Chapter **One**

Strategy and Competition

"However beautiful the strategy, you should occasionally look at the results."
—Winston Churchill

Chapter Overview

Purpose

The purpose of this chapter is to introduce the student to a variety of strategic issues that arise in the manufacturing function of the firm.

Key Points

1. *Manufacturing matters.* This writer contends that the loss of the manufacturing base in the U.S. economy is not healthy and will eventually lead to an overall loss in the standard of living and quality of life in this country. It counters the argument that our evolution into a service economy is a natural and healthy thing.
2. *Strategic dimensions.* Along with cost and/or product differentiation, other dimensions along which firms distinguish themselves include (a) quality, (b) delivery speed, (c) delivery reliability, and (d) flexibility.
3. *Global competition.* How do we measure our success and economic health on a global scale? One way is to examine classical measures of relative economic strength, which include (a) balance of trade, (b) share of world exports, (c) creation of jobs, and (d) cost of labor. However, such macro measures do not adequately explain why certain countries dominate certain industries. National competitive advantage is a consequence of several factors (factor conditions, demand conditions, related and supporting industries, firm strategy structure, and rivalry), although productivity also plays an important role.
4. *Strategic initiatives.* We discuss several strategic initiatives that have allowed many companies to shine in their respective arenas. These include (a) business process reengineering, (b) just-in-time manufacturing and purchasing systems, (c) time-based competition, and (d) competing on quality.
5. *Product and process life cycles.* Most of us understand that products have natural life cycles: start-up, rapid growth, maturation, stabilization, or decline. However, it is rarely recognized that processes too have life cycles. Initially, new manufacturing processes have the characteristics of a job shop. As the process matures, automation is introduced. In the mature phases of a manufacturing process, most major operations are automated. A firm needs to match the phases of product and process life cycles to be the most successful in its arena.

6. *Learning and experience curves.* These are helpful in forecasting the decline in unit cost of a manufacturing process as one gains experience with the process. Learning curves are more appropriate when modeling the learning of an individual worker, and experience curves are more appropriate when considering an entire industry.
7. *Capacity growth planning.* Another important strategic issue in operations is determining the timing and sizing of new capacity additions. Simple models (make or buy problem) and more complex exponential growth models are explored in Section 1.11. In addition, some of the factors that determine appropriate location of new facilities is explored.

Strategy is a long-term plan of action designed to achieve a particular goal, most often winning. Its root is from the Greek *stratēgos*, which referred to a "military commander" during the age of Athenian Democracy. Strategy was originally conceived in the military context. Two famous books dealing with military strategy are *The Prince* by Machiavelli and *The Art of War* by Sun Tzu.

Hence, we can see that business strategy relates closely to military strategy. Companies fight on an economic battlefield, and long-term strategies determine winners and losers. Business strategy is the highest level of corporate activity that bundles together the disparate functional area strategies. Business strategy sets the terms and goals for a company to follow.

Perhaps the reason that chief executive officers (CEOs) are compensated so highly in the United States is the realization that the strategic vision of the CEO is often the difference between the success and failure of a company. The strategic visions of industry giants such as Henry Ford, Jack Welch, and Bill Gates were central to the success of their companies that have, at one time or another, dominated their competition.

Perhaps the most dramatic example is Apple Corporation. With the introduction of the iPod in 2002 and the iPhone is 2007, Apple transformed itself from a failing computer company to a major force in portable computing and telecommunications. The fascinating transformation of the firm is described in the Snapshot Application on the next page.

Success requires a vision, and visions must be articulated so all of the firm's employees can share in that vision. The formal articulation of the vision is known as the company mission statement. A good mission statement should provide a clear description of the goals of the firm and is the first step toward formulating a coherent business strategy. Poor mission statements tend to be wordy and full of generalities. Good mission statements are direct, clear, and concise. In their book, Jones and Kahaner (1995) list the 50 corporate mission statements that they perceive as the best. One example is the Gillette Corporation. Their mission statement is: "Our Mission is to achieve or enhance clear leadership, worldwide, in the existing or new core consumer product categories in which we choose to compete." They then go on to list exactly which areas they perceive as their core competencies. Intel defines their mission as: "Do a great job for our customers, employees, and stockholders by being the preeminent building block supplier to the computing industry." The statement then provides details on "Values and Objectives." In many cases, their objectives are quite specific (e.g., "Lead in LAN products and Smart Network Services"). Certainly, the award for conciseness has to go to the General Electric Corporation, whose mission statement is three words: "Boundaryless . . . Speed . . . Stretch." The commentary following the statement provides an explanation of exactly what these words mean in the context of the corporation.

Snapshot Application

APPLE ADOPTS A NEW BUSINESS STRATEGY AND SHIFTS ITS CORE COMPETENCY FROM COMPUTERS TO PORTABLE ELECTRONICS

Apple Computer was the kind of success story one sees in the movies. Two youngsters, Steve Wozniak and Steve Jobs, grew up with an interest in hobbyist computers. Working from a garage, they founded Apple Computer in April 1976, and soon after, introduced a build your own hobbyist computer called the Apple I. The firm was incorporated a year later with the help of Mike Markula, and the Apple II was introduced in April 1977, ushering in the world of personal computing. Perhaps it was the fact that it was selected as the platform for Visicalc, the first spreadsheet program, that led to its success, as much as the superior capabilities of the hardware.

While personal computers have become a common part of our everyday lives, we forget that they are a relatively new invention. The nature of the personal computer marketplace was dramatically altered by the introduction of the first PC by IBM in 1981. IBM's open architecture allowed for inexpensive clones to enter the market, and crowd out Apple's significantly more expensive products. By the turn of the century, Apple's future looked to be in doubt.

Apple's subsequent transformation and rebirth is a fascinating bit of business history. Around 2001, an independent consultant, Tony Fadell, was shopping around his concept of an MP3 music player linked to a music sales service. At that time, MP3 players were not new; one could "rip" music from one's CD collection, and load the songs on the player. While no one else was interested in Fadell's idea, he was hired by Apple and assigned to a team of 30 people, including designers, programmers, and hardware engineers.

When Apple decided to go ahead with the MP3 player concept, they also decided that they needed a new design to separate themselves from the rest of the marketplace. Apple subcontracted much of the development and design work to PortalPlayer, who devoted all of their resources to the project. Steve Jobs himself was intimately involved with the design and function of the new player. The iPod was a huge success and has been redesigned over several models and generations. The most current reports record worldwide sales of over 350 million units.

While the iPod was a huge success, Apple did not rest on its laurels. In 2007, Apple launched the first iPhone. Again, the concept of a smartphone was not new. Several companies, notably Motorola, Samsung, Palm, and Nokia, had smartphones on the market for several years prior. But as with the iPod, Apple again produced an innovative product with unique features. Apple continues to improve upon the iPhone and introduces a new generation of the product virtually every year. As of this writing, sales have reached over 500 million worldwide.

Apple's most recent product, the iPad, was also an instant success and essentially defined a new market category. This tablet computer, introduced in March 2010, is convenient for web surfing and reading e-books, and again, has become the product the competition measures itself against. Apple registered more than 1 million sales of the iPad in the first three months alone. As a testament to Apple's phenomenal success in portable computing, the market capitalization of Apple surpassed that of the software behemoth Microsoft in 2010.

Source: Various websites and L. Kahney "Inside Look at the Birth of the iPod" July 2004 (http:/www.wired.com).

Once having articulated a vision, the next step is to plan a strategy for achieving that vision. This is the firm's business strategy. The overall business strategy includes defining

1. The market in which the enterprise competes.
2. The level of investment.
3. The means of allocating resources to and the integrating of separate business units.
4. Functional area strategies, including
    - The marketing strategy
    - The financial strategy
    - The operations strategy

Broadly defined, the operations strategy is the means by which the firm deploys its resources to achieve its competitive goals. For manufacturing firms, it is the sum total of all decisions concerning the production, storage, and distribution of goods. Important operations strategy decisions include where to locate new manufacturing facilities, how large these facilities should be, what processes to use for manufacturing

and moving goods through the system, and what workers to employ. Service firms also require an operations strategy. The United States continues to be a leader in financial services, which must be supported by effective and reliable operations departments in order to remain competitive. The Disney theme park's continuing record of success is due in part to its careful attention to detail in every phase of its operations.

Does the American culture place too much emphasis on marketing (selling the product) and finance (leveraged buyouts, mergers, stock prices) and too little on operations (making and delivering the product)? Years ago, this was certainly the case. However, we are quick learners. The enormous success of the Japanese auto industry, for example, provided strong motivation for the American big three to close their inefficient plants and change the way things were done. The dramatic differences that were brought to light by Womack, Jones, and Roos (1990) have largely been eliminated. Today, the best American auto plants rival their Japanese counterparts for quality and efficiency.

Still, a coherent operations strategy is essential. When the Apple Macintosh was introduced, the product was extremely successful. However, the company was plagued with backorders and failed to keep up with consumer demand. According to Debbi Coleman, Apple's former director of worldwide manufacturing:

> Manufacturing lacked an overall strategy which created problems that took nine months to solve . . . we had extremely poor forecasting. Incoming materials weren't inspected for defects and we didn't have a mechanism for telling suppliers what was wrong, except angry phone calls. Forty percent of Mac materials were coming from overseas and no one from Apple was inspecting them before they were shipped. . . . One of the biggest tasks that high-tech manufacturers face is designing a manufacturing strategy that allows a company to be flexible so it can ride with the highs and lows of consumer and business buying cycles. (Fallon, 1985)

Although it is easy to be critical of American management style, we must be aware of the factors motivating American managers and those motivating managers from other cultures. For example, the Japanese have not achieved their dramatic successes without cost. Sixteen-hour work days and a high rate of nervous breakdowns among management are common in Japan.

Measuring a firm's success by the performance of its share price can result in shortsighted management practices. Boards of directors are more concerned with the next quarterly report than with funding major long-term projects. In fact, Hayes and Wheelwright (1984) make a compelling argument that such factors led to a myopic management style in the United States, characterized by the following:

1. Managers' performance is measured on the basis of **return on investment** (ROI), which is simply the ratio of the profit realized by a particular operation or project over the investment made in that operation or project.
2. Performance is measured over short time horizons. There is little motivation for a manager to invest in a project that is not likely to bear fruit until after he or she has moved on to another position.

In order to improve ROI, a manager must either increase the numerator (profits) or decrease the denominator (investment). In the short term, decreasing the denominator by cutting back on the investment in new technologies or new facilities is easier than trying to increase profits by improving efficiency, the quality of the product, or the productivity of the operating unit. The long-term effects of decreasing investment are devastating. At some point, the capital costs required to modernize old factories become more than the firm can bear, and the firm loses its competitive position in the marketplace.

It would be encouraging if the problems of U.S. industries arising from overemphasis on short-term financial performance were decreasing, but sadly, they appear to be worsening. Because of gross mismanagement and questionable auditing practices, two giants of American industry were brought down in 2001: Enron and Arthur Andersen. "Enron went from the No. 7 company on the Fortune 500 to a penny stock in a stunning three weeks because it apparently lied on its financial statements," said Representative John D. Dingell, one-time member of the House Energy Committee. While other parts of the world have experienced spectacular problems as well (such as the Asian financial crisis that hit in the late 1990s), few Americans can understand how a company that had recently expanded and profited from the energy crisis, and an American icon such as Arthur Andersen, could both be brought down so quickly and completely. It is our continual focus on short-term performance and the incentive system we have built up around this objective that led to these crises.

Measuring individual performance over the short term is a philosophy that seems to pervade American life. Politicians are elected for two-, four-, or six-year terms. There is a strong incentive for them to show results in time for the next election. Even university professors are evaluated yearly on their professional performance in many institutions, even though most serious academic projects extend over many years.

1.1 MANUFACTURING MATTERS

A question that is being debated and has been debated by economists for several decades is the importance of a strong manufacturing base. The decline of manufacturing domestically has led to a shift in jobs from the manufacturing sector to the service sector. Because there are major disparities in labor costs in different parts of the world, there are strong incentives for American firms to locate volume manufacturing facilities overseas to reduce labor costs. Is a strong manufacturing base important for the health of the economy?

There is little debate that manufacturing jobs have been steadily declining in the United States. The growth of manufacturing overseas, and in China in particular, is well documented. If we compare the proportion of nonagriculture jobs in the United States in service versus manufacturing in 1950 versus 2002, the change is quite dramatic. In 1950, manufacturing jobs accounted for 34 percent of nonagriculture labor and service jobs accounted for 59 percent. In 2002, however, manufacturing jobs only accounted for 13 percent of nonagriculture jobs, while service jobs soared to 82 percent of the total (Hagenbaugh, 2002).

One mitigating factor in the loss of manufacturing was the dramatic rise in manufacturing productivity during this same period. Average annual manufacturing productivity growth was 2.57 percent annually during the 1980s and 3.51 percent annually during the 1990s (Faux, 2003). This dramatic rise in manufacturing productivity has had the effect of offsetting the loss of high-paying manufacturing jobs at home, thus partially accounting for the success of the U.S. economy in the latter part of the first decade of the century.

An argument put forth by several scholars (e.g., Daniel Bell, 1976) is that we are simply evolving from an industrial to a service economy. In this view, the three stages of economic evolution are (1) agrarian, (2) industrial, and (3) service. In the early years of our country, we were primarily an agrarian economy. With the industrial revolution, a large portion of the labor force shifted from agriculture to manufacturing. In recent years it seems that there is less interest in manufacturing. These scholars would argue

that we are merely entering the third stage of the evolutionary process: moving from an industrial economy to a service economy.

It is comforting to think that the American economy is healthy and simply evolving from an industrial to a service economy. One might even argue that manufacturing is not important for economic well-being. According to economist Gary S. Becker (1986), "Strong modern economies do not seem to require a dominant manufacturing sector."

It is far from clear, however, that we evolved from an agrarian economy to an industrial economy. Although fewer American workers are employed in the agricultural sector of the economy, agricultural production has not declined. Based on U.S. Department of Commerce data, Cohen and Zysman (1987) state that "agriculture has sustained, over the long term, the highest rate of productivity increase of any sector." By utilizing new technologies, agriculture has been able to sustain growth while consuming fewer labor hours. Hence, the figures simply do not bear out the argument that our economy has shifted from an agricultural one to an industrial one.

The argument that the economy is undergoing natural stages of evolution is simply not borne out by the facts. I believe that all sectors of the economy—agricultural, manufacturing, and service—are important and that domestic economic well-being depends upon properly linking the activities of these sectors.

The return on innovations will be lost if new products are abandoned after development. The payoff for research and development (R&D) can come only when the product is produced and sold. If manufacturing is taken offshore, then the "rent on innovation" cannot be recaptured. Furthermore, manufacturing naturally leads to innovation. It will be difficult for the United States to retain its position as a leader in innovation if it loses its position as a leader in manufacturing.

That manufacturing naturally leads to innovation is perhaps best illustrated by the Japanese experience in the video market. After Japan had captured the lion's share of the world market for televisions, the next major innovation in consumer video technology, the videocassette recorder (VCR), (at least, the inexpensive consumer version) was developed in Japan, not the United States. Virtually all VCRs sold were manufactured in Asia.

A more recent book by Pisano and Shih (2012) underscores many of the same themes that appeared in Cohen and Zysman (1987). They point to other products that were invented in the United States but whose base of manufacturing is now overseas. An example is the PV (photovoltaic) cell (more commonly known as the solar cell). PV cells were invented in Bell Labs, but only a very small percentage of the world demand is filled by American companies.

One of the more disturbing trends discussed by Pisano and Shih (2012) is the widening trade deficit in manufactured goods. The foreign trade deficit has continued to increase, resulting in the United States going from the largest creditor nation in the 1970s to the largest debtor nation today. The overwhelming source of this deficit is the continued negative trade balance in manufactured goods (the trade balance in services is actually increasing).

Where to locate manufacturing as well as R and D facilities is one of the key management decisions a firm must make. During the 1990s we saw an exodus of domestic manufacturing to China. The offshoring movement was rampant, not only in the United States but in most developed countries. The primary driver of this exodus is wage rates, but other factors were relevant as well. Favorable tax treatments, proximity to natural resources, and proximity to markets are also reasons companies locate facilities offshore.

In recent years, the advantage of offshoring is decreasing. For example, as the standard of living in China has improved, manufacturing wage rates have risen. When

the disadvantages of offshoring are taken into account, the best course of action is no longer obvious. These disadvantages include longer lead times, infrastructure deficiencies, local politics, and quality problems.

An example cited in Reshoring Manufacturing (2013) is the start-up company ET Water Systems. In 2005 the firm moved manufacturing operations to China in search of lower labor costs. However, the disadvantages of locating offshore became apparent as the company starting suffering losses due to several factors, including the cost of funds tied up in goods in transit, the disconnect between manufacturing and design, and recurring quality problems. When the firm's chief executive, Mark Coopersmith, carefully looked at the total cost difference between manufacturing in China versus California he was amazed to discover that California was only 10% more expensive than China. He concluded that this cost difference was more than offset by the advantages of locating manufacturing domestically. ET Water Systems closed their plant in China and reshored the manufacturing function to General Electronics Assembly in San Jose.

Unfortunately, ET's experience is rare. More companies are continuing to choose the offshoring option. However, that reshoring is occurring at all is a positive step. Perhaps more companies will come to the same conclusion that ET Water Systems did when taking into account the full spectrum of the costs of offshoring.

In order to get some idea of the extent of reshoring compared to offshoring, Porter and Rivkin (2012) conducted an extensive survey of Harvard Business School alumni who made location decisions for their companies in 2011. They found that only 9% of the respondents were considering moving offshore activities back to the United States, 34% were planning on keeping their facilities where they were, and 57% were considering moving existing facilities out of the United States. Their results suggest that offshoring still dominates both staying put and reshoring. The respondents' main reasons for offshoring were lower wages, proximity to customers, better access to skilled labor, higher labor productivity, lower taxes, proximity of suppliers, and proximity to other company operations. The respondents' main reasons for staying put or reshoring were proximity to the U.S. market, less corruption, greater safety, better intellectual property protection, similar language and culture, better infrastructure, and proximity to other company operations.

Has offshoring been a help or a hindrance to the U.S. economy? The answer is not simple. On one hand, offshoring has resulted in loss of jobs domestically, lower average domestic wages which in turn have yielded a lower tax base and a smaller domestic market. On the other hand, offshoring has improved the bottom line for many domestic firms, and have resulted in lower costs of manufactured goods for the American consumer.

Manufacturing Jobs Outlook

The U.S. Bureau of Labor Statistics (a subsidiary of the Department of Labor) provides up-to-date information on the prospects for jobs in the manufacturing sector by industry. According to the *Occupational Outlook Handbook* (OOH), 2010–2011 Edition (http://www.bls.gov/oco/), even though manufacturing jobs are expected to decline overall, there are some areas of growth and opportunity. Consider the individual sectors:

1. *Aerospace products and parts*. This sector is projected to grow, but more slowly than the economy in general. Earnings are higher here than in most other manufacturing industries, as workers must be highly skilled. Opportunities will result from a large number of anticipated retirements.
2. *Chemical (except pharmaceuticals and medicines)*. The chemical industry continues to be a major employer of professionals, producing over 500,000 jobs.

However, employment is projected to decline and competition for better jobs to increase over the coming years.

3. *Computer and Electronic Products*. Employment is projected to decrease nearly 20 percent in the decade 2008–2018 due to productivity improvements and movement of jobs to lower wage countries.
4. *Food Manufacturing*. The jobs picture in this industry is stable, but production workers continue to have the highest incidences of injury and illness among all industry, with seafood product preparation and packaging being the worst sector in this regard.
5. *Machinery*. Productivity improvements will lead to fewer jobs overall, but opportunities will arise as a result of anticipated retirements. Machinery manufacturing has some of the most highly skilled, and highly paid, production jobs in manufacturing.
6. *Motor Vehicles and Parts*. Almost half the jobs are located in Michigan, Indiana, and Ohio, but jobs continue to shift away from this area to the South. Average earnings continue to be high in this sector, but employment is expected to decline in coming years.
7. *Pharmaceuticals and Medicine*. This continues to be a growth area, with earnings higher than in other manufacturing industries. Job prospects are particularly favorable for candidates with advanced degrees.
8. *Printing*. Most printing establishments are very small, with 70 percent employing under 10 people. Traditional printing is a declining industry due to increased computerization, but digital press operators will continue to be in demand.
9. *Steel*. Steel continues to be a declining industry domestically, with fewer jobs projected as a result of consolidation and automation. Opportunities will be best for engineers and skilled production and maintenance workers.
10. *Textiles, Textile Products, and Apparel*. About half the jobs are located in three states: California, North Carolina, and Georgia. Employment is expected to decline rapidly because of technological advances and imports of apparel and textiles from lower wage countries.

1.2 A FRAMEWORK FOR OPERATIONS STRATEGY

Classical literature on competitiveness claims that firms position themselves strategically in the marketplace along one of two dimensions: lower cost or product differentiation (Porter, 1990).

Often new entrants to a market position themselves as the low-cost providers. Firms that have adopted this approach include the Korean automakers (Hyundai, Daewoo, Kia), discount outlets such as Costco, and retailers such as Wal-Mart. While being the low-cost provider can be successful over the near term, it is a risky strategy. Consumers ultimately will abandon products that they perceive as poor quality regardless of cost. For example, many manufacturers of low-cost PC clones popular in the 1980s are long gone.

Most firms that have a long record of success in the marketplace have differentiated themselves from their competitors. By providing uniqueness to buyers, they are able to sustain high profit margins over time. One example is BMW, one of the most profitable auto firms in the world. BMW continues to produce high-performance, well-made cars that are often substantially more expensive than those of competitors in their class. Product differentiation within a firm has also been a successful strategy. Consider the success of General Motors in the early years compared to Ford. GM was able to successfully capture different market segments at the same time by forming five distinct

divisions, while Henry Ford's insistence on providing only a single model almost led the company to bankruptcy (Womack et al., 1990).

Strategic Dimensions

However, cost and product differentiation are not the only two dimensions along which firms distinguish themselves. The following additional factors relate directly to the operations function:

- Quality
- Delivery speed
- Delivery reliability
- Flexibility

What does *quality* mean? It is a word often bandied about, but one that means different things in different contexts. Consider the following hypothetical remarks.

1. "That hairdryer was a real disappointment. It really didn't dry my hair as well as I expected."
2. "I was thrilled with my last car. I sold it with 150,000 miles and hardly had any repairs."
3. "I love buying from that catalogue. I always get what I order within two days."
4. "The refrigerator works fine, but I think the shelves could have been laid out better."
5. "That park had great rides, but the lines were a mess."
6. "Our quality is great. We've got less than six defectives per one million parts produced."

In each case, the speaker is referring to a different aspect of quality. In the first case, the product simply didn't perform the task it was designed to do. That is, its function was substandard. The repair record of an automobile is really an issue of reliability rather than quality, per se. In the third case, it is delivery speed that translates to quality service for that customer. The fourth case refers to a product that does what it is supposed to do, but the consumer is disappointed with the product design. The product quality (the rides) at the amusement park were fine, but the logistics of the park management were a disappointment. The final case refers to the statistical aspects of quality control.

