6 Market Imperfections

We have so far focused on unimpeded markets, and seen that markets may perform efficiently.[64] In this chapter, we examine impediments to the efficiency of markets. Some of these impediments are imposed on otherwise efficiently functioning markets, as occurs with taxes. Others, such as monopoly or pollution, are problems that may arise in some circumstances, and may require correction by the government.

6.1 Taxes

There are a variety of types of taxes, such as income taxes, property taxes, ad valorem (percentage of value) taxes, and excise taxes (taxes on a specific good like cigarettes or gasoline). Here, we are primarily concerned with *sales taxes*, which are taxes on goods and services sold at retail. Our insights into sales taxes translate naturally into some other taxes.

6.1.1 Effects of Taxes

Consider first a fixed tax such as a twenty cent tax on gasoline. The tax could either be imposed on the buyer or the supplier. It is imposed on the buyer if the buyer pays a price for the good, and then also pays the tax on top of that. Similarly, if the tax is imposed on the seller, the price charged to the buyer includes the tax. In the United States, sales taxes are generally imposed on the buyer – the stated price does not include the tax – while in Canada, the sales tax is generally imposed on the seller.

An important insight of supply and demand theory is that it doesn't matter – to anyone – whether the tax is imposed on the supplier or the buyer. The reason is that ultimately the buyer cares only about the total price paid, which is the amount the supplier gets plus the tax, and the supplier cares only about the net to the supplier, which is the total amount the buyer pays minus the tax. Thus, with a twenty cent tax, a price of $2.00 to the buyer is a price of $1.80 to the seller. Whether the buyer pays $1.80 to a seller and additional twenty cents in tax, or pays $2.00, produces the same outcome to both the buyer and seller. Similarly, from the seller's perspective, whether the sellers charge $2.00 and then pay twenty cents to the government, or charges $1.80 and pay no tax, leads to the same profit.[65]

[64] The standard term for an unimpeded market is a *free market*, which is free in the sense of "free of external rules and constraints." In this terminology, eBay is free market, even though it charges for the use of the market.

[65] There are two minor issues here that won't be considered further. First, the party who collects the tax has a legal responsibility and it could be that businesses have an easier time complying with taxes than individual consumers. The transaction costs associated with collecting taxes could create a difference arising from who pays the tax. Such differences will be ignored in this book. Second, if the tax is percentage tax, it won't matter to the outcome but the calculations are more complicated, because a ten percent tax on the seller at a seller's price of $1.80 is different from a ten percent tax on a buyer's price of $2.00. Then the equivalence between taxes imposed on the seller and taxes imposed on the buyer requires different percentages that produce the same effective tax level. In addition, there is a political issue: imposing the tax on buyers makes the presence and size of taxes more transparent to voters.

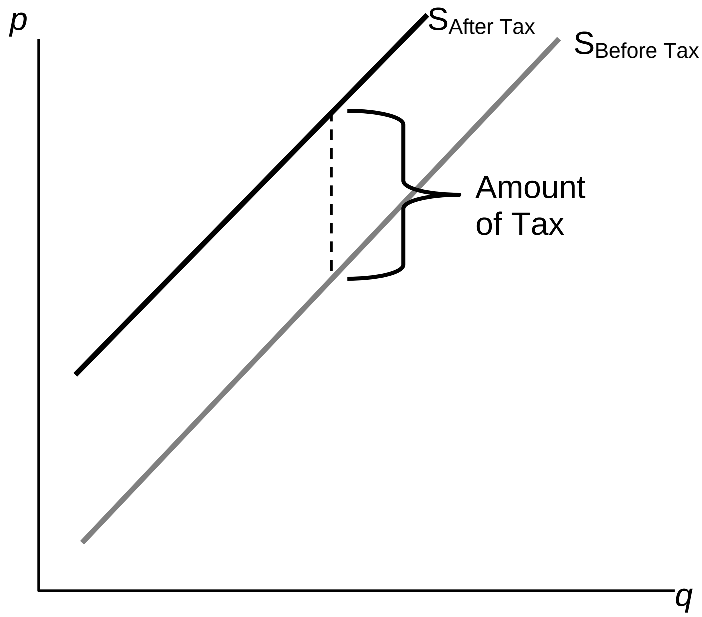

Figure 6-1: Effect of a Tax on Supply

First, consider a tax imposed on the seller. At a given price *p*, and tax *t*, each seller obtains *p*-*t*, and thus supplies the amount associated with this net price. Taking the before tax supply to be $S_{\text{Before Tax}}$, the after tax supply is shifted up by the amount of the tax. This is the amount that covers the marginal value of the last unit, plus providing for the tax. Another way of saying this is that at any lower price, the sellers would reduce the number of units offered. The change in supply is illustrated in Figure 6-1.

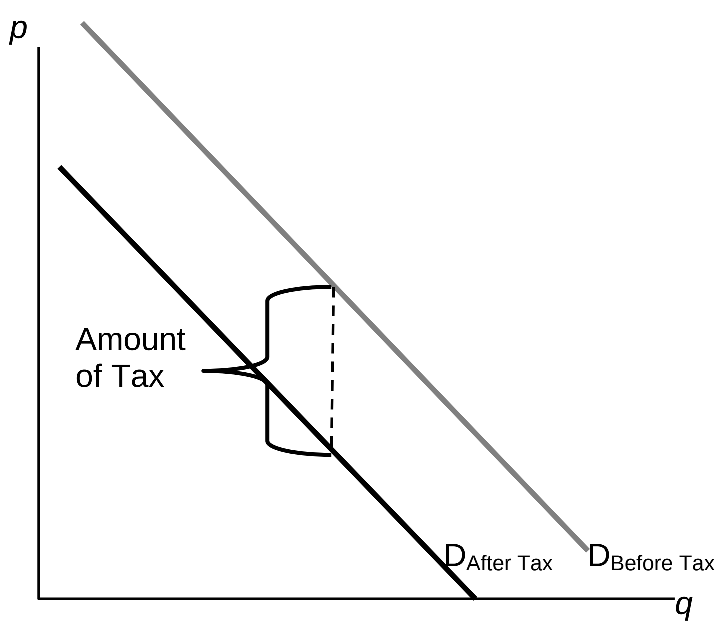

Figure 6-2: Effect of a Tax on Demand

Now consider the imposition of a tax on the buyer, illustrated in Figure 6-2. In this case, the buyer pays the price of the good, *p*, plus the tax, *t*. This reduces the willingness to pay for any given unit by the amount of the tax, thus shifting down the demand curve by the amount of the tax.

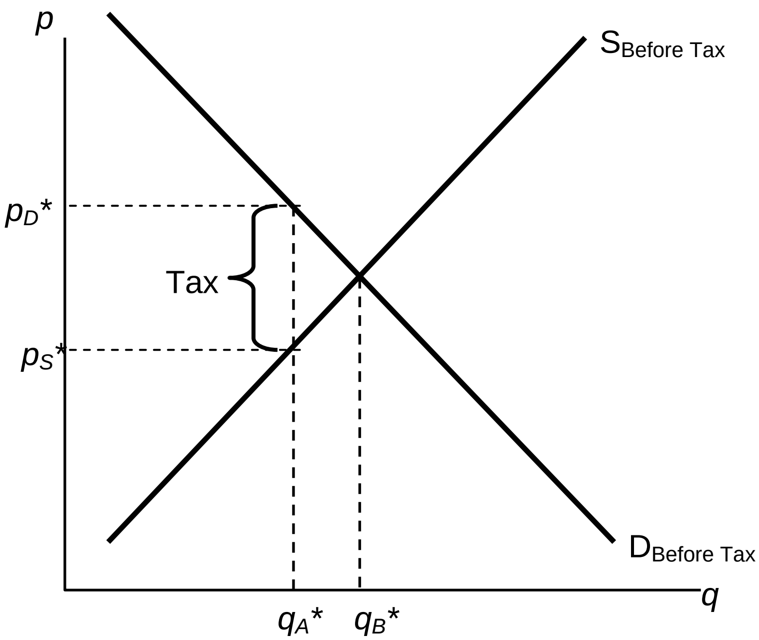

Figure 6-3: Effect of a Tax on Equilibrium

In both cases, the effect of the tax on the supply-demand equilibrium is to shift the quantity toward a point where the before tax demand minus the before tax supply is the amount of the tax. This is illustrated in Figure 6-3. The quantity traded before a tax was imposed was q_B*. When the tax is imposed, the price that the buyer pays must exceed the price the sellers receive, by the amount equal to the tax. This pins down a unique quantity, denoted q_A*. The price the buyer pays is denoted by p_D* and the sellers receive that amount minus the tax, which is noted as p_S*. The relevant quantities and prices are illustrated in Figure 6-3.

Another thing notable from this picture is that the price that buyers pay rises, but generally by less than the tax. Similarly, the price the sellers obtain falls, but by less than the tax. These changes are known as the *incidence* of the tax – is a tax mostly borne by buyers, in the form of higher prices, or by sellers, in the form of lower prices net of taxation?

There are two main effects of a tax: a fall in the quantity traded, and a diversion of revenue to the government. These are illustrated in Figure 6-4. First, the revenue is just the amount of the tax times the quantity traded, which is the area of the shaded rectangle. The tax raised of course uses the after tax quantity q_A* because this is the quantity traded once the tax is imposed.

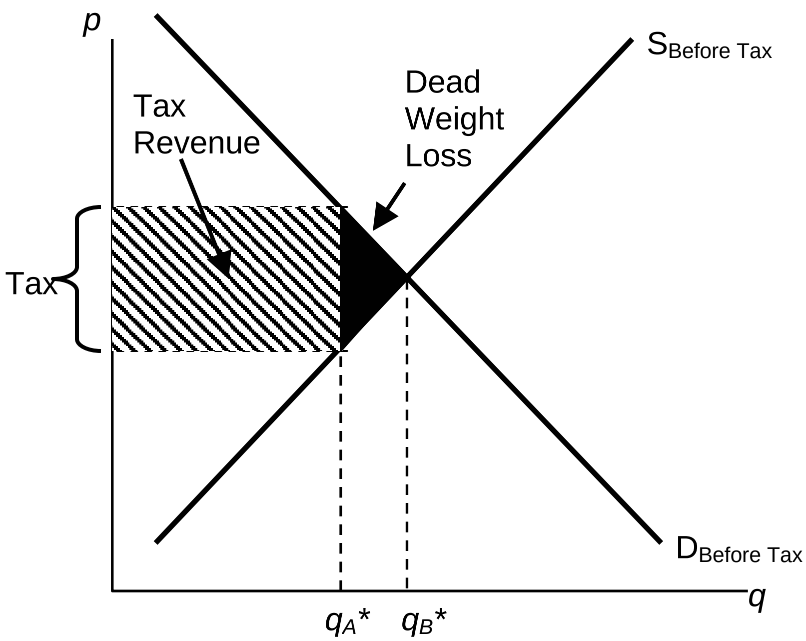

Figure 6-4: Revenue and Dead weight Loss

In addition, a tax reduces the quantity traded, thereby reducing some of the gains from trade. Consumer surplus falls because the price to the buyer rises, and producer surplus (profit) falls because the price to the seller falls. Some of those losses are captured in the form of the tax, but there is a loss captured by no party – the value of the units that would have been exchanged were there no tax. The value of those units is given by the demand, and the marginal cost of the units is given by the supply. The difference, shaded in black in the diagram, is the lost gains from trade of units that aren't traded because of the tax. These lost gains from trade are known as a *dead weight loss.* That is, the dead weight loss is the buyers' values minus the sellers' costs of units that are not economic to trade *only* because of a tax or other interference in the market. The net lost gains from trade, measured in dollars, of these lost units is illustrated by the black triangular region in the diagram.

The dead weight loss is important because it represents a loss to society much the same as if resources were simply thrown away or lost. The dead weight loss is value that people don't enjoy, and in this sense can be viewed as an opportunity cost of taxation. That is, to collect taxes, we have to take money away from people, but obtaining a dollar in tax revenue actually costs society more than a dollar. The costs of raising tax revenues include the money raised (which the taxpayers lose), the direct costs of collection like tax collectors and government agencies to administer tax collection, and the dead weight loss – the lost value created by the incentive effects of taxes, which reduce the gains for trade. The dead weight loss is part of the overhead of collecting taxes. An interesting issue, to be considered in the subsequent section, is the selection of activities and goods to tax in order to minimize the dead weight loss of taxation.

Without more quantification, only a little more can be said about the effect of taxation. First, a small tax raises revenue approximately equal to the tax level times the quantity, or tq. Second, the drop in quantity is also approximately proportional to the size of the tax. Third, this means the size of the dead weight loss is approximately proportional to the tax squared. Thus, small taxes have an almost zero dead weight loss per dollar of revenue raised, and the overhead of taxation, as a percentage of the taxes raised, grows when the tax level is increased. Consequently, the cost of taxation tends to rise in the tax level.

6.1.1.1 (Exercise) Suppose demand is given by $q_d(p) = 1 - p$ and supply $q_s(p) = p$, with prices in dollars. If sellers pay a 10 cent tax, what is the after tax supply? Compute the before tax equilibrium price and quantity, and the after tax equilibrium quantity, and buyer's price and seller's price.

6.1.1.2 (Exercise) Suppose demand is given by $q_d(p) = 1 - p$ and supply $q_s(p) = p$, with prices in dollars. If buyers pay a 10 cent tax, what is the after tax demand? Do the same computations as the previous exercise and show that the outcomes are the same.

6.1.1.3 (Exercise) Suppose demand is given by $q_d(p) = 1 - p$ and supply $q_s(p) = p$, with prices in dollars. Suppose a tax of t cents is imposed, $t \leq 1$. What is the equilibrium quantity traded, as a function of t? What is the revenue raised by the government, and for what level of taxation is it highest?

6.1.2 Incidence of Taxes

How much does the quantity fall when a tax is imposed? How much does the buyer's price rise and the price to the seller fall? The elasticities of supply and demand can be used to answer this question. To do so, we consider a percentage tax t and employ the methodology introduced in Chapter 2 and assume constant elasticity of both demand and supply. Let the equilibrium price to the seller be p_s and the equilibrium price to the buyer be p_b. As before, we will denote the demand function by $q_d(p)=ap^{-\varepsilon}$ and supply function by $q_s(p)=bp^{\eta}$. These prices are distinct because of the tax, and the tax determines the difference:

$p_b = (1+t)\, p_s$.

Equilibrium requires

$$ap_d^{-\varepsilon} = q_d(p_b) = q_s(p_s) = bp_s^{\eta}.$$

Thus,

$$a\big((1+t)p_s\big)^{-\varepsilon} = ap_d^{-\varepsilon} = q_d(p_b) = q_s(p_s) = bp_s^{\eta}.$$

This solves for

$$p_s = \left(\frac{a}{b}\right)^{1/(\eta+\varepsilon)}(1+t)^{-\varepsilon/(\eta+\varepsilon)},$$

and

$$q^* = q_s(p_s) = bp_s^{\eta} = b\left(\frac{a}{b}\right)^{\eta/\eta+\varepsilon}(1+t)^{-\varepsilon\eta/\eta+\varepsilon} = a^{\eta/\eta+\varepsilon}b^{\varepsilon/\eta+\varepsilon}(1+t)^{-\varepsilon\eta/\eta+\varepsilon}.$$

Finally, $p_d = (1+t)p_s = \left(\frac{a}{b}\right)^{1/\eta+\varepsilon}(1+t)^{\eta/\eta+\varepsilon}$.

Recall the approximation $(1+t)^r \approx 1+rt$.

Thus, a small proportional tax increases the price to the buyer by approximately $\frac{\eta t}{\varepsilon+\eta}$, and decreases the price to the seller by $\frac{\varepsilon t}{\varepsilon+\eta}$. The quantity falls by approximately $\frac{\eta\varepsilon t}{\varepsilon+\eta}$. Thus, the price effect is mostly on the "relatively inelastic party." If demand is inelastic, ε is small, then the price decrease to the seller will be small and the price increase to the buyer close to the entire tax. Similarly, if demand is very elastic, ε is very large, and the price increase to the buyer will be small and the price decrease to the seller close to the entire tax.

We can rewrite the quantity change as $\frac{\eta\varepsilon t}{\varepsilon+\eta} = \frac{t}{\frac{1}{\varepsilon}+\frac{1}{\eta}}$. Thus the effect of a tax on quantity is small if either the demand or the supply is inelastic. To minimize the distortion in quantity, it is useful to impose taxes on goods that either have inelastic demand, or inelastic supply.

For example, cigarettes are a product with very inelastic demand and moderately elastic supply. Thus a tax increase will generally increase the price almost the entire amount of the tax. In contrast, travel tends to have relatively elastic demand, so taxes on travel – airport, hotel and rental car taxes – tend not to increase the final prices so much, but have large quantity distortions.

6.1.2.1 (Exercise) For the case of constant elasticity (of both supply and demand), what tax rate maximizes the government's revenue? How does the revenue-maximizing tax rate change when demand becomes more inelastic?

6.1.3 Excess Burden of Taxation

The presence of the dead-weight loss implies that raising $1 in taxes costs society more than $1. But how much more? This idea – that the cost of taxation exceeds the taxes raised – is known as the *excess burden of taxation,* or just the excess burden. We can quantify the excess burden with a remarkably sharp formula.

To start, we will denote the marginal cost of the quantity q by $c(q)$ and the marginal value by $v(q)$. The elasticities of demand and supply are given by the standard formulae:

$$\varepsilon = -\frac{dq/q}{dv/v} = -\frac{v(q)}{qv'(q)} \text{ and } \eta = \frac{dq/q}{dc/c} = \frac{c(q)}{qc'(q)}.$$

Consider an *ad valorem* tax that will be denoted by t. If sellers are charging $c(q)$, the *ad valorem* (at value) tax is $tc(q)$, and the quantity q^* will satisfy

$v(q^*) = (1 + t)c(q^*)$.

