprises is a large restaurant supply firm that dominates the local market. It does however have one rival, Loew Supplies. Because of this competition, Suite earns a profit of \$100,000 per year. It could, however, cut its prices to cost and drive out Loew. To do this, Suite would have to forego all profit for one year and earn zero. After that year, Loew would be gone forever and Suite could earn \$110,000 per year. The interest rate Suite confronts is 12 percent per annum, and so the discount factor $R = 0.8929$.

a. Is driving Loew out of the market a good "investment" for Suite?
b. Consider the alternative strategy in which Suite buys Loew for \$80,000 today and then operates the new combined firm, Suite & Loew, as a monopoly earning \$110,000 in all subsequent periods. Is this a good investment?

2.3 EFFICIENCY, SURPLUS, AND SIZE RELATIVE TO THE MARKET

Now that we have described the perfectly competitive and pure monopoly market outcomes, it is time to try to understand why perfect competition is extolled and pure monopoly is guarded against by law. In both cases firms are driven by profit maximization. Also, in both cases the firms sell to consumers who decide how much they want to buy at any given price. What makes one market good and the other market not? The answer to this question does not reflect any concern about too much profit or firms "ripping off" consumers. The answer instead lies in the economic concept of efficiency. In economics, efficiency has a very precise meaning. Briefly speaking, a market outcome is said to be efficient when it is impossible to find some small change in the allocation of capital, labor, goods, or services that would improve the well-being of one individual in the market without hurting any others.[7] If the only way we can make somebody better off is by making someone else worse off, then there is really no slack or inefficiency in how the market is working. If, on the other hand, we can imagine changes that would somehow allow one person to have more goods and services while nobody else has less, then the current market outcome is not efficient. As it turns out, that is precisely the case for a monopolized market. One can think of changes to the monopoly outcome that would yield more for at least one individual and no less for any other. However, as we'll see, market forces alone will not get us there in the case of the textbook monopolist.

It is readily apparent that, to implement our efficiency criterion, we need some measure of how well off consumers and firms are in any market outcome. For this purpose, we use the notions of consumer surplus and producer surplus. The consumer surplus obtained from consuming one unit of the good is defined as the difference between the maximum amount a consumer is willing to pay for that unit and the amount the consumer actually does pay.

[7] This notion of efficiency is often referred to as Pareto optimality after the great Italian social thinker of the late nineteenth and early twentieth centuries, Vilfredo Pareto.

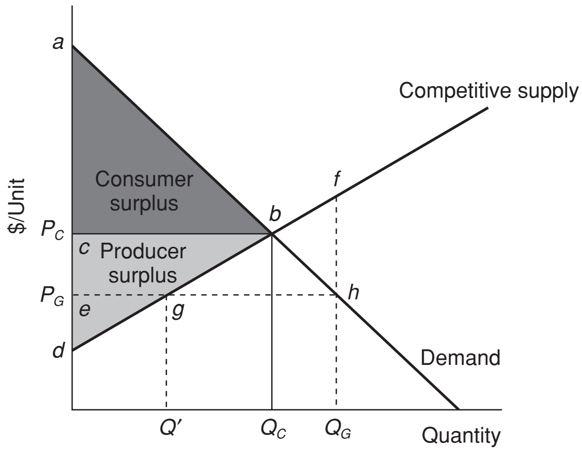

Figure 2.5 Competition maximizes the total surplus
At the competitive price P_C and output Q_C, consumers enjoy a surplus equal to triangle *abc*. Producers enjoy a surplus equal to triangle *cbd*. This is the maximum. Producing less would lose some of the total surplus given by triangle *abd*. Subsidizing production to output Q_G reduces the price to P_G. The required subsidy is *gfh*. Consumers gain additional surplus *cbge*. However, this amount represents a transfer of surplus from producers to consumers and, hence no net gain in total surplus. Consumers also gain the triangle *gbh*, but this is more than offset by the funds required for the needed subsidy. The remaining part of the subsidy equal to triangle *bfh* is a deadweight loss as resources valued more highly in alternative uses are transferred to the industry in question where the marginal value of output is only P_G.

Total consumer surplus in a market is then measured by summing this difference over each unit of the good bought in the market. Analogously, the producer surplus obtained from producing a single unit of the good is the difference between the amount the seller receives for that unit of the good and the cost of producing it. Total producer surplus in a market is then measured by summing up this difference over each unit of the good sold.