Hence the word *quality* means different things in different contexts. A Honda Civic is a quality product and so is a Ferrari Testarosa. Consumers buying these products are both looking for quality cars but have fundamentally different objectives. The fact is that everyone competes on quality. For this reason, Terry Hill (1993) would classify quality as an order qualifier rather than an order winner. An option is immediately eliminated from consideration if it does not meet minimum quality standards. It is the particular aspect of quality on which one chooses to focus that determines the nature of the competitive strategy and the positioning of the firm.

Delivery speed can be an important competitive weapon in some contexts. Some firms base their primary competitive position on delivery speed, such as UPS and Federal Express. Mail-order and Web-based retailers also must be able to deliver products reliably and quickly to remain competitive. Building contractors that complete projects on time will have an edge.

Delivery reliability means being able to deliver products or services when promised. Online brokerages that execute trades reliably and quickly will retain customers. Contract manufacturers are measured on several dimensions, one being whether they can deliver on time. As third-party sourcing of manufacturing continues to grow, the successful contract manufacturers will be the ones that put customers first and maintain a record of delivering high-quality products in a reliable fashion.

Flexibility means offering a wide range of products and being able to adjust to unexpected changes in the demand of the product mix offered. Successful manufacturers in the 21st century will be those that can respond the fastest to unpredictable changes in customer tastes. This writer was fortunate enough to tour Toyota's Motomachi Plant located in Toyoda City, Japan. What was particularly impressive was the ability to produce several different models in the same plant. In fact, each successive car on the assembly line was a different model. A right-hand drive Crown sedan, for the domestic market, was followed by a left-hand drive Lexus coupe, designated for shipment to the United States. Each car carried unique sets of instructions that could be read by both robot welders and human assemblers. This flexibility allowed Toyota to adjust the product mix on a real-time basis and to embark on a system in which customers could order custom-configured cars directly from terminals located in dealer showrooms (Port, 1999).

Hence, one way to think of operations strategy is the strategic positioning the firm chooses along one of the dimensions of cost, quality, delivery speed, delivery reliability, and flexibility. Operations management is concerned with implementing the strategy to achieve leadership along one of these dimensions.

1.3 COMPETING IN THE GLOBAL MARKETPLACE

International competitiveness has become a national obsession. Americans are concerned that their standard of living is eroding while it seems to improve elsewhere. Evidence exists that there is some truth to this perception. Our balance of trade with Japan has been in the red for decades, with no evidence of a reversal. American firms once held a dominant position worldwide in industries that have nearly disappeared domestically. Consumer electronics, steel, and machine tools are some examples. All the news is not bad, however. The American economy is strong and continues to grow. American firms still have the lion's share of the world market in many industries.

In his excellent study of international competitiveness, Porter (1990) poses the following question: Why does one country become the home base for successful international competitors in an industry? That certain industries flourish in certain countries cannot be disputed. Some examples are

1. Germany: printing presses, luxury cars, chemicals.
2. Switzerland: pharmaceuticals, chocolate.
3. Sweden: heavy trucks, mining equipment.
4. United States: personal computers, software, films.
5. Japan: automobiles, consumer electronics, robotics.

What accounts for this phenomenon? One can offer several compelling explanations, but most have counterexamples. Here are a few:

1. *Historical*. Some industries are historically strong in some countries and are not easily displaced. *Counterexample*: The demise of the steel industry in the United States is one of many counterexamples.
2. *Tax structure*. Some countries, such as Germany, have no capital gains tax, thus providing a more fertile environment for industry. *Counterexample*: However, there is no reason that favorable tax treatment should favor certain industries over others.
3. *National character*. Many believe that workers from other countries, particularly from Pacific Rim countries, are better trained and more dedicated than American workers. *Counterexample*: If this is true, why then do American firms dominate in some industry

segments? How does one explain the enormous success Japanese-based corporations have had running plants in the United States with an American workforce?

4. *Natural resources*. There is no question that some industries are highly resource dependent and these industries have a distinct advantage in some countries. One example is the forest products industry in the United States and Canada. *Counterexample*: Many industry sectors are essentially resource independent but still seem to flourish in certain countries.
5. *Government policies*. Some governments provide direct assistance to fledgling industries, such as MITI in Japan. The role of the U.S. government is primarily regulatory. For example, environmental standards in the United States are probably more stringent than almost anywhere else. *Counterexample*: This does not explain why some industries dominate in countries with strict environmental and regulatory standards.
6. *Advantageous macroeconomic factors*. Exchange rates, interest rates, and government debt are some of the macroeconomic factors that provide nations with competitive advantage. For example, in the 1980s when interest rates were much higher in the United States than they were in Japan, it was much easier for Japanese firms to borrow for new projects. *Counterexample*: These factors do not explain why many nations have a rising standard of living despite rising deficits (Japan, Italy, and Korea are some examples).
7. *Cheap, abundant labor*. Although cheap labor can attract new industry, most countries with cheap labor are very poor. On the other hand, many countries (Germany, Switzerland, and Sweden are examples) have a high standard of living, high wage rates, and shortages of qualified labor.
8. *Management practices*. There is evidence that Japanese management practices are more effective in general than Western-style practices. Counterexample: If American management practices are so ineffective, why do we continue to dominate certain industries, such as personal computers, software development, and pharmaceuticals?

Talking about competitiveness is easier than measuring it. What are the appropriate ways to measure one country's success over another? Some possibilities are

- Balance of trade.
- Share of world exports.
- Creation of jobs.
- Low labor costs.

Arguments can be made against every one of these as a measure of international competitiveness. Switzerland and Italy have trade deficits, and at the same time have experienced rising standards of living. Similar arguments can be made for countries that import more than they export. The number of jobs created by an economy is a poor gauge of the health of that economy. More important is the quality of the jobs created. Finally, low labor costs correlate with a low standard of living. These counterexamples show that it is no easy task to develop an effective measure of international competitiveness.

Porter (1990) argues that the appropriate measure to compare national performance is the rate of productivity growth. Productivity is the value of output per unit of input of labor or capital. Porter argues that productivity growth in some industries appears to be stronger in certain countries, and that there are reasons for this. In some cases we can find the reasons in domestic factor advantages. The factor theory says all countries have access to the same technology (an assumption that is not strictly true) and that national advantages accrue from endowments of production factors such as land, labor, natural resources, and capital.

There are some excellent examples of factor theory. Korea has relatively low labor costs, so it exports labor-intensive goods such as apparel and electronic assemblies. Sweden's iron ore is low in impurities, which contributes to a strong Swedish steel industry. As compelling as it is, there are counterexamples to the factor endowment theory as well. For example, after the Korean War, South Korea developed and excelled in several highly capital-intensive industries such as steel and shipbuilding even though the country was cash poor. Also, many countries have similar factor endowments, but some seem to excel in certain industries, nonetheless. These examples suggest that factor endowments do not explain all cases of nations with dominant industry segments.

Porter (1990) suggests the following four determinants of national advantage:

1. Factor conditions (previously discussed).
2. Demand conditions. If domestic consumers are sophisticated and demanding, they apply pressure on local industry to innovate faster, which gives firms an edge internationally. Consumers of electronics in Japan are very demanding, thus positioning this industry competitively in the international marketplace.
3. Related and supporting industries. Having world-class suppliers nearby is a strong advantage. For example, the Italian footwear industry is supported by a strong leather industry and a strong design industry.
4. Firm strategy, structure, and rivalry. The manner in which firms are organized and managed contributes to their international competitiveness. Japanese management style is distinctly different from American. In Germany, many senior executives possess a technical background, producing a strong inclination to product and process improvement. In Italy, there are many small family-owned companies, which encourages individualism.

Even though Porter makes a very convincing argument for national competitive advantage in some industries, there is a debate among economists as to whether the notion of international competitiveness makes any sense at all. Companies compete, not countries. This is the point of view taken by Paul Krugman (1994). According to Krugman, the United States and Japan are simply not competitors in the same way that Ford and Toyota are. The standard of living in a country depends on its own domestic economic performance and not on how it performs relative to other countries.

Krugman argues that too much emphasis on international competitiveness can lead to misguided strategies. Trade wars are much more likely in this case. This was the case in mid-1995 when the Clinton administration was planning to impose high tariffs on makers of Japanese luxury cars. Most economists agree that trade wars and their consequences, such as tariffs, benefit no one in the long run. Another problem that arises from national competitive pride is it can lead to poorly conceived government expenditures. France has spent billions propping up its failing computer industry. (Certainly, not all government investments in domestic industry can be considered a mistake. The Japanese government, for example, played a major role in nurturing the flat-panel display industry. Japanese-based firms now dominate this multibillion dollar industry.)

Another point supporting Krugman's position is that the lion's share of U.S. gross domestic product (GDP) is consumed in the United States, thus making a firm's success in our domestic market more important than its success in the world market. Krugman agrees that productivity growth is a valid concern. He argues, however, that we should be more productive in order to produce more, not to better our international competitors.

The debate over competitive advantage will continue. Policy makers need to be aware of all points of view and weigh each carefully in formulating policy. Although Krugman makes several telling points, there is no question that globalization is a trend

Snapshot Application

GLOBAL MANUFACTURING STRATEGIES IN THE AUTOMOBILE INDUSTRY

Consider the following four foreign automobile manufacturers: Honda, Toyota, BMW, and Mercedes Benz. As everyone knows, Honda and Toyota are Japanese companies and BMW and Mercedes are German companies. The four account for the lion's share of foreign nameplates sold in the U.S. auto market. However, many assume that these cars are manufactured in their home countries. In fact, depending on the model, it could be more likely that a consumer buying a Honda, Toyota, BMW, or Mercedes is buying a car manufactured in the United States.

Honda was the first of the foreign automakers to commit to a significant investment in U.S.-based manufacturing facilities. Honda's first U.S. facility was its Marysville Motorcycle plant, which started production in 1979. Honda must have been pleased with the Ohio-based facility, since an automobile plant followed shortly. Automobile production in Marysville began in 1982. Today, Honda operates four plants in west-central Ohio, producing the Accord sedan and coupe, the Acura TL sedan, the Honda Civic line, and the Honda Element, with the capacity to produce a whopping 440,000 vehicles annually.

Next to make a significant commitment in U.S. production facilities was Toyota. Toyota's plant in Georgetown, Kentucky, has been producing automobiles since 1986 and accounts for all of the Camrys sold in the domestic market. It is interesting to note that the Honda Accord and the Toyota Camry are two of the biggest-selling models in the United States, and are also produced here. They also top almost all reliability surveys.

The two German automakers were slower to commit to U.S.-based manufacturing facilities. BMW launched its Spartanburg, South Carolina, plant in March of 1995. BMW produces both the Z series sports cars and its SUV line in this plant. It is interesting to note that BMW's big sellers, its 3, 5, and 7 series sedans, are still manufactured in Germany.

Mercedes was the last of these four to make a significant commitment to production facilities here. The facility in Tuscaloosa, Alabama, is dedicated to producing the line of Mercedes SUVs. As with BMW, the more popular C, E, and S class sedans are still manufactured in Germany.

(One might ask why Volkswagen is not on this list. In fact, Volkswagen has 45 separate manufacturing facilities located in 18 countries around the world, but no significant manufacturing presence in the mainland United States.)

Sources: Honda's Web site (http://www.ohio.honda.com/), Toyota's Web site (http://www.toyota.com), Autointell's Web site (http://www.autointell-news.com/european_companies/BMW/bmw3.htm), Mercedes Benz's Web site (http://www.mbusi.com/).

that shows no sign of reversing. We cannot stick our heads in the sand and say that foreign markets are not important to us. Economic borders are coming down all across the globe.

Problems for Sections 1.1–1.3

1. Why is it undesirable for the United States to evolve into a service economy?
2. What disadvantages do you see if the chief executive officer (CEO) is primarily concerned with short-term ROI?
3. Can you think of companies that have gone out of business because they focused only on cost and were not able to achieve a minimum quality standard?
4. What are the different quality standards referred to in the example comparing the Honda Civic and the Ferrari?
5. Discuss the pros and cons of trade barriers from the industry point of view and from the consumer point of view.
6. What are the advantages and disadvantages of producing new products in existing facilities?
7. What are the four determinants of national advantage suggested by Porter? Give examples of companies that have thrived as a result of each of these factors.
8. What factor advantage favors the aluminum industry in the United States over Japan and makes aluminum much cheaper to produce here? (Hint: Aluminum

production is very energy intensive. In what part of the country is an inexpensive energy source available?)

9. Paul Krugman argues that because most of our domestic product is consumed domestically, we should not dwell on international competition. What industries in the United States have been hardest hit by foreign competition? What are the potential threats to the United States if these industries fail altogether?
10. Krugman points out some misguided government programs that have resulted from too much emphasis on international competitiveness. What risks are there from too little emphasis on international competitiveness?
11. Consider the Snapshot Application in this section concerning foreign automakers locating manufacturing facilities in the United States. Discuss the advantages and disadvantages of the strategy of locating manufacturing facilities where the product is consumed rather than where the company is located.
12. The North American Free Trade Agreement (NAFTA) was established in 1994 under the Clinton administration.
    a) What was the purpose of NAFTA?
    b) At the time, political opponents characterized "the big sucking sound" as jobs would be lost as a result. Is there any evidence that this was, in fact, the case?

1.4 STRATEGIC INITIATIVES: REENGINEERING THE BUSINESS PROCESS

Seemingly on schedule, every few years a hot new production control method or management technique comes along, almost always described by a three-letter acronym. While it is easy to be skeptical, by and large, the methods are sound and can have substantial value to corporations when implemented intelligently. *Business process reengineering* (BPR) caught on after the publication of the book by Hammer and Champy (1993). BPR is not a specific technique, such as materials requirements planning or a production-planning concept like just-in-time. Rather, it is the idea that entrenched business processes can be changed and can be improved. The process is one of questioning why things are done a certain way, and not accepting the answer, "because that's the way we do it."

Hammer and Champy, who define BPR as "starting over," provide several examples of successful reengineering efforts. The first is the IBM Credit Corporation, a wholly owned subsidiary of IBM, that, if independent, would rank among the *Fortune* 100 service companies. This arm of IBM is responsible for advancing credit to new customers purchasing IBM equipment. The traditional credit approval process followed five steps:

1. An IBM salesperson would call in with a request for financing and the request would be logged on a piece of paper.
2. Someone carried the paper upstairs to the credit department, where someone else would enter the information into a computer system and check the potential borrower's credit rating. The specialist wrote the results of the credit check on a piece of paper, which was sent off to the business practices department.
3. A third person, in the business practices department, modified the standard loan document in response to the customer request. These modifications, which were done on yet another computer system, were then attached to the original request form and the credit department specialist's report.

4. Next the request went to a pricer, who keyed the information into a spreadsheet to determine the appropriate interest rate to charge the customer. The pricer's recommendation was written on a piece of paper and delivered (with the other papers) to the clerical group.
5. The information was turned into a quote letter that would be delivered to the field salesperson by Federal Express.

This process required an average of six days and sometimes as long as two weeks. Sales reps logged endless complaints about this delay: during this time, the customer could find other financing or another vendor. In an effort to see if this process could be streamlined, two senior managers decided to walk a new request through all five steps, asking personnel to put aside what they were doing and process it as they normally would. They found that the entire five-step process required an average of only 90 minutes of work! The rest of the time, requests were either in transit from one department to another or queueing up on somebody's desk waiting to be processed. Clearly the problem did not lie with the efficiency of the personnel but with the design of the credit approval process itself.

The solution was simple: the four specialists handling each loan request were replaced by a single loan generalist who handled each request from beginning to end. Up-to-date software was designed to support the generalist, who had no trouble dealing with most requests. The credit approval process was designed assuming that each request was sufficiently complex to require someone with special knowledge in each area. In truth, most requests were routine, and specialists generally did little more than a simple table lookup to determine the appropriate figure.

What was the result of this change? The six-day turnaround for loan requests was slashed to only four hours! And this was accomplished with fewer personnel and with a hundredfold increase in the number of deals handled.

While each reengineering effort requires careful thought and no two solutions will be exactly alike, Hammer and Champy (1993) suggest that reengineering efforts utilize the following general principles:

1. *Several jobs are combined into one.* Few examples of BPR are as dramatic as that of IBM Credit, but there are other success stories in the literature as well. Many of the successful cases have a common thread: the reduction of a complex process requiring many steps to a simpler one requiring fewer steps. In the case of IBM Credit, a five-step process was reduced to only a single step. This suggests a general principle. The IBM Credit process was a natural evolution of the concept of division of labor. The economist Adam Smith espoused this principle as far back as the 18th century (see the quote from *The Wealth of Nations* at the beginning of Section 1.10 of this chapter). However, a good thing can be carried too far. If one divides a process into too many steps, one eventually reaches the point of diminishing returns. BPR's most dramatic successes have come from complex processes that were simplified by reducing the number of steps required.

2. *Workers make decisions.* One goal is to reduce the number of levels of reporting by allowing workers to make decisions that were previously reserved for management. In the case of IBM Credit, most decisions once reserved for specialists are now done by a single generalist. Giving workers greater decision-making power may pose a threat to management, who might see such a step as encroaching on their prerogatives.

3. *The steps in the process are performed in a natural order*. Process steps should not be performed necessarily in rigid linear sequence, but in an order that makes sense in the context of the problem being solved. In particular, in many cases, some tasks can be

done simultaneously rather than in sequence. (These ideas, of course, are well known and form the basis for the concepts of project management in Chapter 10.)

4. *Processes should have multiple versions.* One should allow for contingencies, not by designing multiple independent processes, but by designing one flexible process that can react to different circumstances. In the case of IBM Credit, for example, the final credit issuance process had three versions: one for straightforward cases (handled by computer), one for cases of medium difficulty (handled by the deal structurer), and one for difficult cases (performed by the deal structurer with help from specialist advisers).

5. *Work is performed where it makes the most sense.* One of the basic principles of reengineering is not to carry the idea of division of labor too far. Another is not to carry the idea of centralization too far. For example, in most companies, purchasing is done centrally. This means that every purchase request is subject to the same minimum overhead in time and paperwork. A consequence might be that the cost of processing a request exceeds the cost of the item being purchased! A great deal can be saved in this case by allowing individual departments to handle their own purchasing for low-cost items. (Hammer and Champy discuss such a case.)

The authors list several other basic principles, involving minimizing checks and reconciliations, having a single point of contact, and being able to employ hybrid centralized/decentralized operations.

It is easier to list the steps one might consider in a reengineering effort than to actually implement one. In the real world, political realities cannot be ignored. For many of the success stories in the literature, not only are the processes simplified, but the headcount of personnel is reduced as well. It is certainly understandable for employees to see BPR as a thinly veiled excuse for downsizing (euphemistically called "right-sizing"). This was exactly the case in one financial services company. When word got out that management was planning a reengineering effort, most assumed that there would be major layoffs. Some even thought the company was on the verge of bankruptcy. In another instance, union leadership saw reengineering as a means for management to throw away the job categories and work rules they had won in hard-fought negotiations over the years, and persuaded the members to strike. In a third case, a senior manager was unhappy with the potential loss of authority that might accompany a reengineering effort. He resigned to start his own company. (These examples are related in a follow-up book by Hammer and Stanton, 1995.)

These stories show that starting a reengineering effort is not without risks. Rarely is the process as simple as IBM's. Reengineering has been described by Ronald Compton, the CEO of Aetna Life and Casualty, as "agonizingly, heartbreakingly, tough." There needs to be some cost–benefit analysis done up front to be sure the potential gains compensate for the risks.

Process optimization is not new. In its early years, the field of industrial engineering dealt with optimal design of processes, setting standards using time and motion studies, and flowcharting for understanding the sequence of events and flow of material in a factory. Why is BPR different? For one, BPR is concerned with business process flows rather than manufacturing process flows. Second, the concept is not one of optimizing an existing process, but one of rethinking how things should be done from scratch. As such, it is more revolutionary than evolutionary. It is likely to be more disruptive but could have larger payoffs. To make BPR work, employees at every level have to buy into the approach, and top management must champion it. Otherwise, the reengineering effort could be a costly failure.

1.5 STRATEGIC INITIATIVES: JUST-IN-TIME

Just-in-time (JIT) is a manufacturing process on one hand and a broad-based operations strategy on the other. The process elements of JIT will be discussed in detail in Chapter 8 as part of a complete analysis of push and pull inventory systems. However, JIT (or lean production, as it is also known) is a philosophy that includes treatment of inventory in the plant, relationships with suppliers, and distribution strategies. The core of the philosophy is to eliminate waste. This is accomplished by efficient scheduling of incoming orders, work-in-process inventories, and finished goods inventories.

JIT is an outgrowth of the **kanban system** introduced by Toyota. Kanban is a Japanese word meaning card or ticket. Originally, kanban cards were the only means of implementing JIT. The kanban system was introduced by Toyota to reduce excess work-in-process (WIP) inventories. Today, JIT is more ambitious. Both quality control systems and relationships with suppliers are part of an integrated JIT system. JIT systems can be implemented in ways other than using kanban cards. Integrating JIT philosophies with sophisticated information systems makes information transfer faster. The speed with which information can be transferred from one part of the firm to another is an important factor in the success of the JIT system.

JIT is a philosophy of operating a company that includes establishing understandings and working relationships with suppliers, providing for careful monitoring of quality and work flow, and ensuring that products are produced only as they are needed. Although JIT can be used simply as it was originally designed by Toyota, namely as a means of moving work-in-process (WIP) from one work center to another, proponents of the method recommend much more. They would have a firm integrate the JIT philosophy into its overall business strategy.

Inventory and material flow systems are classified as either **push** or **pull systems.** A push system is one in which decisions concerning how material will flow through the system are made centrally. Based on these decisions, material is produced and "pushed" to the next level of the system. A typical push system is materials requirements planning (MRP), which is discussed in detail in Chapter 8. In MRP, appropriate production amounts for all levels of the production hierarchy are computed all at once based on forecasts of end-product demand and the relationship between components and end items. In JIT, production is initiated at one level as a result of a request from a higher level. Units are then "pulled" through the system.

JIT has many advantages over conventional systems. Eliminating WIP inventories results in reduced holding costs. Less inventory means less money tied up in inventory. JIT also allows quick detection of quality problems. Since units are produced only as they are needed, the situation in which large amounts of defective WIP inventory are produced before a quality problem is detected should never occur in a properly running JIT system. JIT also means that relationships with suppliers must be tightened up. Suppliers must be willing to absorb some uncertainty and adjust delivery quantities and the timing of deliveries to match the rates of product flows.

Part of what made the kanban system so effective for Toyota was its success in reducing setup times for critical operations. The most dramatic example of setup time reduction is the so-called SMED, or single-minute exchange of dies. Each time a major change in body style is initiated, it is necessary to change the dies used in the process.