From this equation, we immediately deduce

$$\frac{dq^*}{dt} = \frac{c(q^*)}{v'(q^*) - (1+t)c'(q^*)} = \frac{c(q^*)}{-\frac{v(q^*)}{\varepsilon q^*} - (1+t)\frac{c(q^*)}{\eta q^*}} = -\frac{q^*}{(1+t)\left(\frac{1}{\varepsilon} + \frac{1}{\eta}\right)} = -\frac{q^*\varepsilon\eta}{(1+t)(\varepsilon+\eta)}.$$

Tax revenue is given by

$$Tax = tc(q^*)q^*.$$

The effect on taxes collected, Tax, of an increase in the tax rate t is

$$\frac{dTax}{dt} = c(q^*)q^* + t\left(c(q^*) + q^*c'(q^*)\right)\frac{dq^*}{dt} = c(q^*)\left(q^* - t\left(1 + \frac{1}{\eta}\right)\frac{q^*\varepsilon\eta}{(1+t)(\varepsilon+\eta)}\right)$$

$$= \frac{c(q^*)q^*}{(1+t)(\varepsilon+\eta)}\left((1+t)(\varepsilon+\eta) - t(1+\eta)\varepsilon\right) = \frac{c(q^*)q^*}{(1+t)(\varepsilon+\eta)}\left(\varepsilon + \eta - t\eta(\varepsilon - 1)\right).$$

Thus, tax revenue is maximized when the tax rate is $t_{\max}$, given by

$$t_{\max} = \frac{\varepsilon+\eta}{\eta(\varepsilon-1)} = \frac{\varepsilon}{\varepsilon-1}\left(\frac{1}{\eta} + \frac{1}{\varepsilon}\right).$$

The value $\frac{\varepsilon}{\varepsilon-1}$ is the monopoly markup rate, which we will meet in Section 6.5. Here, it is applied to the sum of the inverse elasticities.

The gains from trade (including the tax) is the difference between value and cost for the traded units, and thus is

$$GFT = \int_0^{q^*} v(q) - c(q) \; dq .$$

Thus, the change in the gains from trade as taxes increase is given by

$$\frac{dGFT}{dTax} = \frac{\partial GFT/\partial t}{\partial Tax/\partial t} = \frac{\left(v(q^*) - c(q^*)\right)\frac{dq^*}{dt}}{\frac{c(q^*)q^*}{(1+t)(\varepsilon+\eta)}\left(\varepsilon+\eta-t\eta(\varepsilon-1)\right)} = -\frac{\left(v(q^*) - c(q^*)\right)\frac{q^*\varepsilon\eta}{(1+t)(\varepsilon+\eta)}}{\frac{c(q^*)q^*}{(1+t)(\varepsilon+\eta)}\left(\varepsilon+\eta-t\eta(\varepsilon-1)\right)}$$

$$= -\frac{\left(tc(q^*)\right)\varepsilon\eta}{c(q^*)\left(\varepsilon+\eta-t\eta(\varepsilon-1)\right)} = -\frac{\varepsilon\eta t}{\varepsilon+\eta-t\eta(\varepsilon-1)} = -\frac{\varepsilon}{\varepsilon-1}\,\frac{t}{t_{\max}-t} .$$

The value $t_{\max}$ is the value of the tax rate t that maximizes the total tax take. This remarkable formula permits the quantification of the cost of taxation. The minus sign indicates it is a loss – the dead weight loss of monopoly, as taxes are raised, and it is composed of two components. First, there is the term $\frac{\varepsilon}{\varepsilon-1}$, which arises from the change in revenue as quantity is changed, thus measuring the responsiveness of revenue to a quantity change. The second term provides for the change in the size of the welfare loss triangle. The formula can readily be applied in practice to assess the social cost of taxation, knowing only the tax rate and the elasticities of supply and demand.

The formula for the excess burden is a local formula – it calculates the increase in the dead weight loss associated with raising an extra dollar of tax revenue. All elasticities, including those in $t_{\max}$, are evaluated locally around the quantity associated with the current level of taxation. The calculated value of $t_{\max}$ is value given the local elasticities; if elasticities are not constant, this value will not necessarily be the actual value that maximizes the tax revenue. One can think of $t_{\max}$ as the projected value. It is sometimes more useful to express the formula directly in terms of elasticities rather than in terms of the projected value of $t_{\max}$, in order to avoid the potential confusion between the projected (at current elasticities) and actual (at the elasticities relevant to $t_{\max}$) value of $t_{\max}$. This level can be read directly from the derivation above:

$$\frac{dGFT}{dTax} = -\frac{\varepsilon\eta t}{\varepsilon+\eta-\eta(\varepsilon-1)t} .$$

6.2 Price Floors and Ceilings

A *price floor* is a minimum price at which a product or service is permitted to sell. Many agricultural goods have price floors imposed by the government. For example, tobacco sold in the United States has historically been subject to a quota and a price floor set by the Secretary of Agriculture. Unions may impose price floors as well. For example, the Screen Actors Guild imposes minimum rates for guild members, generally pushing up the price paid for actors above that which would prevail in an unconstrained market.

(The wages of big name stars aren't generally affected by SAG, because these are individually negotiated.) The most important example of a price floor is the *minimum wage,* which imposes a minimum amount that a worker can be paid per hour.

A *price ceiling* is a maximum price that can be charged for a product or service. Rent control imposes a maximum price on apartments (usually set at the historical price plus an adjustment for inflation) in many U.S. cities. Taxi fares in New York, Washington, D.C. and other cities are subject to maximum legal fares. During World War II, and again in the 1970s, the United States imposed price controls to limit inflation, imposing a maximum price for legal sale of many goods and services. For a long time, most U.S. states limited the legal interest rate that could be charged (these are called *usury laws*) and this is the reason so many credit card companies are located in South Dakota. South Dakota was the first state to eliminate such laws. In addition, ticket prices for concerts and sporting events are often set below the equilibrium price. Laws prohibiting scalping then impose a price ceiling. Laws preventing scalping are usually remarkably ineffective in practice, of course.

6.2.1 Basic Theory

The theory of price floors and ceilings is readily articulated with simple supply and demand analysis. Consider a price floor – a minimum legal price. If the price floor is low enough – below the equilibrium price – there are no effects, because the same forces that tend to induce a price equal to the equilibrium price continue to operate. If the price floor is higher than the equilibrium price, there will be a *surplus*, because at the price floor, more units are supplied than are demanded. This surplus is illustrated in Figure 6-5.

In Figure 6-5, the price floor is illustrated with a horizontal line and is above the equilibrium price. Consequently, at the price floor, a larger quantity is supplied than is demanded, leading to a surplus. There are units that are socially efficient to trade but aren't traded – because their value is less than the price floor. The gains from trade associated with these units, which is lost due to the price floor, represent the dead weight loss.

The price increase created by a price floor will increase the total amount paid by buyers when the demand is inelastic, and otherwise will reduce the amount paid. Thus, if the price floor is imposed in order to be a benefit to sellers, we would not expect to see the price increased to the point where demand becomes elastic, for otherwise the sellers receive less revenue. Thus, for example, if the minimum wage is imposed in order to increase the average wages to low-skilled workers, then we would expect to see the total income of low-skilled workers rise. If, on the other hand, the motivation for the minimum wage is primarily to make low-skilled workers a less effective substitute for union workers, and hence allow union workers to increase their wage demands, then we might observe a minimum wage which is in some sense "too high" to be of benefit to low-skilled workers.

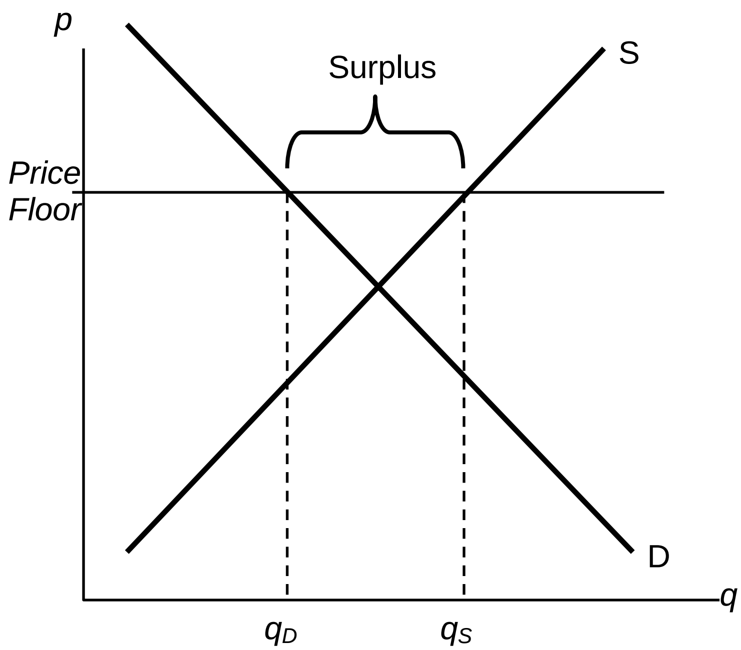

Figure 6-5: A Price Floor

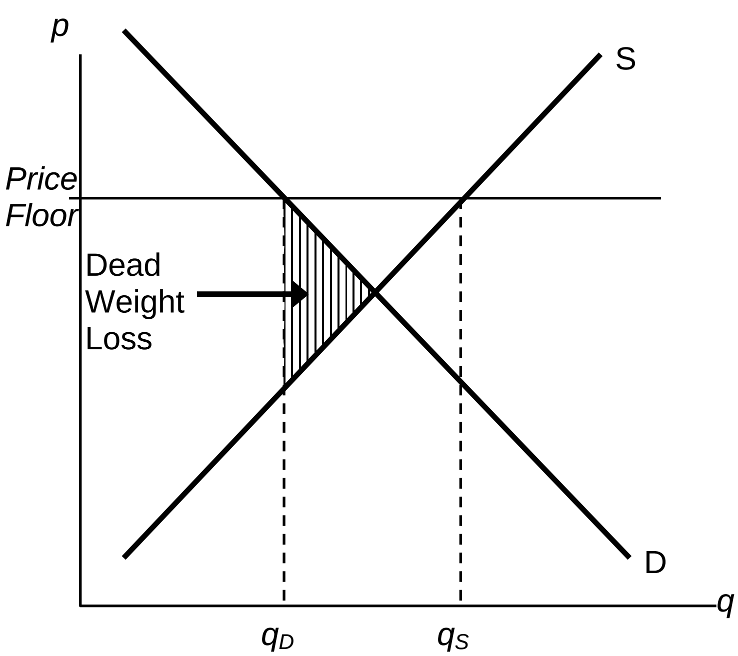

Figure 6-6: Dead weight Loss of a Price Floor

The dead weight loss illustrated in Figure 6-6 is the difference between the value of the units not traded, and value is given by the demand curve, and the cost of producing these units. The triangular shaped region representing the difference between value and cost is illustrated in the above diagram, in the shaded region.

However, this is the *minimum* loss to society associated with a price floor. Generally there will be other losses. In particular, the loss given above assumes that suppliers who don't sell don't produce. As a practical matter, some suppliers who won't in the end sell may still produce because they hope to sell. In this case additional costs are incurred and the dead weight loss will be larger to reflect these costs.

Example: Suppose both supply and demand are linear, with the quantity supplied equal to the price, and the quantity demanded equal to one minus the price. In this case, the equilibrium price, and the equilibrium quantity, are both ½. A price floor of $p > ½$ induces a quantity demanded of $1-p$. How many units will suppliers offer, if a supplier's chance of trading is random? Suppose $q \geq 1-p$ units are offered. A supplier's chance of selling is $\frac{1-p}{q}$. Thus, the marginal supplier (who has a marginal cost of q by assumption) has a probability $\frac{1-p}{q}$ of earning p, and a certainty of paying q. Exactly q units will be supplied when this is a break-even proposition for the marginal supplier, that is,

$$\frac{1-p}{q} p - q = 0 \text{, or } q = \sqrt{p(1-p)}.$$

The dead weight loss then includes not just the triangle illustrated in the previous picture, but also the cost of the $\sqrt{p(1-p)} - (1-p)$ unsold units.

6.2.1.1 (Exercise) In this example, show that the quantity produced is less than the equilibrium quantity, which is ½. Compute the gains from trade, given the overproduction of suppliers. What is the dead weight loss of the price floor?

6.2.1.2 (Exercise) Suppose that units aren't produced until after a buyer has agreed to purchase, as typically occurs with services. What is the dead weight loss in this case? (Hint: what potential sellers will offer their services? What is the average cost of supply of this set of potential sellers?)

The Screen Actors Guild, a union of actors, has some ability to impose minimum prices (a price floor) for work on regular Hollywood movies. If the Screen Actors Guild would like to maximize the total earnings of actors, what price should they set in the linear demand and supply example?

The effects of a price floor include lost gains from trade, because too few units are traded (inefficient exchange), units produced that are never consumed (wasted production), and more costly units produced than necessary (inefficient production)

A price ceiling is a maximum price. Analogous to a low price floor, a price ceiling that is larger than the equilibrium price has no effect. Tell me that I can't charge more than a billion dollars for this book (which is being given away free) and it won't affect the price charged or the quantity traded. Thus, the important case of a price ceiling is a price ceiling less than the equilibrium price.

In this case, which should now look familiar, the price is forced below the equilibrium price, and too few units are supplied, while a larger number are demanded, leading to a shortage. The dead weight loss is illustrated in Figure 6-7, and again represents the loss associated with units that are valued more than they cost but aren't produced.

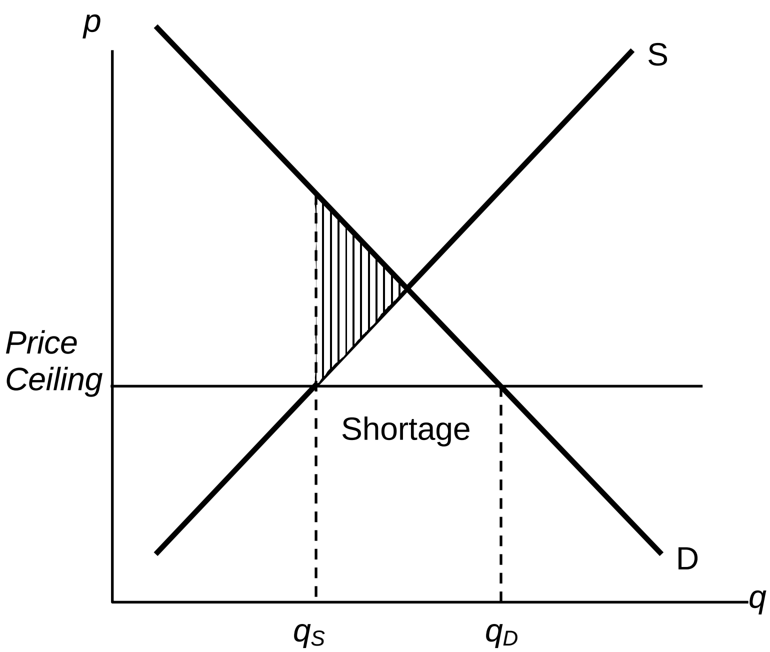

Figure 6-7: A Price Ceiling

Analogous to the case of a price floor, there can be additional losses associated with a price ceiling. In particular, some lower value buyers may succeed in purchasing, denying the higher value buyers the ability to purchase. This effect results in buyers with high values failing to consume, and hence their value is lost.

6.2.1.3 (Exercise) Adapt the price floor example above to the case of a price ceiling, with $p < ½$, and compute the lost gains from trade if buyers willing to purchase are all able to purchase with probability q_S/q_D. (Hint: Compute the value of q_D units; the value realized by buyers collectively will be that amount times the probability of trade.)

In addition to the misallocation of resources (too few units, and units not allocated to those who value them the most), price ceilings tend to encourage illegal trade as people attempt to exploit the prohibited gains from trade. For example, it became common practice in New York to attempt to bribe landlords to offer rent-controlled apartments, and such bribes could exceed $50,000. In addition, potential tenants expended a great deal of time searching for apartments, and a common strategy was to read the obituaries late at night, when the *New York Times* had just come out, hoping to find an apartment that would be vacant and available for rent.

An important and undesirable by-product of price ceilings is discrimination. In a free or unconstrained market, discrimination against a particular group, based on race, religion, or other factors, requires transacting not based on price but on another factor. Thus, in a free market, discrimination is costly – discrimination entails, for instance, not renting an apartment to the highest bidder, but the highest bidder of the favored group. In contrast, with a price ceiling, there is a shortage, and sellers can discriminate at lower cost, or even at no cost. That is, if there are twice as many people seeking apartments as there are apartments at the price ceiling, landlords can "pick and choose" among prospective tenants and still get the maximum legal rent. Thus a price ceiling has the undesirable by-product of reducing the cost of discrimination.

6.2.2 Long- and Short-run Effects

Both demand and supply tend to be more elastic in the long-run. This means that the quantity effects of price floors and ceilings tend to be larger over time. An extreme example of this is rent control, a maximum price imposed on apartments.

Rent control is usually imposed in the following way: as a prohibition or limitation on price increases. For example, New York City's rent control, imposed during World War II, prevented landlords from increasing rent, even when their own costs increased, such as when property taxes increased. This law was softened in 1969 to be gradually replaced by a rent stabilization law that permitted modest rent increases for existing tenants.

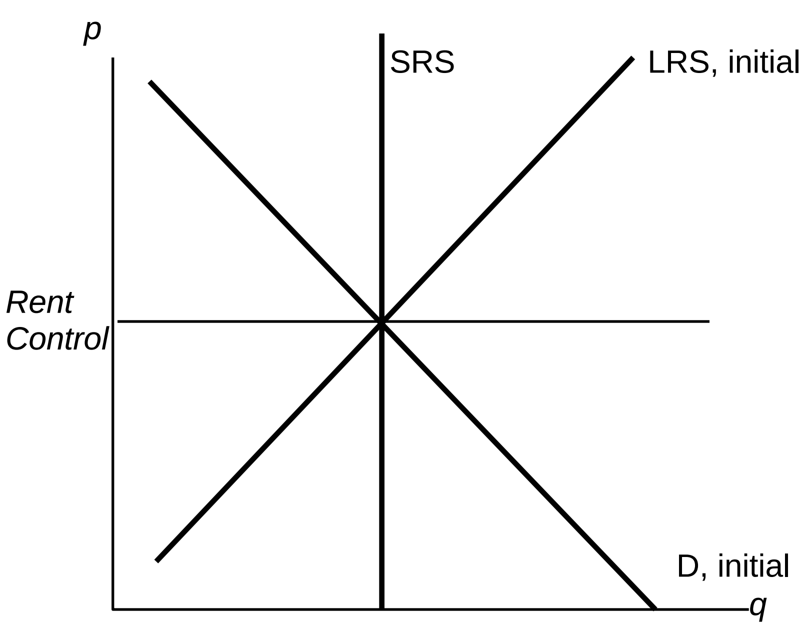

Figure 6-8: Rent Control, Initial Effect

Thus the nature of rent control is that it begins with at most minor effects because it doesn't bind until the equilibrium rent increases. Moreover, the short-run supply of apartments tends to be extremely inelastic, because one doesn't tear down an apartment or convert it to a condominium (there were limitations on this) or abandon it without a pretty significant change in price. Demand also tends to be relatively inelastic, because one has to live somewhere and the alternatives to renting in the city are to live a long distance away or buy (which is relatively expensive), neither of which is a very good substitute for many consumers. Long-run demand and short-run demand are not very different and are treated as being identical. Finally, the long-run supply is much more elastic than the short-run supply, because in the long-run a price increase permits the creation of apartments from warehouses (lofts), rooms rented in houses, etc. Thus, the apartment market in New York is characterized by inelastic short-run supply, much more elastic long-run supply, and inelastic demand. This is illustrated in Figure 6-8.

