I. The Definition of Monopoly

Monopoly: a firm that is the only seller of a good or service with no close substitutes.

This definition is abstract, just as the definition of perfect competition is abstract. And just as it’s hard to find a market that really seems perfectly competitive in all respects, it’s hard to find a firm that is a total monopoly.

The source of ambiguity is the term “close substitutes.” How close is close? For example, is Amtrak a monopoly? Yes, if you’re concerned with long-distance passenger rail service. Not really, if you’re concerned with local rail travel (consider SEPTA, New Jersey Transit, etc.), or with long-distance transportation (consider buslines and airlines). Definitely not if you’re concerned with transportation broadly speaking (consider all of the above plus cars).

What counts as a substitute is ultimately a matter of consumer’s preferences, which are generally not in the form of perfect complements or perfect substitutes. Still, there are various situations that seem to qualify for use of this model: local phone companies in most areas, at least until recently, municipal cable monopolies, etc.

Why might there be only one firm in a market? There can be various reasons, some of which we discussed earlier: high costs of entry, patent protection, government licensing or exclusive grants, etc. Later, we’ll talk about another: economies of scale, which can lead to natural monopoly in some cases.

II. The Single-Price Monopolist

The analysis here is pretty much identical to our generic theory of the firm. The main difference is in interpretation: here, we interpret the firm’s demand curve as the entire market demand, not just a fraction of it. As a result, we expect a monopolist’s demand curve to be steeper than a generic firm’s, because consumers lack the ability to substitute the products of other firms.

The picture is identical to that of the generic firm, except we use capital Q and capital D for quantity and demand, respectively, to indicate that we’re talking about the entire market being served by just one firm. Note that, unlike in perfect competition, we draw only a single diagram, rather than side-by-side market-and-firm.

Just as in the generic analysis, we can find whether the firm is making profits or losses by drawing in the ATC curve. And there’s nothing contradictory about a monopoly making losses. If a monopolist is making losses, it means that it doesn’t make economic sense to produce the good in question at all, because even under the most favorable conditions to
the producer he cannot get revenues sufficient to cover all opportunity costs of production.

In general, economists are hostile to monopolies, and here is why: It is apparent that the firm could produce more units, at a marginal cost lower than what the remaining customers would be willing to pay. As a result, there is a dead-weight loss (DWL) from mutually beneficial trades not made.

Recall that under perfect competition, firms priced at marginal cost (P = MC). If the monopoly did that, its price would be $P^e$ and its quantity $Q^e$, as shown in the diagram. (The “c” is for competitive.) So we say that the monopoly causes losses in wealth relative to the perfectly competitive benchmark. (However, we will later talk about reasons why this may be somewhat misleading.)

But there’s something fishy here. If the units from $Q^m$ to $Q^e$ cost less to produce than consumers will pay for them, why doesn’t the monopolist choose to produce them?

The answer depends crucially on the single-price assumption, which says that the monopolist must charge a single price for all units. This could be because the firm cannot distinguish “high-price” consumers from “low-price” consumers. The single-price assumption is responsible for MR being below demand. The firm would like to be able to sell from $Q^m$ to $Q^e$ if it could charge $P^e$ for those units. But doing that would require selling the units from zero to $Q^m$ at $P^m$ instead of $P^e$, and it’s not worth it to do that.

### III. Welfare Analysis of Monopoly (First Pass)

Welfare analysis means analyzing who gains and who loses from a policy or situation. In this case, it is clear that producers gain from monopolization or increased concentration of an industry – assuming, of course, that they are among those who remain in business.
Under perfect competition, they would make lower profit (and zero profit in the long run). Notice that the competitive price and quantity are not the profit-maximizing profit for the single firm, so it must be true that monopoly is better for the producer. Consumers, on the other hand, end up worse off. Why? Because the quantity they get is smaller, and they pay a higher price for the units they do buy. This higher price is, naturally, the reason monopolization is good for the producer. The difference between the monopoly and competitive prices, multiplied by the monopoly quantity, is called the “transfer” from monopolization.

Note that in this context, the transfer is a shift in wealth relative to what consumers would have had in a perfectly competitive market. If we imagine a previously competitive market being suddenly monopolized, the transfer is the amount of wealth taken from consumers and given to the monopolist by the act of monopolization. It is because the transfer outweighs the monopolist’s portion of DWL (the part below the competitive price) that the monopolist chooses to produce at \( Q_m \) instead of \( Q_c \).

From a pure efficiency perspective, the transfer is irrelevant. What makes monopoly inefficient is the existence of DWL.

IV. Welfare Analysis of Monopoly (Second Pass)

All of the above seems to imply that a monopoly necessarily leads to higher price, lower quantity, and dead-weight losses relative to perfect competition. But this may not be true, because a monopoly may be able to achieve cost savings unavailable to smaller firms. Consolidation to serve larger numbers of consumers can allow the firm to take advantage of greater economies of scale, moving to a lower ATC and corresponding MC.

For example, consider this diagram.
In this situation, one large firm operating at SRATC$_2$ could produce at a much lower cost than three firms operating at SRATC$_1$. The corresponding MC would tend to push down price and increase quantity, hence passing some of the benefits on to the consumers.

Thus, the picture we drew earlier of CS, PS, etc., might not be valid, because the MC under monopoly is not the same as under perfect competition. So we have two effects of greater concentration (fewer firms with greater market share each) in an industry. First, competition is reduced, which tends to drive prices up. Second, cost savings may be realized from economies of scale, which tends to drive prices down. So which effect dominates? This is an empirical question. It is highly relevant to many policy questions, such as whether the FTC will allow a merger to take place.

Obviously, much of the analysis here is useful not just for monopoly, but for any industry with imperfect competition.