The die-changing operation typically took from four to six hours. During the die-changing operation the production line was closed down. Toyota management heard that Mercedes Benz was able to reduce its die-changing operation to less than one hour. Realizing that even more dramatic reductions were possible, Toyota set about focusing on the reduction of the time required for die changing. In a series of dramatic improvements, Toyota eventually reduced this critical operation to only several minutes. The essential idea behind SMED is to make as many changes as possible off-line, while the production process continues.

An important part of JIT is forming relationships with suppliers. What separates JIT purchasing from conventional purchasing practices? Freeland (1991) gives a list of characteristics contrasting the conventional and JIT purchasing behavior. Some of these include

Conventional Purchasing	JIT Purchasing
1. Large, infrequent deliveries.	1. Small, frequent deliveries.
2. Multiple suppliers for each part.	2. Few suppliers; single sourcing.
3. Short-term purchasing agreements.	3. Long-term agreements.
4. Minimal exchange of information.	4. Frequent information exchange.
5. Prices established by suppliers.	5. Prices negotiated.
6. Geographical proximity unimportant.	6. Geographical proximity important.

In his study, Freeland notes that the industries that seemed to benefit most from JIT purchasing were those that typically had large inventories. Companies without JIT purchasing tended to be more job-shop oriented or make-to-order oriented. Vendors that entered into JIT purchasing agreements tended to carry more safety stock, suggesting manufacturers are reducing inventories at the expense of the vendors. The JIT deliveries were somewhat more frequent, but the differences were not as large as one might expect. Geographical separation of vendors and purchasers was a serious impediment to successful implementation of JIT purchasing. The automotive industry was one that reported substantial benefit from JIT purchasing arrangements. In other industries, such as computers, the responses were mixed; some companies reported substantial benefits and some reported few benefits.

Although reducing excess work-in-process inventory can have many benefits, JIT is not necessarily the answer for all manufacturing situations. According to Stasey and McNair (1990),

> Inventory in a typical plant is like insurance, insurance that a problem in one area of a plant won't affect work performed in another. When problems creating the need for insurance are solved, then inventories disappear from the plant floor.

The implication is that we merely eliminate all sources of uncertainty in the plant and the need for inventories disappears. The problem is that there are some sources of variation that can never be eliminated. One is variation in consumer demand. JIT is effective only if final demand is regular. Another may be sources of variation inherent in the production process or in the equipment. Can one simply legislate away all sources of uncertainty in the manufacturing environment? Of course not. Hence, although the underlying principles of JIT are sound, it is not a cure-all and will not necessarily be the right method for every production situation.

1.6 STRATEGIC INITIATIVES: TIME-BASED COMPETITION

Professor Terry Hill of the London School of Business has proposed an interesting way to look at competitive factors. He classifies them into two types: "qualifiers" and "order winners." A product not possessing a qualifying factor is eliminated from consideration. The order winner is the factor that determines who gets the sale among the field of qualifiers.

Two factors about which we hear a great deal are quality and **time to market.** In the past decade, the Japanese and Germans gained a loyal following among U.S. consumers by producing quality products. American firms are catching up on the quality dimension. From the discussion in Section 1.6, we see that successful U.S.–based companies have been able to produce products that match the defect rates of foreign competitors. If this trend continues, product quality will be assumed by the consumer. Quality may become an order qualifier rather than an order winner.

If that is the case, what factors will determine order winners in years to come? Japanese automobile companies provided and continue to provide high-quality automobiles. In recent years, however, the major automobile producers in Japan have begun to focus on aesthetics and consumer tastes. They have branched out from the stolid small cars of the 1970s and 1980s to new markets with cars such as the Toyota-made Lexus luxury line and Mazda's innovative and successful Miata.

The timely introduction of new features and innovative design will determine the order winners in the automobile industry. In the computer industry, Compaq built its reputation partly on its ability to be the first to market with new technology. Time-based competition is a term that we will hear more and more frequently in coming years.

What is **time-based competition**? It is not the time and motion studies popular in the 1930s that formed the basis of the industrial engineering discipline. Rather, according to Blackburn (1991),

> Time-based competitors focus on the bigger picture, on the entire value-delivery system. They attempt to transform an entire organization into one focused on the total time required to deliver a product or service. Their goal is not to devise the best way to perform a task, but to either eliminate the task altogether or perform it in parallel with other tasks so that over-all system response time is reduced. Becoming a time-based competitor requires making revolutionary changes in the ways that processes are organized.

Successful retailers understand time-based competition. The success of the fashion chains The Gap and The Limited is due largely to their ability to deliver the latest fashions to the customer in a timely manner. Part of the success of the enormously successful Wal-Mart chain is its time-management strategy. Each stock item in a Wal-Mart store is replenished twice a week, while the industry average is once every two weeks. This allows Wal-Mart to achieve better inventory turnover rates than its competition and respond more quickly to changes in customer demand. Wal-Mart's strategies have enabled it to become the industry leader, with a growth rate three times the industry average and profits two times the industry average (Blackburn, 1991, Chapter 3).

Time-based management is a more complex issue for manufacturers, and in some industries it is clearly the key factor leading to success or failure. The industry leaders in the dynamic random access memory (DRAM) industry changed four times between 1978 and 1987. In each case, the firm that was first to market with the next-generation

DRAM dominated that market. The DRAM experience is summarized in the following table (Davis, 1989):

Product	Firm	Year Introduced	First Year of Volume Production	Market Leaders in First Year of Volume Production
16 K	Mostek	1976	1978	Mostek (25%) NEC (20%)
64 K	Hitachi	1979	1982	Hitachi (19%) NEC (15%)
256 K	NEC	1982	1984	NEC (27%) Hitachi (24%)
1 MB	Toshiba	1985	1987	Toshiba (47%) Mitsubishi (16%)

I am aware of no other example that shows so clearly and so predictably the value of getting to the market first.

1.7 STRATEGIC INITIATIVES: COMPETING ON QUALITY

What competitive factors do American managers believe will be important in the next decade? Based on a survey of 217 industry participants, the following factors were deemed as the most important for gaining a competitive edge in the coming years; they are listed in the order of importance.

1. Conformance quality
2. On-time delivery performance
3. Quality
4. Product flexibility
5. After-sale service
6. Price
7. Broad line (features)
8. Distribution
9. Volume flexibility
10. Promotion

In this list we see some important themes. **Quality** and **time management** emerge as leading factors. Quality control was brought to public attention with the establishment of the prestigious Malcolm Baldrige Award (modeled after the Japanese Deming Prize, which has been around a lot longer). Quality means different things in different contexts, so it is important to understand how it is used in the context of manufactured goods. A high-quality product is one that performs as it was designed to perform. Products will perform as they are designed to perform if there is little variation in the manufacturing process. With this definition of quality, it is possible for a product with a poor design to be of high quality, just as it is possible for a well-designed product to be of poor quality. Even granting this somewhat narrow definition of quality, what is the best measure? Defect rates are a typical barometer. However, a more appropriate measure might be reliability of the product after manufacture. This measure is typically used to monitor quality of products such as automobiles and consumer electronics.

There has been an enormous groundswell of interest in the quality issue in the United States in recent years. With the onslaught of Japanese competition, many American

industries are fighting for their lives. The business of selling quality is at an all-time high. Consulting companies that specialize in providing quality programs to industry, such as the Juran Institute and Philip Crosby Associates, are doing a booming business. The question is whether American firms are merely paying lip service to quality or are seriously trying to change the way they do business. There is evidence that, in some cases at least, the latter is true.

For example, in a comparison of American and Japanese auto companies, quality as measured by defects reported in the first three months of ownership declined significantly from 1987 to 1990 for U.S. companies, narrowing the gap with Japan significantly. The Buick Division of General Motors, a winner of the Baldrige Award, has made dramatic improvements along these lines. Between 1987 and 1990 Buick decreased this defect rate by about 70 percent, equaling the rate for Hondas in 1990 (*Business Week,* October 22, 1990).

There are many success stories in U.S. manufacturing. Ford Motors achieved dramatic success with the Taurus. Ford improved both quality and innovation, providing buyers with reliable and technologically advanced cars. In 1980, James Harbour reported that Japanese automakers could produce a car for $1,500 less than their American counterparts. That gap has been narrowed by Ford to within a few hundred dollars. Part of Ford's success lies in former CEO Donald Petersen's decision not to invest billions in new plants incorporating the latest technology as GM did in the mid-1980s. This is only part of the story, however. According to Faye Wills (1990),

> If you are looking for surprise answers to Ford's ascendancy, for hidden secrets, forget it. Good solid everyday management has turned the trick—textbook planning and execution, common-sense plant layouts and procedures, intelligent designs that not only sell cars, but also cut costs and bolster profit margins. It's that simple.

We can learn from our successes. The machine tool industry was one in which the Japanese made dramatic inroads in the 1980s. Many American firms fell to the onslaught of Asian competition, but not the Stanley Works of New Britain, Connecticut. In 1982, the firm's president was considering whether Stanley should remain in the hardware business as Asian firms flooded the U.S. market with low-priced hammers, screwdrivers, and other tools. Stanley decided to fight back. It modernized its plants and introduced new quality control systems. Between 1982 and 1988 scrap rates dropped from 15 percent to only 3 percent at New Britain. Stanley not only met the competition head-on here at home, but also competed successfully in Asia. Stanley now runs a profitable operation selling its distinctive yellow tape measures in Asia.

Where are most PC clones made? Taiwan? Korea? Guess again. The answer may surprise you: Texas. Two Texas firms have been extremely successful in this marketplace. One is Compaq Computer (now part of HP), which entered the market in the early 1980s with the first portable PC. It continued to build well-designed and high-quality products, and rose to command 20 percent of the world's PC market. Compaq established itself as a market leader. The other successful PC maker, Dell Computer, is also from Texas. The sudden rise of Dell is an interesting story. Michael Dell, a former University of Texas student, started reselling IBM PCs in the early 1980s. He later formed PC's Limited, which marketed one of the first mail-order PC clones. Dell is now a market leader in the PC marketplace, offering a combination of state-of-the-art designs, high-quality products, and excellent service.

Another American firm that has made a serious commitment to quality is Motorola. Motorola, winner of the Baldrige Award in 1987, has been steadily driving down the

rate of defects in its manufactured products. Defects were reported to be near 40 parts per million at the end of 1991, down from 6,000 parts per million in 1986. Motorola has announced that its goal is to reach six-sigma (meaning six standard deviations away from the mean of a normal distribution), which translates to 3.4 parts per million. Motorola feels that the process of applying for the Baldrige Award was so valuable that it now requires all its suppliers to apply for the award as well.

Success stories like these show that the United States can compete successfully with Japan and other overseas rivals on the quality dimension. However, total quality management must become ingrained into our culture if we are going to be truly world class. The fundamentals must be there. The systems must be in place to monitor the traditional quality measures: conformance to specifications and defect-free products. However, quality management must expand beyond statistical measures. Quality must pervade the way we do business, from quality in design, quality in manufacture, and quality in building working systems with vendors, to quality in customer service and satisfaction.

Problems for Sections 1.4–1.7

13. What is an operational definition of quality? Is it possible for a 13-inch TV selling for $100 to be of superior quality to a 35-inch console selling for $1,800?
14. Studies have shown that the defect rates for many Japanese products are much lower than for their American-made counterparts. Speculate on the reasons for these differences.
15. What does "time-based competition" mean? Give an example of a product that you purchased that was introduced to the marketplace ahead of its competitors.
16. Consider the old maxim, "Build a better mousetrap and the world will beat a path to your door." Discuss the meaning of this phrase in the context of time-based competition. In particular, is getting to the market first the only factor in a product's eventual success?
17. What general features would you look for in a business process that would make that process a candidate for reengineering? Discuss a situation from your own experience in which it was clear that the business process could have been improved.
18. In what ways might the following techniques be useful as part of a reengineering effort?
    - Computer-based simulation
    - Flowcharting
    - Project management techniques
    - Mathematical modeling
    - Cross-functional teams
19. What problems can you foresee arising in the following situations?
    a. Top management is interested in reengineering to cut costs, but the employees are skeptical.
    b. Line workers would like to see a reengineering effort undertaken to give them more say-so in what goes on, but management is uninterested.
20. Just-in-time has been characterized as a system whose primary goal is to eliminate waste. Discuss how waste can be introduced in (*a*) relationships with vendors,

(*b*) receipt of material into the plant, and (*c*) movement of material through the plant. How do JIT methods cut down on these forms of waste?

21. In what ways can JIT systems improve product quality?

1.8 STRATEGIC INITIATIVES: SERVICIZATION

Hyundai is a huge multinational company based in South Korea, known as a chaebol, meaning a collection of diverse companies under a single umbrella. Prior to spinning off several of its businesses as separate companies following the Asian financial crisis in 1997, Hyundai was the largest chaebol in South Korea. The Hyundai Motor Company was established in 1967 and first began selling cars in the United States in 1986. At that time, Japanese, American, and European carmakers were firmly entrenched in the lucrative U.S. market. One way in which Hyundai sought to differentiate itself from its competitors was by offering an exceptional warranty. Today Hyundai offers a comprehensive warranty package, including a seven year/100,000 mile powertrain warranty. Sales of Hyundai models have risen steadily since 1986, with the company now offers high-end luxury vehicles, along with its low-cost entry models.

While there is no question that competitive pricing and improving reliability and performance of its products account for much of the company's success, one cannot deny that their exceptional warranties played a role as well. We will use the term "servicization" to describe the trend of manufacturing companies to bundle additional services with their products. Adding services is a means for firms to gain an edge over their competitors, and to provide an alternative to inexpensive Asian labor. (Note that the Europeans have coined the term "servitization". The term appears to have been first used by Vandermerwe and Rada (1988).

Cost-only considerations have driven much of the worldwide manufacturing to China. The Chinese economy has benefitted enormously from the investments from foreign economies, led by the United States and Japan. However, there is more to being an effective manufacturer than low labor rates. The quality of service in many overseas factories can be disappointing. Quality problems are common and turnaround times can be crippling. Such factors have lead many firms to rethink their decision to subcontract their manufacturing to Chinese factories.

Consider the case of Sleek Audio, a producer of high-end earphones. Following the trend of his competitors, the CEO, Mark Krywko, decided to have his product manufactured in a factory in Dongguan, China. Unfortunately, Krywko and his son Jason found it necessary to travel to Dongguan every few months because of persistent quality problems. For example, an entire shipment of 10,000 earphones had to be scrapped because of improper welding. Delivery delays resulted in emergency air freighting to meet promised deadlines. Furthermore, design changes took months to be implemented. As problems continued to mount, the Krywko's finally decided they'd had enough. Once the decision to come back to the states was made, Sleek Audio had no problem finding a suitable domestic partner. The earphones are now manufactured by Dynamic Innovations, whose facility is located only 15 minutes away from the company's headquarters in Palmetto, Florida. And how has the decision panned out? After more than a year producing in the United States, Sleek Audio is projecting 2011 to be the most profitable year in the company's history (Koerner, 2011).

Servicization: Moving Downstream

Wise and Baumgartner (1999) have noted that many manufacturing companies have moved their energies downstream to remain competitive. As the installed base of products increase, and the demand for new products decrease, firms are changing their business model in order to remain competitive. The focus is no longer on the manufacturing

function alone, but on those services required to operate and maintain products. During the 40-year period of 1960–2000, the service sector share of the U.S. economy grew 16 percentage points while manufacturing's share declined by 10 percentage points. What are some of the downstream activities that manufacturers are becoming more involved with? The answer is financing, after sales parts and services, and possibly training. The profit margins on service typically exceed those of manufacturing, thus providing an incentive for firms to move in this direction. Of course, the model of vertical integration is far from new. Part of Henry Ford's success was the complete vertical integration of the firm. Ford not only produced the cars, but owned the dealerships where the cars were sold and even owned the stands of rubber trees used to make tires.

As firms became more specialized, they moved away from vertical integration. However, changing patterns of demands and profits are leading many companies back in this direction. As an example, the Boeing Company, the world's foremost manufacturer of commercial aircraft, has significantly broadened its view of the value chain. The company now offers financing, local parts supply, ground maintenance, logistics management, and even pilot training. Servicization can be the key to maintaining competitiveness.

The IBM Story

IBM has become almost synonymous with computing, but few realize that IBM had its roots in mechanical tabulating machines dating to the late 1800s. The company's start came in 1890 when the German immigrant, Herman Hollerith, developed a new process to track the U.S. census. Hollerith's concept involved the use of punched cards, which persisted into the 1960s. The original firm established by Hollerith was the Tabulating Machine Company. In 1924, 10 years after T.J. Watson joined the firm, the name was changed to International Business Machines, or IBM. IBM continued to innovate mechanical computing machines, but was actually a relative late comer into the electronic computer business. In fact, the first commercial computer was produced by Engineering Research Associates of Minneapolis in 1950 and sold to the U.S. Navy. Remington Rand produced the Univac one year later, which was the first commercially viable machine. They sold 46 machines that year at a cost of over $1 million each. IBM entered the fray in 1953 when it shipped its first computer, the 701. During three years of production, IBM only sold 19 machines. In 1955, AT&T Bell Laboratories announced the first fully transistorized computer, the TRADIC, and a year later the Burroughs Corporation threw its hat into the ring and later became a major player in the computer business. Eventually, Burroughs merged with Sperry Rand to form Unisys.

In 1959, IBM produced its first transistorized-based mainframe computer, the 7000 series. However, it was not until 1964 that the firm became a leader in computer sales. That was the year that IBM announced the system 360. This was a family of six mutually compatible computers and 40 peripherals that worked seamlessly together. There is little question that the system 360 computers were state of the art at the time. Within two years IBM was shipping 1000 systems per month. IBM has been dominant in the mainframe business ever since. While all of us are familiar with personal computers of various types and configurations, the mainframe business has not gone away. Even today, IBM continues to be a presence in the mainframe business with their newest system z architecture.

While the quality of their hardware was an important factor in IBM's success, it was not the only factor. What really sealed IBM's domination of the business market was the total customer solution. IBM's industry-specific software and the 360 operating system were a large part of attracting customers away from its competitors. IBM not only had one of the most successful sales forces in the business, but also was a master of after-sales service. Each client would have an SE (systems engineer) assigned to make sure that their needs were met. As much as anything else, it was IBM's commitment to after-sales service that locked in their position as market leader in mainframe

computing. Clearly, the idea of edging out competitors by bundling services with products is not new, however, today the servicization concept is becoming an increasingly important means of gaining competitive advantage. (Information was gathered from www.computerhistory.org and the IBM company website for this section).

Performance Based Contracts

The impetus behind performance based contracting (PBC) was to reduce excessive costs in government contracting. Costs were being driven up by unnecessary provisions that specified exactly how each contract was to be carried out. According to Jon Desenberg of the Washington-based Performance Institute, a think tank dedicated to improving government performance, the idea "is to let the contracted group come up with the best possible solution and only pay them based on solving the problem . . . not on the individual steps and minutia that we have for so many years required."

Government contracts are a "why," "how," and "what" proposition. The "why" is established by the funding agency. Under PBC the "how" is shifted from the government to the contractor to determine the best way to achieve the "what". This is not always a win-win for the contractor. The task of pricing a contract now becomes much more onerous. It can be difficult to estimate all costs in advance, and who pays for contract changes can be problematic.

PBC's are typical for consumers seeking professional services. A plumber may quote a fixed price in advance for a simple job, but if there's uncertainty about the time required, might want to be compensated on a time plus materials basis. Anyone who has had a major remodeling of their homes is likely to have entered into a PBC with their contractor. However, it is rare that the final cost matches the quoted number for a variety of reasons: weather delays, poor estimation of material costs, difficulty in finding subcontractors, and more often than not, changes made in the original project by the homeowner.

While PBC's sound like a good solution for government contracting, it can lead to the wrong kind of behavior. As an example, Shen (2003) examines the result of PBC's in the Maine Addiction Treatment System. Because the system was being measured on the success of curing addicts, the center had a strong incentive to only treat the less severe cases. This, of course, runs counter to the purpose of a treatment center; it is the most severe cases of abuse that need attention. We can conclude that a PBC is not appropriate in all circumstances. For any contract it is important that incentives be properly aligned with desired outcomes.

Leasing Versus Buying

Leases may be viewed as a service provided to the consumer by the seller. In most cases, leasing is simply another way for the consumer to finance a purchase. It can also be viewed as a means for the consumer to reduce risk. The trend towards leasing of goods and services has increased substantially in recent years for several reasons.

Car leasing has long been an option for consumers. Leases are more popular when financing is difficult or expensive to obtain. The prospect of a low monthly payment attracts consumers, who may ultimately pay more in the long run. Consider the buyer that likes to drive a relatively new car and trades in their automobile every three years. This buyer faces the risk of not being able to accurately predict the trade-in value of the car three years down the road. In a lease situation, this residual risk is assumed by the seller. The terms of the lease are predicated on an assumption about the residual value of the car at the end of the lease. If the manufacturer overestimates the residual value of the car, the consumer benefits by simply turning the car in at the end of the lease period. If the manufacturer underestimates the residual value, the leaser wins by purchasing the car at the end of the lease and selling it for a higher market price. Hence, the manufacturer absorbs the risk of estimating the depreciation, which provides an incentive to the consumer to lease rather than buy. Of course, since automobiles tend to depreciate most in the first several years, buying and holding a

car will be a less expensive alternative in the long run for most vehicles—especially those that are more durable and more reliable. Since auto leases are typically two to four years, the buyer that keeps cars for a long time will not choose to lease.

Leasing (that is, renting) versus buying is also an important choice for the consumer when it comes to choosing how and where to live. While owning a home has been touted as the "American Dream," there are clearly many who should not be homeowners. In the United States, most homeowners have a mortgage on their primary property. A mortgage is simply a loan provided to homeowner, with the home itself as the collateral. In periods when housing prices are rising faster than inflation, homeowners have done very well. However, as we've seen in recent years, rising housing prices is not a certainty. There have been several periods in which housing prices have dropped precipitously, including the Great Depression of the 1930s. In fact, as of this writing, the housing industry has been in a slump since 2008.

The mortgage crisis of 2008 led to a total collapse of many major financial institutions, and nearly caused a worldwide depression. What precipitated this crisis? In a nutshell, it was awarding mortgages to individuals that did not qualify for them. Why did this happen? There were a variety of factors including government deregulation, unscrupulous mortgage brokers, and naive consumers. The stringent standards that banks traditionally applied before awarding mortgages were thrown out the window. When real estate prices were on an apparent never-ending upward spiral, homeowners borrowed well beyond their means to finance other houses, or large capital expenditures such as boats or cars. People who had never owned a home in their lives were cajoled by unscrupulous mortgage brokers into taking loans they had no chance of paying off. Housing prices climbed far beyond reasonable levels, and when the bubble burst, millions of homeowners found themselves underwater (meaning their houses were worth less than the balance of their mortgage loans). Since many loans were granted with little or no down payment, many just walked away from their homes leaving entire neighborhoods vacant. Homes were looted for valuable materials, such as copper piping, thus assuring that these homes would never be sellable.