We start with a rent control law that has little or no immediate effect because it is set at current rents. Thus, in the near term, tenants' fears of price increases are eased and there is little change in the apartment rental market. This is not to say there is zero effect – some companies considering building an apartment on the basis of an expectation of higher future rents may be deterred, and a few marginal apartments may be converted to other uses because the upside potential for the owner has been removed, but such effects are modest at best.

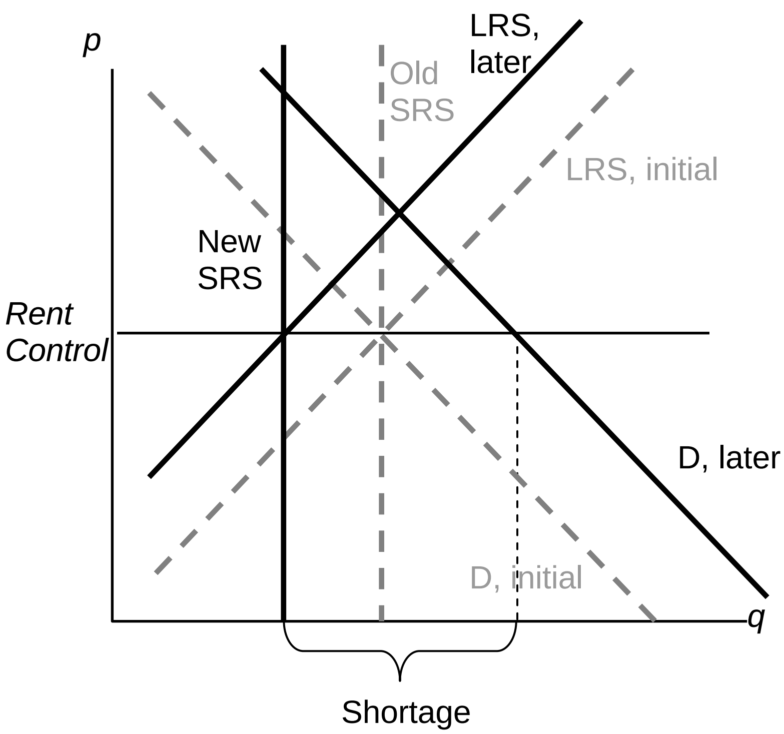

Figure 6-9: Rent Control, Long-Run Effect

Over time, however, the demand for apartments grows as the city population and incomes grow. Moreover, as the costs of operating an apartment rise due to property tax increases, wage increases and cost of maintenance increases, the supply is reduced. This has little effect on the short-run supply but a significant effect on the long-run supply. The supply reduction and demand increases cause a shortage, but results in few apartments being lost because the short-run supply is very inelastic. Over time, however, apartments are withdrawn from the market and the actual quantity falls, even as the demand rises, and the shortage gets worse and worse. These changes are illustrated in Figure 6-9. The old values of demand, short-run supply and long-run supply are illustrated in dashed grey lines. The new values, reflecting an increase in demand, a fall in long-run supply, and a reduction in the number available set of apartments (where the rent control covers the long-run cost) are given in dark black lines.

The shortage is created by two separate factors – demand is increasing as incomes and population rise, and supply is decreasing as costs rise. This reduces the quantity of available housing units supplied and increases the demand for those units.

How serious is the threat that units will be withdrawn from the market? In New York City, over 200,000 apartment units were abandoned by their owners, usually because the legal rent didn't cover the property taxes and legally mandated maintenance. In some cases, tenants continued to inhabit the buildings even after the electricity and water were shut off. It is fair to say that rent control devastated large areas of New York City, such as the Bronx. So why would New York, and so many other communities, impose rent control on itself?

6.2.3 Political Motivations

The politics of rent control are straightforward. First, rent control involves a money transfer from landlords to tenants, because tenants pay less than they would absent the law, and landlords obtain less revenue. In the short-run, due to the inelastic short-run supply, the effect on the quantity of apartments is small, so rent control is primarily just a transfer from landlords to tenants.

In a city like New York, the majority of people rent. A tiny fraction of New Yorkers are landlords. Thus, it is easy to attract voters to support candidates who favor rent control – most renters will benefit, while landlords don't. The numbers of course don't tell the whole story, because while landlords are small in numbers, they are wealthier on average, and thus likely have political influence beyond the number of votes they cast. However, even with their larger economic influence, the political balance favors renters. In the 100*ab* zip codes of Manhattan (first three digits are 100), 80% of families were renters in the year 2000. Thus, a candidate who runs on a rent control platform appeals to large portion of the voters.

Part of the attraction of rent control is that there is little economic harm in the short-run, and most of that harm falls on new residents to New York. As new residents generally haven't yet voted in New York, potential harm to them has only a small effect on most existing New Yorkers, and thus isn't a major impediment to getting voter support for rent control. The slow rate of harm to the city is important politically because the election cycle encourages a short time horizon – if successful at lower office, a politician hopes to move on to higher office, and is unlikely to be blamed for the long-run damage to New York by rent control.

Rent control is an example of a political situation sometimes called the *tyranny of the majority*, where a majority of people have an incentive to confiscate the wealth of a minority. But there is another kind of political situation that is in some sense the reverse, where a small number of people care a great deal about something, and the majority are only slightly harmed on an individual basis. No political situation appears more extreme in this regard than that of refined sugar. There are few U.S. cane sugar producers (nine in 1997), yet the U.S. imposes quotas that raise domestic prices much higher than world prices, in some years tripling the price Americans pay for refined sugar. The domestic sugar producers benefit, while consumers are harmed. But consumers are harmed by only a small amount each, perhaps twelve to fifteen cents per pound – which is not enough to build a consensus to defeat politicians who accept donations from sugar producers. This is a case where *concentrated benefits and diffused costs* determine the political outcome. Because there aren't many sugar producers, it is straightforward for them to act as a single force. In contrast, it is pretty hard for consumers to become passionate about twelve cents per pound increase in the domestic sugar price when they consume about 60 pounds per year of sugar.

6.2.4 Price Supports

A price support is a combination of two programs – a minimum price or price floor, and government purchase of any surplus. Thus, a price support is different from a price floor, because with a price floor, any excess production by sellers was a burden on the sellers. In contrast, with a price support, any excess production is a burden on the government.

The U.S. Department of Agriculture operates a price support for cheese, and has possessed warehouses full of cheese in the past. There are also price supports for milk and other agricultural products.

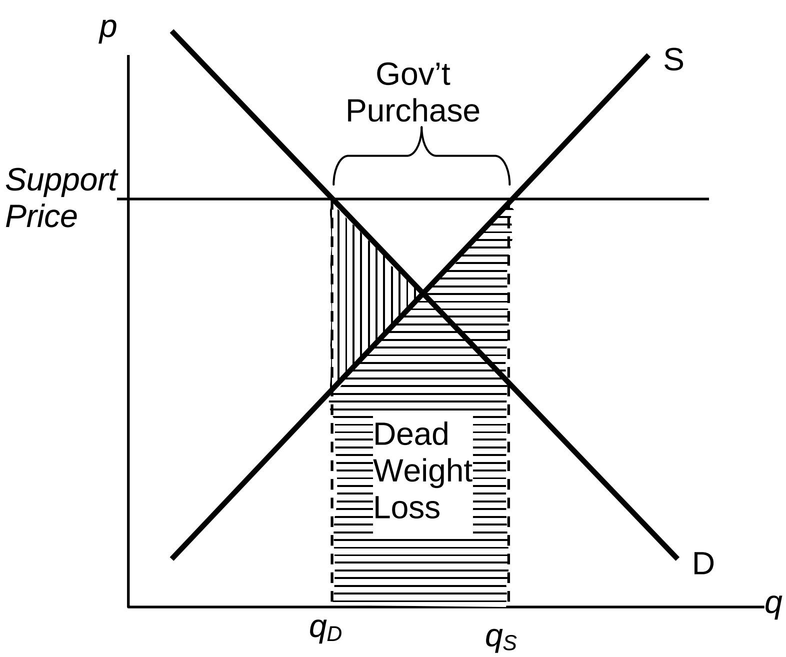

Figure 6-10: Price Supports

Figure 6-10 illustrates the effect of a support program. The government posts a price, called the *support price*, and purchases any excess production offered on the market.

The government purchases, which are the difference between the quantity supplied and quantity demanded, are illustrated on the diagram. The cost of the program to the government is the support price times this quantity purchased, which is the area of the rectangle directly underneath the words “Gov’t Purchases.”

There are two kinds of dead weight loss in a price support program. First, consumers who would like to buy at the equilibrium price are deterred by the higher prices, resulting in the usual dead weight loss, illustrated with the vertical shading. In addition, however, there are goods produced that are then either destroyed or put in warehouses and not consumed, which means the costs of production of those goods is also lost, resulting in a second dead weight loss. That loss is the cost of production, which is given by the supply curve, and thus is the area under the supply curve, for the government purchases. It is shaded in a horizontal fashion. The total dead weight loss of the price support is the sum of these two individual losses. Unlike the case of a price floor or ceiling, a price support creates no ambiguity about what units are produced, or which consumers are willing and able to buy, and thus the rationing aspect of a price floor or ceiling is not present for a price support, nor is the incentive to create a black market other than that created by selling the warehouse full of product.

6.2.5 Quantity Restrictions and Quotas

The final common way that governments intervene in market transactions is to impose a quota. A quota is a maximal production quantity, usually set based on historical production. In tobacco, peanuts, hops, California oranges, and other products, producers have production quotas based on their historical production. Tobacco quotas were established in the 1930s and today a tobacco farmer’s quota is a percentage of the 1930s level of production. The percentage is set annually by the Secretary of Agriculture. Agricultural products are not the only products with quotas. The right to drive a tax in New York requires a medallion issued by the city, and there are a limited number of medallions. This is a quota. Is it a restrictive quota? The current price of a New York taxi medallion – the right to drive a taxi legally in New York City – is \$300,000 (2004 number). This adds approximately \$30,000-\$40,000 annually to the cost of operating a taxi in New York, using a risk adjusted interest rate.

What are the effects of a quota? A quota restricts the quantity below that which would otherwise prevail, forcing the price up, which is illustrated in Figure 6-11. It works like a combination of a price floor and a prohibition on entry.

Generally, the immediate effects of a quota involve a transfer of money from buyers to sellers. The inefficient production and surplus of the price floor are avoided because a production limitation created the price increase. This transfer has an undesirable and somewhat insidious attribute. Because the right to produce is a capital good, it maintains a value, which must be captured by the producer. For example, an individual who buys a taxi medallion today, and pays \$300,000, makes no economic profits – he captures the foregone interest on the medallion through higher prices but no more than that. The individuals who received the windfall gain were those who were driving taxis and were grandfathered in to the system, and issued free medallions. Those people – who were driving taxis 70 years ago and thus are mostly dead at this point – received a windfall gain from the establishment of the system. Future generations pay for the program, which provides no net benefits to the current generation; all the benefits were captured by people long since retired.

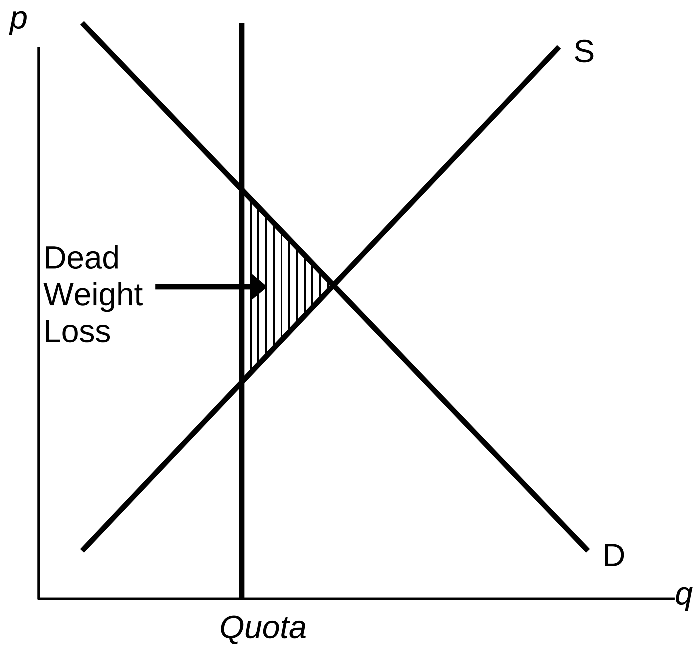

Figure 6-11: A Quota

Does this mean it is harmless to eliminate the medallion requirement? Unfortunately not. The current medallion owners, who if they bought recently paid a lot of money for their medallions, would see the value of these investments destroyed. Thus, elimination of the program would harm current medallion owners.

If the right to produce is freely tradable, the producers will remain the efficient producers, and the taxi medallions are an example of this. Taxi medallions can be bought and sold. Moreover, a medallion confers the right to operate a taxi, but doesn't require that the owner of the medallion actually drive the taxi. Thus, a "medallion owning company" can lease the right to drive a taxi to an efficient driver, thereby eliminating any inefficiency associated with who drives a taxi.

In contrast, because tobacco farming rights aren't legally tradable across county lines, tobacco is very inefficiently grown. The average size of a burley tobacco farm is less than five acres, so some are much smaller. There are tobacco farms in Florida and Missouri, which only exist because of the value of the quota – if they could trade their quota to a farm in North Carolina or Kentucky, which are much better suited to producing cigarette tobacco, it would pay to do so. In this case, the quota, which locked in production rights, also locked in production which gets progressively more inefficient as the years pass.

Quotas based on historical production have the problem that they don't evolve as production methods and technology evolve, thus tending to become progressively more inefficient. Tradable quotas eliminate this particular problem, but continue to have the problem that future generations are harmed with no benefits.

6.2.5.1 (Exercise) Suppose demand for a product is $q_d = 1 - p$, and the marginal cost of production is c. A quota at level $Q \leq 1 - c$ is imposed. What is the value of the quota, per unit of production? Use this to derive the demand for the quota as a function of the level of quota released to the market. If the government wishes to sell the quota, how much should they sell to maximize the revenue on the product?

6.3 Externalities

When the person sitting next to you lights up a cigarette, he gets nicotine, and the cigarette company gets some of his money. You just suffer, with no compensation. If your neighbor's house catches fire because he fell asleep with that cigarette burning in his hand, your house may burn to the ground. The neighbor on the other side who plays very loud music late into the night before your big economics test enjoys the music, and the record company and stereo component companies get his money. You flunk out of college and wind up borrowing $300,000 to buy a taxi medallion. Drunk drivers, cell phones ringing in movies, loud automobiles, polluted air, and rivers polluted to the point that they catch fire like Cleveland's Cuyahoga did, are all examples where a transaction between two parties harmed other people. These are "external effects."

But external effects are not necessarily negative. The neighbor who plants beautiful flowers in her yard brightens your day. Another's purchase of an electric car reduces the smog you breathe. Your neighbor's investment in making his home safe from fire conveys a safety advantage to you. Indeed, even your neighbor's investment in her own education may provide an advantage to you – you may learn useful things from your neighbor. Inventions and creations, whether products or poetry, produce value for others. The creator of a poem, or a mathematical theorem, provides a benefit to others.

These effects are called *external effects*, or *externalities*. An externality is any effect on people not involved in a particular transaction. Pollution is the classic example. When another person buys and smokes cigarettes, there is a transaction between the cigarette company and the smoker. But if you are sitting near the smoker, you are an affected party not directly compensated from the transaction, at least before taxes were imposed on cigarettes. Similarly, you pay nothing for the benefits you get from viewing your neighbor's flowers, nor is there a direct mechanism to reward your neighbor for her efforts.

Externalities will generally cause competitive markets to behave inefficiently from a social perspective, absent a mechanism to involve all the affected parties. Without such a mechanism, the flower-planter will plant too few beautiful flowers, for she has no reason to take account of your preferences in her choices. The odious smoker will smoke too much, and too near others, and the loud neighbor will play music much too late into the night. Externalities create a *market failure*, that is, a competitive market does not yield the socially efficient outcome.

Education is viewed as creating an important positive externality. Education generates many externalities, including more and better employment, less crime, and fewer negative externalities of other kinds. It is widely believed that educated voters elect better politicians.[66] Educated individuals tend to make a society wealthy, an advantage to all of society's members. As a consequence, most societies subsidize education, in order to promote it.

A major source of externalities arises in communicable diseases. Your vaccination not only reduces the likelihood that you contract a disease, but also makes it less likely that you infect others with the disease.

6.3.1 Private and Social Value, Cost

Let's consider pollution as a typical example. A paper mill produces paper, and a bad smell is an unfortunate by-product of the process. Each ton of paper produced increases the amount of bad smells produced. The paper mill incurs a marginal cost, associated with inputs like wood and chemicals and water. For the purposes of studying externalities, we will refer to the paper mill's costs as a *private cost*, the cost to the paper mill itself. In addition, there are *external costs*, which are the costs borne by others, which arise in this case from the smell. Adding the private costs and the external costs yields the social costs. These costs, in their marginal form, are illustrated in Figure 6-12.

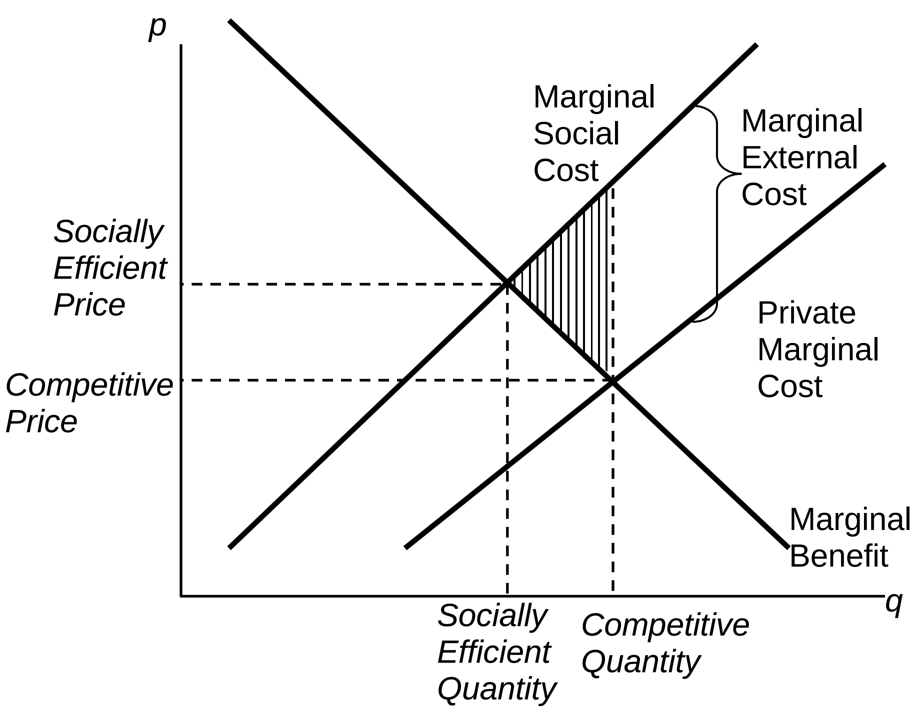

Figure 6-12: A Negative Externality

In Figure 6-12, the demand has been labeled "marginal benefit," for reasons that will become apparent, but it is at this point just the standard demand, the marginal value of

[66] This is a logical proposition, but there is scant evidence in favor of it. There is evidence that educated voters are more likely to vote, but little evidence that they vote for better candidates.

the product. The paper mill's costs have been labeled marginal private cost to reflect the fact that these costs are only the mill's costs and don't include the cost of the bad smell imposed on others. The marginal social cost is obtained by adding the marginal external cost to the marginal private cost. The marginal external cost isn't graphed on the diagram, but the size of it is illustrated at one quantity, and it is generally the difference between marginal social cost and marginal private cost.