We illustrate these concepts in Figure 2.5. In the competitive outcome, Q_C units of the good are bought and sold. The maximum amount a consumer is willing to pay for the last unit, the Q_Cth unit, is just the equilibrium price P_C. However, the maximum amount a consumer is willing to pay for the first, the second, the third and so on, up to the Q_Cth unit is greater than P_C. We know this because, at a given sales volume, the demand curve is a precise measure of the maximum amount any consumer is willing to pay for one more unit. Hence, the area under the demand curve but above the market equilibrium price P_C is surplus to consumers. It is a measure of how much they were willing to pay less what they actually did pay in the competitive outcome. This is shown in Figure 2.5 as area *abc*.

For competitive producers, the supply curve tells us the marginal cost of producing each unit.[8] Similar to consumer surplus, we can construct a measure of producer surplus. For each unit of the good sold, producer surplus is measured by the difference between market price P_C and the corresponding reservation supply price on the supply curve. By adding up this difference for each value of output up to the competitive output, we obtain total producer

[8] Again, remember that the market supply curve is the horizontal summation of each competitive firm's marginal cost curve, and so the supply curve tells us exactly what is the opportunity cost to the firm of producing and selling each unit of the good.

surplus. This is illustrated by the area cbd in Figure 2.5. Note that when the equilibrium quantity, Q_C, of the good is produced and sold at price P_C, the total surplus or welfare to consumers and producers is given by the area abd.[9]

Suppose that an output greater than Q_C, say Q_G was produced in this market. For consumers to buy this quantity of the good, the price must fall to P_G. This rise in production and sales results in an increase in consumer surplus. Specifically, consumer surplus increases to aeh. Producer surplus, however, falls. Moreover, it falls by more than the increase in consumer surplus. Much of the rise in consumer surplus that results from moving to output Q_G—in particular, the shaded area cbge—is not an increase in total surplus. It simply reflects a transfer of surplus from producers to consumers. As for the additional increase in consumer surplus—the triangle *gbh*—this is clearly less than the additional decrease in producer surplus—the triangle *gfh*. Producers now receive a positive surplus only on the first Q' units produced. Because the gain in consumer surplus is less than the loss in producer surplus, the overall surplus at output Q_G is less than that at output Q_C. It is easy to repeat this analysis for any output greater than Q_C. In short, we cannot increase total surplus by raising output beyond the competitive level; we can only decrease it.

A similar thought experiment can be performed to show that output levels below Q_C also reduce the total surplus (see Practice Problem 2.4). This is because restricting output to be less than Q_C reduces consumer surplus by more than it raises producer surplus. Accordingly, the overall surplus at an output below Q_C must be smaller than the surplus under perfect competition. Note that saying that neither an increase nor a decrease in output from Q_C can increase the total surplus but only decrease it is equivalent to saying that the surplus is maximized at Q_C. Yet if we cannot increase the total surplus then we cannot make anyone better off without making someone worse off. That is, if we cannot make the size of the pie bigger, we can only give more to some individuals by giving less to others. Since this is the case under perfect competition, the perfectly competitive output level is efficient.[10]

Practice Problem 2.4

Let's return to the cellular phone industry when it was organized as a perfectly competitive industry. Use the information in Practice Problem 2.1 to work out consumer surplus and producer surplus in a competitive equilibrium.

a. Show that when $Q^C = 500$ units and $P^C = \$30$ per unit then consumer surplus is equal to \$22,500 and producer surplus is equal to \$5,000. This results in a total surplus equal to \$27,500.
b. Show that when an output of 275 units is produced in this industry the sum of consumer and producer surplus falls to \$21,931.25.

9 Observe that the unit of measurement of the areas of consumer and producer surplus is the dollar. To work out the areas, you must take \$/unit as measured on the vertical axis times units on the horizontal axis. This gives you a measure in dollars, which is a money measure of the welfare created by having this good produced at output level Q_C and sold at price P_C.

10 We focus here on the concept of allocational or static efficiency in which we examine the best way to allocate resources for the production of a given set of goods and services with a given technology. Dynamic efficiency, which considers the allocation of resources so as to promote the development of new goods and new production techniques, is addressed explicitly in Chapter 22.