V. Price Discrimination (the Multi-Price Monopoly)

As noted earlier, the monopolist would like to sell to consumers who won’t buy at the monopoly price, if it could do so without lowering price for everyone else.

So if the monopolist can distinguish between different groups of consumers, it may wish to charge differential prices. This requires, of course, being able to prevent resale of the product by low-price buyers to high-price buyers. Incidentally, this is not a phenomenon limited to monopolists; it can be done by any firm that faces a downward-sloping demand curve and has the ability to distinguish between customer groups.

Price discrimination is defined as charging different prices to different consumers on the basis of non-cost-related characteristics of the consumers. Note that it must be the same product or similar product with the same cost structure. For instance, charging a different price for first-class seats on an airplane is not an instance of (pure) price discrimination; nor is charging different prices for gasoline with different octane levels.

In general, the most important factor for price discrimination is distinguishing different elasticities, to isolate people who are more or less sensitive to price. People with more elastic demand will tend to be charged a lower price; people with more inelastic demand will tend to be charged a higher price.

Examples of price discrimination:
- Movie prices (for students and seniors)
- Airline prices (for business and personal)
- Grocery coupons (“smart” versus casual shoppers)
- 10-10-321 and other long-distance services (price sensitive v. not price sensitive)

In order for a firm to be able to price discriminate, it needs to meet three conditions. First, it has to have some degree of market power; that is, it must not be in a perfectly competitive industry. This is equivalent to the condition that the firm must face a
What are the effects of price discrimination?

- It tends to transfer wealth from consumers to producers. It does this because some inelastic consumers, who would have paid a lower price under a single-price monopoly, end up paying more under price discrimination. This is equivalent to their experiencing a larger transfer of wealth (see discussion on transfers above).
- It tends to decrease dead-weight loss for both consumers and producers. This is because the firm can charge a lower price to more elastic consumers, for units that it would otherwise not have sold.

So consumers as a whole may or may not benefit from price discrimination; it depends on which consumers one has in mind, and how the firm discriminates. (Note again that from an efficiency perspective, the DWL is what’s important. To the extent that price discrimination reduces the size of DWL, it’s more efficient than monopoly.)

Price discrimination is categorized into three main types.

- First-degree price discrimination, also known as perfect price discrimination, occurs when the firm can charge a different price for each consumer, with the price exactly equal to the maximum price the consumer will pay. A firm that perfectly price discriminates can extract the maximum possible transfer from consumers. Note that a perfectly price-discriminating monopolist has a MR that’s identical to the demand curve, and it chooses the competitive quantity. There is no DWL under perfect price discrimination. Probably the best example of perfect price discrimination is tuition at universities and colleges, which is adjusted using scholarships and grants to give each student an individualized, take-it-or-leave-it offer.
- Second-degree price discrimination, also known as quantity discrimination, the firm charges different prices based on the amount that a customer buys (e.g., lower prices for bulk purchases), but all customers face the same price schedule. Note that not all bulk discounts reflect price discrimination – the discount may reflect the firm’s ability to produce larger amounts at lower cost.
- Third-degree price discrimination, also known as multimarket price discrimination, prices are set for customer groups, but each customer faces a single price. This is the most common form of price discrimination. Most of the examples we’ve discussed are third-degree price discrimination, although long-distance service is possibly an example of second-degree price discrimination (since everyone has the option of using the 10-10-321 service).

VI. Market Concentration
When does it make sense to use a monopoly model instead of a competitive model, or vice versa? It depends on the characteristics of the market. Probably the two most important characteristics of the market are the number of firms in the market and the market share of each firm. By looking at these two factors, we can get a hint as to which model will be more useful in analyzing the industry.

One measure of market concentration is the N-firm concentration ratio. This tells you the total market share of the N largest firms in the industry. The 4-firm and 8-firm concentration ratios are the most commonly used. Market share is usually measured in terms of sales revenue, but others measures (such as production or production capacity) can also be used. Example: Suppose there is an industry who largest firm has 30% of the market, second largest 20%, third largest 10%, and fourth largest 5%. Then the 4-firm concentration ratio is 65%. The remaining percentage (35% in this case) can be interpreted as the “competitive fringe” of the market.

Another measure of market concentration is the Hirschman-Herfindahl Index (HHI). This index is constructed by taking each firm’s market share, squaring it, and adding it to all the other firms’ squared market shares. For example, if there are two firms, one with a 60% share and the other with a 40% share, the HHI is $(.6)^2 + (.4)^2 = .52$. It turns out that if all N firms have the same market share, the HHI equals $1/N$. For example, if there are five firms, the HHI is $1/5 = .2$; if there are 100 firms, the HHI is $1/100 = .01$. It is therefore possible for us to talk about the “numbers-equivalent of firms” in an industry, which is $1/HHI$. For example, if there are five firms with market shares of 50, 30, 10, and 10, the HHI is $.25 + .09 + .01 + .01 = .36$. The numbers-equivalent of firms is $1/.36 = 2.78$, or almost three firms. In other words, the industry concentration described is about a competitive as an industry with three equal-sized firms.

As a rule of thumb, regulators usually consider an industry competitive with an HHI less than .2, oligopolistic with HHI between .2 and .6, and monopolistic with HHI above .6.

There are problems both N-firm concentration ratios and HHI indices as measures of the competitiveness of an industry. First, the threat of entry can make a big difference in the competitiveness of an industry; but these measures only consider existing firms. Second, there are strategic factors that can make a big difference; sometimes, even just two firms are enough to generate intense rivalry. We will be exploring some of these issues when we begin studying game theory. Third, the degree of product differentiation can make a difference. Highly differentiated products can shield firms in low-concentration industry from competitive pressures, while homogeneous products can expose firms in a high-concentration industry to greater competitive pressures.