Left to its own devices, the paper market would equate the marginal private cost and the marginal benefit, to produce the competitive quantity sold at the competitive price. Some of these units – all of those beyond the quantity labeled "Socially Efficient Quantity," are bad from a social perspective – they cost more to society than they provide in benefits. This is because the social cost of these units includes pollution, but paper buyers have no reason to worry about pollution or even to know it is being created in the process of manufacturing paper.

The dead weight loss of these units is a shaded triangle. The loss arises because the marginal social cost of the units exceeds the benefit, and the difference between the social cost and the benefits yields the loss to society. This is a case where too much is produced because the market has no reason to account for all the costs; some of the costs are borne by others.

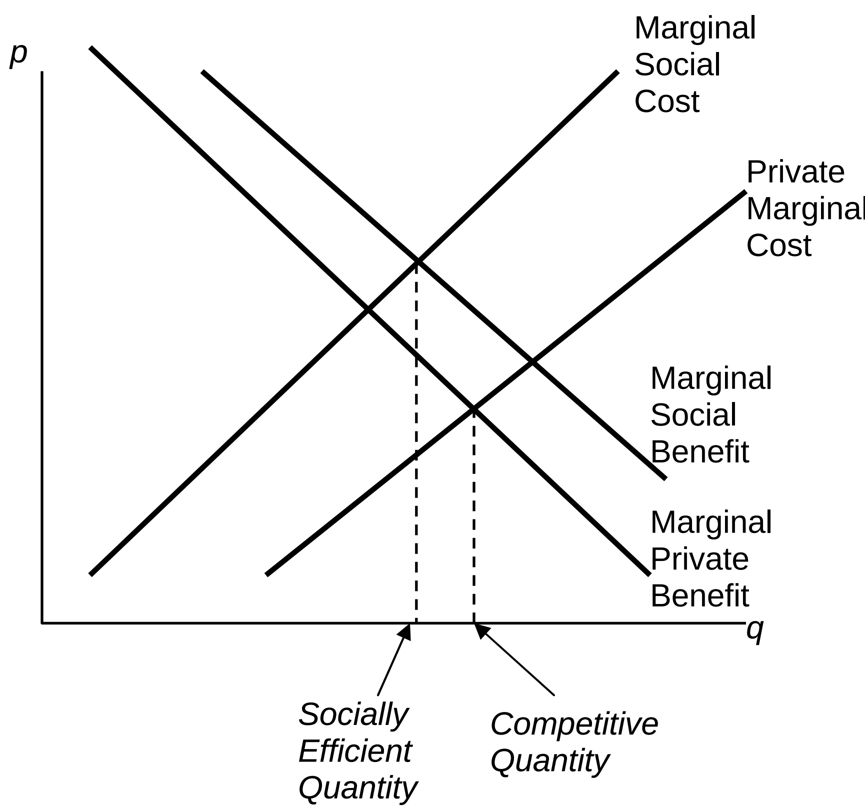

Figure 6-13: External Costs and Benefits

Generally, a negative externality like pollution creates a marginal social cost higher than the marginal private cost. Similarly, a positive externality like beautification creates a higher marginal social benefit than the marginal private benefit (demand). These are to some extent conventions – one could have incorporated a positive externality by a reduction in cost – but the convention remains. An example of a product that produces both positive and negative externalities is illustrated in Figure 6-13. Street lights are an example of a product that produces both externalities – most of us like lit streets, but they are terrible for astronomers. Similarly, large highways produce benefits for commuters and harm to nearby residents.

The marginal private benefit and the marginal private cost give the demand and supply of a competitive market, and hence the competitive quantity results from the intersection of these two. The marginal social benefit and the marginal social cost gives the value and cost from a social perspective; equating these two generates the socially efficient outcome. This can be either greater or less than the competitive outcome depending on which externality is larger.

Example (Tragedy of the Commons): Consider a town on a scenic bay filled with lobsters. The town members collect and eat lobsters, and over time the size of the lobsters collected falls, until they are hardly worth searching for. This situation persists indefinitely; few large lobsters are caught and it is barely worth one's time attempting to catch them.

The tragedy of the commons is a problem with a *common resource*, in this case the lobster bay. Catching lobsters creates an externality, by lowering the productivity of other lobster catchers. The externality leads to over-fishing, since individuals don't take into account the negative effect they have on each other, ultimately leading to a nearly useless resource and potentially driving the lobsters into extinction. As a consequence, the lobster catch is usually regulated.

6.3.1.1 (Exercise) A child who is vaccinated against polio is more likely to contract polio (from the vaccine) than an unvaccinated child. Does this fact imply that programs forcing vaccination on schoolchildren are ill-advised? Include with your answer with a diagram illustrating the negative marginal benefit of vaccination, and a horizontal axis representing the proportion of the population vaccinated.

6.3.1.2 (Exercise) The total production from an oil field generally depends on the rate at which the oil is pumped, with faster rates leading to lower total production but earlier production. Suppose two different producers can pump from the field. Illustrate, using an externality diagram where the horizontal axis is the rate of production for one of the producers, the difference between the socially efficient outcome and the equilibrium outcome. Like many other states, Texas' law requires that when multiple people own land over a single oil field, the output is shared among the owners, with each owner obtaining a share equal to proportion of the field under their land. This process is called *unitization*. Does it solve the problem of externalities in pumping and yield an efficient outcome? Why or why not?

6.3.2 Pigouvian Taxes

Arthur Cecil Pigou, 1877-1959, proposed a solution to the problem of externalities that has become a standard approach. This simple idea is to impose a per-unit tax on a good generating negative externalities equal to the marginal externality at the socially efficient quantity. Thus, if at the socially efficient quantity, the marginal external cost is a dollar, then a one dollar per unit tax would lead to the right outcome. This is illustrated in Figure 6-14.

The tax that is added is the difference, at the socially efficient quantity, between the marginal social cost and the marginal private cost, which equals the marginal external cost. The tax level need not equal the marginal external cost at other quantities, and the diagram reflects a marginal external cost that is growing as the quantity grows. Nevertheless, the new supply curve created by the addition of the tax intersects demand (the marginal benefit) at the socially efficient quantity. As a result, the new competitive equilibrium, taking account of the tax, is efficient.

6.3.2.1 (Exercise) Identify the tax revenue produced by a Pigouvian tax in Figure 6-14. What is the relationship between the tax revenue and the damage produced by the negative externality? Is the tax revenue sufficient to pay those damaged by the external effect an amount equal to their damage? Hint: Is the marginal external effect increasing or decreasing.

The case of a positive externality is similar. In this case, a subsidy is needed to induce the efficient quantity. It is left as an exercise.

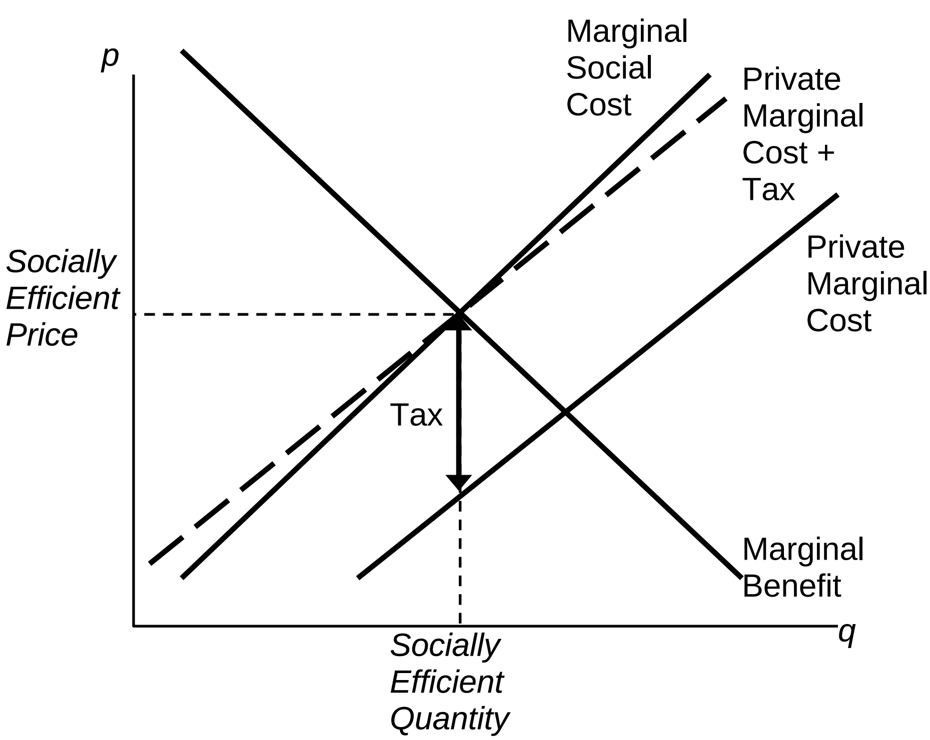

Figure 6-14: The Pigouvian Tax

6.3.2.2 (Exercise) Identify on a diagram the Pigouvian subsidy needed to induce the efficient quantity in the case of a positive externality. When is the subsidy expended smaller than the total external benefit?

6.3.2.3 (Exercise) Use the formulae for estimating the effect of a tax on quantity to deduce the size of the tax needed to adjust for an externality when the marginal social cost is twice the marginal private cost.

Taxes and subsidies are fairly common instruments to control externalities. We subsidize higher education with state universities, and the federal government provides funds for research and limited funds for the arts. Taxes on cigarettes and alcoholic beverages are used to discourage these activities, perhaps because smoking and drinking alcoholic beverages create negative externalities. (Cigarettes and alcohol also have inelastic demands, which make them good candidates for taxation since there is only a small distortion of the quantity.) However, while important in some arenas, taxes and subsidies are not the most common approach to regulation of externalities.

6.3.3 Quotas

The Pigouvian tax and subsidy approach to dealing with externalities has several problems. First, it requires knowing the marginal value or cost of the external effect, and this may be a challenge to estimate. Second, it requires the imposition of taxes and permits the payment of subsidies, which encourages what might be politely termed as "misappropriation of funds." That is, once a government agency is permitted to tax some activities and subsidize others, there will be a tendency to tax things people in the agency don't like, and subsidize "pet" projects, using the potential for externalities as an excuse rather than a real reason. U.S. politicians have been especially quick to see positive externalities in oil, cattle and the family farm, externalities that haven't been successfully articulated. (The Canadian government, in contrast, sees externalities in film-making and railroads.)

An alternative to the Pigouvian tax or subsidy solution is to set a quota, which is a limit on the activity. Quotas can be maxima or minima, depending on whether the activity generates negative or positive externalities. We set maximum levels of many pollutants rather than tax them, and ban some activities, like lead in gasoline or paint, or chlorofluorocarbons (CFCs) outright (a quota equal to zero). We set maximum amounts of impurities, like rat feces, in foodstuffs. We impose minimum educational attainment (eighth grade or age 16, whichever comes first), minimum age to drive, minimum amount of rest time for truck drivers and airline pilots. A large set of regulations govern electricity and plumbing, designed to promote safety, and these tend to be "minimum standards." Quotas are a much more common regulatory strategy for dealing with externalities than taxes and subsidies.

The idea behind a quota is to limit the quantity to the efficient level. If a negative externality in pollution means our society pollutes too much, then impose a limit or quantity restriction on pollution. If the positive externality of education means individuals in our society receive too little education from the social perspective, force them to go to school.

As noted, quotas have the advantage that they address the problem without letting the government spend more money, limiting the government's ability to misuse funds. On the other hand, quotas have the problem of identifying who should get the quota; quotas will often misallocate the resource. Indeed, a small number of power plants account for almost half of the man-made sulfur dioxide pollution emitted into the atmosphere, primarily because these plants historically emitted a lot of pollution and their pollution level was set by their historical levels. Quotas tend to harm new entrants compared to existing firms, and discourage the adoption of new technology. Indeed, the biggest polluters must stay with old technology in order to maintain their right to pollute.

6.3.3.1 (Exercise) If a quota is set to the socially efficient level, how does the value of a quota right compare to the Pigouvian tax?

6.3.3.2 (Exercise) Speeding (driving fast) creates externalities by increasing the likelihood and severity of automobile accidents, and most countries put a limit on speed, but one could instead require fast drivers to buy a permit to speed. Discuss the advantages and disadvantages of "speeding permits."

6.3.4 Tradable Permits and Auctions

A solution to inefficiencies in the allocation of quota rights is to permit trading them. Tradable permits for pollution create a market in the right to pollute, and thereby create a tax on polluting: the emission of pollution requires the purchase of permits to pollute, and the price of these permits represents a tax on pollution. Thus, tradable permits represent a hybrid of a quota system and a Pigouvian taxation system – a quota determines the overall quantity of pollution as in a quota system, determining the supply of pollution rights, but the purchase of pollution rights acts like a tax on pollution, a tax whose level is determined by the quota supply and demand.

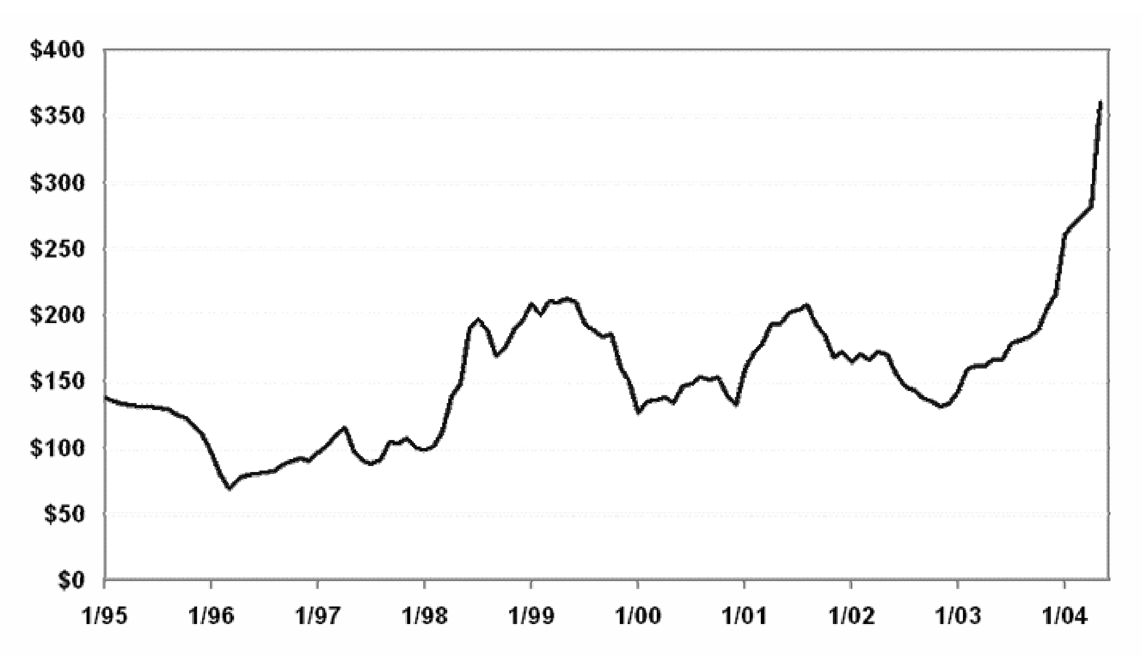

Figure 6-15: SO_2 Permit Prices

The United States has permitted the trading of permits for some pollutants, like sulfur dioxide. Figure 6-15 shows the price of sulfur dioxide permits over the past decade.[67] Each permit conveys the right to emit one ton of sulfur dioxide into the air. The overall pollution level is being reduced over time, which accounts for some of the increase in prices. These prices represent significant taxes on large polluters, as a coal-fired power plant, using coal with high sulfur content, can annually produce as much as 200,000 tons of sulfur dioxide.

The major advantage of a tradable permits system is that it creates the opportunity for efficient exchange – one potential polluter can buy permits from another, leaving the total amount of pollution constant. Such exchange is efficient because it uses the pollution in the manner creating the highest value, eliminating a bias toward "old" sources. Indeed, a low value polluter might sell its permits and just shut down, if the price of pollution were high enough.

A somewhat unexpected advantage of tradable permits was the purchase of permits by environmental groups like the Sierra Club. Environmental groups can buy permits and then not exercise them, as a way of cleaning the air. In this case, the purchase of the permits creates a major positive externality on the rest of society, since the environmental group expends its own resources to reduce pollution of others.

Tradable permits offer the advantages of a taxation scheme – efficient use of pollution – without needing to estimate the social cost of pollution directly. This is especially valuable when the strategy is to set a quantity equal to the current quantity, and then gradually reduce the quantity to reduce the effects of the pollution. The price of permits can be a very useful instrument is assessing the appropriate time to reduce the quantity, since high permit prices, relative to likely marginal external costs, suggests that the quantity of the quota is too low, while low prices suggest that the quantity is too large and should be reduced.

6.3.5 Coasian Bargaining

The negative externality of a neighbor playing loud music late at night is not ordinarily solved with a tax or with a quota, but instead though an agreement. When there aren't many individuals involved, the individuals may be able to solve the problem of externalities without involving a government, but through negotiation. This insight was developed by Nobel laureate Ronald Coase (1910 –).

Coase offered the example of a cattle ranch next to a farm. There is a negative externality, in that the cattle tend to wander over to the farm and eat the crops, rather than staying on the ranch. What happens next depends on *property rights*, which are the rights that come with ownership.