2.3.1 The Monopolist and Producer Surplus

Now consider the monopoly outcome. We have suggested that this is inefficient. If this is the case, then it must be possible to show that by producing an output level different from the monopoly output Q_M, one individual can be made better off and no one else worse off. The way to show this is similar to the solution to Practice Problem 2.4 and is shown in Figure 2.6. This figure shows the competitive output and price, Q_C and P_C respectively, much as in Figure 2.5. However, in Figure 2.6 we also show what happens when the industry is monopolized. The monopolist produces output Q_M and sets price P_M. Consumer surplus is then the triangle jax. The monopolist's profit at Q_M is measured by area *jxzk*. The sum of these two surpluses is *axzk*. This is clearly smaller than the area *ayk*, which measures the total surplus obtained in the perfectly competitive outcome.

It is worth noting that while the total surplus is greater under perfect competition than it is under monopoly, the opposite holds true for producer surplus. True, a move from monopoly to competition gains the producer surplus *wyz*. But to achieve this gain requires setting the competitive price P_C and the consequent loss of the firm's surplus, *hjxw*. The loss is obviously greater than the gain.

Note that the reduction in consumer surplus that monopoly causes is not purely the result of an increase in the monopolist's surplus. Quite to the contrary, the decline in total surplus alerts us to the fact that the monopolist's gain is less than the consumer's loss. In other words, as a result of moving from a competitive industry to one of monopoly, consumers lose more than the profit that the monopolist earns. They also lose an additional amount—the area *xwy* in Figure 2.6—beyond that part of their surplus that is transferred to the monopolist.

The area of the shaded triangle *xyz* is an exact measure of inefficiency under monopoly. The upper boundary of this triangle is comprised of points that lie on the consumers' demand curve. Every point on this boundary indicates the marginal value that consumers place on successive increases in output beyond Q_M. The lower boundary of this triangle traces out the

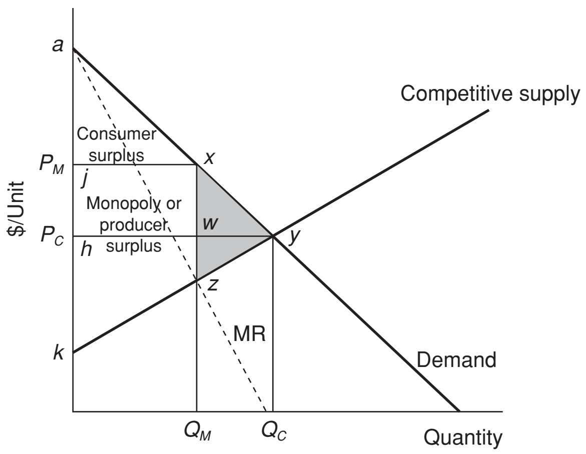

Figure 2.6 The deadweight loss of monopoly
The monopolist produces Q_M units and sells each at price P_M. A competitive industry produces Q_C units and each sells at a price of P_C. The deadweight loss caused by a move from competition to monopoly is triangle *xyz*.

marginal cost of producing this additional output. The triangle *xyz* thus reflects all the trades that generate a surplus which do not take place under monopoly. Within this triangle, the price consumers would willingly pay exceeds the cost of producing extra units and this difference is the surplus lost—that is, earned by no one—due to monopolization of the industry. If this additional output were produced, there would be a way to distribute it and make one person better off without lowering the profit of the monopolist or the welfare of any other individual. The triangle *xyz* is often referred to as the deadweight loss of monopoly. It is also a good approximation of the gains to be had by restructuring the industry to make it a competitive one.

The deadweight loss in Figure 2.6 is not due to the excess profit of the monopolist. From the viewpoint of economic efficiency, we do not care whether the surplus generated in a market goes to consumers—as it does under perfect competition—or to producers. The welfare triangle in Figure 2.6 is a loss because it reflects the potential surplus that would have gone to someone—consumers or producers—had the efficient output been produced. It is not the division of the surplus but its total amount that is addressed by economic efficiency.

Efficiency is a powerful concept both because of its underlying logic and because it is open to explicit computation. With appropriate statistical techniques, economists can try to calculate the deadweight loss of Figure 2.6 for a given industry. Hence, they can estimate the potential gains from moving to a more competitively structured market.

2.5 Practice Problem

Water is produced and sold by the government. Demand for water is represented by the linear function $Q = 50 - 2P$. The total cost function for water production is also a linear function: $TC(Q) = 100 + 10Q$. You will also need to work out both the average cost of production, denoted by $AC(Q)$, equal to the total cost of producing a quantity of output divided by that quantity of output, $TC(Q)/Q$, and the marginal cost of production, denoted by $MC(Q)$, which is the additional cost incurred to produce one more unit.

a. How much should the government charge per unit of water in order to reach the efficient allocation?
b. How much should it charge if it wishes to maximize profit from the sale of water?
c. What is the value of the efficiency loss that results from charging the price in part b rather than the price determined in part a?

