Table 1.1 Index of antitrust laws around the world; maximum score possible is 32

United States	21	Italy	15	Denmark	12
Ukraine	20	Czech Republic	14	Brazil	11
Turkey	19	Israel	14	Costa Rica	11
Belgium	18	Korea	14	Finland	11
Latvia	18	Slovenia	14	Norway	11
Poland	18	Taiwan	14	Germany	10
Romania	18	Venezuela	14	Jamaica	10
Argentina	17	Zambia	14	New Zealand	10
Lithuania	17	Australia	13	Panama	10
South Africa	17	Canada	13	Sri Lanka	10
Uzbekistan	17	Indonesia	13	Tunisia	10
France	16	Macedonia	13	Japan	9
Ireland	16	Mexico	13	United Kingdom	9
Kenya	16	Peru	13	Yugoslavia	8
Slovakia	16	Spain	13	Netherlands	7
Sweden	16	Thailand	13	Chile	4
Croatia	15	Armenia	12	Malta	4
Estonia	15				

Summary

Industrial organization is the study of imperfect competition. Industrial economists are interested in markets that we actually encounter in the real world. However, these real world markets come in many shapes and flavors. For example, some are comprised of a few large firms; some have one large firm and many smaller ones. In some, the products are greatly differentiated while in others they are nearly identical. Some firms compete largely by trying to keep prices as low as possible. In other markets, advertising and other forms of non-price competition are the dominant tactics. This range of possibilities has meant that over time, industrial organization has become a field rich with practical insights regarding real business behavior and public policy. This book is all about these developments.

Firms in imperfectly competitive industries need to make strategic decisions—that is, decisions that will have an identifiable impact on other participants in the market, be they rival firms, suppliers, or distributors. As a result, making any such choice must inevitably involve some consideration of how these other players in the game will react. Examples of such strategic choice variables include price, product design, decisions to expand capacity, and whether or not to invest heavily in research and development of a new product. This book presents the modern analysis of market situations involving such strategic interaction—an analysis that is rooted in noncooperative game theory. We use this analysis to examine such issues as why there are so many varieties of cereals, or how firms maintain a price-fixing agreement, or how advertising and product innovation affect the nature of competition. We also describe how the predictions of these models have been tested.

Our interest is more than just determining the profit-maximizing strategies that firms in a particular market context should adopt. As economists we are interested in the market outcomes that result when firms adopt such strategies, and whether those outcomes are close to those of the competitive ideal. If not, we then need to ask whether and how public policy can improve market allocations. Our hope is to convey the value of economic research and the gains from learning "to think like an economist." More generally, we hope to demonstrate the vitality and relevance of industrial organization, both in theory and in practice.

Problems

1. List three markets that you think are imperfectly competitive. Explain your reasoning.
2. Explain why a perfectly competitive market does not reflect a setting of strategic interaction.
3. The Appendix to this chapter lists the current, major antitrust laws of the US. Review Sections 2 and 7 of the Clayton Act. What potential threats to competition do these sections address?
4. Suppose that sophisticated statistical research provides clear evidence that, all else equal, worker productivity increases as industrial concentration increases. How would you interpret this finding?
5. Why do you think that the US courts have consistently disallowed any form of price-fixing agreements among different firms but have been more tolerant of market dominance by one firm?

References

Bain, Joseph. 1956. *Barriers to New Competition*. Cambridge: Harvard University Press.

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

Berki, Sylvester, ed. 1996. *Antitrust Policy: Economics and Law*. Boston: D.C. Heath and Company.

Chamberlin, E. H. 1933. *The Theory of Monopolistic Competition*. Cambridge: Harvard University Press.

Keynes, J. M. 1935. *The General Theory of Employment, Interest and Money*. New York: Harcourt Brace and Company.

Kovacic, W. E., and C. Shapiro. 2000. "Antitrust Policy: A Century of Legal and Economic Thinking." *Journal of Economic Perspectives*, 14:43–60.

Kwoka, J. E., and L. J. White. 2004. *The Antitrust Revolution: Economics, Competition, and Policy*. 4th Edition. Oxford: Oxford University Press.

Mason, E. S. 1939. "Price and Production Policies of Large Scale Enterprise." *American Economic Review*, 29:61–74.

Mueller, Dennis C. 1996. "Lessons from the United State's Antitrust History." *International Journal of Industrial Organization*, 14:415–45.