Renting is a sensible choice for many. While there appears to be a stigma associated with being a renter, it became clear from the subprime loan debacle that, in fact, many folks have no business owning a home. Landlords must absorb the risks of repairs and price fluctuations. It is no longer obvious that real estate is necessarily a safe investment. Perhaps the mortgage crises can be viewed as a case of servicization gone out of control.

Green Leasing

A recent trend has been towards green leasing. According to the U.S. Green Building Council, buildings account for more than one-third of all energy use, carbon dioxide emissions, waste output, and use of raw materials in the United States. While green leasing is a relatively new concept to American companies, it has been a practice in other parts of the world for years. Several foreign governments have promulgated environmental-based rules for their properties. In America, however, the movement is far less centralized.

Most green leasing initiatives in the United States have been proposed by state and local governments. Government agencies and academic institutions have been the front runners in green building technology, representing approximately 26 percent of all LEED certified buildings. LEED, developed by the Green Building Counsel, is an internationally recognized certification system that measures building performance in categories such as: energy savings, water efficiency, carbon dioxide emissions, indoor environmental quality, and resource management. LEED also provides a framework for stages throughout a building's lifecycle, from design and construction to operation and management. LEED provides several levels of certification. Most green building initiatives provide either incentives or penalties based upon a building's LEED certification level.

Problems for Section 1.8

22. Define "servicization" and provide an example from your own experience of a case where services were the deciding factor in a purchase.
23. What are some of the services that IBM provided for its mainframe customers during its meteoric rise in sales in the 1960s?
24. Why can car leasing be viewed as a service? What are the advantages and disadvantages of car leasing from the buyer's point of view? Why do manufacturers offer leases?

1.9 MATCHING PROCESS AND PRODUCT LIFE CYCLES

The Product Life Cycle

The demand for new products typically undergoes cycles that can be identified and mapped over time. Understanding the nature of this evolution helps to identify appropriate strategies for production and operations at the various stages of the product cycle. A typical product life cycle is pictured in Figure 1–3. The product life cycle consists of four major segments:

1. Start-up
2. Rapid growth
3. Maturation
4. Stabilization or decline

During the start-up phase, the market for the product is developing, production and distribution costs are high, and competition is generally not a problem. During this phase the primary strategy concern is to apply the experiences of the marketplace and of manufacturing to improve the production and marketing functions. At this time, serious design flaws should be revealed and corrected.

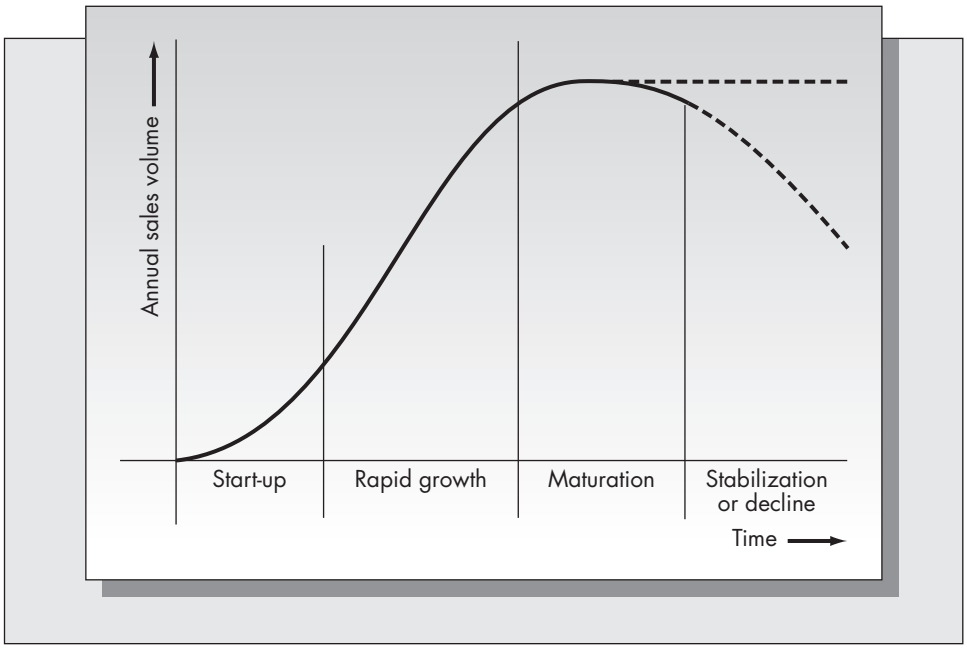


FIGURE 1–3
The product life-cycle curve

The period of rapid growth sees the beginning of competition. The primary strategic goal during this period is to establish the product as firmly as possible in the marketplace. To do this, management should consider alternative pricing patterns that suit the various customer classes and should reinforce brand preference among suppliers and customers. The manufacturing process should be undergoing improvements and standardization as product volume increases. Flexibility and modularization of the manufacturing function are highly desirable at this stage.

During the maturation phase of the product life cycle, the objective should be to maintain and improve the brand loyalty that the firm cultivated in the growth phase. Management should seek to increase market share through competitive pricing. Cost savings should be realized through improved production control and product distribution. During this phase the firm must listen to the messages of the marketplace. Most problems with product design and quality should have been corrected during the start-up and growth phases, but additional improvements should also be considered during this phase.

The appropriate shape of the life-cycle curve in the final stage depends on the nature of the product. Many products will continue to sell, with the potential for annual growth continuing almost indefinitely. Examples of such products are commodities such as household goods, processed food, and automobiles. For such products the company's primary goals in this phase would be essentially the same as those described previously for the maturation phase. Other products will experience a natural decline in sales volume as the market for the product becomes saturated or as the product becomes obsolete. If this is the case, the company should adopt a strategy of squeezing out the most from the product or product line while minimizing investment in new manufacturing technology and media advertising.

Although a useful concept, the product life-cycle curve is not accurate in all circumstances. Marketing departments that base their strategies on the life-cycle curve may make poor decisions. Dhalla and Yuspeh (1976) report an example of a firm that shifted advertising dollars from a successful stable product to a new product. The assumption was that the new product was entering the growth phase of its life cycle and the stable product was entering the declining phase of its life cycle. However, the new product never gained consumer acceptance, and because of a drop in the advertising budget, the sales of the stable product went into a decline and never recovered. They suggest that in some circumstances it is more effective to build a model that is consistent with the product's history and with consumer behavior than to blindly assume that all products follow the same pattern of growth and decline. Although we believe that the life-cycle concept is a useful way of looking at customer demand patterns in general, a carefully constructed model for each product will ultimately be a far more effective planning tool.

The Process Life Cycle

Abernathy and Townsend (1975) have classified three major stages of the **manufacturing process life cycle:** early, middle, and mature. These phases do not necessarily coincide exactly with the stages of the product life cycle, but they do provide a conceptual framework for planning improvements in the manufacturing process as the product matures.

In the first phase of the process life cycle, the manufacturing function has the characteristics of a job shop. It must cope with a varied mix of relatively low-volume orders and be responsive to changes in the product design. The types and quality of the inputs may vary considerably, and the firm has little control over suppliers.

In the middle phase of the process life cycle, automation begins to play a greater role. The firm should be able to exert more control over suppliers as the volume of production

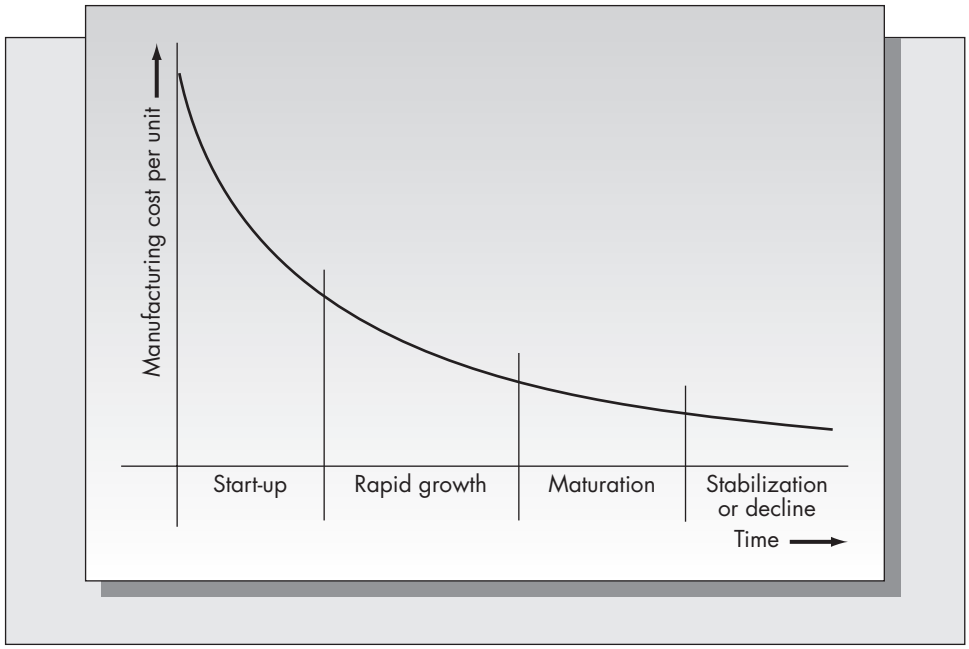


FIGURE 1–4
The process life cycle and the experience curve

increases. Unit production costs decline as a result of learning effects. The production process may involve batch processing and some transfer lines (assembly lines).

In the last phase of the process life cycle, most of the major operations are automated, the production process is standardized, and few manufacturing innovations are introduced. The production process may assume the characteristics of a continuous flow operation.

This particular evolutionary scenario is not appropriate for all new manufacturing ventures. Companies that thrive on small one-of-a-kind orders will maintain the characteristics of a job shop, for example. The process life-cycle concept applies to new products that eventually mature into high-volume items. The issue of matching the characteristics of the product with the characteristics of the process is discussed subsequently.

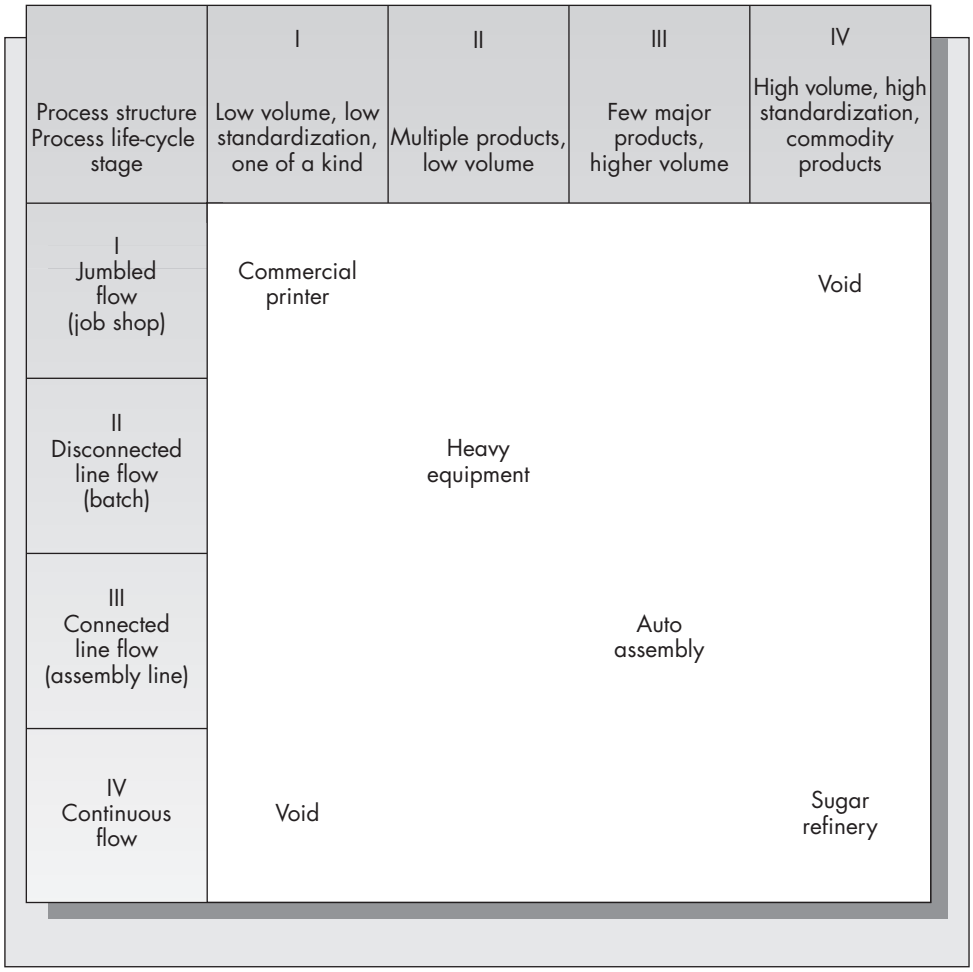
Experience curves show that unit production costs decline as the cumulative number of units produced increases. One may think of the experience curve in terms of the process life cycle shown in Figure 1–4. An accurate understanding of the relationship between the experience curve and the process life cycle can be very valuable. By matching the decline in unit cost with the various stages of the process life cycle, management can gain insight into the consequences of moving from one phase of the process life cycle into another. This insight will assist management in determining the proper timing of improvements in the manufacturing process.

The Product–Process Matrix

Hayes and Wheelwright (1979) consider linking the product and process life cycles using the **product–process matrix** pictured in Figure 1–5. The matrix is based on four phases in the evolution of the manufacturing process: (1) jumbled flow, (2) disconnected line flow, (3) connected line flow, and (4) continuous flow. This matrix may be viewed in two ways. One is to match the appropriate industry in its mature phase with

FIGURE 1–5
The product–process matrix

Source: Robert H. Hayes and Steven C. Wheelwright, "Link Manufacturing Process and Product Life Cycles" in the *Harvard Business Review* (January–February 1979). © 1979 by the President and Fellows of Harvard College; all rights reserved. Reprinted by permission.



the appropriate process. This point of view recognizes that not all industries necessarily follow the process evolution described in the previous section on the process life cycle. Certain companies or certain products could remain in an early phase of the process life cycle indefinitely. However, even firms that do not evolve to a position in the lower right-hand corner of the matrix should, in most cases, be located somewhere on the diagonal of the matrix.

Located in the upper left-hand corner of this matrix are companies that specialize in "one of a kind" jobs in which the manufacturing function has the characteristics of a jumbled flow shop. A commercial printer is an example of a jumbled flow shop. Production is in relatively small lots, and the shop is organized for maximum flexibility.

Farther down the diagonal are firms that still require a great deal of flexibility but produce a limited line of standardized items. Manufacturers of heavy equipment would fall into this category because they would produce in somewhat higher volumes. A disconnected line would provide enough flexibility to meet custom orders while still retaining economies of limited standardization.

The third category down the diagonal includes firms that produce a line of standard products for a large-volume market. Typical examples are producers of home appliances or electronic equipment, and automobile manufacturers. The assembly line or transfer line would be an appropriate process technology in this case.

Finally, the lower right-hand portion of the matrix would be appropriate for products involving continuous flow. Chemical processing, gasoline and oil refining, and sugar refining are examples. Such processes are characterized by low unit costs, standardization of the product, high sales volume, and extreme inflexibility of the production process.

What is the point of this particular classification scheme? It provides a means of assessing whether a firm is operating in the proper portion of the matrix; that is, if the process is properly matched with the product structure. Firms choosing to operate off the diagonal should have a clear understanding of the reasons for doing so. One example of a successful firm that operates off the diagonal is Rolls-Royce. Another is a company producing handmade furniture. The manufacturing process in these cases would have the characteristics of a jumbled flow shop, but competitors might typically be located in the second or third position on the diagonal.

There is another way to look at the product–process matrix. It can be used to identify the proper match of the production process with the phases of the product life cycle. In the start-up phase of product development, the firm would typically be positioned in the upper left-hand corner of the matrix. As the market for the product matures, the firm would move down the diagonal to achieve economies of scale. Finally, the firm would settle at the position on the matrix that would be appropriate based on the characteristics of the product.

Problems for Section 1.9

25. *a.* What are the four phases of the manufacturing process that appear in the product–process matrix?
    b. Discuss the disadvantages of operating off the diagonal of the matrix.
26. Give an example of a product that has undergone the four phases of the product life cycle and has achieved stability.
27. Discuss the following: "All firms should evolve along the diagonal of the product–process matrix."
28. Locate the following operations in the appropriate position on the product–process matrix.
    a. A small shop that repairs musical instruments.
    b. An oil refinery.
    c. A manufacturer of office furniture.
    d. A manufacturer of major household appliances such as washers, dryers, and refrigerators.
    e. A manufacturing firm in the start-up phase.

1.10 LEARNING AND EXPERIENCE CURVES

As experience is gained with the production of a particular product, either by a single worker or by an industry as a whole, the production process becomes more efficient. As noted by the economist Adam Smith as far back as the 18th century in his landmark work, *The Wealth of Nations:*

> The division of labor, by reducing every man's business to some one simple operation, and by making this operation the sole employment of his life, necessarily increases very much the dexterity of the worker.

By quantifying the relationship that describes the gain in efficiency as the cumulative number of units produced increases, management can accurately predict the eventual capacity of existing facilities and the unit costs of production. Today we recognize that many other factors besides the improving skill of the individual worker contribute to this effect. Some of these factors include the following:

- Improvements in production methods.
- Improvements in the reliability and efficiency of the tools and machines used.
- Better product design.
- Improved production scheduling and inventory control.
- Better organization of the workplace.

Studies of the aircraft industry undertaken during the 1920s showed that the direct-labor hours required to produce a unit of output declined as the cumulative number of units produced increased. The term **learning curve** was adopted to explain this phenomenon. Similarly, it has been observed in many industries that marginal production costs also decline as the cumulative number of units produced increases. The term **experience curve** has been used to describe this second phenomenon.

Learning Curves

As workers gain more experience with the requirements of a particular process, or as the process is improved over time, the number of hours required to produce an additional unit declines. The learning curve, which models this relationship, is also a means of describing dynamic economies of scale. Experience has shown that these curves are accurately represented by an exponential relationship. Let $Y(u)$ be the number of labor hours required to produce the uth unit. Then the learning curve is of the form

$$Y(u) = au^{-b},$$

where a is the number of hours required to produce the first unit and b measures the rate at which the marginal production hours decline as the cumulative number of units produced increases. Traditionally, learning curves are described by the percentage decline of the labor hours required to produce item $2n$ compared to the labor hours required to produce item n, and it is assumed that this percentage is independent of n. That is, an 80 percent learning curve means that the time required to produce unit $2n$ is 80 percent of the time required to produce unit n for any value of n. For an 80 percent learning curve

$$\frac{Y(2u)}{Y(u)} = \frac{a(2u)^{-b}}{au^{-b}} = 2^{-b} = .80.$$

It follows that

$$-b \ln(2) = \ln(.8)$$

$$\text{or } b = -\ln(.8)/\ln(2) = .3219. \text{ (ln is the natural logarithm.)}$$

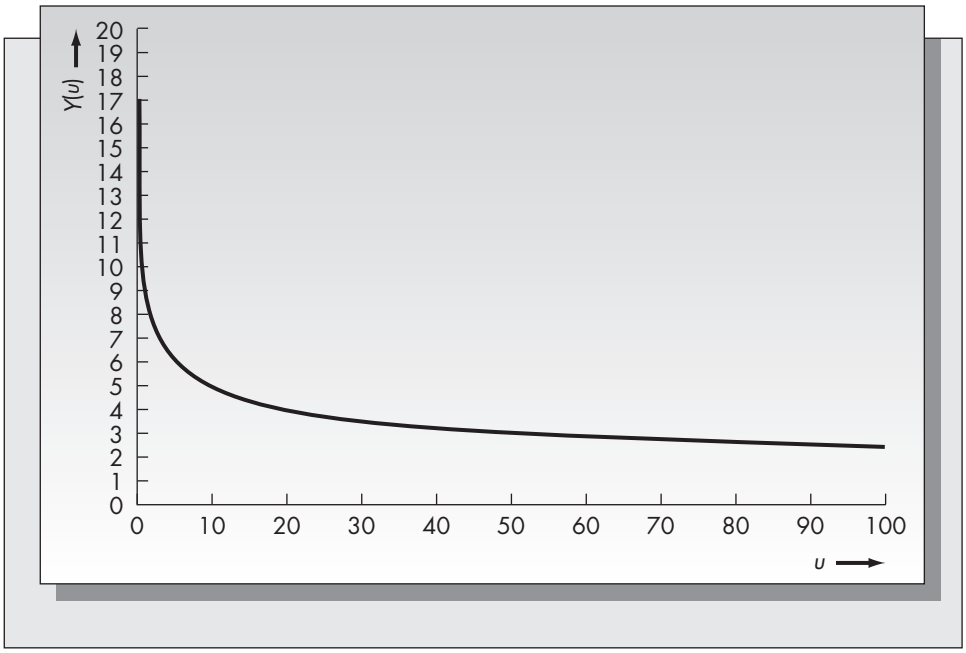
More generally, if the learning curve is a $100L$ percent learning curve, then

$$b = -\ln(L)/\ln(2).$$

Figure 1–6 shows an 80 percent learning curve. When graphed on double-log paper, the learning curve should be a straight line if the exponential relationship we have assumed is accurate. If logarithms of both sides of the expression for $Y(u)$ are taken, a linear relationship results, since

$$\ln(Y(u)) = \ln(a) - b \ln(u).$$

FIGURE 1–6
An 80 percent learning curve



Linear regression is used to fit the values of a and b to actual data after the logarithm transformation has been made. (General equations for finding least squares estimators in linear regression appear in Appendix 2–B.)

Example *1.1*

XYZ has kept careful records of the average number of labor hours required to produce one of its new products, a pressure transducer used in automobile fuel systems. These records are represented in the following table.