2.3.2 The Nonsurplus Approach to Economic Efficiency[11]

In considering the deadweight loss of monopoly, it is useful to pursue the question as to why the monopolist fails to earn that lost triangle of surplus. If it is there for the taking, why doesn't she go out and get it? After all, the monopolist is the only seller in the market. Shouldn't she be able to use her power to extract this additional profit?

Our concept of surplus provides a useful tool with which to consider this question. Suppose that the monopolist expands output from Q_M to the competitive level of Q_C. By doing so,

[11] This section and the previous one make extensive use of the nonsurplus approach developed in Makowski and Ostroy (1995). It has had an important influence on our understanding of market participation. It also plays a central role in the business strategies advocated by Brandenburger and Nalebuff (1996).

the monopolist will indeed generate an increase in the total surplus exactly equal to the deadweight loss. That's the good news. The bad news is that the monopolist cannot appropriate all of this gain for herself. To begin with, some of the surplus generated by selling an additional $Q_C - Q_M$ units at price P_C will flow to those consumers lucky enough to buy these goods at this lower price. Yet many of these consumers were willing to pay more than P_C for this additional consumption. The surplus that these individuals enjoy as a result of acquiring the good while paying only P_C is surplus that the monopolist cannot claim. Moreover, the monopolist must confront a second problem, as well. The monopolist cannot sell the same good at two different prices. If she tried to do so, she would find it very difficult to get anyone to buy at the higher price, P_M. Those who buy the product at the lower price P_C can make an easy economic profit by reselling the good to anyone to whom the monopolist tries to charge P_M. What this means is that selling the additional $Q_C - Q_M$ units requires that the price fall to P_C on every unit sold and not just on the extra $Q_C - Q_M$ units. Yet this price cut lowers the monopolist's profit on the initial Q_M units. It thereby further reduces the surplus that flows to the monopolist as a result of selling the extra $Q_M - Q_C$ units.

Indeed, even in our original equilibrium with output at Q_M, the monopoly firm was generating more total surplus than it was actually reaping as profit. To see this, just observe what would happen if the monopoly closed shop and left the market entirely. Not only would the monopoly profit be lost, but—and this is the crucial point—consumer surplus would vanish as well. Viewed in this light, we see that the monopolist always creates surplus that she does not get. If the monopolist could appropriate the entire surplus created in the market, then she would have an incentive to produce the output that maximizes that surplus—the efficient production level. It is a monopoly firm's inability to appropriate the surplus its production creates that leads it to choose an inefficient output level.

It may seem strange to say that a monopoly firm, which earns some surplus, underproduces just because it does not get the entire surplus when, by comparison, a competitive industry, in which each firm gets no surplus, achieves the efficient higher level of output. Remember, though, that we are making our comparisons at the firm level, not the industry level. The monopoly firm is a large producer relative to the market. Its choice of output materially alters the market supply and hence the market price. It thereby alters the surplus of consumers as well. This is not the case for the competitive firm. A perfectly competitive firm's supply is tiny relative to the market. Indeed, it is so small that its output decision has no effect on market price. Drop any one competitive firm from the market and nothing happens to either the market price or the industry's total output. That is what we mean in calling a competitive firm a "price-taker." But if the competitive firm cannot change the market price it also cannot change anyone's surplus. Again, this is not the case for the competitive industry overall. Taken together, all the firms in that industry do affect the total surplus. If we drop them all from the market, that total surplus will decline.

However, decisions are made at the level of the individual firm. So, we must look at the incentives facing a single competitive producer. Here we see that such a firm does capture the entire surplus its actions generate. It earns zero profit from its market participation and, as we have just seen, this is an exact measure of the contribution the firm makes to the total surplus. So, the perfectly competitive firm gets out of the market exactly what that firm puts in.

In contrast, the monopoly firm does not get the entire surplus that its participation in the market generates, even though it does earn a positive profit. As shown earlier, that profit is less than the surplus the monopolist generates. Since the monopoly firm gets less than what it puts in, it should not be surprising that its output choice is inefficiently small. We hasten to add that this approach to monopoly is not presented to garner sympathy for the monopoly firm. Our aim is rather to clarify the source of inefficiency under monopoly. If the monopolist could collect as profit the entire surplus its production generates, it would have every incentive to produce the efficient level of output.