Nash, J. 1951. "Noncooperative Games." *Annals of Mathematics*, 54:286–95.

Nicholson, M. W. 2008. "An Antitrust Law Index for Empirical Analysis of International Competition Policy." *Journal of Competition Law and Economics*, 4 (December): 1009–1029.

Posner, Richard. 1970. "A Statistical Study of Antitrust Enforcement." *Journal of Law and Economics*, 13 (October): 365–419.

Ritter, G. 1997. *Goldbugs and Greenbacks: The AntiMonopoly Tradition and the Politics of Finance in America, 1865–1896*. Cambridge: Cambridge University Press.

Schelling, T. 1959. *The Strategy of Conflict*. New York: Oxford University Press.

Schmalensee, R. 1988. "Industrial Economics: An Overview." *Economic Journal*, 98:643–81.

Smith, A. 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. (Reprint). Oxford: Oxford University Press, 1993.

Von Neumann, J., and O. Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Appendix

Excerpts from Key Antitrust Statutes

THE SHERMAN ACT

Sec. 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. Every person who shall make any contract or engage in any combination or conspiracy hereby declared to be illegal shall be deemed guilty of a felony, and, on conviction thereof, shall be punished by fine not exceeding $100,000,000 if a corporation, or, if any other person, $350,000, or by imprisonment not exceeding three years, or by both said punishments, in the discretion of the court.

Sec. 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, and, on conviction thereof, shall be punished by fine not exceeding $100,000,000 if a corporation, or, if any other person, $350,000, or by imprisonment not exceeding three years, or by both said punishments, in the discretion of the court.

THE CLAYTON ACT, INCLUDING KEY AMENDMENTS OF THE ROBINSON-PATMAN ACT AND CELLER-KEFAUVER ACT

Sec. 2.

(a) Price; selection of customers

It shall be unlawful for any person engaged in commerce, in the course of such commerce, either directly or indirectly, to discriminate in price between different purchasers of commodities of like grade and quality, where either or any of the purchases involved in such discrimination are in commerce, where such commodities are sold for use, consumption, or resale within the United States or any Territory thereof or the District of Columbia or any insular possession or other place under the jurisdiction of the United States, and where the effect of such discrimination may be substantially to lessen competition or tend to create a monopoly in any line of commerce, or to injure, destroy, or prevent competition with any person who either grants or knowingly receives the benefit of such discrimination, or with customers of either of them: Provided, That nothing herein contained shall prevent differentials which make only due allowance for differences in the cost of manufacture, sale, or delivery resulting from the differing methods or quantities in which such commodities are to such purchasers sold or delivered: Provided, however, That the Federal Trade Commission may, after due investigation and hearing to all interested parties, fix and establish quantity limits, and revise the same as it finds necessary, as to particular commodities or classes of commodities, where it finds that available purchasers in greater quantities are so few as to render differentials on account thereof unjustly discriminatory or promotive of monopoly in any line of commerce; and the foregoing shall then not be construed to permit differentials based on differences in quantities greater than those so fixed and established: And provided further, That nothing herein contained shall prevent persons engaged in selling goods, wares, or merchandise in commerce from selecting their own customers in bona fide transactions and not in restraint of trade: And provided further, That nothing herein contained shall prevent price changes from time to time where in response to changing conditions affecting the market for or the marketability of the goods concerned, such as but not limited to actual or imminent deterioration of perishable goods, obsolescence of seasonal goods, distress sales under court process, or sales in good faith in discontinuance of business in the goods concerned.

(b) Burden of rebutting prima-facie case of discrimination

Upon proof being made, at any hearing on a complaint under this section, that there has been discrimination in price or services or facilities furnished, the burden of rebutting the prima-facie case thus made by showing justification shall be upon the person charged with a violation of this section, and unless justification shall be affirmatively shown, the Commission is authorized to issue an order terminating the discrimination: Provided, however, That nothing herein contained shall prevent a seller rebutting the prima-facie case thus made by showing that his lower price or the furnishing of services or facilities to any purchaser or purchasers was made in good faith to meet an equally low price of a competitor, or the services or facilities furnished by a competitor.