Cumulative Number of Units Produced (A)	Ln (Column A)	Hours Required for Next Unit (B)	Ln (Column B)
10.00	2.30	9.22	2.22
25.00	3.22	4.85	1.58
100.00	4.61	3.80	1.34
250.00	5.52	2.44	0.89
500.00	6.21	1.70	0.53
1,000.00	6.91	1.03	0.53
5,000.00	8.52	0.60	−0.51
10,000.00	9.21	0.50	−0.69

According to the theory, there should be a straight-line relationship between the logarithm of the cumulative number of units produced and the logarithm of the hours required for the last unit of production. The graph of the logarithms of these quantities for the data above appears in Figure 1–7. The figure suggests that the exponential learning curve is fairly accurate in this case. Using the methods outlined in Appendix 2–B, we have obtained estimators for the slope and the intercept of the least squares fit of the data in Figure 1–7. The values of the least squares estimators are

$$\text{Intercept} = 3.1301,$$
$$\text{Slope} = -.42276.$$

Since the intercept is $\ln(a)$, the value of a is $\exp(3.1301) = 22.88$. Hence, it should have taken about 23 hours to produce the first unit. The slope term is the constant $-b$. From the

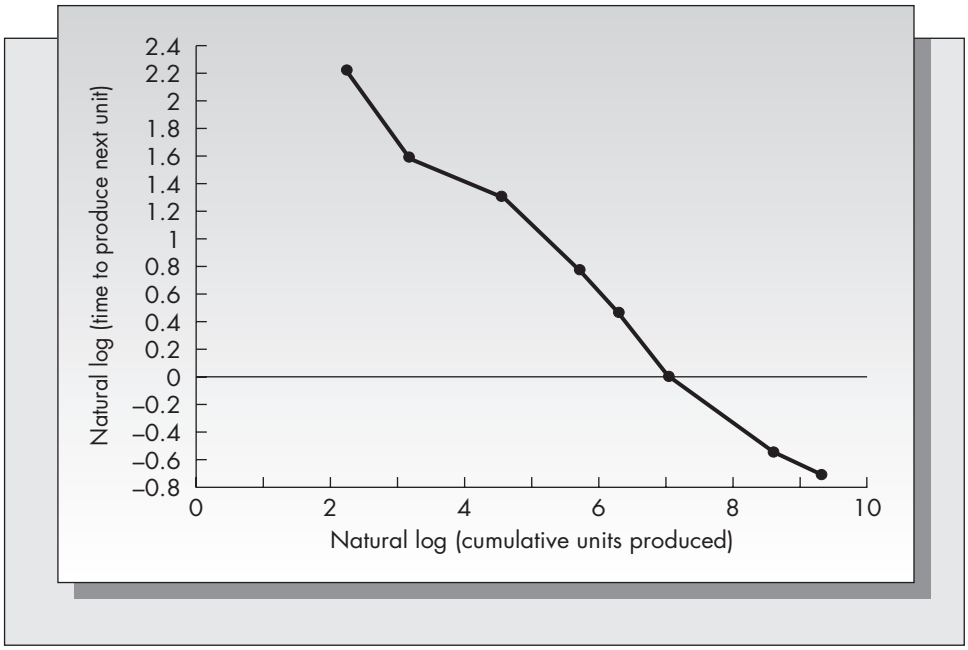


FIGURE 1–7
Log–log plot of XYZ data

equation for b on page 33 we have that

$$\ln(L) = -b \ln(2) = (-.42276)(.6931) = -.293.$$

It follows that $L = \exp(-.293) = .746$.

Hence, these data show that the learning effect for the production of the transducers can be accurately described by a 75 percent learning curve. This curve can be used to predict the number of labor hours that will be required for continued production of these particular transducers. For example, substituting $u = 50{,}000$ into the relationship

$$Y(u) = 22.88u^{-.42276}$$

gives a value of $Y(50{,}000) = .236$ hour. One must interpret such results with caution, however. A learning curve relationship may not be valid indefinitely. Eventually the product will reach the end of its natural life cycle, which could occur before 50,000 units have been produced in this example. Alternatively, there could be some absolute limit on the number of labor hours required to produce one unit that, because of the nature of the manufacturing process, can never be improved. Even with these limitations in mind, learning curves can be a valuable planning tool when properly used.

Experience Curves

Learning curves are a means of calibrating the decline in marginal labor hours as workers become more familiar with a particular task or as greater efficiency is introduced into the production process. Experience curves measure the effect that accumulated experience with production of a product or family of products has on overall cost and price. Experience curves are most valuable in industries that are undergoing major changes, such as the microelectronics industry, rather than very mature industries in which most radical changes have already been made, such as the automobile industry. The steady decline in the prices of integrated circuits (ICs) is a classic example of an experience curve. Figure 1–8 (Noyce, 1977) shows the average price per unit as a function of the industry's accumulated experience, in millions of units of production, during the period 1964 to 1972. This graph is shown on log–log scale and the points fall very close to a straight line. This case represents a 72 percent experience curve. That

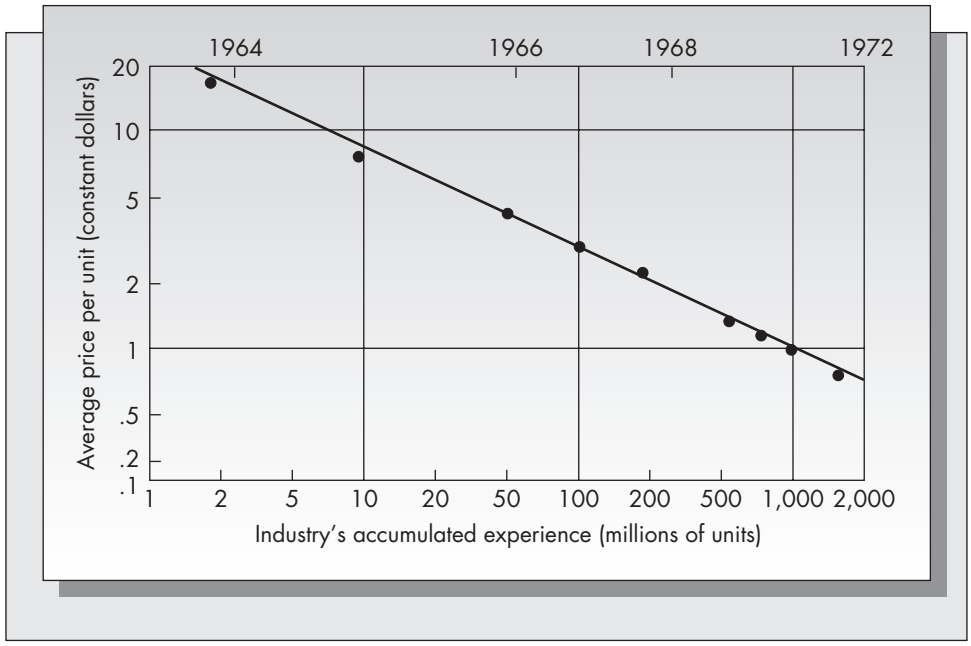


FIGURE 1–8
Prices of integrated circuits during the period 1964–1972

Source: Robert N. Noyce, "Microelectronics," in *Scientific American,* September 1977. © 1977 by Scientific American, Inc. All rights reserved. Reprinted with permission of the publisher.

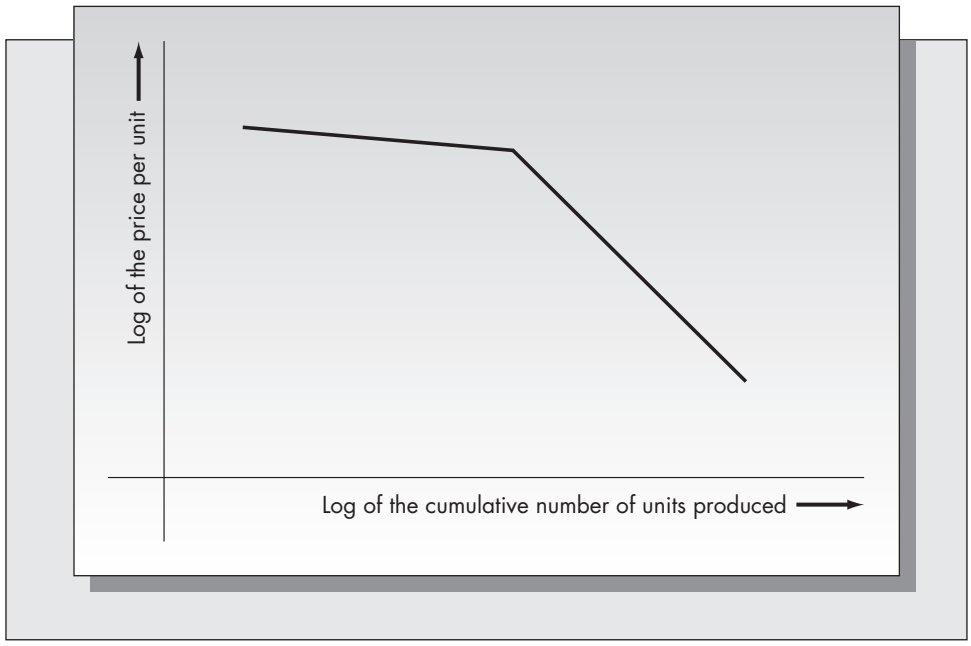


FIGURE 1–9
"Kinked" experience curve due to umbrella pricing

is, the average price per unit declines to about 72 percent of its previous value for each doubling of the cumulative production of ICs throughout the industry.

Experience curves are generally measured in terms of cost per unit of production. In most circumstances, the price of a product or family of products closely tracks the cost of production. However, in some cases umbrella pricing occurs. That is, prices remain fairly stable during a period in which production costs decline. Later, as competitive pressures of the marketplace take hold, prices decline more rapidly than costs until they catch up. This can cause a kink in the experience curve when price rather than cost is measured against cumulative volume. This type of phenomenon is pictured in Figure 1–9. Hayes

and Wheelwright (1984, p. 243) give two examples of umbrella pricing and its effect on the experience curve. The experience curves for the pricing of free-standing gas ranges and polyvinyl chloride are examples of this phenomenon.

Learning curves have been the subject of criticism in the literature recently on a number of grounds: (*a*) they lack theoretical justification; (*b*) they confuse the effects of learning, economies of scale, and other technological improvements; and (*c*) they focus on cost rather than profit (Devinney, 1987). However, it is clear that such curves are accurate descriptors of the way that marginal labor hours and costs decline as a function of the cumulative experience gained by the firm or industry.

Learning and Experience Curves and Manufacturing Strategy

We define a **learning curve strategy** as one in which the primary goal is to reduce costs of production along the lines predicted by the learning curve. Ford Motors adopted a learning curve strategy in seeking cost reductions in the Model T during the period 1909 to 1923. Abernathy and Wayne (1974) showed that the selling price of the Model T during this period closely followed an 85 percent experience curve. Ford's strategy during this time was clearly aimed at cost cutting; the firm acquired or built new facilities including blast furnaces, logging operations and saw mills, a railroad, weaving mills, coke ovens, a paper mill, a glass plant, and a cement plant. This allowed Ford to vertically integrate operations, resulting in reduced throughput time and inventory levels—a strategy similar in spirit to the just-in-time philosophy discussed earlier in this chapter.

A learning curve strategy may not necessarily be the best choice over long planning horizons. Abernathy and Wayne (1974) make the argument that when manufacturing strategy is based on cost reduction, innovation is stifled. As consumer tastes changed in the 1920s, Ford's attention to cost cutting and standardization of the Model T manufacturing process resulted in its being slow to adapt to changing patterns of customer preferences. Ford's loss was General Motors's gain. GM was quick to respond to customer needs, recognizing that the open car design of the Model T would soon become obsolete. Ford thus found itself fighting for survival in the 1930s after having enjoyed almost complete domination of the market. Survival meant a break from the earlier rigid learning curve strategy to one based on innovation.

Another example of a firm that suffered from a learning curve strategy is Douglas Aircraft. The learning curve concept was deeply rooted in the airframe industry. Douglas made several commitments in the 1960s for delivery of jet aircraft based on extrapolation of costs down the learning curve. However, because of unforeseen changes in the product design, the costs were higher than anticipated, and commitments for delivery times could not be met. Douglas was forced into a merger with McDonnell Company as a result of the financial problems it experienced.

We are not implying by these examples that a learning curve strategy is wrong. Standardization and cost reduction based on volume production have been the keys to success for many companies. Failure to achieve quick time-to-volume can spell disaster in a highly competitive marketplace. What we are saying is that the learning curve strategy must be balanced with sufficient flexibility to respond to changes in the marketplace. Standardization must not stifle innovation and flexibility.

Problems for Section 1.10

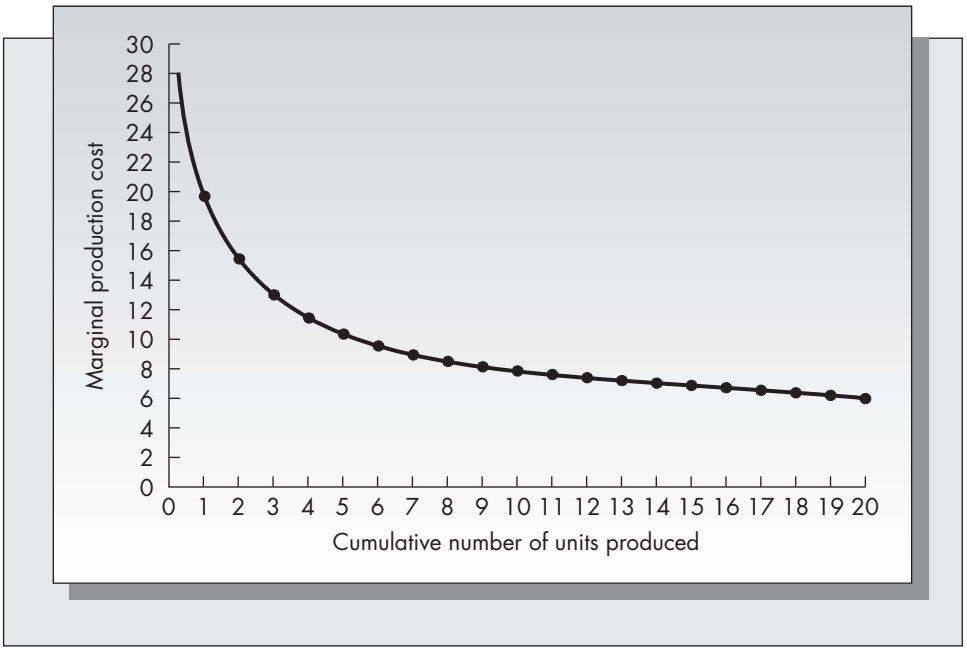
29. What are the factors that contribute to the learning curve/experience curve phenomenon?
30. What is a "learning curve strategy"? Describe how this strategy led to Ford's success up until the mid-1920s and Ford's problems after that time.

31. What are some of the pitfalls that can occur when using learning curves and experience curves to predict costs? Refer to the experience of Douglas Aircraft.
32. Consider the example of XYZ Corporation presented in this section. If the learning curve remains accurate, how long will it take to produce the 100,000th unit?
33. A start-up firm has kept careful records of the time required to manufacture its product, a shutoff valve used in gasoline pipelines.

Cumulative Number of Units Produced	Number of Hours Required for Next Unit
50	3.3
100	2.2
400	1.0
600	0.8
1,000	0.5
10,000	0.2

   a. Compute the logarithms of the numbers in each column. (Use natural logs.)
   b. Graph the ln(hours) against the ln(cumulative units) and eyeball a straight-line fit of the data. Using your approximate fit, estimate *a* and *b*.
   c. Using the results of part (*b*), estimate the time required to produce the first unit and the appropriate percentage learning curve that fits these data.
   d. Repeat parts (*b*) and (*c*), but use an exact least squares fit of the logarithms computed in part (*a*).
34. Consider the learning curve derived in Problem 30. How much time will be required to produce the 100,000th unit, assuming the learning curve remains accurate?
35. Consider the experience curve plotted in Figure 1–10. What percentage experience curve does this represent?

FIGURE 1–10
(for Problem 32)



36. Discuss the limitations of learning and experience curves.
37. An analyst predicts that an 80 percent experience curve should be an accurate predictor of the cost of producing a new product. Suppose that the cost of the first unit is $1,000. What would the analyst predict is the cost of producing the
    a. 100th unit?
    b. 10,000th unit?

1.11 CAPACITY GROWTH PLANNING: A LONG-TERM STRATEGIC PROBLEM

The capacity of a plant is the number of units that the plant can produce in a given time. Capacity policy plays a key role in determining the firm's competitive position in the marketplace. A capacity strategy must take into account a variety of factors, including

- Predicted patterns of demand.
- Costs of constructing and operating new facilities.
- New process technology.
- Competitors' strategy.

Capacity planning is an extremely complex issue. Each time a company considers expanding existing productive capacity, it must sift through a myriad of possibilities. First, the decision must be made whether to increase capacity by modifying existing facilities. From an overhead point of view, this is an attractive alternative. It is cheaper to effect major changes in existing processes and plants than to construct new facilities. However, such a strategy ultimately could be penny wise and pound foolish. There is substantial evidence that plants that have focus are the most productive. Diminishing returns quickly set in if the firm tries to push the productive capacity of a single location beyond its optimal value.

Given the decision to go ahead with construction of a new plant, many issues remain to be resolved. These include

1. *When.* The timing of construction of new facilities is an important consideration. Lead times for construction and changing patterns of demand are two factors that affect timing.

2. *Where.* Locating new facilities is a complex issue. Consideration of the logistics of material flows suggests that new facilities be located near suppliers of raw materials and market outlets. If labor costs were the key issue, overseas locations might be preferred. Tax incentives are sometimes given by states and municipalities trying to attract new industry. Cost of living and geographical desirability are factors that would affect the company's ability to hire and keep qualified employees.

3. *How much.* Once management has decided when and where to add new capacity, it must decide on the size of the new facility. Adding too much capacity means that the capacity will be underutilized. This is an especially serious problem when capital is scarce. On the other hand, adding too little capacity means that the firm will soon be faced with the problem of increasing capacity again.

Economies of Scale and Economies of Scope

Economies of scale are generally considered the primary advantages of expanding existing capacity. Panzer and Willig (1981) introduced the concept of **economies of scope,** which they defined as the cost savings realized from combining the production

of two or more product lines at a single location. The idea is that the manufacturing processes for these product lines may share some of the same equipment and personnel so that the cost of production at one location could be less than at two or more different locations.

The notion of economies of scope extends beyond the direct cost savings that the firm can realize by combining the production of two or more products at a single location. It is often necessary to duplicate a variety of support functions at different locations. These functions include information storage and retrieval systems and clerical and support staff. Such activities are easier to coordinate if they reside at the same location. The firm also can realize economies of scope by locating different facilities in the same geographic region. In this way employees can, if necessary, call upon the talents of key personnel at a nearby location.

Goldhar and Jelinek (1983) argue that considerations of economies of scope support investment in new manufacturing technology. Flexible manufacturing systems and computer-integrated manufacturing result in "efficiencies wrought by variety, not volume." These types of systems, argue the authors, allow the firm to produce multiple products in small lot sizes more cheaply using the same multipurpose equipment. (Flexible manufacturing systems are discussed in greater detail in Chapter 11.)

Management must weigh the benefits that the firm might realize by combining product lines at a single location against the disadvantages of lack of focus discussed previously. Too many product lines produced at the same facility could cause the various manufacturing operations to interfere with each other. The proper sizing and diversity of the functions of a single plant must be balanced so that the firm can realize economies of scope without allowing the plant to lose its essential focus.

Make or Buy: A Prototype Capacity Expansion Problem

A classic problem faced by the firm is known as the **make-or-buy decision.** The firm can purchase the product from an outside source for c_1 per unit, but can produce it internally for a lower unit price, $c_2 < c_1$. However, in order to produce the product internally, the company must invest $\$K$ to expand production capacity. Which strategy should the firm adopt?

The make-or-buy problem contains many of the elements of the general capacity expansion problem. It clarifies the essential trade-off of investment and economies of scale. The total cost of the firm to produce x units is $K + c_2x$. This is equivalent to $K/x + c_2$ per unit. As x increases, the cost per unit of production decreases, since K/x is a decreasing function of x. The cost to purchase outside is c_1 per unit, independent of the quantity ordered. By graphing the total costs of both internal production and external purchasing, we can find the point at which the costs are equal. This is known as the break-even quantity. The break-even curves are pictured in Figure 1–11.

The break-even quantity solves

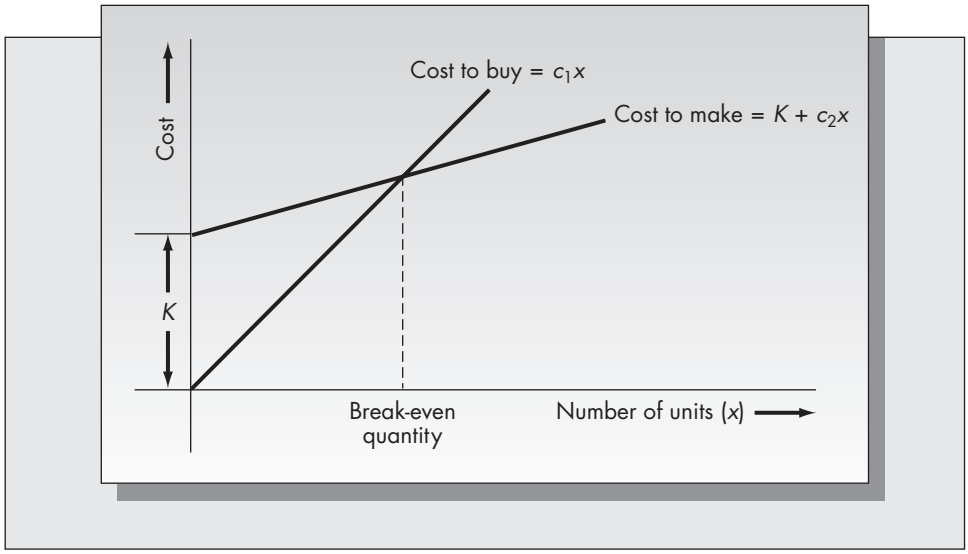
$$K + c_2x = c_1x,$$

giving $x = K/(c_1 - c_2)$.

Example *1.2*

A large international computer manufacturer is designing a new model of personal computer and must decide whether to produce the keyboards internally or to purchase them from an outside supplier. The supplier is willing to sell the keyboards for $50 each, but the manufacturer estimates that the firm can produce the keyboards for $35 each. Management estimates that expanding the current plant and purchasing the necessary equipment to make the keyboards would cost $8 million. Should they undertake the expansion?

FIGURE 1–11
Break-even curves



The break-even quantity is

$$x = 8{,}000{,}000/(50 - 35) = 533{,}333.$$

Hence, the firm would have to sell at least 533,333 keyboards in order to justify the $8 million investment required for the expansion.

Break-even curves such as this are useful for getting a quick ballpark estimate of the desirability of a capacity addition. Their primary limitation is that they are static. They do not consider the dynamic aspects of the capacity problem, which cannot be ignored in most cases. These include changes in the anticipated pattern of demand and considerations of the time value of money. Even as static models, break-even curves are only rough approximations. They ignore the learning effects of production; that is, the marginal production cost should decrease as the number of units produced increases. (Learning curves are discussed in detail in Section 1.10.) Depending on the structure of the production function, it may be economical to produce some units internally and purchase some units outside. Manne (1967) discusses the implications of some of these issues.

Dynamic Capacity Expansion Policy

Capacity decisions must be made in a dynamic environment. In particular, the dynamics of the changing demand pattern determine when the firm should invest in new capacity. Two competing objectives in capacity planning are

1. Maximizing market share
2. Maximizing capacity utilization

A firm that bases its long-term strategy on maximization of capacity utilization runs the risk of incurring shortages in periods of higher-than-anticipated demand. An alternative strategy to increasing productive capacity is to produce to inventory and let the inventory absorb demand fluctuations. However, this can be very risky. Inventories can become obsolete, and holding costs can become a financial burden.