One of three things might be efficient from a social perspective. It might be efficient to erect a fence to keep the cows away from the crops. It might be efficient to close down

[67] Source: Environmental Protection Agency, July 22, 2004, http://www.epa.gov/airmarkets/trading/so2market/alprices.html

the farm. Finally, it might be efficient to close down the ranch, if the farm is valuable enough, and the fence costs more than the value of the ranch.

If the farmer has a right not to have his crops eaten, and can confiscate the cows if they wander onto the farm, then the rancher will have an incentive to erect a fence to keep the cows away, if that is the efficient solution. If the efficient solution is to close down the ranch, then the rancher will do that, since the farmer can confiscate the cows if they go to the farm and it isn't worth building the fence by hypothesis. Finally, if the efficient solution to the externality is to close down the farm, the rancher will have an incentive to buy the farm in order to purchase the farm's rights, so that he can keep the ranch in operation. Since it is efficient to close down the farm only if the farm is worth less than the ranch, there is enough value in operating the ranch to purchase the farm at its value and still have money left over – that is there are gains from trade from selling the farm to the rancher. In all three cases, if the farmer has the property rights, the efficient outcome is reached.

Now suppose instead that the rancher has the rights, and that the farmer has no recourse if the cows eat his crops. If shutting down the farm is efficient, the farmer has no recourse but to shut down. Similarly, if building the fence is efficient, the farmer will build the fence to protect his crops. Finally, if shutting down the ranch is efficient, the farmer will buy the ranch from the rancher, in order to be able to continue to operate the more valuable farm. In all cases, the efficient solution is reached through negotiation.

Coase argued that bargaining can generally solve problems of externalities, and that the real problem is ill-defined property rights. If the rancher and the farmer can't transfer their property rights, then the efficient outcome may not arise. In the Coasian view of externalities, if an individual owned the air, air pollution would not be a problem, because the owner would charge for the use and wouldn't permit an inefficient level of pollution. The case of air pollution demonstrates some of the limitations of the Coasian approach, because ownership of the air, or even the more limited right to pollute into the air, would create an additional set of problems, a case where the cure is likely worse than the disease.

Bargaining to solve the problem of externalities is often feasible when a small number of people are involved. When a large number of people are potentially involved, as with air pollution, bargaining is unlikely to be successful in addressing the problem of externalities, and a different approach required.

6.3.6 Fishing and Extinction

Consider an unregulated fishing market like the lobster market considered above, and let S be the stock of fish. The purpose of this example is illustrative of the logic, rather than an exact accounting of the biology of fish populations, but is not unreasonable. Let S be the stock of a particular species of fish. Our starting point is an environment without fishing: how does the fish population change over time? Denote the change over time in the fish population by $\dot{S}$ ($\dot{S}$ is notation for the derivative with respect to time, notation that dates back to Sir Isaac Newton.) We assume that population growth follows the logistic equation $\dot{S} = rS(1-S)$. This equation reflects two underlying assumptions. First, mating and reproduction is proportional to the stock of fish S. Second, survival is proportional to the amount of available resources 1-S, where 1 is set to be the maximum sustainable population. (Set the units of the number of fish so that 1 is the full population.)

The dynamics of the number of fish is illustrated in Figure 6-16. On the horizontal axis is the number of fish, and on the vertical axis is the change in S. When $\dot{S} > 0$, S is increasing over time, and the arrows on the horizontal axis reflect this. Similarly, if $\dot{S} < 0$, S is decreasing.

Absent fishing, the value 1 is a *stable steady state* of the fish population. It is a steady state because, if S=1, $\dot{S} = 0$, that is, there is no change in the fish population. It is stable because the effect of a small perturbation – S near but not exactly equal to 1 – is to return to 1. (In fact, the fish population is very nearly globally stable – start with any population other than zero and the population returns to 1.)[68]

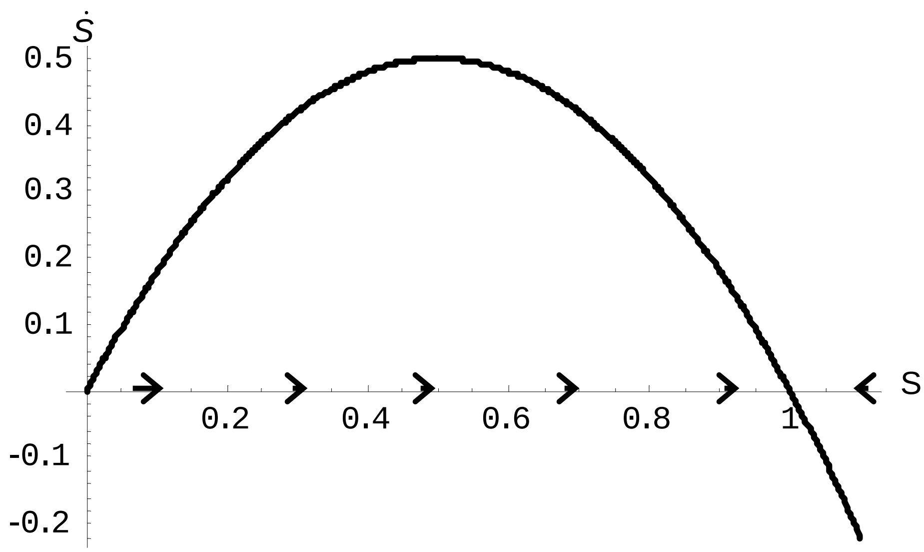

Figure 6-16: Fish Population Dynamics

Now we introduce a human population and turn to the economics of fishing. Suppose that a boat costs b to launch and operate, and that it captures a fixed fraction a of the total stock of fish S, that is, each boat catches aS. Fish sell for a price $p = Q^{-1/\varepsilon}$, where the price arises from the demand curve, which in this case has constant elasticity ε, and Q is the quantity of fish offered for sale. Suppose there be n boats launched; then the quantity of fish caught is Q=naS. Fishers enter the market as long as profits are positive, which leads to zero profits for fishers, that is, $b = \left(Q/n\right) p(Q)$. This equation

[68] It turns out that there is a closed form solution for the fish population: $S(t) = \frac{S(0)}{S(0) + (1 - S(0))e^{-rt}}$.

makes a company just indifferent to launching an additional boat, because the costs and revenues are balanced. These two equations yield two equations in the two unknowns n and Q:

$$n = \frac{Qp(Q)}{b} = \frac{1}{b}Q^{\frac{\varepsilon-1}{\varepsilon}}, \text{ and}$$

$Q=naS$. These two equations solve for the number of fish caught:

$$Q = \left(\frac{aS}{b}\right)^{\varepsilon}$$

and the number of boats $n = \frac{a^{\varepsilon-1}}{b^{\varepsilon}} S^{\varepsilon-1}$.

Subtracting the capture by humans from the growth in the fish population yields:

$$\dot{S} = rS(1-S) - \left(\frac{aS}{b}\right)^{\varepsilon}.$$

Thus, a steady state satisfies $0 = \dot{S} = rS(1-S) - \left(\frac{aS}{b}\right)^{\varepsilon}$.

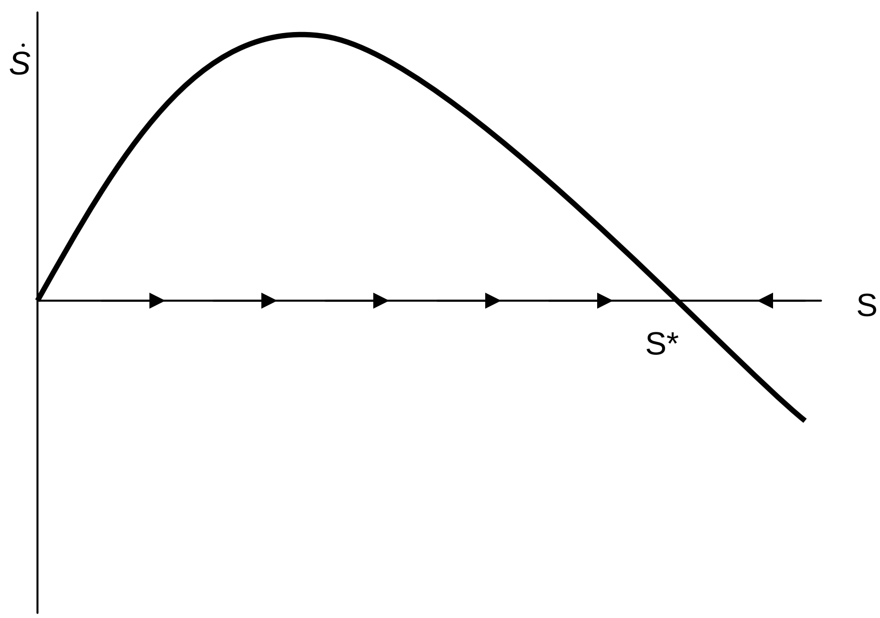

Figure 6-17: Fish Population Dynamics with Fishing

Will human fishing drive the fish to extinction? Extinction must occur when the only stable solution to the stock of fish is zero. Consider first the case when demand is elastic (ε>1). In this case, for S near zero but positive, $\dot{S} \approx rS > 0$, because the other terms are small relative to the linear term. Thus, with elastic demand, there is always a steady state without extinction. (Extinction is also an equilibrium, too, but over-fishing won't get the system there.) This equilibrium is illustrated in Figure 6-17.

The dark curve represents $\dot{S}$, and thus for S between 0 and the point labeled S^*, $\dot{S}$ is positive and thus S is increasing over time. Similarly, to the right of S^*, S is decreasing. Thus, S^* is stable under small perturbations in the stock of fish and is an equilibrium.

We see that if demand for fish is elastic, fishing will not drive the fish to extinction. Even so, fishing will reduce the stock of fish below the efficient level, because individual fishers don't take account of the externality they impose – their fishing reduces the stock for future generations. The level of fish in the sea converges to S^* satisfying

$$0 = rS^*(1-S^*) - \left(\frac{aS^*}{b}\right)^{\varepsilon}.$$

In contrast, if demand is inelastic, fishing may drive the fish to extinction. For example, if r=2 and a=b=1, and ε=0.7, extinction is necessary, as is illustrated in Figure 6-18.

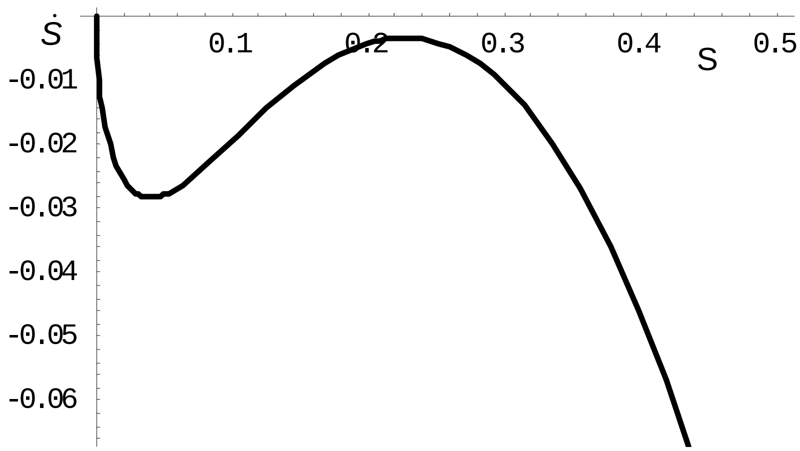

Figure 6-18: Fish Population Dynamics: Extinction

Figure 6-18 shows that, for the given parameters, the net growth of the fish population is negative for every value of the stock S. Thus the population of fish consistently dwindles. This is a case when the fishing externality (overfishing today reduces the stock of fish tomorrow) has particularly dire consequences. The reason why the elasticity of demand matters is that, with inelastic demand, the fall in the stock of fish increases the price by a large amount (enough so that total revenue rises). This, in turn, increases the number of fishing boats, in spite of the fall in the catch. In contrast, with elastic demand, the number of fishing boats falls as the stock falls, reducing the proportion of fish caught, and thus preventing extinction. We see this for the equation for the number of fishing boats

$$n = \frac{a^{\varepsilon-1}}{b^{\varepsilon}} S^{\varepsilon-1}$$

which reflects the fact that *fishing effort rises as the stock falls if and only if demand is inelastic.*

It is possible, even with inelastic demand, for there to be a stable fish population: not all parameter values lead to extinction. Using the same parameters as before, but with ε=0.9, we obtain a stable outcome illustrated in Figure 6-19.

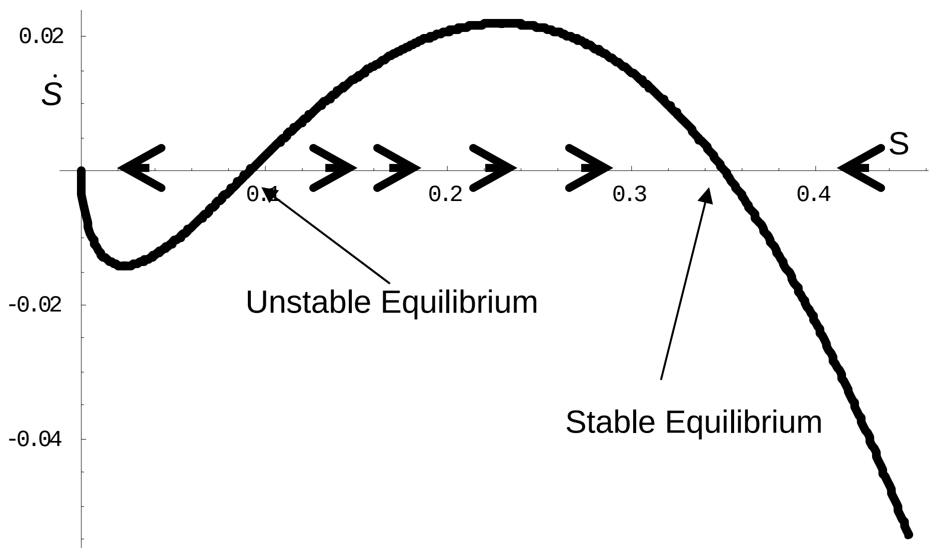

Figure 6-19: Possibility of Multiple Equilibria

In addition to the stable equilibrium outcome, there is an unstable steady state, which might either converge upward or downward. It is a feature of fishing with inelastic demand that there is a region where extinction is inevitable, for when the stock is near zero, the high demand price induced by inelasticity forces sufficient fishing to insure extinction.

As a consequence of the fishing externality, nations attempt to regulate fishing, both by extending their own reach 200 miles into the sea, and by treaties limiting fishing in the open sea. These regulatory attempts have met with only modest success at preventing over-fishing.

What is the efficient stock of fish? This is a challenging mathematical problem, but some insight can be gleaned via a steady state analysis. A steady state arise when $\dot{S} = 0$. If a constant amount Q is removed, a steady state in the stock must occur at $0 = \dot{S} = rS(1-S) - Q$. This maximum catch then occurs at S = ½, and Q = ¼ r. This is not the efficient level, for it neglects the cost of boats, and the efficient stock will actually be larger. More generally, it is never efficient to send the population below the maximum point on the survival curve plotted in Figure 6-16.

Conceptually, fishing is an example of the tragedy of the commons externality already discussed. However, the threat of a permanent extinction and alluring possibility of solving dynamic models make it a particularly dramatic example.

6.3.6.1 (Exercise) Suppose $\varepsilon = 1$. For what parameter values are fish necessarily driven to extinction? Can you interpret this condition to say that the demand for caught fish exceeds the production via reproduction?

6.4 Public Goods

A public good has two attributes: *nonexcludability*, which means the producer can't prevent the use of the good by others, and *nonrivalry*, which means that many people can use the good simultaneously.

6.4.1 Examples

Consider a company offering a fireworks display. Pretty much anyone nearby can watch the fireworks, and people with houses in the right place have a great view of them. The company that creates the fireworks can't compel those with nearby homes to pay for the fireworks, and so a lot of people get to watch them without paying. This will make it difficult or impossible for the fireworks company to make a profit. In addition, fireworks offer nonrivalry, in that one person's viewing of the display doesn't impinge significantly on another's viewing. Nonrivalry has the implication that the efficient price is zero, since the marginal cost of another viewer is zero.

The classic example of a public good is national defense. National defense is clearly non-excludable, for if we spend the resources necessary to defend our national borders, it isn't going to be possible to defend everything except one apartment on the second floor of a three story apartment on East Maple Street. Once we have kept our enemies out of our borders, we've protected everyone within the borders. Similarly, the defense of the national borders exhibits a fair degree of nonrivalry, especially insofar as the strategy of defense is to deter an attack in the first place. That is, the same expenditure of resources protects all.

It is theoretically possible to exclude some from the use of a poem, or a mathematical theorem, but exclusion is generally quite difficult. Both poems and theorems are nonrivalrous. Similarly, technological and software inventions are non-rivalrous, even though a patent grants the right to exclude the use by others. Another good that permits exclusion at a cost is a highway. A toll highway shows that exclusion is possible on the highways. Exclusion is quite expensive, partly because the tollbooths require staffing, but mainly because of the delays imposed on drivers associated with paying the tolls – the time costs of toll roads are high. Highways are an intermediate case where exclusion is possible only at a significant cost, and thus should be avoided if possible. Highways are also rivalrous at high congestion levels, but nonrivalrous at low congestion levels. That is, the marginal cost of an additional user is essentially zero for a sizeable number of users, but then marginal cost grows rapidly in the number of users. With fewer than 700 cars per lane per hour on a four lane highway, generally the flow of traffic is unimpeded.[69] As congestion grows beyond this level, traffic slows down and congestion sets in. Thus, west Texas interstate highways are usually nonrivalrous, while Los Angeles freeways are usually very rivalrous.

Like highways, recreational parks are nonrivalrous at low use levels, becoming rivalrous as they become sufficiently crowded. Also like highways, it is possible but expensive to exclude potential users, since exclusion requires fences and a means for admitting some but not others. (Some exclusive parks provide keys to legitimate users, while others use gatekeepers to charge admission.)

6.4.2 Free-Riders

Consider a neighborhood association which is considering buying land and building a park in the neighborhood. The value of the park is going to depend on the size of the park, and we suppose for simplicity that the value in dollars of the park to each household in the neighborhood is $S^b n^{-a}$, where n is the number of park users, S is the size of the park and a and b a are parameters satisfying $0<a\leq b<1$. This functional form builds in the property that larger parks provide more value at a diminishing rate, but there is an effect from congestion. The functional form gives a reason for parks to be public – it is more efficient for a group of people to share a large park than for each individual to possess a small park, at least if $b > a$, because the gains from a large park exceed the congestion effects. That is, there is a scale advantage – doubling the number of people and the size of the park increases each individual's enjoyment.