Indeed, the real source of the monopoly problem is not the fact that only one firm is active in the market. The true cause of the inefficiency is that the firm is large relative to the market size. To see this, consider a simple example in which the monopolist is a reproducer of classic cars. Suppose that in particular, the monopolist in question is the only maker of reproductions of the classic 1939 Rolls-Royce Wraith. Suppose further that because of limited supplies of parts and materials, the reproduction artist can only produce two such cars—each at a cost of $80,000. Demand however is not so limited. There are 50,000 classic car collectors in the world. Of these, the 200 who value the cars the most are each willing to pay a price of $150,000—but not a penny more—to own precisely one of these autos. The next 40,000 are each willing to pay $130,000 to own one car. The remaining 9,800 will willingly pay $100,000 to own a reproduction Rolls-Royce Wraith. In short, the market is characterized by some variety in consumer tastes.

The key point to note is that monopoly does not result in inefficiency. This is because whether he produces and sells none or one or the maximum of two cars, the market price of the reproductions will remain at $150,000 apiece. If the monopolist sells both cars, he will sell them to two different buyers, each of whom is among the 200 collectors willing to pay $150,000. If he decides to sell just one car, he again sells it for $150,000, this time dealing with only one buyer. Finally, if he sells no reproduced autos, no price will be recorded, but there will be an implicit opportunity cost of $150,000 incurred for each car not produced and sold. In short, the antique car producer cannot move the market price for one car away from $150,000 even though he is a monopolist.

Note that any buyer who pays $150,000 for one of the cars enjoys a zero surplus from the deal. The fact that $150,000 is exactly the maximum price that such a buyer is willing to pay indicates that the buyer is essentially indifferent between purchasing the car at that price and not buying it at all. In other words, such a buyer gets no surplus, implying that the car builder appropriates the entire surplus that building and selling a reproduced Rolls generates. Alternatively, if the monopolist were to leave—or, equivalently, not sell any cars—the surplus enjoyed by all other market participants would be unchanged. So, whether the monopolist sells both cars at the market-clearing price of $150,000, or whether he does not participate in the market at all, his actions leave unchanged the surplus of each and every other antique car market participant.

Obviously, the above story is a little contrived. Still, it serves to make the point that monopoly, per se, is not the source of market inefficiency. The car firm owner has a monopoly, but its supply of cars is small relative to the potential market. His situation is therefore similar to the one that describes a perfectly competitive firm and not the standard monopolist. Just like the perfect competitor, the car sellers's decision on how many cars to sell has no effect on the price. Matters would have been quite different had we assumed that there was only one collector willing to pay $150,000 to own a classic Rolls-Royce, while all other collectors were willing to pay only $20,000 apiece for such an automobile. In this second case, the example is more like the standard monopoly case. The car owner's choice of how many cars to sell affects the equilibrium price and the surplus of others as well.

The foregoing analysis that focuses on market actions and the surplus that they generate is called the nonsurplus approach to understanding economic efficiency. It makes the important connection between the incentive to trade in a market and the efficiency of market trading. Firms are motivated by profit to trade. Under perfect competition, a single firm's (zero) profit is equal to that firm's contribution to the surplus or welfare created by market trading. So, profit-maximizing behavior leads to an efficient market outcome. By contrast, the (large) monopoly firm's profit is less than the surplus created by market trading. Consequently, profit maximization under monopoly does not lead to an efficient market outcome.

Summary

We have formally presented the basic microeconomic analysis of markets characterized by either perfect competition or perfect monopoly. In both cases, the goal of any firm is assumed to be to maximize profit. The necessary condition for profit maximization is that the firm produce where marginal revenue equals marginal cost. Because firms in competitive markets take market price as given, price equals marginal revenue for the competitive firm. As a result, the competitive market equilibrium is one in which price is set equal to marginal cost. In turn, this implies that the competitive market equilibrium is efficient in that it maximizes the sum of producer and consumer surplus.

The pure monopoly case does not yield an efficient outcome. The monopoly firm understands that it can affect the market price and this implies that marginal revenue will be less than the price for a monopoly firm. If the market demand curve is linear, this difference is reflected in the fact that the monopolist's marginal revenue curve has the same price intercept but is twice as steeply sloped as that demand curve. For the monopoly firm equating marginal revenue with marginal cost as required for profit maximization yields an output inefficiently below that of the competitive equilibrium. Resources are misallocated because too few resources are employed in the production of the monopolized commodity. The inefficiency that results is often called the deadweight or welfare loss of monopoly.