Alternatively, a firm may assume the strategy of maintaining a "capacity cushion." This capacity cushion is excess capacity that the firm can use to respond to sudden demand surges; it puts the firm in a position to capture a larger portion of the marketplace if the opportunity arises.

Consider the case where the demand exhibits an increasing linear trend. Two policies, (*a*) and (*b*), are represented in Figure 1–12. In both cases the firm is acquiring new capacity at equally spaced intervals x, $2x$, $3x$, . . . , and increasing the capacity by the same amount at each of these times. However, in case (*a*) capacity leads demand, meaning that the firm maintains excess capacity at all times; whereas in case (*b*), the capacity lags the demand, meaning that the existing capacity is fully utilized at all times. Following policy (*a*) or (*b*) results in the time path of excess capacity (or capacity shortfall, if appropriate) given in Figure 1–13.

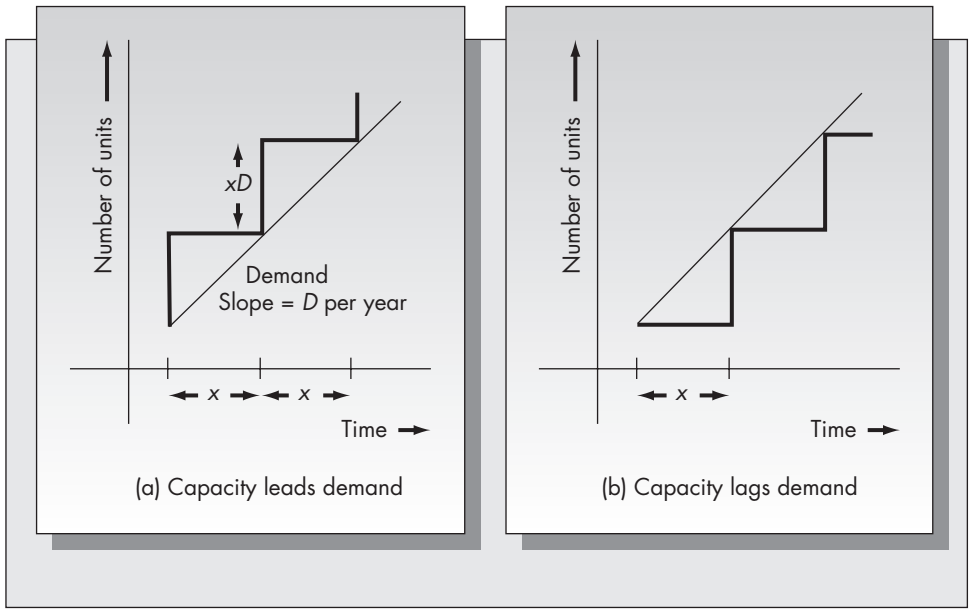


FIGURE 1–12
Capacity planning strategies

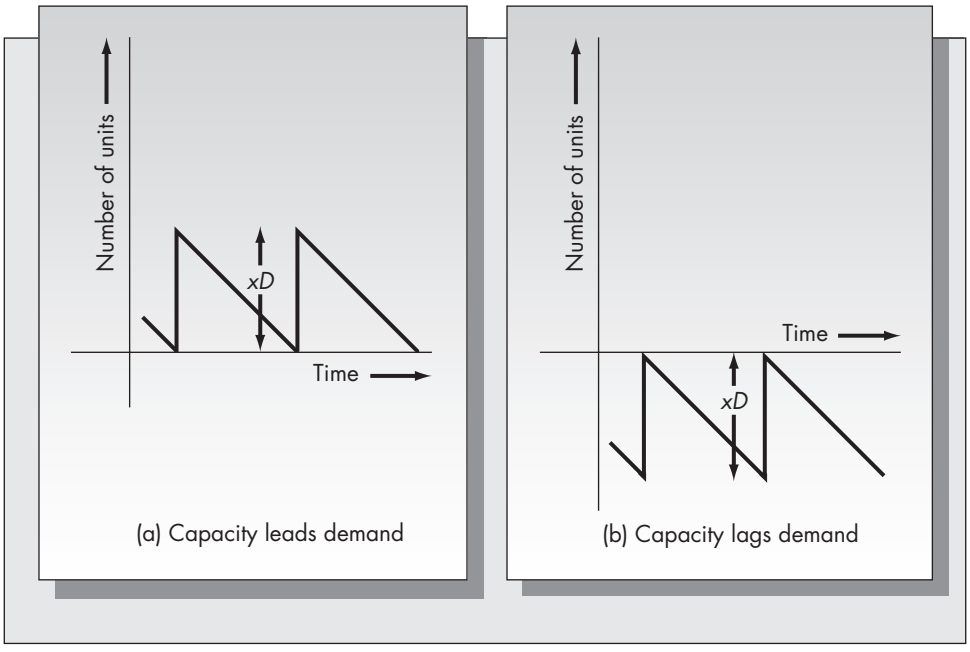


FIGURE 1–13
Time path of excess or deficient capacity

Consider the following specific model that appears in Manne (1967). Define

D = Annual increase in demand.
x = Time interval between introduction of successive plants.
r = Annual discount rate, compounded continuously.
$f(y)$ = Cost of opening a plant of capacity y.

From Figure 1–12*a* (which is the strategy assumed in the model), we see that if the time interval for plant replacement is x, it must be true that the plant size at each replacement is xD. Furthermore, the present value of a cost of \$1 incurred t years into the future is given by e^{-rt}. (A discussion of discounting and the time value of money appears in Appendix 1–A.)

Define $C(x)$ as the sum of discounted costs for an infinite horizon given a plant opening at time zero. It follows that

$$\begin{aligned} C(x) &= f(xD) + e^{-rx}f(xD) + e^{-2rx}f(xD) + \cdots \\ &= f(xD)[1 + e^{-rx} + (e^{-rx})^2 + (e^{-rx})^3 + \cdots] \\ &= \frac{f(xD)}{1 - e^{-rx}}. \end{aligned}$$

Experience has shown that a representation of $f(y)$ that explains the economies of scale for plants in a variety of industries is

$$f(y) = ky^a,$$

where k is a constant of proportionality. The exponent a measures the ratio of the incremental to the average costs of a unit of plant capacity. A value of 0.6 seems to be common (known as the six-tenths rule). As long as $a < 1$, there are economies of scale in plant construction, since a doubling of the plant size will result in less than a doubling of the construction costs. To see this, consider the ratio

$$\frac{f(2y)}{f(y)} = \frac{k(2y)^a}{k(y)^a} = 2^a.$$

Substituting $a = 0.6$, we obtain $2^a = 1.516$. This means that if $a = 0.6$ is accurate, the plant capacity can be doubled by increasing the dollar investment by about 52 percent. Henceforth, we assume that $0 < a < 1$ so that there are economies of scale in the plant sizing.

Given a specific form for $f(y)$, we can solve for the optimal timing of plant additions and hence the optimal sizing of new plants. If $f(y) = ky^a$, then

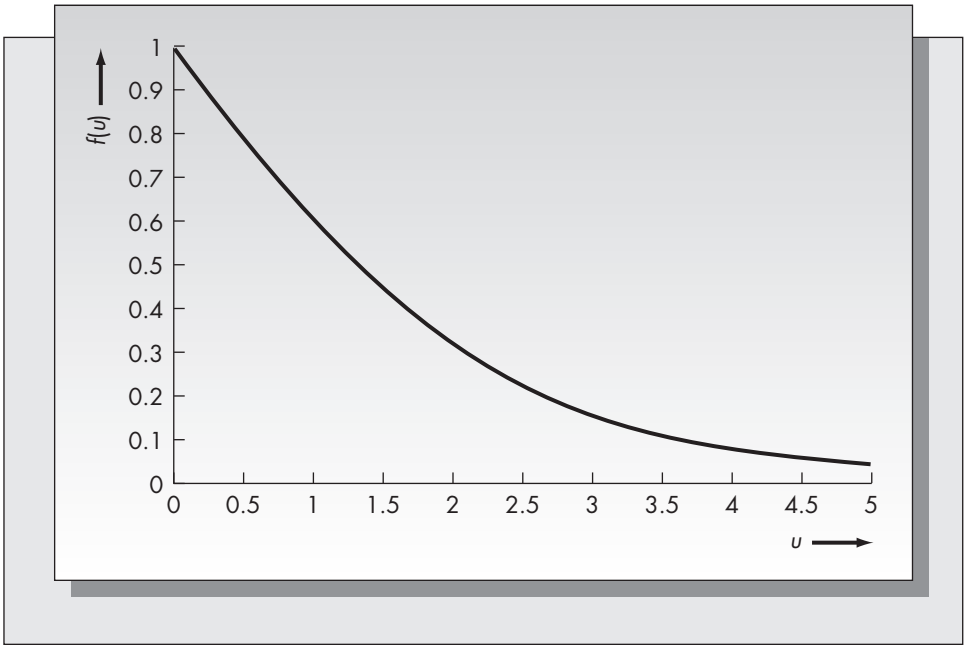
$$C(x) = \frac{k(xD)^a}{1 - e^{-rx}}.$$

Consider the logarithm of $C(x)$:

$$\begin{aligned} \log[C(x)] &= \log[k(xD)^a] - \log[1 - e^{-rx}] \\ &= \log(k) + a\log(xD) - \log[1 - e^{-rx}]. \end{aligned}$$

It can be shown that the cost function $C(x)$ has a unique minimum with respect to x and furthermore that the value of x for which the derivative of $\log[C(x)]$ is zero is the

FIGURE 1–14
The function $u/(e^u - 1)$



value of x that minimizes $C(x)$. It is easy to show[1] that the optimal solution satisfies

$$\frac{rx}{e^{rx} - 1} = a.$$

The function $f(u) = u/(e^u - 1)$ appears in Figure 1–14, where $u = rx$. By locating the value of a on the ordinate axis, one can find the optimal value of u on the abscissa axis.

Example *1.3*

A chemicals firm is planning for an increase of production capacity. The firm has estimated that the cost of adding new capacity obeys the law

$$f(y) = .0107y^{.62},$$

where cost is measured in millions of dollars and capacity is measured in tons per year. For example, substituting $y = 20{,}000$ tons gives $f(y) = \$4.97$ million plant cost. Furthermore, suppose that the demand is growing at a constant rate of 5,000 tons per year and future costs are discounted using a 16 percent interest rate. From Figure 1–14 we see that, if $a = .62$, the value of u is approximately 0.9. Solving for x, we obtain the optimal timing of new plant openings:

$$x = u/r = .9/.16 = 5.625 \text{ years.}$$

The optimal value of the plant capacity should be $xD = (5.625)(5{,}000) = 28{,}125$ tons. Substituting $y = 28{,}125$ into the equation for $f(y)$ gives the cost of each plant at the optimal solution as \$6.135 million.

Much of the research into the capacity expansion problem consists of extensions of models of this type. This particular model could be helpful in some circumstances

[1] $$\frac{d \log[C(x)]}{dx} = \frac{aD}{xD} - \frac{(-e^{-rx})(-r)}{1 - e^{-rx}} = \frac{a}{x} - \frac{r}{e^{rx} - 1} = 0,$$

which gives $\dfrac{rx}{e^{rx} - 1} = a.$

but ignores a number of fundamental features one would expect to find in the real world:

1. *Finite plant lifetime.* The assumption of the model is that, once constructed, a plant has an infinite lifetime. However, companies close plants for a variety of reasons: Equipment becomes obsolete or unreliable and cannot be replaced easily. Labor costs or requirements may dictate moving either to less expensive locations domestically or to overseas locations. Major changes in the process technology may not be easily adaptable to existing facilities.
2. *Demand patterns.* We have assumed that demand grows at a constant rate per year. Models have been proposed to account for more complex growth patterns of demand. In truth, demand uncertainty is a key factor. In many industries, foreign competition has made significant inroads into established markets, thus forcing rethinking of earlier strategies.
3. *Technological developments.* The model assumes that the capacity of all new plants constructed remains constant and that the cost of building a plant of given size remains constant as well. This is obviously unreasonable. Major changes in process technology occur on a regular basis, changing both the maximum size of new plants and the costs associated with a fixed plant size.
4. *Government regulation.* Environmental and safety restrictions may limit choices of plant location and scale.
5. *Overhead costs.* Most capacity expansion and location models do not explicitly account for the costs of overhead. During the energy crunch in the late 1970s, costs of energy overhead soared, wreaking havoc with plant overhead budgets.
6. *Tax incentives.* The financial implications of the sizing and location of new facilities must be considered in the context of tax planning. Tax incentives are offered by local or state municipalities to major corporations considering sites for the construction of new facilities.

An interesting question is whether models of this type really capture the way that companies have made capacity expansion decisions in the past. There is some preliminary evidence that they do not. Lieberman (1987) attempted to assess the factors that motivated firms to construct new chemical plants during the period 1957 to 1982. He found that the size of new plants increased by about 8 percent per year independent of market conditions. In periods of high demand, firms constructed more plants. This preliminary study indicates that the rational thinking that leads to models such as the one developed in this section does not accurately reflect the way that companies make plant-sizing decisions. His results suggest that firms build the largest plants possible with the existing technology. (Given sufficient economies of scale, however, this policy may theoretically be optimal.)

Issues in Plant Location

This section has been concerned with capacity expansion decisions, specifically, determining the amount and timing of new capacity additions. A related issue is the **location of the new facility.** Deciding where to locate a plant is a complex problem. Many factors must be carefully considered by management before making the final choice.

The following information about the plant itself is relevant to the location decision:

1. *Size of the plant.* This includes the required acreage, the number of square feet of space needed for the building structure, and constraints that might arise as a result of special needs.

2. *Product lines to be produced.*
3. *Process technology to be used.*
4. *Labor force requirements.* These include both the number of workers required and the specification of the particular skills needed.
5. *Transportation needs.* Depending on the nature of the product produced and the requirements for raw materials, the plant may have to be located near major interstate highways or rail lines.
6. *Utilities requirements.* These include special needs for power, water, sewage, or fossil fuels such as natural gas. Plants that have unusual power needs should be located in areas where energy is less expensive or near sources of hydroelectric power.
7. *Environmental issues.* Because of government regulations, there will be few allowable locations if the plant produces significant waste products.
8. *Interaction with other plants.* If the plant is a satellite of existing facilities, it is likely that management would want to locate the new plant near the others.
9. *International considerations.* Whether to locate a new facility domestically or overseas is a very sensitive issue. Although labor costs may be lower in some locations, such as the Far East, tariffs, import quotas, inventory pipeline costs, and market responsiveness also must be considered.
10. *Tax treatment.* Tax consideration is an important variable in the location decision. Favorable tax treatment is given by some countries, such as Ireland, to encourage new industry. There are also significant differences in state tax laws designed to attract domestic manufacturers.

Mathematical models are useful for assisting with many operational decisions. However, they generally are of only limited value for determining a suitable location for a new plant. Because so many factors and constraints enter into the decision process, such decisions are generally made based on the inputs of one or more of the company's divisions, and the decision process can span several years. (Mathematical techniques for making location decisions will be explored in Chapter 11, on facilities layout and location.) Schmenner (1982) has examined the decision process at a number of the *Fortune* 500 companies. His results showed that in most firms, the decision of where to locate a new facility was made either by the corporate staff or by the CEO, even though the request for new facilities might have originated at the division level. The degree of decision-making autonomy enjoyed at the division level depended on the firm's management style.

Based on a sample survey, Schmenner reported that the major factors that influenced new facilities location decisions were the following:

1. *Labor costs.* This was a primary concern for industries such as apparel, leather, furniture, and consumer electronics. It was less of a concern for capital-intensive industries.
2. *Unionization.* A motivating factor for a firm considering expanding an existing facility, as opposed to one considering building a new facility, is the potential for eliminating union influence in the new facility. A fresh labor force may be more difficult to organize.
3. *Proximity to markets.* When transportation costs account for a major portion of the cost of goods sold, locating new plants near existing markets is essential.
4. *Proximity to supplies and resources.* The decision about where to locate plants in certain industries is based on the location of resources. For example, firms producing wood or paper products must be located near forests and firms producing processed food, near farms.

5. *Proximity to other facilities.* Many companies tend to place manufacturing divisions and corporate facilities in the same geographic area. For example, IBM originated in Westchester County in New York, and located many of its divisions in that state. By locating key personnel near each other, the firm has been able to realize economies of scope.
6. *Quality of life in the region.* When other issues do not dictate the choice of a location, choosing a site that will be attractive to employees may help in recruiting key personnel. This is especially true in high-tech industries that must compete for workers with particular skills.

Problems for Section 1.11

38. A start-up company, Macrotech, plans to produce a device to translate Morse code to a written message on a home computer and to send written messages in Morse code over the airwaves. The device is primarily of interest to ham radio enthusiasts. The president, Ron Lodel, estimates that it would require a $30,000 initial investment. Each unit costs him $20 to produce and each sells for $85.
    a. How many units must be sold in order for the firm to recover its initial investment?
    b. What is the total revenue at the break-even volume?
    c. If the price were increased to $100 each, find the break-even volume.
39. For Problem 35, suppose that sales are expected to be 100 units in the first year and increase at a rate of 40 percent per year. How many years will it take to recoup the $30,000 initial investment? Assume that each unit sells for $85.
40. A domestic producer of baby carriages, Pramble, buys the wheels from a company in the north of England. Currently the wheels cost $4 each, but for a number of reasons the price will double. In order to produce the wheels themselves, Pramble would have to add to existing facilities at a cost of $800,000. It estimates that its unit cost of production would be $3.50. At the current time, the company sells 10,000 carriages annually. (Assume that there are four wheels per carriage.)
    a. At the current sales rate, how long would it take to pay back the investment required for the expansion?
    b. If sales are expected to increase at a rate of 15 percent per year, how long will it take to pay back the expansion?
41. Based on past experience, a chemicals firm estimates that the cost of new capacity additions obeys the law

$$f(y) = .0205y^{.58}$$

    where y is measured in tons per year and $f(y)$ in millions of dollars. Demand is growing at the rate of 3,000 tons per year, and the accounting department recommends a rate of 12 percent per year for discounting future costs.
    a. Determine the optimal timing of plant additions and the optimal size of each addition.
    b. What is the cost of each addition?
    c. What is the present value of the cost of the next four additions? Assume an addition has just been made for the purposes of your calculation. (Refer to Appendix 1–A for a discussion of cost discounting.)

42. A major oil company is considering the optimal timing for the construction of new refineries. From past experience, each doubling of the size of a refinery at a single location results in an increase in the construction costs of about 68 percent. Furthermore, a plant size of 10,000 barrels per day costs $6 million. Assume that the demand for the oil is increasing at a constant rate of two million barrels yearly and the discount rate for future costs is 15 percent.
    a. Find the values of k and a assuming a relationship of the form $f(y) = ky^a$. Assume that y is in units of barrels per day.
    b. Determine the optimal timing of plant additions and the optimal size of each plant.
    c. Suppose that the largest single refinery that can be built with current technology is 15,000 barrels per day. Determine the optimal timing of plant additions and the optimal size of each plant in this case. (Assume 365 days per year for your calculations.)

1.12 Summary

This chapter discussed the importance of **operations strategy** and its relationship to the overall business strategy of the firm. Operations continues to grow in importance in the firm. While a larger portion of direct manufacturing continues to move off-shore, the importance of the manufacturing function should not be underestimated. The success of the operations strategy can be measured along several dimensions. These include the obvious measures of cost and product characteristics, but also include quality, delivery speed, delivery reliability, and flexibility.

The classical view of manufacturing strategy, due primarily to Wickham Skinner, considers the following four dimensions of strategy: **time horizon, focus, evaluation,** and **consistency.** Different types of decisions relate to different time frames. A plant should be designed with a specific focus in mind, whether it be to minimize unit cost or to maximize product quality. Several evaluation criteria may be applied to analyze the effectiveness of a strategy.

We hear more and more frequently that we are part of a global community. When buying products today, we are less concerned with the country of origin than with the characteristics of the product. How many consumers of cell phones are even aware that Nokia is headquartered in Finland, Ericsson in Sweden, and Motorola in the United States? An interesting question explored by Michael Porter is: Why do some industries seem to thrive in some countries? While the answer is complex, Porter suggests that the following four factors are most important: *factor conditions; demand conditions; related and supporting industries;* and *firm strategy, structure, and rivalry.*

Changing the way that one does things can be difficult. Even more difficult is changing the way that a company does things. For that reason, **business process engineering (BPR)** is a painful process, even when it works. The most dramatic successes of BPR have come in service functions, but the concept can be applied to any environment. It is the process of rethinking how and why things are done in a certain way. Intelligently done, BPR can lead to dramatic improvements. However, it can also be a time-consuming and costly process.

Just-in-time (JIT) is a philosophy that grew from the kanban system developed by Toyota. At the heart of the approach is the elimination of waste. Systems are put in place to reduce material flows to small batches to avoid large buildups of work-in-process inventories. While JIT developed on the factory floor, it is a concept that has been applied to the purchasing function as well. Successful application of JIT purchasing requires the development of long-term relationships, and usually requires close proximity to suppliers. The mechanics of JIT are discussed in more detail in Chapter 8.

Being able to get to the market quickly with products that people want in the volumes that the marketplace requires is crucial if one wants to be a market leader. **Time-based competition** means that the time from product conception to its appearance in the marketplace must be reduced. To do so, one performs as many tasks concurrently as possible. In many instances, time to market is less important than time to volume. Being the first to the market may not mean much if one cannot meet product demand.

The dramatic successes of the Japanese during the 1970s and 1980s were to a large extent due to the outstanding **quality** of their manufactured products. Two Americans, Deming and Juran, visited Japan in the early 1950s and played an important role in making the Japanese aware of the importance of producing quality products. The quality movement in the United States has resulted in a much greater awareness of the importance of quality, recognition for outstanding achievement in this arena via the Malcolm Baldrige Award, and initiation of important programs such as quality circles and the six-sigma program at Motorola. (Both the statistical and the organizational issues concerning quality are discussed in detail in Chapter 12.)

It is important to understand both **product and process life cycles.** Both go through the four cycles of start-up, rapid growth, maturation, and stabilization or decline. It is also important to understand which types of processes are appropriate for which types of products and industries. To this end, Hayes and Wheelwright have developed the concept of the product–process matrix.

Learning and experience curves are useful in modeling the decline in labor hours or the decline in product costs as experience is gained in the production of an item or family of items. These curves have been shown to obey an exponential law, and can be useful predictors of the cost or time required for production. (Moore's Law, due to Gordon Moore, a founder of Intel, predicted the doubling of chip performance every 18 months. This is an example of an experience curve, and the prediction has continued to be accurate to the present day.)

We discussed two methods for assisting with **capacity expansion** decisions. Break-even curves provide a means of determining the sales volume necessary to justify investing in new or existing facilities. A simple model for a dynamic expansion policy is presented that gives the optimal timing and sizing of new facilities assuming constant demand growth and discounting of future costs. We also discussed issues that arise in trying to decide where to locate new facilities. This problem is very complex in that there are many factors that relate to the decision of where to locate production, design, and management facilities.