(c) Payment or acceptance of commission, brokerage, or other compensation

It shall be unlawful for any person engaged in commerce, in the course of such commerce, to pay or grant, or to receive or accept, anything of value as a commission, brokerage, or other compensation, or any allowance or discount in lieu thereof, except for services rendered in connection with the sale or purchase of goods, wares, or merchandise, either to the other party to such transaction or to an agent, representative, or other intermediary therein where such intermediary is acting in fact for or in behalf, or is subject to the direct or indirect control, of any party to such transaction other than the person by whom such compensation is so granted or paid.

(d) Payment for services or facilities for processing or sale

It shall be unlawful for any person engaged in commerce to pay or contract for the payment of anything of value to or for the benefit of a customer of such person in the course of such commerce as compensation or in consideration for any services or facilities furnished by or through such customer in connection with the processing, handling, sale, or offering for sale of any products or commodities manufactured, sold, or offered for sale by such person, unless such payment or consideration is available on proportionally equal terms to all other customers competing in the distribution of such products or commodities.

(e) Furnishing services or facilities for processing, handling, etc.

It shall be unlawful for any person to discriminate in favor of one purchaser against another purchaser or purchasers of a commodity bought for resale, with or without processing, by contracting to furnish or furnishing, or by contributing to the furnishing of, any services or facilities connected with the processing, handling, sale, or offering for sale of such commodity so purchased upon terms not accorded to all purchasers on proportionally equal terms.

(f) Knowingly inducing or receiving discriminatory price

It shall be unlawful for any person engaged in commerce, in the course of such commerce, to be a party to, or assist in, any transaction of sale, or contract to sell, which discriminates to his knowledge against competitors of the purchaser, in that, any discount, rebate, allowance, or advertising service charge is granted to the purchaser over and above any discount, rebate, allowance, or advertising service charge available at the time of such transaction to said competitors in respect of a sale of goods of like grade, quality, and quantity; to sell, or contract to sell, goods in any part of the United States at prices lower than those exacted by said person elsewhere in the United States for the purpose of destroying competition, or eliminating a competitor in such part of the United States; or, to sell, or contract to sell, goods at unreasonably low prices for the purpose of destroying competition or eliminating a competitor.

Sec. 3.

Sale, etc., on agreement not to use goods of competitor

It shall be unlawful for any person engaged in commerce, in the course of such commerce, to lease or make a sale or contract for sale of goods, wares, merchandise, machinery, supplies, or other commodities, whether patented or unpatented, for use, consumption, or resale within the United States or any Territory thereof or the District of Columbia or any insular possession or other place under the jurisdiction of the United States, or fix a price charged therefore, or discount from, or rebate upon, such price, on the condition, agreement, or understanding that the lessee or purchaser thereof shall not use or deal in the goods, wares, merchandise, machinery, supplies, or other commodity of a competitor or competitors of the lessor seller, where the effect of such lease, sale, or contract for sale or such condition, agreement, or understanding may be to substantially lessen competition or tend to create a monopoly in any line of commerce.

Sec. 7.

No person engaged in commerce or in any activity affecting commerce shall acquire, directly or indirectly, the whole or any part of the stock or other share capital and no person subject to the jurisdiction of the Federal Trade Commission shall acquire the whole or any part of the assets of another person engaged also in commerce or in any activity affecting commerce, where in any line of commerce or in any activity affecting commerce in any section of the country, the effect of such acquisition may be substantially to lessen competition, or to tend to create a monopoly.

2

Basic Microeconomics

Considerable time passed before Adam Smith's original and intuitive insights regarding the nature of market behavior and market outcomes were translated into formal models. It then took even more time for that formal understanding to make its way into a standard professional literature. Yet by the late nineteenth century, a rigorous understanding of the benefits of competition versus monopoly had been established as evidenced in particular by the publication of Alfred Marshall's ***Principles of Economics, Vol. 1*** (1890). While we are ultimately interested in modeling the gray area that lies between competition and monopoly, a sound understanding of the perfectly competitive and pure monopolized markets is nevertheless 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 brief review of the basic models of perfect competition and monopoly.

2.1 COMPETITION VERSUS MONOPOLY: THE POLES OF MARKET PERFORMANCE

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 that 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

[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.