How much will selfish individuals voluntarily contribute to the building of the park? That of course depends on what they think others will contribute. Consider a single household, and suppose that household thinks the others will contribute S_{-1} to the building of the park. Should the household contribute, and if so, how much? If the household contributes s, the park will have size $S = S_{-1} + s$, which the household values at $(S_{-1}+s)^b n^{-a}$. Thus, the net gain to a household that contributes s when the others contribute S_{-1} is $(S_{-1}+s)^b n^{-a} - s$.

6.4.2.1 (Exercise) Verify that individual residents gain from contributing to the park if $S < (bn^{-a})^{\frac{1}{1-b}}$ and gain from reducing their contributions if $S > (bn^{-a})^{\frac{1}{1-b}}$.

The previous exercise shows that individual residents gain from their marginal contribution if and only if the park is smaller than $S_0 = (bn^{-a})^{\frac{1}{1-b}}$. Consequently, under voluntary contributions, the only equilibrium park size is S_0. That is, for any park size smaller than S_0, citizens will voluntarily contribute to make the park larger. For any larger size, no one is willing to contribute.

[69] The effect of doubling the number of lanes from 2 to 4 is dramatic. A two lane highway generally flows at 60 mph or more provided there are fewer than 200 cars per lane per hour, while a four lane highway can accommodate 700 cars per lane per hour at the same speed.

Under voluntary contributions, as the neighborhood grows in number, the size of the park shrinks. This makes sense – the benefits of individual contributions to the park mostly accrue to others, which reduces the payoff to any one contributor.

How large *should* the park be? The total value of the park of size S to the residents together is n times the individual value, which gives a collective value of $S^b n^{1-a}$, and the park costs S, so from a social perspective the park should be sized to maximize $S^b n^{1-a} - S$, which yields an optimal park of size $S^* = (bn^{1-a})^{\frac{1}{1-b}}$. Thus, as the neighborhood grows, the park should grow, but as we saw the park would shrink if the neighborhood has to rely on voluntary contributions. This is because people contribute individually as if they were building the park for themselves, and don't account for the value they provide to their neighbors when they contribute. Under individual contributions, the hope that others contribute leads individuals not to contribute. Moreover, use of the park by others reduces the value of the park to each individual, so that the size of the park shrinks as the population grows under individual contributions. In contrast, the park ought to grow *faster* than the number of residents grows, as the per capita park size is $S^*/n = b^{\frac{1}{1-b}} n^{\frac{b-a}{1-b}}$, which is an increasing function of n.[70]

The lack of incentive for individuals to contribute to a social good is known as a *free-rider problem*. The term refers to the individuals who don't contribute, who are said to *free-ride* on the contributions of others. There are two aspects of the free-rider problem apparent in this simple mathematical model. First, the individual incentive to contribute to a public good is reduced by the contributions of others, and thus individual contributions tend to be smaller when the group is larger. Put another way, the size of the free-rider problem grows as the community grows larger. Second, as the community grows larger, the optimal size of the public good grows. The market failure under voluntary contributions is greater the larger is the community. In the theory presented, the optimal size of the public good is $S^* = (bn^{1-a})^{\frac{1}{1-b}}$, and the actual size under voluntary contributions is $S_0 = (bn^{-a})^{\frac{1}{1-b}}$, a gap that gets very large as the number of people grows.

The upshot is that people will voluntarily contribute too little from a social perspective, by free-riding on the contributions of others. A good example of the provision of public goods is a co-authored term paper. This is a public good because the grade given to the paper is the same for each author, and the quality of the paper depends on the sum of the efforts of the individual authors. Generally, with two authors, both work pretty hard on the manuscript in order to get a good grade. Add a third author and it is a virtual

[70] Reminder: In making statements like should and ought, there is no conflict in this model because every household agrees about the optimal size of the park, so that a change to a park size of S*, paid with equal contributions, maximizes every household's utility.

certainty that two of the authors think the third didn't work as hard and was a free-rider on the project.

The term paper example also points to the limitations of the theory. Many people are not as selfish as the theory assumed and will contribute more than would be privately optimal. Moreover, with small numbers, bargaining between the contributors and the division of labor (each works on a section) may help reduce the free-rider problem. Nevertheless, even with these limitations, the free-rider problem is very real and it gets worse the more people are involved. The theory shows that if some individuals contribute more than their share in an altruistic way, the more selfish individuals contribute even less, undoing some of the good done by the altruists.

6.4.2.2 (Exercise) For the model presented in this section, compute the elasticity of the optimal park size with respect to the number of residents, that is, the percent change in S^* for a small percentage change in n. [Hint: use the linear approximation trick $(1+\Delta)^r \approx r\Delta$ for Δ near zero.]

6.4.2.3 (Exercise) For the model of this section, show that an individual's utility when the park is optimally sized and the expenses are shared equally among the n individuals is $u=\left(b^{\frac{b}{1-b}}-b^{\frac{1}{1-b}}\right)n^{\frac{b-a}{1-b}}$. Does this model predict an increase in utility from larger communities?

6.4.2.4 (Exercise) Suppose two people, person 1 and person 2, want to produce a playground to share between them. The value of the playground of size S to each person is $\sqrt{S}$, where S is the number of dollars spent building it. Show that under voluntary contributions, the size of the playground is ¼ and that the efficient size is 1.

6.4.2.5 (Exercise) For the previous exercise, now suppose person 1 offers "matching funds," that is, offers to contribute an equal amount to the contributions of the person 2. How large a playground will person 2 choose?

6.4.3 Provision with Taxation

Faced with the fact that voluntary contributions produce an inadequate park, the neighborhood turns to taxes. Many neighborhood associations or condominium associations have taxing authority, and can compel individuals to contribute. Clearly in the example from the previous section, and indeed a solution is to require each resident to contribute the amount 1, resulting in a park that is optimally sized at n. Generally it is possible in principle to provide the correct size of the public good using taxes to fund it. However, it will be a challenge in practice, which can be illustrated with a slight modification of the example.

Let individuals have different strengths of preferences, so that individual i values the public good of size S at $v_i S^b n^{-a}$ in dollars. (It is useful to assume that no two people have the same v values to simplify arguments.) The optimal size of the park for the neighborhood is $n^{\frac{-a}{1-b}}\left(b\sum_{i=1}^{n} v_i\right)^{\frac{1}{1-b}} = (b\bar{v})^{\frac{1}{1-b}} n^{\frac{1-a}{1-b}}$, where $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$ is the average value. Again, taxes can be assessed to pay for an optimally-sized park, but some people (those with small v values) will view that as a bad deal, while others (with large v) view it as a good deal. What will the neighborhood choose?

If there are an odd number of voters in the neighborhood, the prediction is that the park will serve the *median voter* the best.[71] With equal taxes, an individual obtains $v_i S^b n^{-a} - S/n$. If there are an odd number of people, n can be written as $2k+1$. The median voter is the person for whom k have values v_i larger than hers, and k have values smaller. Consider increasing S. If the median voter likes it, then so do all the people with higher v's, and the proposition to increase S passes. Similarly, a proposal to decrease S will get a majority if the median voter likes it. If the median voter likes reducing S, all the individuals with smaller v_i will vote for it as well. Thus, we can see that voting maximizes the preferences of the median voter, and simple calculus shows that entails $S = (bv_k)^{\frac{1}{1-b}} n^{\frac{1-a}{1-b}}$.

Unfortunately, voting does not result in an efficient outcome generally, and only does so when the average value equals the median value. On the other hand, voting generally performs much better than voluntary contributions. The park size can either be larger or smaller under median voting than is efficient.[72]

6.4.3.1 (Exercise) For the model of this section, show that, under voluntary contributions, only one person contributes, and that person is the person with the largest v_i. How much do they contribute? [Hint: which individual i is willing to contribute at the largest park size? Given the park this individual desires, can anyone else benefit from contributing at all?]

6.4.3.2 (Exercise) Show that if all individuals value the public good equally, voting on the size of the good results in the efficient provision of the public good.

6.4.4 Local Public Goods

The example in the previous section showed that there are challenges to a neighborhood's provision of public goods created by differences in the preferences of the public good. Voting does not generally lead to the efficient provision of the public good, and does so only in special circumstances, like agreement of preferences.

[71] The voting model used is that there is a status quo, which is a planned size of S. Anyone can propose to change the size of S, and the neighborhood votes yes or no. If an S exists such that no replacement gets a majority vote, that S is an equilibrium under majority voting.
[72] The general principle here is that the median voting will do better when the distribution of values is such that the average of n values exceeds the median, which in turn exceeds the maximum divided by n. This is true for most empirically relevant distributions.

A different solution was proposed by Tiebout[73] in 1956. This solution works only when the public goods are local in nature – people living nearby may or may not be excludable, but people living further away can be excluded, and such goods are called "local public goods." Schools are local – more distant people can readily be excluded. Parks are harder to exclude from, but are still local in nature; few people will drive 30 miles to use a park.

Suppose that there are a variety of neighborhoods, some with high taxes, better schools, big parks, beautifully maintained trees on the streets, frequent garbage pickup, a first-rate fire department, extensive police protection and spectacular fireworks displays, and others with lower taxes and more modest provision of public goods. People will tend to move to the neighborhood that fits their preferences. The result is neighborhoods that are relatively homogeneous with respect to the desire for public goods. That homogeneity, in turn, makes voting work better. That is, the ability of people to choose their neighborhoods to suit their preferences over taxes and public goods will make the neighborhood provision of public goods more efficient. The "Tiebout theory" shows that local public goods will tend to be efficiently provided. In addition, even private goods like garbage collection and schools can be efficiently provided publicly if they are local goods, and there are enough distinct localities to offer a broad range of services.

6.4.4.1 (Exercise) Consider a baby-sitting cooperative, where parents rotate supervision of the children of several families. Suppose that, if the sitting service is available with frequency Y, the value placed by person i is $v_i Y$ and the costs of contribution y is ½ ny^2, where Y is the sum of the individual contributions and n is the number of families. Rank $v_1 \geq v_2 \geq \ldots \geq v_n$. (i) What is the size of the service under voluntary contributions? (Hint: Let y_i be the contribution of family i. Identify the payoff of family j as

$$v_j\left(y_j + \sum_{i \neq j} y_i\right) - ½\, n\left(y_j\right)^2$$

What value of y_j maximizes this expression?)

(ii) What contributions maximize the total social value

$\left(\sum_{j=1}^{n} v_j\right)\left(\sum_{j=1}^{n} y_j\right) - ½\, n \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(y_j\right)^2$? [Hint: Are the values of y_i different for different i?]

(iii) Let $\mu = \frac{1}{n}\sum_{j=1}^{n} v_j$ and $\sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(v_j - \mu)^2$. Conclude that, under voluntary contributions, the total value generated by the cooperative is $\frac{n}{2}\left(\mu^2 - \sigma^2\right)$. (Hint:

[73] Charles Tiebout, 1919-1962. His surname is pronounced "tee-boo."

It helps to know that

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(v_j - \mu)^2 = \frac{1}{n}\sum_{j=1}^{n} v_j^2 - \frac{2}{n}\sum_{j=1}^{n}\mu v_j + \frac{1}{n}\sum_{j=1}^{n}\mu^2 = \frac{1}{n}\sum_{j=1}^{n} v_j^2 - \mu^2 .)$$

6.5 Monopoly

We have spent a great deal of time on the competitive model, and we now turn to the polar opposite case, that of monopoly. A monopoly is a firm that faces a downward sloping demand, and has a choice about what price to charge – an increase in price doesn't send most or all of the customers away to rivals.

There are very few pure monopolies. The U.S. post office has a monopoly in first-class mail, but faces competition by FedEx and other express mail companies, as well as by faxes and email, in the broader "send documents to others" market. Microsoft has a great deal of market power, but a small percentage of personal computer users choose Apple or Linux operating systems. There is only one U.S. manufacturer of aircraft carriers.

However, there are many firms that have *market power* or *monopoly power*, which means that they can increase their price above marginal cost and sustain sales for a long period of time.[74] The theory of monopoly is applicable to such firms, although they may face an additional and important constraint: a price increase may affect the behavior of rivals. The behavior of rivals is the subject of the next chapter.

A large market share is not a proof of monopoly, nor is a small market share proof that a firm lacks monopoly power. U.S. Air dominated air traffic to Philadelphia and Pittsburgh, but still lost money. Porsche has a small share of the automobile market, or even the high-end automobile market, but still has monopoly power in that market.

6.5.1 Sources of Monopoly

There are three basic sources of monopoly. The most common source is to be granted a monopoly by the government, either through patents, in which case the monopoly is temporary, or through a government franchise. Intelsat was a government franchise that was granted a monopoly on satellite communications, a monopoly that ultimately proved lucrative indeed. Many cities and towns license a single cable TV company or taxi company, although usually basic rates and fares are set by the terms of the license agreement. New drugs are granted patents that provide a monopoly for a period of time. (Patents generally last twenty years, but pharmaceutical drugs have their own patent laws.) Copyright also confers a monopoly for a supposedly limited period of time. Thus, the Disney Corporation owns copyrights on Mickey Mouse, copyrights which by law should have expired, but have been granted an extension by Congress each time they were due to expire. Copyrights create monopoly power over music as well as cartoon characters, and Time-Warner owns the rights to the song "Happy Birthday to

[74] These terms are used somewhat differently by different authors. Both require downward sloping demand, and usually some notion of sustainability of sales. Some distinguish the terms by whether they are "large" or not, others by how long the price increase can be sustained. We won't need such distinctions here.

You," and receives royalties every time it is played on the radio or other commercial venue.[75] Many of the Beatles songs which McCartney co-authored were purchased by Michael Jackson. This book is copyrighted under terms that expressly prohibit commercial use but permit most other uses.

A second source of monopoly is a large economy of scale. The scale economy needs to be large relative to the size of demand. If the average cost when a single firm serves the entire market is lower than when two or more firms serve the market, a monopoly can be the result. For example, long distance telephone lines were expensive to install, and the first company to do so, A.T. & T., wound up being the only provider of long distance service in the United States. Similarly, scale economies in electricity generation meant that most communities had a single electricity provider prior to the 1980s, when new technology made relatively smaller scale generation more efficient.

The demand-side equivalent of an economy of scale is a *network externality*. A network externality arises when others' use of a product makes it more valuable to each consumer. Standards are a common source of network externality. That AA batteries are standardized makes them more readily accessible, helps drive down their price through competition and economies of scale, and thus makes the AA battery more valuable. AA batteries are available everywhere, unlike proprietary batteries. Fax machines are valuable only if others have similar machines. In addition to standards, a source of network externality is third-party products. Choosing Microsoft Windows as a computer operating system means that there is more software available than for Macintosh or Linux, as the widespread adoption of Windows has led a large variety of software to be written for it. The JVC Video Home System of VCRs came to dominate the Sony Beta system, primarily because there were more movies to rent in the VHS format than in the Beta format at the video rental store. In contrast, recordable DVD has been hobbled by incompatible standards of DVD+R and DVD-R, a conflict not resolved even as the next generation – 50GB discs such as Sony's Blu-ray – start to reach the market. DVDs themselves were slow to be adopted by consumers, because few discs were available for rent at video rental stores, which is a consequence of few adoptions of DVD players. As DVD players became more prevalent, and the number of discs for rent increased, the market *tipped* and DVDs came to dominate VHS.

The third source of monopoly is control of an essential, or a sufficiently valuable, input to the production process. Such an input could be technology that confers a cost advantage. For example, software is run by a computer operating system, and needs to be designed to work well with the operating system. There have been a series of allegations that Microsoft kept secret some of the "application program interfaces" used by Word as a means of hobbling rivals. If so, access to the design of the operating system itself is an important input.

6.5.2 Basic Analysis

Even a monopoly is constrained by demand. A monopoly would like to sell lots of units at very high prices, but higher prices necessarily lead to a loss in sales. So how does a monopoly choose its price and quantity?

[75] Fair use provisions protect individuals with non-commercial uses of copyrighted materials.

A monopoly can choose price, or a monopoly can choose quantity and let the demand dictate the price. It is slightly more convenient to formulate the theory in terms of quantity rather than price, because costs are a function of quantity. Thus, we let $p(q)$ be the demand price associated with quantity q, and $c(q)$ be the cost of producing q. The monopoly's profits are

$$\pi = p(q)q - c(q).$$

The monopoly earns the revenue pq and pays the cost c. This leads to the first order condition, for the profit-maximizing quantity q_m:

$$0 = \frac{\partial \pi}{\partial q} = p(q_m) + q_m p'(q_m) - c'(q_m).$$

The term $p(q) + qp'(q)$ is known as *marginal revenue*. It is the derivative of revenue pq with respect to quantity. Thus, a monopoly chooses a quantity q_m where marginal revenue equals marginal cost, and charges the maximum price $p(q_m)$ the market will bear at that quantity. Marginal revenue is below demand $p(q)$ because demand is downward sloping. That is, $p(q) + qp'(q) < p(q)$.

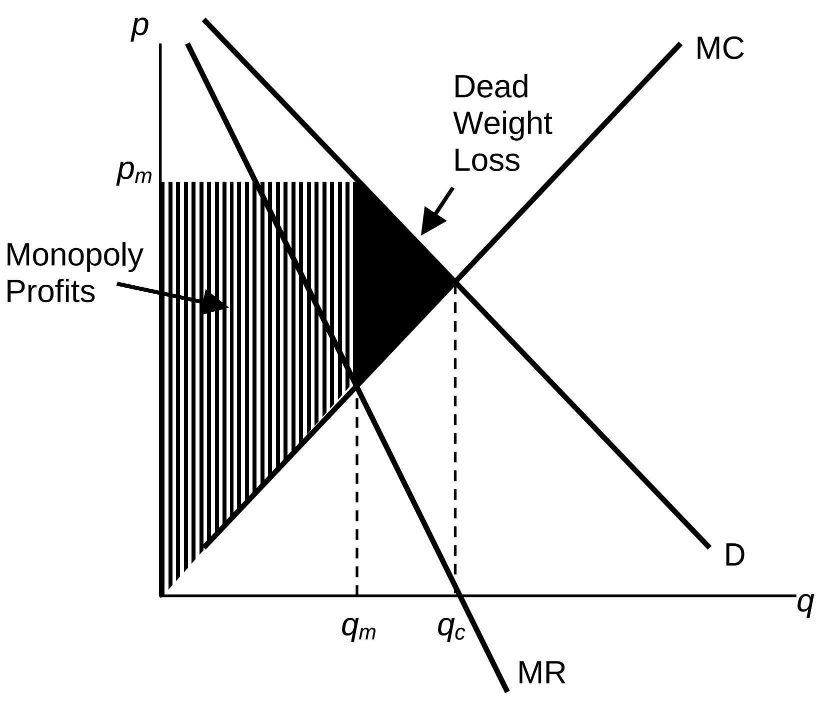

Figure 6-20: Basic Monopoly Diagram

6.5.2.1 (Exercise) If demand is linear, $p(q)=a - bq$, what is marginal revenue? Plot demand and marginal revenue, and total revenue $qp(q)$ as a function of q.