Additional Problems for Chapter 1

43. What is a production and operations strategy? Discuss the elements in common with marketing and financial strategies and the elements that are different.
44. What is the difference between the product life cycle and the process life cycle? In what way are these concepts related?
45. Suppose that the Mendenhall Corporation, a producer of women's handbags, has determined that a 73 percent experience curve accurately describes the evolution of its production costs for a new line. If the first unit costs $100 to produce, what should the 10,000th unit cost based on the experience curve?
46. Delon's Department Store sells several of its own brands of clothes and several well-known designer brands as well. Delon's is considering building a plant in

Malaysia to produce silk ties. The plant will cost the firm $5.5 million. The plant will be able to produce the ties for $1.20 each. On the other hand, Delon's can subcontract to have the ties produced and pay $3.00 each. How many ties will Delon's have to sell worldwide to break even on its investment in the new plant?

47. A Japanese steel manufacturer is considering expanding operations. From experience, it estimates that new capacity additions obey the law

$$f(y) = .00345y^{.51},$$

where the cost $f(y)$ is measured in millions of dollars and y is measured in tons of steel produced. If the demand for steel is assumed to grow at the constant rate of 8,000 tons per year and future costs are discounted using a 10 percent discount rate, what is the optimal number of years between new plant openings?

The following problems are designed to be solved by spreadsheet.

48. Consider the following break-even problem: the cost of producing Q units, $c(Q)$, is described by the curve

$$c(Q) = 48Q[1 - \exp(-.08Q)],$$

where Q is in hundreds of units of items produced and $c(Q)$ is in thousands of dollars.

a. Graph the function $c(Q)$. What is its shape? What economic phenomenon gives rise to a cumulative cost curve of this shape?

b. At what production level does the cumulative production cost equal $1,000,000?

c. Suppose that these units can be purchased from an outside supplier at a cost of $800 each, but the firm must invest $850,000 to build a facility that would be able to produce these units at a cost $c(Q)$. At what cumulative volume of production does it make sense to invest in the facility?

49. Maintenance costs for a new facility are expected to be $112,000 for the first year of operation. It is anticipated that these costs will increase at a rate of 8 percent per year. Assuming a rate of return of 10 percent, what is the present value of the stream of maintenance costs over the next 30 years?

50. Suppose the supplier of keyboards described in Example 1.2 is willing to offer the following incremental quantity discount schedule:

Cost per Keyboard	Order Quantity
$50	$Q \leq 100{,}000$
$45	$100{,}000 < Q \leq 500{,}000$
$40	$500{,}000 < Q$

Determine the cost to the firm for order quantities in increments of 20,000 for $Q = 200{,}000$ to $Q = 1{,}000{,}000$, and compare that to the cost to the firm of producing internally for these same values of Q. What is the break-even order quantity?

Appendix 1–A

PRESENT WORTH CALCULATIONS

Including the time value of money in the decision process is common when considering alternative investment strategies. The idea is that a dollar received today has greater value than one received a year from now. For example, if a dollar were placed in a simple passbook account paying 5 percent, it would be worth $1.05 in one year. More generally, if it were invested at a rate of return of r (expressed as a decimal), it would be worth $1 + r$ in a year, $(1 + r)^2$ in two years, and so on.

In the same way, a cost of $1 incurred in a year has a present value of less than $1 today. For example, at 5 percent interest, how much would one need to place in an account today so that the total principal plus interest would equal $1 in a year? The answer is $1/(1.05) = 0.9524$. Similarly, the present value of a $1 cost incurred in two years at 5 percent is $1/(1.05)^2 = 0.9070$. In general, the present value of a cost of $1 incurred in t years assuming a rate of return r is $(1 + r)^{-t}$.

These calculations assume that there is no compounding. Compounding means that one earns interest on the interest, so to speak. For example, 5 percent compounded semiannually means that one earns 2.5 percent on $1 after six months and 2.5 percent on the original $1 plus interest earned in the first six months. Hence, the total return is

$$(1.025)(1.025) = \$1.050625$$

after one year, or slightly more than 5 percent. If the interest were compounded quarterly, the dollar would be worth

$$(1 + .05/4)^4 = 1.0509453$$

at the end of the year. The logical extension of this idea is continuous compounding. One dollar invested at 5 percent compounded continuously would be worth

$$\begin{aligned}\lim_{n\to\infty} (1 + .05/n)^n &= e^{.05} \\ &= 1.05127\end{aligned}$$

at the end of a year. The number $e = 2.7172818\ldots$ is defined as

$$e = \lim_{n\to\infty} (1 + 1/n)^n.$$

Notice that continuous compounding only increases the effective simple interest rate from 5 percent to 5.127 percent.

More generally, C invested at a rate of r for t years compounded continuously is worth Ce^{rt} at the end of t years.

Reversing the argument, the present value of a cost of C incurred in t years assuming continuous compounding at a discount rate r is Ce^{-rt}. A stream of costs $C_1, C_2, \ldots, C_n$ incurred at times $t_1, t_2, \ldots, t_n$ has present value

$$\sum_{i=1}^{n} C_i e^{-rt_i}.$$

A comprehensive treatment of discounting and its relationship to the capacity expansion problem can be found in Freidenfelds (1981).

Bibliography

Abernathy, W. J., and P. L. Townsend. "Technology, Productivity, and Process Change." *Technological Forecasting and Social Change* 7 (1975), pp. 379–96.

Abernathy, W. J., and K. Wayne. "Limits of the Learning Curve." *Harvard Business Review* 52 (September–October 1974), pp. 109–19.

Becker, G. S. Quoted in *Business Week*, January 27, 1986, p. 12.

Bell, D. *The Coming of the Post-Industrial Society: A Venture in Social Forecasting*. New York: Basic Books, 1987.

Blackburn, J. D. *Time-Based Competition: The Next Battleground in American Manufacturing*. New York: McGraw-Hill/Irwin, 1991.

Business Week. October 22, 1990, pp. 94–95.

Business Week. "The Quality Imperative." October 25, 1991.

Cohen, S. S., and J. Zysman. *Manufacturing Matters: The Myth of the Post-Industrial Economy*. New York: Basic Books, 1987.

Davis, D. "Beating the Clock." *Electronic Business*, May 1989.

Devinney, T. M. "Entry and Learning." *Management Science* 33 (1987), pp. 102–12.

Dhalla, N. K., and S. Yuspeh. "Forget about the Product Life Cycle Concept." Harvard Business Review 54 (January–February 1976), pp. 102–12.

The Economist. "Reshoring Manufacturing: Coming Home." January 2013. Retrieved from http://www.economist.com/news/special-report/21569570-growing-number-american-companies-are-moving-their-manufacturing-back-united

Fallon, M. *San Jose Mercury News*, September 30, 1985, p. 11D.

Faux, J. "Manufacturing Key to America's Future." Presentation to the Industrial Union Council Legislative Conference, Economic Policy Institute, 7 pages, February 4, 2003.

Freeland, J. R. "A Survey of Just-in-Time Purchasing Practices in the United States." *Production and Inventory Management Journal*, Second Quarter 1991, pp. 43–50

Freidenfelds, J. *Capacity Expansion: Analysis of Simple Models with Applications*. New York: Elsevier North Holland, 1981.

Goldhar, J. P., and M. Jelinek. "Plan for Economies of Scope." *Harvard Business Review* 61 (1983), pp. 141–48.

Hagenbaugh, B. "U.S. Manufacturing Jobs Fading Away Fast." *USA Today*, December 12, 2002.

Hammer, M. S., and J. Champy. *Reengineering the Corporation: A Manifesto for Business Resolution*. New York: Harper Business, 1993.

Hammer, M. S., and S. A. Stanton. *The Reengineering Revolution*. New York: Harper Business, 1995.

Hayes, R. H., and S. Wheelwright. "Link Manufacturing Profess and Product Life Cycles." *Harvard Business Review* 57 (January–February 1979), pp. 133–40.

Hayes, R. H., and S. Wheelwright. *Restoring Our Competitive Edge: Competing through Manufacturing*. New York: John Wiley & Sons, 1984.

Hutzel, T. and D. Lippert. Reshoring 101: Rebuilding U.S. Manufacturing through Right Sizing and Right Shoring. Retrieved from http://www.mainstreammanagement.com/pdf/rebuilding-us-manufacturing.pdf, 2012.

Jones, P., and L. Kahaner. *Say It and Live It. The Fifty Corporate Mission Statements That Hit the Mark*. New York: Currency Doubleday, 1995.

Koerner, B. I. "Made in America: Small Businesses Buck the Offshoring Trend." *Wired*, March 2011, pp. 104–111.

Krugman, P. *Peddling Prosperity: Economic Sense and Nonsense in the Age of Diminished Expectations*. New York: W. W. Norton and Company, 1994.

Lieberman, M. "Scale Economies, Factory Focus, and Optimal Plant Size." Paper present at Stanford Graduate School of Business, July 22, 1987.

Manne, A. S., ed. *Investments for Capacity Expansion: Size, Location, and Time Phasing*. Cambridge, MA: MIT Press, 1967.

Noyce, R. N. "Microelectronics." *Scientific American*, September 1977.

Panzer, J. C., and R. O. Willing. "Economies of Scope." *American Economic Review*, 1981, pp. 268–72.

Pisano, G.P. and W.C. Shih. *Producing Prosperity: Why America Needs a Manufacturing Renaissance*. Harvard Business Review Press, Boston, Massachusetts, 2012.

Port, O. "Customers Move into the Driver's Seat." *Business Week*. October 4, 1999, pp. 103–06.

Porter, M. E. and J. W. Rivkin. "Choosing the United States." *Harvard Business Review*, pp. 80–93, March 2012.

Shen, Y. "Selection Incentives in a Performance-Based Contracting System." *Health Services Research*. 38 (2003), pp. 535–552.

Skinner, W. *Manufacturing in the Corporate Strategy*. New York: John Wiley & Sons, 1978.

Stasey, R., and C. J. McNair. *Crossroads: A JIT Success Story*. New York: McGraw-Hill/Irwin, 1990.

Vandermerwe, S. and J. Rada. "Servitization of Business: Adding Value by Adding Services." *European Management Journal* 6 (1988), pp. 314–324.

Wills, F. "Bully for Taurus." *Business Month*, February 1990, p. 13.

Wise, R. and P. Baumgartner. "Go Downstream: The New Profit Imperative in Manufacturing." *Harvard Business Review* 77 (1999), pp. 133–141.

Womack, J. P., D. T. Jones, and D. Roos. *The Machine That Changed the World*. New York: Harper Perennial, 1990.

Chapter **Two**

Forecasting

"It's hard to make predictions, especially about the future."
—Neils Bohr

Chapter Overview

Purpose

To present and illustrate the most important methods for forecasting demand in the context of operations planning.

Key Points

1. *Characteristics of forecasts.*
   - They are almost always going to be wrong.
   - A good forecast also gives some measure of error.
   - Forecasting aggregate units is generally easier than forecasting individual units.
   - Forecasts made further out into the future are less accurate.
   - A forecasting technique should not be used to the exclusion of known information.
2. *Subjective forecasting.* Refers to methods that measure either individual or group opinion. The better known subjective forecasting methods include:
   - Sales force composites.
   - Customer surveys.
   - Jury of executive opinion.
   - The Delphi method.
3. *Objective forecasting methods (time series methods and regression).* Using *objective forecasting* methods, one makes forecasts based on past history. *Time series* forecasting uses only the past history of the series to be forecasted, while *regression models* often incorporate the past history of other series. In time series forecasting, the goal is to find predictable and repeatable patterns in past data. Based on the identified pattern, different methods are appropriate. Time series methods have the advantage of easily being incorporated into a computer program for automatic forecasting and updating. Repeatable patterns that we consider include increasing or decreasing linear trend, curvilinear trend (including exponential growth), and seasonal fluctuations. When using regression, one constructs a causal model that predicts one phenomenon (the dependent variable) based on the evolution of one or more other phenomenon (the independent variables). An example would be predicting the start or end of a recession based on housing starts (housing starts are considered to be a leading economic indicator of the health of the economy).

4. *Evaluation of forecasting methods.* The forecast error in any period, e_t, is the difference between the forecast for period t and the actual value of the series realized for period t ($e_t = F_t - D_t$). Three common measures of forecast error are MAD (average of the absolute errors over n periods), MSE (the average of the sum of the squared errors over n periods), and MAPE (the average of the percentage errors over n periods).
5. *Methods for forecasting stationary time series.* We consider two forecasting methods when the underlying pattern of the series is stationary over time: moving averages and exponential smoothing. A *moving average* is simply the arithmetic average of the N most recent observations. *Exponential smoothing* forecasts rely on a weighted average of the most recent observation and the previous forecast. The weight applied to the most recent observation is α, where $0 < \alpha < 1$, and the weight applied to the last forecast is $1 - \alpha$. Both methods are commonly used in practice, but the exponential smoothing method is favored in inventory control applications—especially in large systems—because it requires much less data storage than does moving averages.
6. *Methods for forecasting series with trend.* When there is an upward or downward linear trend in the data, two common forecasting methods are *linear regression* and double exponential smoothing via *Holt's method.* Linear regression is used to fit a straight line to past data based on the method of least squares, and Holt's method uses separate exponential smoothing equations to forecast the intercept and the slope of the series each period.
7. *Methods for forecasting seasonal series.* A seasonal time series is one that has a regular repeating pattern over the same time frame. Typically, the time frame would be a year, and the periods would be weeks or months. The simplest approach for forecasting seasonal series is based on multiplicative seasonal factors. A multiplicative seasonal factor is a number that indicates the relative value of the series in any period compared to the average value over a year. Suppose a season consists of 12 months. A seasonal factor of 1.25 for a given month means that the demand in that month is 25 percent higher than the mean monthly demand. *Winter's method* is a more complex method based on triple exponential smoothing. Three distinct smoothing equations are used to forecast the intercept, the slope, and the seasonal factors each period.
8. *Box-Jenkins models*. George Box and Gwilym Jenkins developed forecasting methods based largely on a statistical analysis of the autoregressive function of a time series. Autoregression seeks to discover repeating patterns in data by considering the correlation of observations of the series with other observations separated by fixed number of periods. These models have proven to be very powerful for forecasting some economic time series, but they require large data sets (at least 72 observations) and a knowledgeable user. We provide a brief review of these powerful methods.
9. *Other considerations*. In addition to Box-Jenkins methods, when large data sets are available, filtering methods borrowed from electrical engineering can often provide excellent forecasts for economic time series. Two of the better known filters are Kalman Filters and Wiener Filters. Neither of these methods are amenable to automatic forecasting. Monte Carlo simulation is another technique that can be useful for building a forecasting model. Finally, we discuss forecasting demand in the context of a lost sales inventory system.

Families plan vacations around their schedules, and for that reason America's theme parks tend to be very crowded during holidays and school breaks. On June 18, 2011, Universal Studios Theme Parks opened its new Harry Potter section at Universal Orlando Resort in Orlando, FL, which continues to be a major attraction for tourists. If you were planning on taking your family to a theme park such as Universal Studios, when would be the best time to go to avoid the long lines? One might think that Thanksgiving and Christmas would be good choices, as most people are home with families on these holidays. Not so! In fact, these two days (along with New Year's Day) are the busiest days of the year. The period between Thanksgiving and Christmas as well as the periods after New Year's Day and before Halloween are the slowest times, and Universal Orlando Resort recommends that visitors come at these times to avoid longer lines.

As was so eloquently stated by Charles F. Kettering, "My concern is with the future since I plan to spend the rest of my life there." But the future can never be known, so we make forecasts. We forecast traffic patterns and plan routes accordingly. We forecast which foods will be best in a particular restaurant, and order accordingly. We choose universities to attend based on forecasting our experiences there and the doors that a degree from that university will open. We make hundreds of forecasts every day, some carefully thought out, some made almost unconsciously. Forecasting plays a central role in all of our lives.

In the same way, forecasting plays a central role in the operations function of a firm. All business planning is based on forecasts. Sales of existing and new products, requirements and availabilities of raw materials, changing skills of workers, interest rates, capacity requirements, and international politics are only a few of the factors likely to affect the future success of a firm.

The functional areas of the firm that make the most use of forecasting methods are marketing and production. Marketing is responsible for forecasting sales of both new and existing product lines. Sales forecasts are the primary driver for the S&OP (sales and operations planning) function, which will be discussed in detail in Chapter 3. In some circumstances, the forecasts prepared for marketing purposes may not be appropriate or sufficient for operations planning. For example, to determine suitable stocking levels for spare parts, one must know schedules for planned replacements and be able to forecast unplanned replacements. Also it could be that the S&OP planning function might be producing forecasts for aggregate units, while forecasts for individual SKU's (stock-keeping units) might be required.

We have seen firms benefit from good forecasting and pay the price for poor forecasting. During the 1960s, consumer tastes in automobiles slowly shifted from large, heavy gas guzzlers to smaller, more fuel efficient automobiles. Detroit, slow to respond to this change, suffered when the OPEC oil embargo hit in the late 1970s and tastes shifted more dramatically to smaller cars. Compaq Computer became a market leader in the early 1980s by properly forecasting consumer demand for a portable version of the IBM PC, which gained a popularity that far exceeded expectations. Forecasting played a role in Ford Motors's early success and later demise. Henry Ford saw that the consumer wanted a simpler, less expensive car that was easier to maintain than most manufacturers were offering in the early 1900s. His Model T dominated the industry. However, Ford did not see that consumers would tire of the open Model T design. Ford's failure to forecast consumer desires for other designs nearly resulted in the end of a firm that had monopolized the industry only a few years before.

Seeing trends is the first step towards profiting from those trends. As an example, consider the trend towards greater use of renewable energy. Renewable energy sources include wind power, sun power, tidal power, geothermal power, etc. If energy can be generated and stored using renewable methods, this energy can be used to power electric cars, thus cutting down on gasoline consumption.

Some companies were able to see this trend and take advantage of it. In particular, the use of solar cells has grown dramatically in recent years. While Apple Corporation has received a great deal of publicity for its fantastic successes in mobile computing, solar cell installations have actually been growing at a comparable rate. The residential use of solar cells in the United States grew 33 percent in Q1 2013 compared to Q1 2012. Markets in Asia are expected to grow more rapidly than in the United States, thus assuring a steady market growth in this segment. Manufacturers that saw this trend developing are now reaping the rewards of their foresight.

Can all events be accurately forecasted? The answer is clearly no. Consider the experiment of tossing a coin. Assuming that it is a fair coin and the act of tossing does not introduce bias, the best you can say is that the probability of getting heads is 50 percent on any single toss. No one has been able to consistently top the 50 percent prediction rate for such an experiment over a long period of time. Many real phenomena are accurately described by a type of coin-flipping experiment. Games of chance played at casinos are random. By tipping the probabilities in its favor, the house is always guaranteed to win over the long term. There is evidence that daily prices of stocks follow a purely random process, much like a coin-flipping experiment. Studies have shown that professional money managers rarely outperform stock portfolios generated purely at random.

In production and operations management, we are primarily interested in forecasting product demand. Because demand is likely to be random in most circumstances, can forecasting methods provide any value? In most cases, the answer is yes. Although some portions of the demand process may be unpredictable, other portions may be predictable. Trends, cycles, and seasonal variation may be present, all of which give us an advantage over trying to predict the outcome of a coin toss. In this chapter we consider methods for predicting future values of a series based on past observations.

2.1 THE TIME HORIZON IN FORECASTING

We may classify forecasting problems along several dimensions. One is the time horizon. Figure 2–1 is a schematic showing the three time horizons associated with forecasting and typical forecasting problems encountered in operations planning associated with each. Short-term forecasting is crucial for day-to-day planning. Short-term forecasts, typically measured in days or weeks, are required for inventory management, production plans that may be derived from a materials requirements planning system (to be discussed in detail in Chapter 8), and resource requirements planning. Shift scheduling may require forecasts of workers' availabilities and preferences.

The intermediate term is measured in weeks or months. Sales patterns for product families, requirements and availabilities of workers, and resource requirements are typical intermediate-term forecasting problems encountered in operations management.

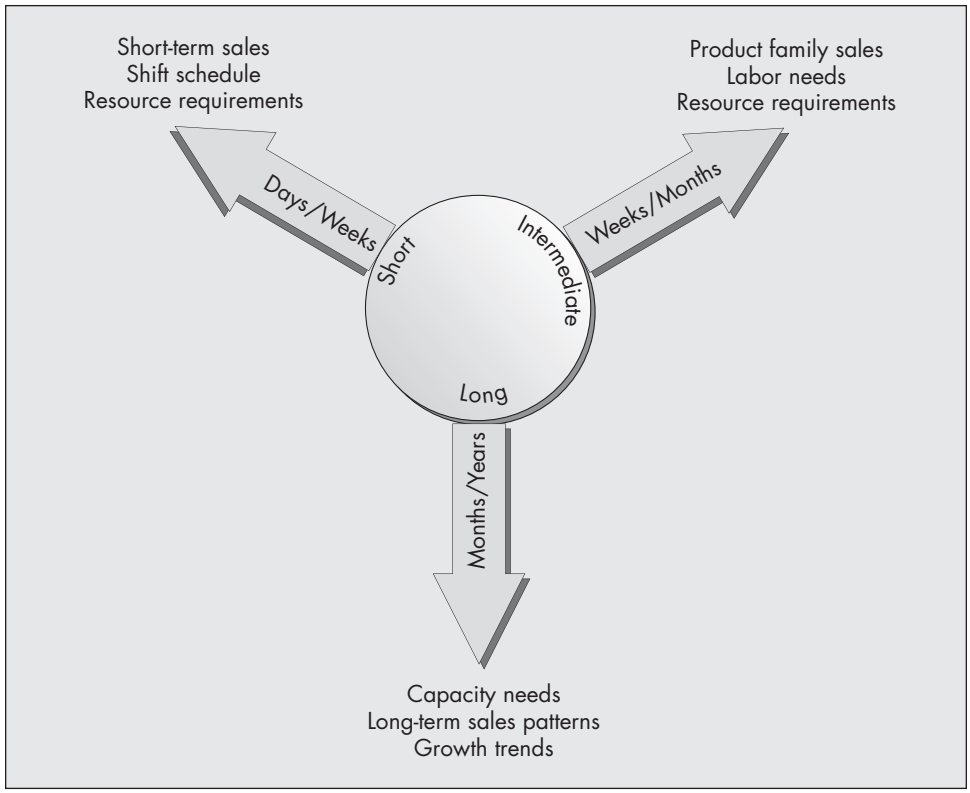


FIGURE 2–1
Forecast horizons in operation planning

Long-term production and manufacturing decisions, discussed in Chapter 1, are part of the overall firm's manufacturing strategy. One example is long-term planning of capacity needs. When demands are expected to increase, the firm must plan for the construction of new facilities and/or the retrofitting of existing facilities with new technologies. Capacity planning decisions may require downsizing in some circumstances. For example, General Motors Corporation historically commanded about 45 percent of the domestic car market. However, in the 1990s that percentage dropped to 35 percent. As a result, GM was forced to significantly curtail its manufacturing operations to remain profitable.