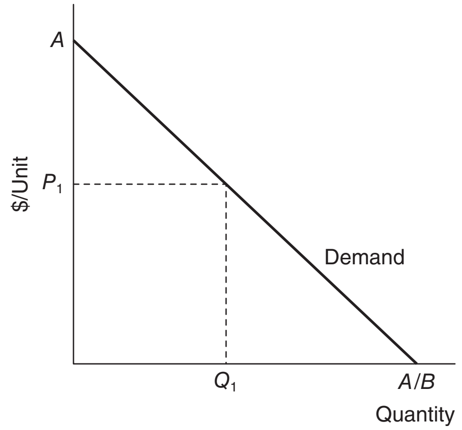

Figure 2.1 Market demand curve
The price P_1 is the marginal consumer valuation of an additional unit of output when current output is Q_1.

the maximum willingness to pay, or maximum reservation demand price that any consumer is willing to pay to have this good. At a market price greater than A, no one in this market wants to buy any of the product. As the 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. Alternatively, we may view P_1 as consumer willingness to pay at the margin. That is, P_1 is the most any consumer would pay for the last or the Q_1th unit of the good.

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, 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 altering its production. Following the tradition in microeconomics, we distinguish between two general time periods: the short run and the long run. 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. To be sure, the precise meaning of this definition may vary depending on whether we consider the short run or the long run. Yet 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 their 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. Because this would imply that the firm has some power over the market price, such a firm would not be a perfect competitor. The fact that the firm cannot affect the market price also means that its actions do not affect other producers. In turn, this means that perfectly competitive firms do not strategically interact with any of their numerous rivals.

Because the output decision of a perfectly competitive firm does not affect the going price, a graph of the inverse demand curve facing such a firm appears as a horizontal line at the current market price. No matter how much or how little the firm produces, that price remains unchanged. Note that this is the case even though the market demand curve describing the demand faced by the entire industry is downward-sloping.[2] That market demand curve indicates that if the aggregate output of all firms increases in any material way, the market price will fall. Again, though, the distinguishing feature of a perfectly competitive firm is that it is so small that its output choice has no such material impact on total industry output—or where the industry is on the market demand curve.

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. It reflects net revenue above what is necessary to pay all of the firm's inputs at least what they could

[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.

earn in alternative employment. This point is important 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 first order condition for 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 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, the marginal cost function also depends on q, according to the function $MC(q)$. This function describes the additional 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 necessary 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. Because 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 industry 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). Because 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

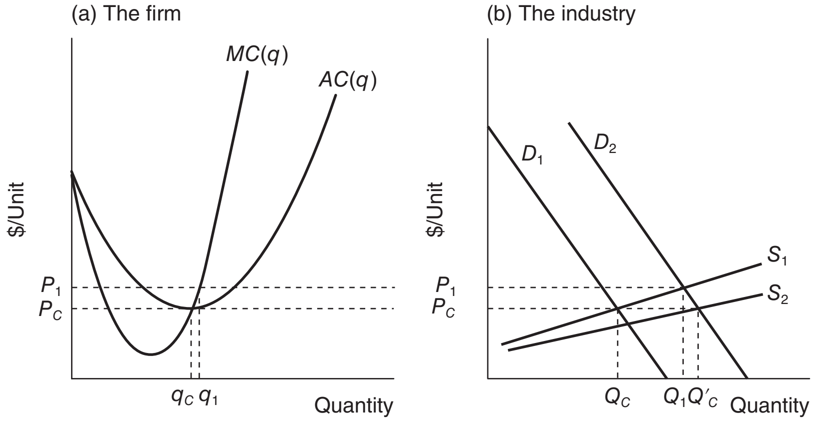

Figure 2.2 The long-run competitive equilibrium
Price P_1 is consistent with a short-run equilibrium in which each firm produces at a point where its marginal cost is equal to P_1. However, at P_1 price exceeds average cost and each firm earns a positive economic profit. This will encourage entry by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

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 or less, 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 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).

[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.

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 in excess of 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 basic requirement that they each produce where $P = MC(q)$. This leads 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 covers 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. These examples illustrate a central element in our definition of an equilibrium—namely, that no firm has the incentive to change its production plan. In the long run, this includes the idea that no firm wishes either to leave or to enter the market.

2.1 **Practice Problem**

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 fifty 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.