6.5.2.2 (Exercise) For the case of constant elasticity of demand, what is marginal revenue?

6.5.2.3 (Exercise) If both demand and supply have constant elasticity, compute the monopoly quantity and price.

The choice of monopoly quantity is illustrated in Figure 6-20. The key points of this diagram are, first, that marginal revenue lies below the demand curve. This occurs because marginal revenue is the demand $p(q)$ plus a negative number. Second, the monopoly quantity equates marginal revenue and marginal cost, but the monopoly price is higher than the marginal cost. Third, there is a dead weight loss, for the same reason that taxes create a dead weight loss: the higher price of the monopoly prevents some units from being traded that are valued more highly than they cost. Fourth, the monopoly profits from the increase in price, and the monopoly profit is shaded. Fifth, since under competitive conditions supply equals marginal cost, the intersection of marginal cost and demand corresponds to the competitive outcome. We see that the monopoly restricts output and charges a higher price than would prevail under competition.

We can rearrange the monopoly pricing formula to produce an additional insight.

$$p(q_m) - c'(q_m) = -q_m p'(q_m)$$

or

$$\frac{p(q_m) - c'(q_m)}{p(q_m)} = \frac{-q_m p'(q_m)}{p(q_m)} = \frac{1}{\varepsilon}.$$

The left hand side of this equation is known as the *price-cost margin* or *Lerner Index*.[76] The right hand side is one over the elasticity of demand. This formula relates the markup over marginal cost to the elasticity of demand. It is important because perfect competition forces price to equal marginal cost, so this formula provides a measure of the deviation from competition, and in particular says that the deviation from competition is small when the elasticity of demand is large, and vice versa.

Marginal cost will always be at least zero or larger. If marginal cost is less than zero, the least expensive way to produce a given quantity is to produce more and throw some away. Thus, the price-cost margin is no greater than one, and as a result, *a monopolist produces in the elastic portion of demand.* One implication of this observation is that if demand is everywhere inelastic (e.g. $p(q) = q^{-a}$ for a>1), the optimal monopoly quantity is essentially zero, and in any event would be no more than one molecule of the product.

---

[76] Abba Lerner, 1903-1982. Note that $\frac{1}{\frac{-q_m p'(q_m)}{p(q_m)}} = -\frac{1/q_m}{p'(q_m)/p(q_m)} = -\frac{dq/q}{dp/p} = \varepsilon,$ which is used in the derivation.

In addition, the effects of monopoly are related to the elasticity of demand. If demand is very elastic, the effect of monopoly on prices is quite limited. In contrast, if the demand is relatively inelastic, monopolies will increase prices by a large margin.

We can rewrite the formula to obtain

$$p(q_m) = \frac{\varepsilon}{\varepsilon - 1} c'(q_m).$$

Thus, a monopolist marks up marginal cost by the factor $\frac{\varepsilon}{\varepsilon - 1}$, at least when $\varepsilon>1$. This formula is sometimes used to justify a "fixed markup policy," which means a company adds a constant percentage markup to its products. This is an ill-advised policy not justified by the formula, because the formula suggests a markup which depends on the demand for the product in question and thus not a fixed markup for all products a company produces.

6.5.3 Effect of Taxes

A tax imposed on a seller with monopoly power performs differently than a tax imposed on a competitive industry. Ultimately a perfectly competitive industry must pass on all of a tax to consumers, because in the long-run the competitive industry earns zero profits. In contrast, a monopolist might absorb some portion of a tax even in the long-run.

To model the effect of taxes on a monopoly, consider a monopolist who faces a tax rate *t* per unit of sales. This monopolist earns

$$\pi = p(q)q - c(q) - tq \, .$$

The first order condition for profit maximization yields

$$0 = \frac{\partial \pi}{\partial q} = p(q_m) + q_m p'(q_m) - c'(q_m) - t \, .$$

Viewing the monopoly quantity as a function of *t*, we obtain:

$$\frac{dq_m}{dt} = \frac{1}{2p'(q_m) + q_m p''(q_m) - c''(q_m)} < 0 \, ,$$

with the sign following from the second order condition for profit maximization. In addition, the change in price satisfies

$$p'(q_m)\frac{dq_m}{dt} = \frac{p'(q_m)}{2p'(q_m) + q_m p''(q_m) - c''(q_m)} > 0.$$

Thus, a tax causes a monopoly to increase its price. In addition, the monopoly price rises by less than the tax if $p'(q_m)\frac{dq_m}{dt} < 1$, or

$$p'(q_m) + q_m p''(q_m) - c''(q_m) < 0.$$

This condition need not be true, but is a standard regularity condition imposed by assumption. It is true for linear demand and increasing marginal cost. It is false for constant elasticity of demand, ε>1 (which is the relevant case, for otherwise the second order conditions fail) and constant marginal cost. In the latter case (constant elasticity and marginal cost), a tax on a monopoly increases price by more than the amount of the tax.

6.5.3.1 (Exercise) Use a revealed preference argument to show that a per unit tax imposed on a monopoly causes the quantity to fall. That is, hypothesize quantities q_b before the tax, and q_a after the tax, and show that two facts – the before tax monopoly preferred q_b to q_a and the taxed monopoly made higher profits from q_b together imply the $q_b \leq q_a$.

6.5.3.2 (Exercise) When both demand and supply have constant elasticity, use the results of 6.5.2.3 (Exercise) to compute the effect of a proportional tax (i.e. a portion of the price paid to the government).

6.5.4 Price Discrimination

Pharmaceutical drugs for sale in Mexico are generally priced substantially below their U.S. counterparts. Pharmaceutical drugs in Europe are also cheaper than in the U.S., although not as inexpensive as in Mexico, with Canadian prices usually between the U.S. and European prices. (The comparison is between identical drugs produced by the same manufacturer.)

Pharmaceutical drugs differ in price across countries primarily because demand conditions vary. The formula

$$p(q_m) = \frac{\varepsilon}{\varepsilon - 1} c'(q_m).$$

shows that a monopoly seller would like to charge a higher markup over marginal cost to customers with less elastic demand than to customers with more elastic demand, because $\frac{\varepsilon}{\varepsilon - 1}$ is a decreasing function of ε, for ε>1. Charging different prices for the same product to different customers is known as *price discrimination*. In business settings, it is sometimes known as *value-based pricing*, which is a more palatable term to tell to customers.

Computer software vendors often sell a "student" version of their software, usually at substantially reduced prices, and requiring proof of being a student to qualify for the lower price. Such student discounts are examples of price discrimination, and students have more elastic demand than business users. Similarly, the student and senior citizen discounts at movies and other venues sell the same thing – a ticket to the show – for different prices, and thus qualify as price discrimination.

In order for a seller to price-discriminate, the seller must be able to

- identify (approximately) the demand of groups of customers
- prevent arbitrage

Arbitrage is also known as "buying low and selling high," and represents the act of being an intermediary. Since price discrimination requires charging one group a higher price than another, there is potentially an opportunity for arbitrage, arising from members of the low price group buying at the low price and selling at the high price. If the seller can't prevent arbitrage, arbitrage essentially converts a two-price system to sales at the low price.

Why offer student discounts at the movies? You already know the answer to this – students have lower incomes on average than others, and lower incomes translate into a lower willingness to pay for normal goods. Consequently a discount to a student makes sense from a demand perspective. Arbitrage can be mostly prevented by requiring a student identification card to be presented. Senior citizen discounts are a bit more subtle. Generally seniors aren't poorer than other groups of customers (in the United States, at least). However, seniors have more free time, and thus are able to substitute to matinee showings[77] or drive to more distant locations should those offer discounts. Thus seniors have relatively elastic demand more because of their ability to substitute than because of their income.

Airlines commonly price discriminate, using "Saturday night stay-overs" and other devices. To see that such charges represent price discrimination, consider a passenger who lives in Dallas but needs to spend Monday through Thursday in Los Angeles two weeks in a row. This passenger could buy two round-trip tickets:

Trip One:
First Monday: Dallas → Los Angeles
First Friday: Los Angeles → Dallas

Trip Two:
Second Monday: Dallas → Los Angeles
Second Friday: Los Angeles → Dallas

At the time of this writing, the approximate combined cost of these two flights was US$2,000. In contrast, another way of arranging exactly the same travel is to have two round-trips, one of which originates in Dallas, while the other originates in Los Angeles:

[77] Matinee showings are those early in the day, which are usually discounted. These discounts are not price discrimination because a show at noon isn't the same product as a show in the evening.

Trip One:
First Monday: Dallas → Los Angeles
Second Friday: Los Angeles → Dallas

Trip Two:
First Friday: Los Angeles → Dallas
Second Monday: Dallas → Los Angeles

This pair of round trips involves exactly the same travel as the first pair, but costs less than $500 for both (at the time of this writing). The difference is that the second pair involves staying over Saturday night for both legs, and that leads to a major discount for most U.S. airlines. (American Airlines quoted the fares.)

How can airlines price discriminate? There are two major groups of customers: business travelers and leisure travelers. Business travelers have the higher willingness to pay overall, and the nature of their trips tends to be that they come home for the weekend. In contrast, a leisure traveler will usually want to be away for a weekend, so a weekend stay-over is an indicator of a leisure traveler. It doesn't work perfectly as an indicator – some business travelers must be away for the weekend – but it is sufficiently correlated with leisure travel that it is profitable for the airline to price discriminate.

These examples illustrate an important distinction. Senior citizen and student discounts are based on the identity of the buyer, and qualifying for the discount requires showing an identity card. In contrast, airline price discrimination is not based on the identity of the buyer but on the choices by the buyer. The former is known as *direct price discrimination*, while the latter is known as *indirect price discrimination*.[78]

Two common examples of indirect price discrimination are coupons and quantity discounts. Coupons offer discounts for products and are especially common in grocery stores, where they are usually provided in a newspaper section available free at the front of the store. Coupons discriminate on the basis of the cost of time. It takes time to find the coupons for the products one is interested in buying, and thus those with a high value of time won't find it worthwhile spending twenty minutes to save $5 (effectively a $15 per hour return), while those with a low value of time will find that return worthwhile. Since those with a low value of time tend to be more price sensitive (more elastic demand), coupons offer a discount available to all but used primarily by customers with a more elastic demand, and thus increase the profits of the seller.

Quantity discounts are discounts for buying more. Thus, the large size of milk, laundry detergent and other items often cost less per unit than smaller sizes, and the difference is greater than the savings on packaging costs. In some cases, the larger sizes entail greater packaging costs; some manufacturers "band together" individual units, incurring additional costs to create a larger size which is then discounted. Thus, the "twenty-four pack" of paper towels sells for less per roll than the individual rolls; such large volumes appeal primarily to large families, who are more price-sensitive on average.

[78] The older and incoherent language for these concepts called direct price discrimination "third degree price discrimination," while indirect price discrimination was called second degree price discrimination. In the older language, first degree price discrimination meant perfect third degree price discrimination.

6.5.5 Welfare Effects

Is price discrimination a good thing, or a bad thing? It turns out that there is no definitive answer to this question. Instead, it depends on circumstances. We illustrate this conclusion with a pair of exercises.

6.5.5.1 (Exercise) Let marginal cost be zero for all quantities. Suppose there are two equal-sized groups of customers, group 1 with demand $q(p)=12-p$, group 2 with demand $q(p)=8-p$. Show that a non-discriminating monopolist charges a price of 5 and the discriminating monopolist charges group 1 the price 6 and group 2 the price 4. Then calculate the gains from trade, with discrimination and without, and show that price discrimination reduces the gains from trade.

This exercise illustrates a much more general proposition: if a price-discriminating monopolist produces less than a non-discriminating monopolist, then price discrimination reduced welfare. This proposition has an elementary proof. Consider the price discriminating monopolist's sales, and then allow arbitrage. The arbitrage increases the gains from trade, since every transaction has gains from trade. Arbitrage, however, leads to a common price like that charged by a non-discriminating monopolist. Thus, the only way price discrimination can increase welfare is if it leads a seller to sell more output than she would otherwise. This is possible, as the next exercise shows.

6.5.5.2 (Exercise) Let marginal cost be zero for all quantities. Suppose there are two equal-sized groups of customers, group 1 with demand $q(p)=12-p$, group 2 with demand $q(p)=4-p$. Show that a non-discriminating monopolist charges a price of 6 and the discriminating monopolist charges group 1 the price 6 and group 2 the price 2. Then calculate the gains from trade, with discrimination and without, and show that price discrimination increases the gains from trade.

In this exercise, we see that price discrimination that brings in a new group of customers may increase the gains from trade. Indeed, this example involves a Pareto improvement: the seller and group 2 are better off, and group 1 no worse off, than without price discrimination. (A Pareto improvement requires that no one is worse off and at least one person is better off.)

Whether price discrimination increases the gains from trade overall depends on circumstances. However, it is worth remembering that people with lower incomes tend to have more elastic demand, and thus get lower prices under price discrimination than absent price discrimination. Consequently, a ban on price discrimination tends to hurt the poor and benefit the rich no matter what the overall effect.

6.5.6 Two-Part Pricing

A common form of price discrimination is known as *two-part pricing*. Two-part pricing usually involves a fixed charge and a marginal charge, and thus offers an ability for a seller to capture a portion of the consumer surplus. For example, electricity often comes with a fixed price per month and then a price per kilowatt-hour, which is two-part pricing. Similarly, long distance and cellular telephone companies charge a fixed fee per month, with a fixed number of "included" minutes, and a price per minute for additional minutes. Such contracts really involve three parts rather than two-parts, but are similar in spirit.

From the seller's perspective, the ideal two-part price is to charge marginal cost plus a fixed charge equal to the customer's consumer surplus, or perhaps a penny less. By setting price equal to marginal cost, the seller maximizes the gains from trade. By setting the fixed fee equal to consumer surplus, the seller captures the entire gains from trade. This is illustrated in Figure 6-21.

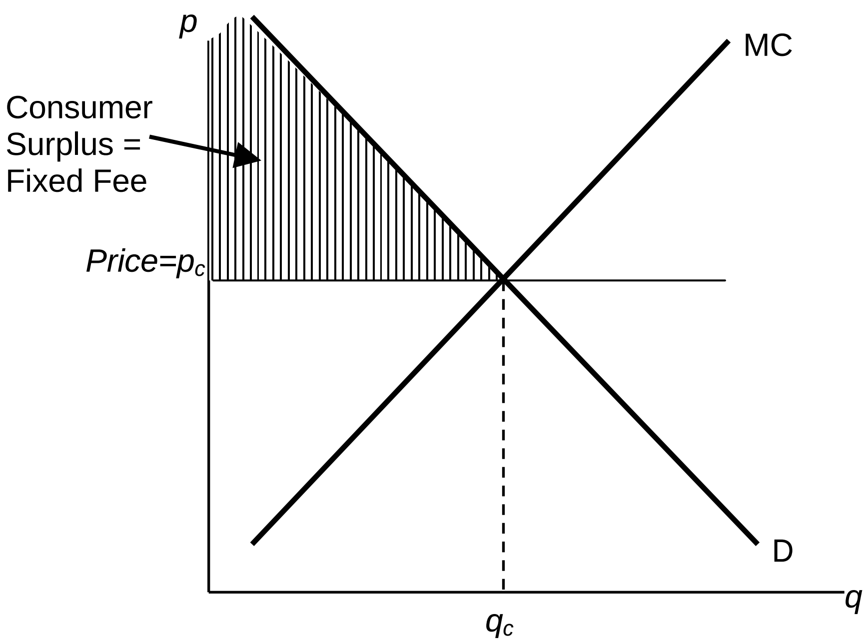

Figure 6-21: Two-Part Pricing

6.5.7 Natural Monopoly

A natural monopoly arises when a single firm can efficiently serve the entire market because average costs are lower with one firm than with two firms. An example is illustrated in Figure 6-22. In this case, the average total cost of a single firm is lower than if two firms operate, splitting the output between them. The monopolist would like to price at p_m, which maximizes profits.[79]

[79] The monopoly price may or may not be *sustainable*. A monopoly price is not sustainable if it would lead to entry, thereby undercutting the monopoly. The feasibility of entry, in turn, depends on whether the costs of entering are not recoverable ("*sunk*"), and how rapidly entry can occur. If the monopoly price is not sustainable, the monopoly may engage in *limit pricing*, which is jargon for pricing to deter (limit) entry.

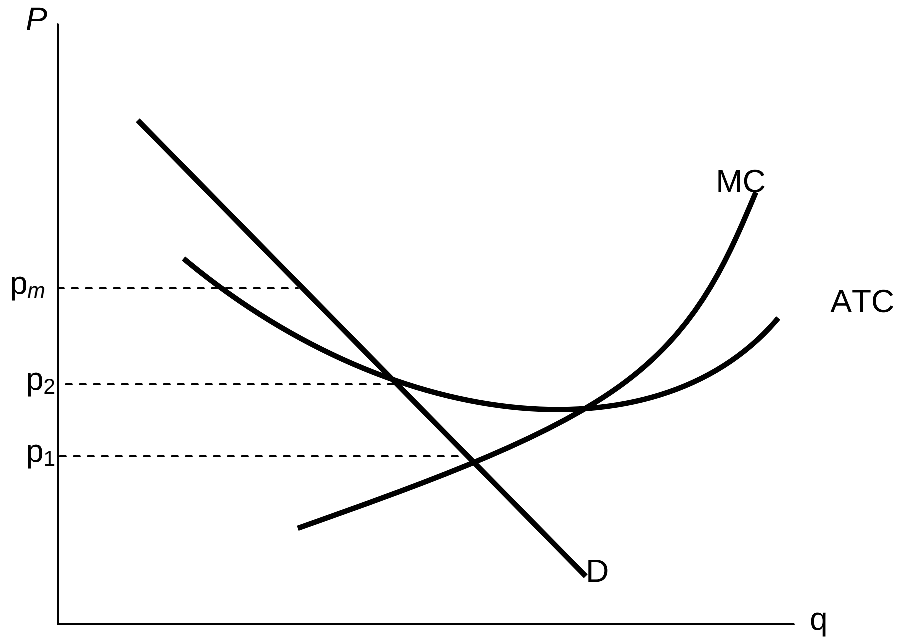

Figure 6-22: Natural Monopoly

Historically, the United States and other nations have regulated natural monopolies like those found in electricity, telephony and water service. An immediate problem with regulation is that the efficient price, that is, the price that maximizes the gains from trade, requires a subsidy from outside the industry. We see the need for a subsidy in Figure 6-22 because the price that maximizes the gains from trade is p_1, which sets the demand (marginal value) equal to the marginal cost. At this price, however, the average total cost exceeds the price, so that a firm with such a regulated price would lose money. There are two alternatives. The product could be subsidized, and subsidies are used with postal service and passenger rail in the United States, and historically for many more products in Canada and Europe including airlines and airplane manufacture. Alternatively, regulation could be imposed that aims to limit the price to p_2, the lowest break-even price. This is the more common strategy in the United States.