Pure competition and pure monopoly are useful market concepts. Whether they are also useful as a description of actual industries is another question. To answer that question we need some way to determine if a market is monopolized; or if it is quite competitive. That is, we need to develop some way to identify or to measure monopoly power. It is that issue that we address in the next chapter.

Problems

1. Suppose that the annual demand for prescription antidepressants such as Prozac, Paxil, and Zoloft is, in inverse form, given by: $P = 1{,}000 - 0.025Q$. Suppose that the competitive supply curve is given by: $P = 150 + 0.033Q$.
   a. Calculate the equilibrium price and annual quantity of antidepressants.
   b. Calculate (i) producer surplus; and (ii) consumer surplus in this competitive equilibrium.

2. Assume that the dairy industry is initially in a perfectly competitive equilibrium. Assume that, in the long run, the technology is such that average cost is constant at all levels of output. Suppose that producers agree to form an association and behave as a profit-maximizing monopolist. Explain clearly in a diagram the effects on (a) market price, (b) equilibrium output, (c) economic profit, (d) consumer surplus, and (e) efficiency.

3. Suppose that the total cost of producing pizzas for the typical firm in a local town is given by: $C(q) = 2q + 2q^2$. In turn, marginal cost is given by: $MC = 2 + 4q$. (If you know calculus, you should be able to derive this expression for marginal cost.)
   a. Show that the competitive supply behavior of the typical pizza firm is described by: $q = \frac{P}{4} - \frac{1}{2}$.
   b. If there are 100 firms in the industry each acting as a perfect competitor, show that the market supply curve is, in inverse form, given by: $P = 2 + Q/25$.

4. Let the market demand for widgets be described by $Q = 1{,}000 - 50P$. Suppose further that widgets can be produced at a constant average and marginal cost of \$10 per unit.
   a. Calculate the market output and price under perfect competition and under monopoly.

b. Define the point elasticity of demand ε_D at a particular price and quantity combination as the ratio of price to quantity times the slope of the demand curve, $\Delta Q/\Delta P$, all multiplied by -1. That is, $\eta_D = -\frac{P}{Q}\frac{\Delta Q}{\Delta P}$. What is the elasticity of demand in the competitive equilibrium? What is the elasticity of demand in the monopoly equilibrium?

c. Denote marginal cost as MC. Show that in the monopoly equilibrium, the following condition is satisfied: $\frac{P - MC}{P} = -\frac{1}{n_D}$.

5. We mentioned Tyco International and its control of the plastic hangar market in the chapter. Suppose that the inverse demand for hangars is given by: $P = 3 - Q/16{,}000$. Suppose further that the marginal cost of producing hangars is constant at \$1.
   a. What is the equilibrium price and quantity of hangars if the market is competitive?
   b. What is the equilibrium price and quantity of hangars if the market is monopolized?
   c. What is the deadweight or welfare loss of monopoly in this market?

6. A single firm monopolizes the entire market for single-lever, ball-type faucets which it can produce at a constant average and marginal cost of $AC = MC = 10$. Originally, the firm faces a market demand curve given by $Q = 60 - P$.
   a. Calculate the profit-maximizing price and quantity combination for the firm. What is the firm's profit?
   b. Suppose that the market demand curve shifts outward and becomes steeper. Market demand is now described as $Q = 45 - 0.5P$. What is the firm's profit-maximizing price and quantity combination now? What is the firm's profit?
   c. Instead of the demand function assumed in part b, assume instead that market demand shifts outward and becomes flatter. It is described by $Q = 100 - 2P$. Now what is the firm's profit-maximizing price and quantity combination? What is the firm's profit?
   d. Graph the three different situations in parts (a), (b), and (c). Based on what you observe, explain why there is no supply curve for a firm with monopoly power.

References

Baumol, W. J., J. C. Panzar, and R. D. Willig. 1982. *Contestable Markets and the Theory of Market Structure*. New York: Harcourt, Brace, Jovanovich.

Brandenburger, A. and B. Nalebuff. 1996. *Co-opetition*. Cambridge: Harvard University Press.

Makowski, L., and J. Ostroy. 1995. "Appropriation and Efficiency: A Revision of the First Theorem of Welfare Economics." *American Economic Review* 85 (September): 808–27.