2.1.2 Monopoly

In a perfectly competitive market, each firm’s production of the good is tiny relative to the market total. What would happen, though, if all of these tiny sellers become consolidated into one firm, i.e., what would happen if the market were monopolized? Because this

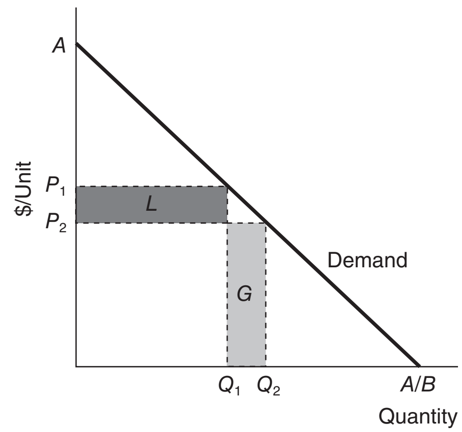

Figure 2.3 The marginal revenue from increased production for a monopolist
An increase in production from Q_1 to Q_2 causes a gain in revenues approximated by area G, and a loss in revenues approximated by area L. The net change or marginal revenue is therefore $G - L$. Note, because the firm is a monopolist,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

single firm would be the only supplier of the good, its demand curve is identical with the market demand curve. In complete contrast to the competitive firm, this monopoly firm can decisively influence the price in this market because its choices decisively alter the total supply. Recognizing this effect, the monopolist's profit-maximizing behavior will differ from that of a competitive firm and so will change the market equilibrium.

Because the monopolist's demand curve slopes downward, any increased production by the monopolist will lead to a price reduction. Consider, for instance, Figure 2.3. Here, a monopolist who is considering selling either Q_1 or Q_2 units faces the following dilemma. If production is restricted to Q_1 units, the market price will be P_1. If instead production is set at Q_2 however, the market price will fall to P_2. 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.

In Figure 2.3, the components of marginal revenue for a monopolist are illustrated by the shaded areas G and L. 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 as $\frac{\Delta P}{\Delta Q}$. 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 $-B\Delta Q$. Because total revenue is defined 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 very 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. It follows that 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 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 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 monopolist therefore 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 cost of producing output 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

[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.

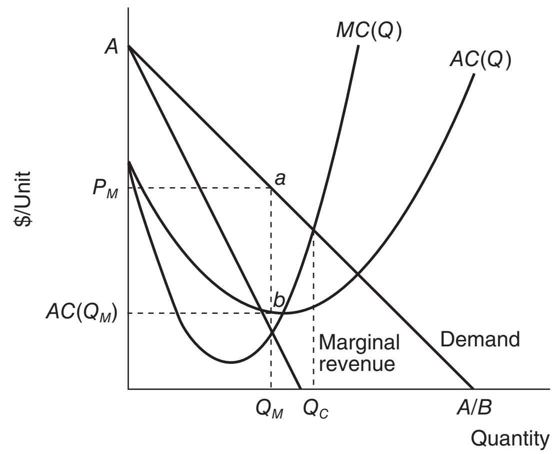

Figure 2.4 The textbook monopoly case
The monopolist maximizes profit by choosing the output, Q_M, at which marginal revenue equals marginal cost. The price at which this output can be sold is identified by the demand curve as P_M, which exceeds marginal cost. Profit is $P_M abAC(Q_M)$. The competitive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

total revenue exceeds total cost, and the monopolist earns a positive economic profit. This 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 demand conditions, the monopolist has no incentive to sell more or to sell less.

Practice Problem 2.2

Suppose that the cellular phone market described in Practice Problem 2.1, is monopolized. The monopolist has 50 identical plants to run, each with the same cost function as described in that problem. Hence, the 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 $Q^D = \dfrac{6000 - 50P}{9}$

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?

[5] Cost minimization for any multiplant firm requires that the marginal unit be produced at the plant with the lowest marginal cost. With divisible output, this implies that the profit-maximizing monopolist will want to allocate total production across the fifty plants in such a way that the marginal cost of producing the last unit of output is the same in each plant. Therefore, the monopolist derives the overall marginal cost function in a manner similar to how we constructed the supply function for the competitive industry. See Chapter 4, footnote 3 and the Appendix.

2.2 ECONOMIC EFFICIENCY AND THE SOCIAL SURPLUS

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 judged harshly by economists. 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 without hurting any others.[6] 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.

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. 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.

2.2.1 Economic Efficiency and Surplus in a Competitive Market

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*.

[6] 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.