There are two strategies toward limiting the price: *price-cap regulation*, which directly imposes a maximum price, and *rate of return regulation*, that limits the profitability of firms. Both of these approaches induce some inefficiency of production. In both cases, an increase in average cost may translate into additional profits for the firm, causing regulated firms to engage in unnecessary activities.

6.5.8 Peak Load Pricing

Fluctuations in demand often require holding capacity which is used only a fraction of the time. Hotels have off-seasons when most rooms are empty. Electric power plants are designed to handle peak demand, usually hot summer days, with some of the capacity standing idle on other days. Demand for trans-Atlantic airline flights is much higher in the summer than the rest of the year. All of these examples have the similarity that an amount of capacity – hotel space, airplane seats, electric generation capacity – will be used over and over, which means it is used in both high demand and low demand states. How should pricing be accomplished when demand fluctuates? This can be thought of as a question of how to allocate the cost of capacity across several time periods when demand systematically fluctuates.

Consider a firm that experiences two kinds of costs – a capacity cost and a marginal cost. How should capacity be priced? This issue is applicable to a wide variety of industries, including pipelines, airlines, telephone networks, construction, electricity, highways, and the internet.

The basic peak-load pricing problem, pioneered by Marcel Boiteux (1922 –), considers two periods. The firm's profits are given by

$$\pi = p_1 q_1 + p_2 q_2 - \beta \max\{q_1, q_2\} - mc(q_1 + q_2).$$

Setting price equal to marginal costs is not sustainable, because a firm selling with price equal to marginal cost would not earn a return on the capacity, and thus would lose money and go out of business. Consequently, a capacity charge is necessary. The question of peak load pricing is how the capacity charge should be allocated. This question is not trivial because some of the capacity is used in both periods.

For the sake of simplicity, we will assume demands are independent, that is, q_1 is independent of p_2 and vice versa. This assumption is often unrealistic, and generalizing it actually doesn't complicate the problem too much. The primary complication is in computing the social welfare when demands are functions of two prices. Independence is a convenient starting point.

Social welfare is

$$W = \int_0^{q_1} p_1(x)dx + \int_0^{q_2} p_2(x)dx - \beta \max\{q_1, q_2\} - mc(q_1 + q_2).$$

The Ramsey problem is to maximize W subject to a minimum profit condition. A technique for accomplishing this maximization is to instead maximize

$$L = W + \lambda\pi.$$

By varying λ, we vary the importance of profits to the maximization problem, which will increase the profit level in the solution as λ increases. Thus, the correct solution to the constrained maximization problem is the outcome of the maximization of L, for some value of λ.

A useful notation is 1_A, which is known as the *characteristic function of the set A*. This is a function which is 1 when A is true, and zero otherwise. Using this notation, the first order condition for the maximization of L is:

$$0 = \frac{\partial L}{\partial q_1} = p_1(q_q) - \beta 1_{q_1 \geq q_2} - mc + \lambda\left(p_1(q_q) + q_1 p_1{}'(q_1) - \beta 1_{q_1 \geq q_2} - mc\right)$$

or,

$$\frac{p_1(q_1) - \beta 1_{q_1 \geq q_2} - mc}{p_1} = \frac{\lambda}{\lambda + 1} \frac{1}{\varepsilon_1}$$

where $1_{q_1 \geq q_2}$ is the characteristic function of the event $q_1 \geq q_2$.

Similarly,

$$\frac{p_2(q_2) - \beta 1_{q_1 \leq q_2} - mc}{p_2} = \frac{\lambda}{\lambda + 1} \frac{1}{\varepsilon_2}$$

Note as before that $\lambda \to \infty$ yields the monopoly solution.

There are two potential types of solution. Let the demand for good 1 exceed the demand for good 2. Either $q_1 > q_2$, or the two are equal.

Case 1: $q_1 > q_2$.

$$\frac{p_1(q_1) - \beta - mc}{p_1} = \frac{\lambda}{\lambda + 1} \frac{1}{\varepsilon_1} \text{ and } \frac{p_2(q_2) - mc}{p_2} = \frac{\lambda}{\lambda + 1} \frac{1}{\varepsilon_2}.$$

In case 1, with all of the capacity charge allocated to good 1, quantity for good 1 still exceeds quantity for good 2. Thus, the peak period for good 1 is an extreme peak. In contrast, case 2 arises when assigning the capacity charge to good 1 would reverse the peak – assigning all of the capacity charge to good 1 would make period 2 the peak.

Case 2: $q_1 = q_2$.

The profit equation can be written

$p_1(q) - mc + p_2(q) - mc = \beta$

This equation determines q, and prices are determined from demand.

The major conclusion from peak load pricing is that either the entire cost of capacity is allocated to the peak period, or there is no peak period in the sense that the two periods have the same quantity demanded given the prices. That is, either the prices equalize the quantity demanded, or the prices impose the entire cost of capacity only on one peak period.

Moreover, the price (or, more properly, the markup over marginal cost) is proportional to the inverse of the elasticity, which is known as Ramsey pricing.

6.6 Information

An important advantage of the price system is that it economizes on information. A typical consumer needs to know only the prices of goods and their own personal preferences in order to make a sensible choice of purchases, and manufacturers only need to know the prices of goods in order to decide what to produce. Such economies of information are an advantage over centrally-planned economies, which attempt to direct production and consumption decisions using something other than prices, and centrally-planned economies typically experience chronic shortages and occasional surpluses. Shortages of important inputs to production may have dramatic effects and the shortages aren't remedied by the price of the input rising in a centrally planned economy, and thus often persist for long periods of time.

There are, however, circumstances where the prices are not the only necessary information required for firms and consumers to make good decisions. In such circumstances, information itself can lead to market failures.

6.6.1 Market for Lemons

Nobel laureate George Akerlof (1940 –) examined the market for used cars and considered a situation where the sellers are better informed than the buyers. This is quite reasonable, as sellers have owned the car for a while and are likely to know its quirks and potential problems. Akerlof showed that this differential information may cause the used car market to collapse; that is, the information possessed by sellers of used cars destroys the market.

To understand Akerlof's insight, suppose that the quality of used cars lies on a 0 to 1 scale and that the population of used cars is uniformly distributed on the interval from 0 to 1. In addition, let that quality represent the value a seller places on the car, and suppose buyers put a value that is 50% higher than the seller. Finally, the seller knows the actual quality, while the buyer does not.

Can a buyer and seller trade in such a situation? First, note that trade is a good thing, because the buyer values the car more than the seller. That is, both the buyer and seller know that they should trade. But can they agree on a price? Consider a price p. At this price, any seller who values the car less than p will be willing to trade. But because of our uniform distribution assumption, this means the distribution of qualities of cars offered for trade at price p will be uniform on the interval 0 to p. Consequently, the average quality of these cars will be ½ p, and the buyer values these cars 50% more which yields ¾ p. Thus, the buyer is not willing to pay the price p for the average car offered at price p.

The effect of the informed seller, and uninformed buyer, produces a "lemons" problem. At any given price, all the lemons and only a few of the good cars are offered, and the buyer – not knowing the quality of the car – isn't willing to pay as much as the actual value of a high value car offered for sale. This causes the market to collapse; and only the worthless cars trade at a price around zero. Economists call the differential information an *informational asymmetry*.

In the real world, of course, the market has found partial or imperfect solutions to the *lemons* problem identified by Akerlof. First, buyers can become informed and regularly hire their own mechanic to inspect a car they are considering. Inspections reduce the informational asymmetry but are costly in their own right. Second, intermediaries offer warranties and certification to mitigate the lemons problem. The existence of both of these solutions, which involve costs in their own right, is itself evidence that the lemons problem is a real and significant problem, even though competitive markets find ways to ameliorate the problems.

An important example of the lemons problem is the inventor who creates an idea that is difficult or impossible to patent. Consider an innovation that would reduce the cost of manufacturing computers. The inventor would like to sell it to a computer company, but can't tell the computer company what the innovation entails prior to price negotiations, because then the computer company could just copy the innovation. Similarly, the computer company can't possibly offer a price for the innovation in advance of knowing what the innovation is. As a result, such innovations usually require the inventor to enter the computer manufacturing business, rather than selling to an existing manufacturer, entailing many otherwise unnecessary costs.

6.6.1.1 (Exercise) In Akerlof's market for lemons model, suppose it is possible to certify cars, verifying that they are better than a particular quality q. Thus, a market for cars "at least as good as q" is possible. What price or prices are possible in this market? [Hint: sellers offer cars only if $q \leq$ quality $\leq p$.] What quality maximizes the expected gains from trade?

6.6.2 Myerson-Satterthwaite Theorem

The lemons problem is a situation where the buyers are relatively uninformed and care about the information held by sellers. Lemons problems are limited to situations where the buyer isn't well-informed and can be mitigated by making information public. In many transactions, the buyer knows the quality of the product, so lemons concerns aren't a significant issue. There can still be a market failure, however, if there are a limited number of buyers and sellers.

Consider the case of one buyer and one seller bargaining over the sale of a good. The buyer knows his own value v for the good, but not the seller's cost. The seller knows her own cost c for the good, but not the buyer's value. The buyer views the seller's cost as uniformly distributed on the interval [0,1], and similarly the seller views the buyer's value as uniformly distributed on [0,1].[80] Can efficient trade take place? Efficient trade requires that trade occurs whenever $v>c$, and the remarkable answer is that it is impossible to arrange efficient trade if the buyer and seller are to trade voluntarily. This

[80] The remarkable fact proved by Roger Myerson and Mark Satterthwaite (Efficient Mechanisms for Bilateral Trade, *Journal of Economic Theory*, 28, 1983, 265-281) is that the distributions don't matter; the failure of efficient trade is a fully general property. Philip Reny and Preston McAfee (Correlated Information and Mechanism Design, *Econometrica* 60, No. 2, March 1992, 395-421) show the nature of the distribution of information matters, and Preston McAfee (Efficient Allocation with Continuous Quantities, *Journal of Economic Theory* 53, no. 1, February 1991: 51-74.) showed that continuous quantities can overturn the Myerson-Satterthwaite theorem.

is true even if a third party is used to help arrange trade, provided the third party doesn't subsidize the transaction.

The total gains from trade under efficiency are

$$\int_0^1 \int_0^v v - c \, dc \, dv = \int_0^1 \frac{v^2}{2} dv = \frac{1}{6}.$$

A means of arranging trade, or a *mechanism*, asks the buyer and seller for their value and cost, respectively, and then orders trade if the value exceeds the cost, and dictates a payment p by the buyer to the seller. Buyers need not make honest reports to the mechanism, however, and the mechanisms must be designed to induce the buyer and seller to report honestly to the mechanism, so that efficient trades can be arranged.[81]

Consider a buyer who actually has value v but reports a value r. The buyer trades with the seller if the seller has a cost less than r, which occurs with probability r.

$u(r,v) = vr - E_c p(r, c)$.

The buyer gets the actual value v with probability r, and makes a payment that depends on the buyer's report and the seller's report, but we can take expectations over the seller's report to eliminate it (from the buyer's perspective), and this is denoted $E_c p(r, c)$, which is just the expected payment given the report r. In order for the buyer to choose to be honest, u must be maximized at $r=v$ for every v, for otherwise some buyers would lie and some trades would not be efficiently arranged. Thus, we can conclude[82]

$$\frac{d}{dv} u(v,v) = u_1(v,v) + u_2(v,v) = u_2(v,v) = r\Big|_{r=v} = v.$$

The first equality is just the total derivative of u(v,v), because there are two terms; the second equality because u is maximized over the first argument r at $r=v$, and the first order condition insures $u_1 = 0$. Finally, u_2 is just r, and we are evaluating the derivative at the point $r = v$. A buyer who has a value $v + \Delta$, but who reports v, trades with probability v and makes the payment $E_c p(v, c)$. Such a buyer gets Δv more in utility than the buyer with value v. Thus a Δ increase in value produces an increase in utility of at least Δv, showing that $u(v+\Delta, v+\Delta) \geq u(v,v) + \Delta v$ and hence that $\frac{d}{dv} u(v,v) \geq v$. A similar argument considering a buyer with value v who reports $v + \Delta$ shows that equality occurs.

---

[81] Inducing honesty is without loss of generality. Suppose that the buyer of type v reported the type $z(v)$. Then we can add a stage to the mechanism, where the buyer reports a type, which is converted via the function z to a report, and then that report given to the original mechanism. In the new mechanism, reporting v is tantamount to reporting $z(v)$ to the original mechanism.

[82] We maintain an earlier notation that the subscript refers to a partial derivative, so that if we have a function f, f_1 is the partial derivative of f with respect to the first argument of f.

The value $u(v,v)$ is the gain accruing to a buyer with value v who reports having value v. Since the buyer with value 0 gets zero, the total gain accruing to the average buyer can be computed by integrating by parts

$$\int_0^1 u(v,v)dv = -(1-v)u(v,v)\Big|_{v=0}^{1} + \int_0^1 (1-v)\left(\frac{du}{dv}\right)dv = \int_0^1 (1-v)v\,dv = \frac{1}{6}.$$

In the integration by parts, dv = d –(1–v) is used. The remarkable conclusion is that, if the buyer is induced to truthfully reveal the buyer's value, the buyer must obtain the entire gains from trade! This is actually a quite general proposition. If you offer the entire gains from trade to a party, they are induced to maximize the gains from trade. Otherwise, they will want to distort away from maximizing the entire gains from trade, which will result in a failure of efficiency.

The logic with respect to the seller is analogous: the only way to get the seller to report her cost honestly is to offer her the entire gains from trade.

6.6.2.1 (Exercise) Let $h(r, c)$ be the gains of a seller who has cost c and reports r,

$h(r, c) = p(v, r) - (1\text{-}r)c$.

Noting that the highest cost seller (c=1) never sells and thus obtains zero profits, show that honesty by the seller implies the expected value of h is $\frac{1}{6}$.

The Myerson-Satterthwaite theorem shows that the gains from trade are insufficient to induce honesty by both parties. (Indeed, they are half the necessary amount!) Thus, any mechanism for arranging trades between the buyer and the seller must suffer some inefficiency. Generally this occurs because buyers act like they value the good less than they do, and sellers act like their costs are higher than they truly are.

It turns out that the worst case scenario is a single buyer and a single seller. As markets get "thick," the per capita losses converge to zero, and markets become efficient. Thus, informational problems of this kind are a "small numbers" issue. However, many markets do in fact have small numbers of buyers or sellers. In such markets, it seems likely that informational problems will be an impediment to efficient trade.

6.6.3 Signaling

An interesting approach to solving informational problems involves *signaling*.[83] Signaling, in economic jargon, means expenditures of time or money whose purpose is to convince others of something. Thus, people signal wealth by wearing Rolex watches, driving expensive cars or sailing in the America's Cup. They signal erudition by tossing out quotes from Kafka or Tacitus into conversations. They signal being chic by wearing the right clothes and listening to cool music. Signaling is also rampant in the animal

[83] Signaling was introduced by Nobel laureate Michael Spence in his dissertation, part of which was reprinted in "Job Market Signaling," *Quarterly Journal of Economics* 87, August 1973, 355-74.

world, from peacock feathers to elk battles and the subject of a vibrant and related research program.

A university education serves not just to educate, but also to signal the ability to learn. Businesses often desire employees who are able to adapt to changing circumstances, and who can easily and readily learn new strategies and approaches. Education signals such abilities because it will easier for quick learners to perform well in university. A simple model suffices to illustrate the point. Suppose there are two types of people. Type A has a low cost c_A of learning, and type B has a higher cost c_B of learning. It is difficult to determine from an interview whether someone is type A or not. Type A is worth more to businesses, and the competitive wage w_A (expressed as a present value of lifetime earnings) for type A's is higher than the wage w_B for type B's.

A person can signal that they are a type A by taking a sufficient amount of education. Suppose the person devotes an amount of time x to learning in university, thus incurring the cost $c_A x$. If x is large enough so that

$$w_A - c_A x > w_B > w_A - c_B x,$$

it pays the type A to obtain the education, but not the type B, if education in fact signals that the student is type A. Thus, a level of education x in this case signals a trait (ease of learning) that is valued by business, and it does so by voluntary choice – those with a high cost of learning choose not to obtain the education, even though they could do it. This works as a signal because only type A would voluntarily obtain the education in return for being perceived to be a type A.

There are several interesting aspects to this kind of signaling. First, the education embodied in x need not be valuable in itself; the student could be studying astronomy or ancient Greek, neither of which are very useful in most businesses, but are nevertheless strong signals of the ability to learn. Second, the best subject matter for signaling is that in which the difference in cost between the type desired by employers and the less desirable type is greatest, that is, where $c_B - c_A$ is greatest. Practical knowledge is somewhat unlikely to make this difference great; instead, challenging abstract problem-solving may be a better separator. Clearly, it is desirable to have the subject matter be useful, if it can still do the signaling job. But interpreting long medieval poems could more readily signal the kind of flexible mind desired in management than studying accounting, not because the desirable type is good at it, or that it is useful, but because the less desirable type is so much worse at it.

Third, one interprets signals by asking "what kinds of people would make this choice?" while understanding that the person makes the choice hoping to send the signal. Successful law firms have very fine offices, generally much finer than the offices of their clients. Moreover, there are back rooms at most law firms, where much of the real work is done, that aren't nearly so opulent. The purpose of the expensive offices is to signal success, essentially making the statement that

> "we couldn't afford to waste money on such expensive offices if we weren't very successful. Thus, you should believe we are successful."

The law firm example is similar to the education example. Here, the cost of the expenditures on fancy offices is different for different law firms because more successful firms earn more money and thus value the marginal dollar less. Consequently, more successful firms have a lower cost of a given level of office luxury. What is interesting about signaling is that it is potentially quite wasteful. A student spends four years studying boring poems and dead languages in order to demonstrate a love of learning, and a law firm pays $75,000 for a conference table that it rarely uses and gets no pleasure out of, in order to convince a client that the firm is extremely successful. In both cases, it seems like a less costly solution should be available. The student can take standardized tests, and the law firm could show its win-loss record to the potential client. But standardized tests may measure test-taking skills rather than learning ability, especially if what matters is the learning ability over a long time horizon. Win-loss records can be "massaged," and in the majority of all legal disputes, the case settles and both sides consider themselves "the winner." Consequently, statistics may not be a good indicator of success, and the expensive conference table a better guide.