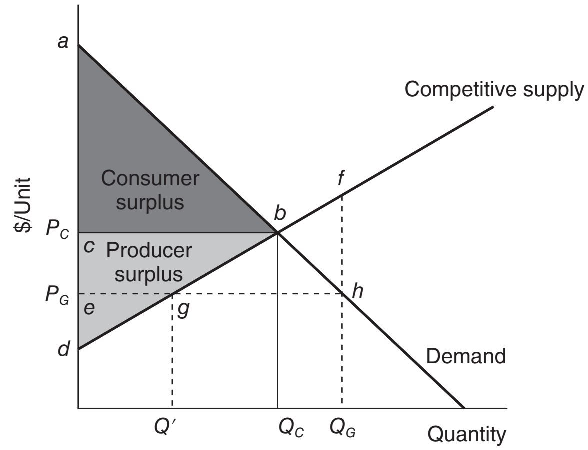

Figure 2.5 Competition maximizes the total surplus
At the competitive price P_C and output Q_C, consumers enjoy a surplus equal to triangle *abc*. Producers enjoy a surplus equal to triangle *cbd*. This is the maximum. Producing less would lose some of the total surplus given by triangle *abd*. Subsidizing 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 thus, 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.

For competitive producers, the supply curve tells us the marginal cost of producing each unit.[7] 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 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*.[8]

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

7 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.

8 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.

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 will show that output levels below Q_C also reduce the total surplus (see Practice Problem 2.3). 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 total 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. Because this is the case under perfect competition, the perfectly competitive output is efficient.[9]

2.3 Practice Problem

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.

2.2.2 The Monopolist and the Social Surplus

Now consider the monopoly outcome. If this is inefficient, it must be possible to show that by producing an output level different from the monopoly output Q_M, the total surplus rises. The way to show this is similar to the solution to Practice Problem 2.3 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

9 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 20.

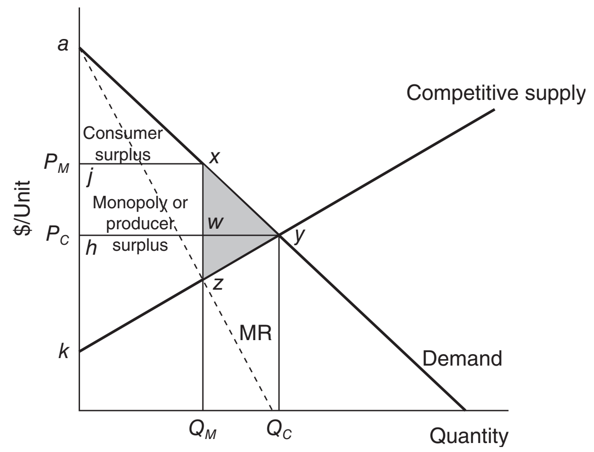

Figure 2.6 The deadweight loss of monopoly
The monopolist produces Q_M units and sells each at price P_M. A competitive industry produces Q_C units and each sells at a price of P_C. The deadweight loss caused by a move from competition to monopoly is triangle *xyz*.

setting the competitive price P_C and the consequent loss of the firm's surplus, *hjxw*. The loss is 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 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 comprises 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 marginal cost of producing this additional output. The triangle *xyz* thus reflects all the trades that generate a surplus but that do not take place under monopoly. Within this triangle, the price that 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.

Again, it is worth repeating that 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 competitive market.

2.4 **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. What fee 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 INTERTEMPORAL CONSIDERATIONS AND CONSTRAINTS ON MONOPOLY POWER

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. 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 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.

2.3.1 Discounting, Present Value, and Intertemporal Trades

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 every day of the week. These investors are thus sacrificing some of their current wealth—which could alternatively be used to purchase a Caribbean vacation, 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 later in the form of dividends and capital gains.

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 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 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. Therefore, it makes sense for your friend to agree that you give her \$1,000 today and that she give you \$1,030 in twelve 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 it would 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. Because 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 twelve 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 twenty-four 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; 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?

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 $Z_1 = Z_2 = \ldots = Z_T = \overline{Z}$. In that case, the present value of the total stream is:

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

An even more special case occurs when not only is the income receipt constant at $Z = \overline{Z}$, but also the stream persists into the indefinite future so that the terminal period T approaches infinity. In that case, because 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:

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

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.[10] 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. 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, research, and development, which often have a multiperiod dimension.

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

Reality Checkpoint

Ticket Discounts

A recent study by the National Highway Traffic Safety Administration reveals that nearly 80 percent of drivers admit to speeding within the last month and 25 percent admit to speeding that day. Given attitudes regarding what truly constitutes speeding and willingness to admit to illegal behavior, a prudent guess is that something like 50 percent of the drivers are driving above the posted speed limits on multi-lane interstate highways at any time.

Of course, the prospect of being pulled over and ticketed by the police is a deterrent to overly fast driving. Fines vary by state but a fine of $200 for those going 15 miles over the posted limit is common, and that fine is just part of the cost. Another major part is the impact of a speeding conviction on one's auto insurance. In Massachusetts, for example, one speeding ticket adds 30 percent to the cost of insurance in four areas: 1) bodily injury, 2) property damage, 3) personal injury, and 4) collision for a total of $300 extra in annual insurance cost for at least three years. Assuming a discount factor of $R = 0.97$ (an interest rate of about 3 percent), the cost of a speeding ticket is: $\$200 + \$300 + R\$300 + R^2\$300 = \$1073$.

Not all speeders are caught. Rough evidence from Massachusetts and Virginia suggests that a typical driver averaging 10–15 miles over the speed limit has about a 15 percent chance of being ticketed every year. Hence, over a three-year period, the average driver, for whom the probability of a second speeding ticket is basically zero, has an expected speeding cost of: $0.15^*\$1073 + 0.15\text{R}\$1073 + 0.15\text{R}^2\$1073 = \$469$. This cost can be avoided though by purchasing a high quality radar (and lidar) detector such as Escort's Passport 9500i, Cobra's XR5-R9G, or the Beltronics Bel STi Driver. These retail for $450 to $500 and last about three years before they either become overly sensitive and give so many false warnings that drivers turn them off or police update their technologies to make that generation of detectors obsolete.

So, the net benefit of buying a detector for a typical driver is small. For the true super-speeders, though, who regularly push the pedal to the metal and who therefore face substantially higher probabilities of getting a ticket, purchasing a detector may well seem a good investment . . . but slowing down may be cheaper.

Source: J. Welsh "No Radar Detectors Give Speed Freaks a Rush," *Wall Street Journal*, January 10, 2008, p. D1. National Highway Traffic Safety Administration, *National Survey of Speeding and Unsafe Driving Attitudes & Behaviors*, (2003).

2.5 **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 thus 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 newly combined firm, Suite & Loew, as a monopoly earning $110,000 in all subsequent periods. Is this a good investment?

2.3.2 Time and the Evolution of Industry Structure

Considerations of time also introduce some potential constraints on the exercise of monopoly power and, consequently, on the amount of welfare loss that accompanies such power. One such mechanism reflects the dynamics of industrial structure. In our textbook analysis above, we simply assumed that the market had either a competitive or monopoly structure. We did not ask how that structure was achieved or whether it is stable. In this sense, the analysis is short run in character. This is not to say that it is inaccurate. The welfare loss associated with monopoly pricing is real. There may be some question, though, as to whether it will be long lasting. Whether or not this is the case depends in part on whether the economic profit or surplus that the monopoly firm earns will attract entry and competition over time.

As noted earlier, a sensible requirement for a long-run equilibrium is that there is no incentive for the industry structure to change. In other words, there should be no incentive for any firm to exit or enter the industry. There are many ways to impose this condition. The most common one however is to assume that it requires that each firm in the industry earn zero economic profit.[11] In other words, for each firm, total revenue Pq_i equals total cost $C(q_i)$. Dividing each side by the firm's output, $q_{i,}$ reveals that this is equivalent to imposing the constraint that firms exit or enter until the market price equals average cost, where average cost is defined as the ratio of total cost to total production or cost per unit.

We will return to a more complete discussion of average (and marginal) cost again in Chapter 4 when we discuss the production technology and its implications for the firm's costs in some detail. For now, we simply note that imposing the long-run equilibrium condition that price equals average cost provides a way to determine the equilibrium structure of the industry. To the extent that that structure has more than one firm in it, that is, to the extent that the market permits entry by rivals over time, the monopolist's ability to sustain a price far above marginal cost will be lessened. Indeed, even if the market only has one firm, the ability of that firm to price above marginal cost may be sharply curtailed by the threat of potential entrants who would come into the market readily and undercut such pricing. This in fact is the fundamental idea behind the idea of contestable markets.[12]

At the same time, it is important to recognize that a firm with monopoly power may be able to take actions that protect that power against rival entrants over time, and that push the evolution of market structure toward the monopoly pole. In short, market structure at any point in time may not be an accurate gauge of the intensity of competition in an industry. Even a monopolist can be constrained by the prospect of actual or potential entry. It is equally important to recognize that because both entry and entry deterrence typically require that a cost be incurred today in return for higher profits in the future, analysis of these issues inevitably involves comparison of profits at different points of time and the use of the discounting techniques just discussed.

2.3.3 Durable Goods and the Coase Conjecture

A second way in which attention to time considerations may reveal a constraint on monopoly power arises in connection with durable goods. Unlike, say, food, many goods such as household appliances and automobiles last a number of years. In turn, this means that a

11 An alternative approach would be to require that each firm within the industry earns nonnegative profits (and so have no incentive to leave) while any firm not in the industry would earn nonpositive profit if it entered (and so have no incentive to do so).

12 See, for example, Baumol, Panzar, and Willig (1982).

monopolist has to think carefully about the price initially set and the volume sold today. That supply will still be around to influence market outcomes one or two periods later. Nobel laureate Ronald Coase argued nearly 40 years ago (Coase 1972) that this durability might greatly reduce if not eliminate the ability of the monopolist to set prices above the efficient level even in the current period.

Consider the following example. The time frame has two discrete periods. A monopolist has two units of a durable good that will provide services to an owner in each of the two periods with no loss due to depreciation. There are two such potential consumers. One values the services of the good at \$50 per period. The other values these services at \$30 per period. Thus, for the high-value consumer the present value of services from the good is $\$50 + R\$50 = (1 + R)\$50$, while for the low-value consumer it is $(1 + R)\$30$. Because the monopolist has the two units already, marginal production cost is zero. In turn, this implies that the maximum total surplus available in this market is $(1 + R)\$80$. Any outcome that yields this surplus is efficient, but any outcome that yields a total surplus less than $(1 + R)\$80$ is inefficient.

Of course, the monopolist seeks to maximize his own surplus, not the total surplus. In doing so, he will recognize the following constraints. First, any first period price below $(1 + R)\$30$ will result in both consumers purchasing the good in the first period and the market process will effectively end, i.e., there will be no more sales in period 2. The monopolist has nothing left to sell in the second period and both consumers continue to enjoy the services of the good then because it is durable. Second, any price above $(1 + R)\$30$ will result in the sale of either one unit in the first period, if the price is less than $(1 + R)\$50$, or no units if it is higher. In the first case, the monopolist will enter the second period with one unit left, which can then be sold to the low-value consumer for \$30. In the second case, the monopolist will face the choice between selling either both units at a price of \$30 or one unit at a price of \$50. Clearly, the best option in this case is to sell both at \$30 for a total second-period profit of \$60.

An important implication of the foregoing analysis is that any and all second period sales must take place at a price of \$30. There is simply no way that the monopolist can make a credible first-period commitment to any second-period price above this amount. Given this, we can now consider the strategy of making one first-period sale to the high-value consumer at a price of $(1 + R)\$50$ from the viewpoint of that consumer. The difficulty with this strategy is then immediately apparent. Confronted with a first-period price of $(1 + R)\$50$, the high-value consumer can either buy in the first period—and earn zero surplus—or defer the purchase to the second-period knowing that the price will then be \$30, and that there will therefore be a surplus of $R\$20$ in present value terms. The high-value consumer will obviously choose the latter. It follows that the monopolist cannot hope to make a first-period sale at $(1 + R)\$50$. The fact that the second-period price must be \$30 makes this impossible.

What about a price below $(1 + R)\$50$ but above $(1 + R)\$30$? Consider the first-period price, $(1 + R)\$(30 + \varepsilon)$, where ε is a small positive constant. Such a price gives the high-value consumer a surplus of $(1 + R)\$(20 - \varepsilon)$. If the consumer instead waits and purchases the good next period at \$30, there will be a surplus with present value $R\$20$. Thus, in order for the monopolist to sell to this consumer in the first period, it must be the case that the monopolist offers a price such that

$$(1 + R)\$(20 - \varepsilon) > R\$20 \rightarrow \varepsilon \leq \frac{\$20}{1 + R} \tag{2.6}$$