2.2 CHARACTERISTICS OF FORECASTS

1. *They are usually wrong*. As strange as it may sound, this is probably the most ignored and most significant property of almost all forecasting methods. Forecasts, once determined, are often treated as known information. Resource requirements and production schedules may require modifications if the forecast of demand proves to be inaccurate. The planning system should be sufficiently robust to be able to react to unanticipated forecast errors.

2. *A good forecast* is more than a single number. Given that forecasts are generally wrong, a good forecast also includes some measure of the anticipated forecast error. This could be in the form of a range, or an error measure such as the variance of the distribution of the forecast error.

3. *Aggregate forecasts are more accurate*. Recall from statistics that the variance of the average of a collection of independent identically distributed random variables is lower than the variance of each of the random variables; that is, the variance of the sample mean is smaller than the population variance. This same phenomenon is true in forecasting as well. On a percentage basis, the error made in forecasting sales for an entire product line is generally less than the error made in forecasting sales for an individual item. This phenomenon, known as risk pooling, will be discussed in the inventory control context in Chapter 5.

4. *The longer the forecast horizon, the less accurate the forecast will be*. This property is quite intuitive. One can predict tomorrow's value of the Dow Jones Industrial Average more accurately than next year's value.

5. *Forecasts should not be used to the exclusion of known information*. A particular technique may result in reasonably accurate forecasts in most circumstances. However, there may be information available concerning the future demand that is not presented in the past history of the series. For example, the company may be planning a special promotional sale for a particular item so that the demand will probably be higher than normal. This information must be manually factored into the forecast.

2.3 SUBJECTIVE FORECASTING METHODS

We classify forecasting methods as either subjective or objective. A subjective forecasting method is based on human judgment. There are several techniques for soliciting opinions for forecasting purposes:

1. *Sales force composites*. In forecasting product demand, a good source of subjective information is the company sales force. The sales force has direct contact with consumers and is therefore in a good position to see changes in their preferences. To develop a sales force composite forecast, members of the sales force submit sales estimates of the products they will sell in the coming year. These estimates might be individual numbers or several numbers, such as pessimistic, most likely, and optimistic estimates. Sales managers would then be responsible for aggregating individual estimates to arrive at overall forecasts for each geographic region or product group. Sales force composites may be inaccurate when compensation of sales personnel is based on meeting a quota. In that case, there is clearly an incentive for the sales force to lowball its estimates.

2. *Customer surveys*. Customer surveys can signal future trends and shifting preference patterns. To be effective, however, surveys and sampling plans must be carefully designed to guarantee that the resulting data are statistically unbiased and representative of the customer base. Poorly designed questionnaires or an invalid sampling scheme may result in the wrong conclusions.

3. *Jury of executive opinion*. When there is no past history, as with new products, expert opinion may be the only source of information for preparing forecasts. The approach here is to systematically combine the opinions of experts to derive a forecast. For new product planning, opinions of personnel in the functional areas of marketing, finance, and operations should be solicited. Combining individual forecasts may be done in several ways. One is to have the individual responsible for preparing the forecast interview the executives directly and develop a forecast from the results of the interviews. Another is to require the executives to meet as a group and come to a consensus.

4. *The Delphi method*. The Delphi method, like the jury of executive opinion method, is based on soliciting the opinions of experts. The difference lies in the manner in which individual opinions are combined. (The method is named for the Delphic oracle of ancient Greece, who purportedly had the power to predict the future.) The Delphi method attempts to eliminate some of the inherent shortcomings of group dynamics, in which the personalities of some group members overshadow those of other members. The method requires a group of experts to express their opinions, preferably by individual sample survey. The opinions are then compiled and a summary of the results is returned to the experts, with special attention to those opinions that are significantly different from the group averages. The experts are asked if they wish to reconsider their original opinions in light of the group response. The process is repeated until (ideally) an overall group consensus is reached.

As with any particular technique, the Delphi method has advantages and disadvantages. Its primary advantage is that it provides a means of assessing individual opinion without the usual concerns of personal interactions. On the negative side, the method is highly sensitive to the care in the formulation of the questionnaire. Because discussions are intentionally excluded from the process, the experts have no mechanism for resolving ambiguous questions. Furthermore, it is not necessarily true that a group consensus will ever be reached. An interesting case study of a successful application of the Delphi method can be found in Helmer and Rescher (1959).

2.4 OBJECTIVE FORECASTING METHODS

Objective forecasting methods are those in which the forecast is derived from an analysis of data. A **time series** method is one that uses only past values of the phenomenon we are predicting. **Causal models** are ones that use data from sources other than the series being predicted; that is, there may be other variables with values that are *linked* in some way to what is being forecasted. We discuss these first.

Causal Models

Let Y represent the phenomenon we wish to forecast and $X_1, X_2, \ldots, X_n$ be n variables that we believe to be related to Y. Then a causal model is one in which the forecast for Y is some function of these variables, say,

$$Y = f(X_1, X_2, \ldots, X_n).$$

Econometric models are special causal models in which the relationship between Y and $(X_1, X_2, \ldots, X_n)$ is linear. That is,

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n$$

for some constants $(\alpha_1, \ldots, \alpha_n)$. The method of least squares is most commonly used to find estimators for the constants. (We discuss the method in Appendix 2–B for the case of one independent variable.)

Let us consider a simple example of a causal forecasting model. A realtor is trying to estimate his income for the succeeding year. In the past he has found that his income is close to being proportional to the total number of housing sales in his territory. He also

has noticed that there has typically been a close relationship between housing sales and interest rates for home mortgages. He might construct a model of the form

$$Y_t = \alpha_0 + \alpha_1 X_{t-1},$$

where Y_t is the number of sales in year t and X_{t-1} is the interest rate in year $t - 1$. Based on past data he would then determine the least squares estimators for the constants α_0 and α_1. Suppose that the values of these estimators are currently $\alpha_0 = 385.7$ and $\alpha_1 = -1{,}878$. Hence, the estimated relationship between home sales and mortgage rates is

$$Y_t = 385.7 - 1{,}878X_{t-1},$$

where X_{t-1}, the previous year's interest rate, is expressed as a decimal. Then if the current mortgage interest rate is 10 percent, the model would predict that the number of sales the following year in his territory would be $385.7 - 187.8 = 197.9$, or about 198 houses sold.

Causal models of this type are common for predicting economic phenomena such as the gross national product (GNP) and the gross domestic product (GDP). Both MIT and the Wharton School of Business at the University of Pennsylvania have developed large-scale econometric models for making these predictions. Econometric prediction models are typically used by the economics and finance arms of the firm to forecast values of macroeconomic variables such as interest rates and currency exchange rates. Time series methods are more commonly used for operations planning applications.

Time Series Methods

Time series methods are often called naive methods, as they require no information other than the past values of the variable being predicted. *Time series* is just a fancy term for a collection of observations of some economic or physical phenomenon drawn at discrete points in time, usually equally spaced. The idea is that information can be inferred from the pattern of past observations and can be used to forecast future values of the series.

In time series analysis we attempt to isolate the patterns that arise most often. These include the following:

1. *Trend*. Trend refers to the tendency of a time series to exhibit a stable pattern of growth or decline. We distinguish between linear trend (the pattern described by a straight line) and nonlinear trend (the pattern described by a nonlinear function, such as a quadratic or exponential curve). When the pattern of trend is not specified, it is generally understood to be linear.

2. *Seasonality*. A seasonal pattern is one that repeats at fixed intervals. In time series we generally think of the pattern repeating every year, although daily, weekly, and monthly seasonal patterns are common as well. Fashion wear, ice cream, and heating oil exhibit a yearly seasonal pattern. Consumption of electricity exhibits a strong daily seasonal pattern.

3. *Cycles*. Cyclic variation is similar to seasonality, except that the length and the magnitude of the cycle may vary. One associates cycles with long-term economic variations (that is, business cycles) that may be present in addition to seasonal fluctuations.

Snapshot Application

ADVANCED FORECASTING, INC., SERVES THE SEMICONDUCTOR INDUSTRY

Advanced Forecasting, Inc. (AFI), is a Cupertino-based firm that specializes in providing forecasts for semiconductor sales and sales of related industries, such as semiconductor equipment and suppliers. The firm has had a history of accurately predicting the turning points in the sales patterns of semiconductors for more than a decade. Forecasts are determined from quantitative models (such as the ones discussed in this chapter). Although the actual models used are proprietary, forecasts are based on basic economic factors related to the semiconductor industry. According to the firm's founder Dr. Moshe Handelsman, the problem with most forecaster's predictions is that they are based on subjective opinions and qualitative data. AFI uses a mathematical model to derive its forecasts, which are not second guessed. While the firm is only a small player in the semiconductor forecasting arena, their success has been dramatic. They have consistently been able to predict major shifts in the market for semiconductors, which is a fundamental need for management. According to Jean-Philippe Dauvin, vice president and chief economist for SGS-Thomson Microelectronics: "Our top management pays more attention to Advanced Forecasting's predictions than to any other industry source." Accurate forecasts allow management to deal with important strategic issues such as when production capacity should be expanded, what personnel needs will be, and what the demands will be on marketing and sales. The success of AFI demonstrates that quantitative-based forecasting can provide consistently accurate forecasts and, over the long term, are far more reliable than subjective methods.

Source: Advanced Forecasting, Inc., Website, http://www.adv-forecast.com/afi/.

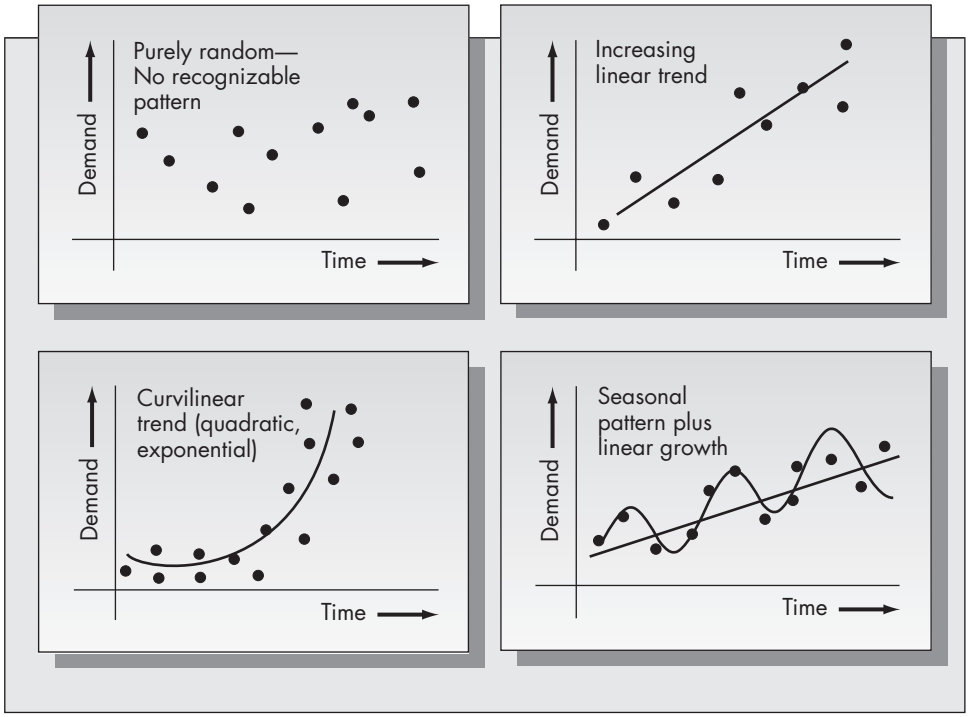
4. *Randomness.* A pure random series is one in which there is no recognizable pattern to the data. One can generate patterns purely at random that often appear to have structure. An example of this is the methodology of stock market chartists who impose forms on random patterns of stock market price data. On the other side of the coin, data that appear to be random could have a very definite structure. Truly random data that fluctuate around a fixed mean form what is called a horizontal pattern.

Examples of time series exhibiting some of these patterns are given in Figure 2–2.

Problems for Sections 2.1–2.4

1. Name the four components of time series (i.e., the four distinct patterns exhibited by time series).
2. What distinguishes seasonality from cycles in time series analysis?
3. What is the appropriate type of forecasting method to use in each of the following scenarios:
   a. Holiday Inn, Inc., is attempting to predict the demand next year for motel rooms, based on a history of demand observations.
   b. Standard Brands has developed a new type of outdoor paint. The company wishes to forecast sales based on new housing starts.
   c. IBM is trying to ascertain the cost of a stock-out of a critical tape drive component. They do so by sample survey of managers at various national spare parts centers. The surveys are sent back to the managers for a reassessment, and the process is repeated until a consensus is reached.

FIGURE 2–2
Time series patterns



4. Discuss the role of forecasting for the following functions of the firm:
   a. Marketing
   b. Accounting
   c. Finance
   d. Production
5. Distinguish between the following types of forecasts:
   a. Aggregate versus single item.
   b. Short-term versus long-term.
   c. Causal versus naive.
6. What is the advantage of the Delphi method over the jury of executive opinion method? What do these methods have in common?
7. Consider the problem of choosing an appropriate college to attend when you were a high school senior. What forecasting concerns did you have when you made that decision? In particular, list the short-term, intermediate-term, and long-term forecasts you might have considered in making your final decision. What objective sources of data might you have used to provide you with better forecasts in each case?
8. Discuss the following quotation from an inventory control manager: "It's not my fault we ran out of those parts. The demand forecast was wrong."
9. Discuss the following statement: "Economists are predicting that interest rates will continue to be under 10 percent for at least 15 years."

2.5 NOTATION CONVENTIONS

The following discussion deals with time series methods. Define $D_1, D_2, \ldots, D_t, \ldots$ as the observed values of demand during periods $1, 2, \ldots, t, \ldots$. We will assume throughout that $\{D_t, t \geq 1\}$ is the time series we would like to predict. Furthermore, we will assume that if we are forecasting in period t, then we have observed $D_t, D_{t-1}, \ldots$ but have not observed D_{t+1}.

Define $F_{t-\tau, t}$ as the forecast made in period t-τ for the demand in period t, where $\tau = 1,2,\ldots$. For the special case of $\tau = 1$, define $F_t = F_{t-1, t}$. That is, F_t is the forecast made in period t —1 for the demand in period t, after having observed $D_{t-1}, D_{t-2} \ldots$, but before having observed D_t . For the time being we will assume that all forecasts are one step ahead forecasts; that is, they are made for the demand in the next period. Multiple step ahead forecasts will be discussed later.

Finally, note that a time series forecast is obtained by applying some set of weights to past data. That is,

$$F_t = \sum_{n=1}^{\infty} a_n D_{t-n} \qquad \text{for some set of weights } a_1, a_2, \ldots .$$

Most of the time series methods discussed in this chapter are distinguished only by the choice of weights.

2.6 EVALUATING FORECASTS

Define the forecast error in period t, e_t, as the difference between the forecast value for that period and the actual demand for that period. For multiple-step-ahead forecasts,

$$e_t = F_{t-\tau, t} - D_t,$$

and for one-step-ahead forecasts,

$$e_t = F_t - D_t.$$

Let $e_1, e_2, \ldots, e_n$ be the forecast errors observed over n periods. Two common measures of forecast accuracy during these n periods are the mean absolute deviation (MAD) and the mean squared error (MSE), given by the following formulas:

$$\text{MAD} = (1/n)\sum_{i=1}^{n} |e_i|$$

$$\text{MSE} = (1/n)\sum_{i=1}^{n} e_i^2$$

Note that the MSE is similar to the variance of a random sample. The MAD is often the preferred method of measuring the forecast error because it does not require squaring. Furthermore, when forecast errors are normally distributed, as is generally assumed, an estimate of the standard deviation of the forecast error, σ_e, is given by 1.25 times the MAD.

Although the MAD and the MSE are the two most common measures of forecast accuracy, other measures are used as well. One that is not dependent on the magnitude of the values of demand is known as the mean absolute percentage error (MAPE) and is given by the formula

$$\text{MAPE} = \left[(1/n)\sum_{i=1}^{n} |e_i / D_i|\right] \times 100.$$

Example *2.1*

Artel, a manufacturer of static random access memories (SRAMs), has production plants in Austin, Texas, and Sacramento, California. The managers of these plants are asked to forecast production yields (measured in percent) one week ahead for their plants. Based on six weekly forecasts, the firm's management wishes to determine which manager is more successful at predicting his plant's yields. The results of their predictions are given in the following spreadsheet.

Week	P1	O1	|E1|	E1^2	|E1/O1|	P2	O2	|E2|	E2^2	|E2/O2|
1	92	88	4	16	0.04545	96	91	5	25	0.05495
2	87	88	1	1	0.01136	89	89	0	0	0
3	95	97	2	4	0.02062	92	90	2	4	0.02222
4	90	83	7	49	0.08434	93	90	3	9	0.03333
5	88	91	3	9	0.03297	90	86	4	16	0.04651
6	93	93	0	0	0	85	89	4	16	0.04494

Cell Formulas

Cell	Formula	Copied to
D2	=ABS(B2-C2)	D3:D7
E2	=ABS(D2/C2)	E3:E7
(Similar formulas and copies for cells H2 and I2)		
B10	=AVERAGE(D2:D7)	
B11	=AVERAGE(I2:I7)	
B13	=AVERAGE(E2:E7)	
B14	=AVERAGE(J2:J7)	
B16	=AVERAGE(F2:F7)	
B17	=AVERAGE(K2:K7)	

Interpret P1 as the forecast made by the manager of plant 1 at the beginning of each week, O1 as the yield observed at the end of each week in plant 1, and E1 as the difference between the predicted and the observed yields. The same definitions apply to plant 2.

Let us compare the performance of these managers using the three measures MAD, MSE, and MAPE as defined previously. To compute the MAD we simply average the observed absolute errors:

$$MAD_1 = 17/6 = 2.83$$

$$MAD_2 = 18/6 = 3.00.$$

Based on the MADs, the first manager has a slight edge. To compute the MSE in each case, square the observed errors and average the results to obtain

$$MSE_1 = 79/6 = 13.17$$

$$MSE_2 = 70/6 = 11.67.$$

The second manager's forecasts have a lower MSE than the first, even though the MADs go the other way. Why the switch? The reason that the first manager now looks worse is that the MSE is more sensitive to one large error than is the MAD. Notice that the largest observed error of 7 was incurred by manager 1.

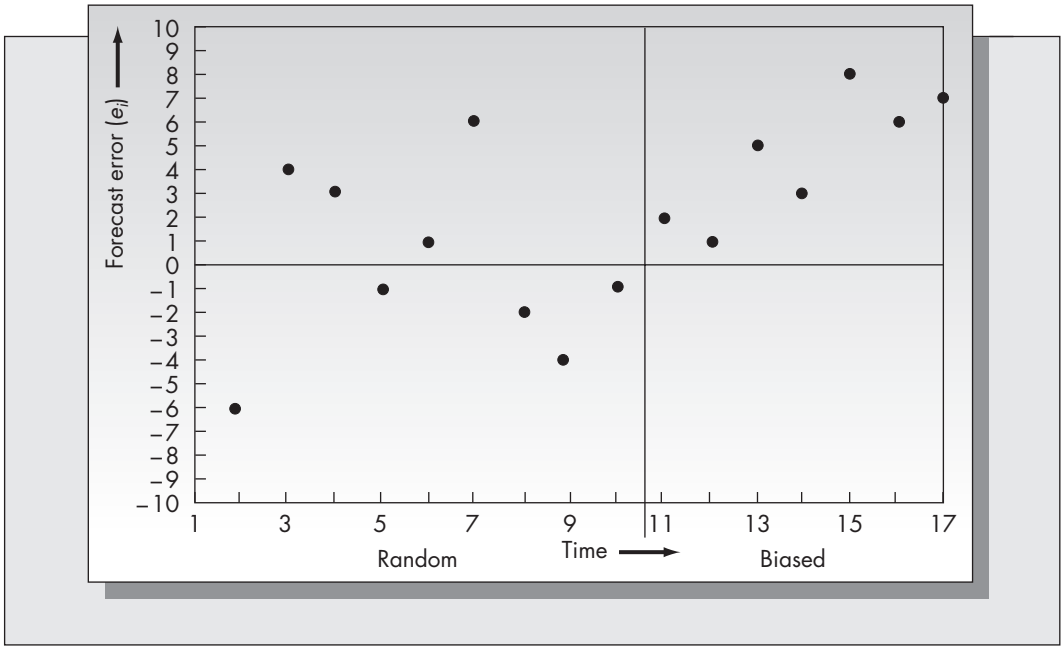
Let us now compare their performances based on the MAPE. To compute the MAPE we average the ratios of the errors and the observed yields:

$$MAPE_1 = .0325$$

$$MAPE_2 = .0336.$$

Using the MAD or the MAPE, the first manager has a very slight edge, but using the MSE, the second manager looks better. The forecasting abilities of the two managers would seem to be very similar. Who is declared the "winner" depends on which method of evaluation management chooses.

FIGURE 2–3
Forecast errors over time



A desirable property of forecasts is that they should be unbiased. Mathematically, that means that $E(e_i) = 0$. One way of tracking a forecast method is to graph the values of the forecast error e_i over time. If the method is unbiased, forecast errors should fluctuate randomly above and below zero. An example is presented in Figure 2–3.

An alternative to a graphical method is to compute the cumulative sum of the forecast errors, Σe_i. If the value of this sum deviates too far from zero (either above or below), it is an indication that the forecasting method is biased. At the end of this chapter we discuss a smoothing technique that also can be used to signal a bias in the forecasts. Statistical control charts also are used to identify unusually large values of the forecast error. (Statistical control charts are discussed in Chapter 12.)

Problems for Section 2.6

10. In *Dave Pelz's Short Game Bible*, the author attempts to characterize the skill of a golfer from a specific distance in terms of the ratio of the error in the shot and the intended shot distance. For example, a five iron that is hit 175 yards and is 20 yards off target would have an accuracy rating of 20/175 = .114, while a sand wedge hit 60 yards that is 10 yards off target would have an accuracy rating of 10/60 = .1667 (the lower the rating, the better). To what evaluation method discussed in this section is this most similar? Why does this evaluation method make more sense in golf than absolute or squared errors?

Snapshot Application

PFIZER BETS BIG ON FORECASTS OF DRUG SALES

One of Pfizer's most successful products is Lipitor. Lipitor, the first statin for reducing blood cholesterol, generated annual gross sales in excess of $12 billion. Not only was Lipitor the best selling statin, it was the most profitable pharmaceutical ever sold until its patent expired in November 2011. At that point, generic versions of the drug that cost less entered the market, and Lipitor's sales dropped dramatically. Pfizer was forced to lower prices to compete with generics and saw their profit margin sink considerably. To counteract this loss of a valuable revenue stream, a common strategy for drug companies is something called "evergreening." Evergreening means that the company plans on bringing a new and more effective drug to the market to counter the anticipated losses when the patent runs out. In Pfizer's case, this drug was torcetrapib. Torcetrapib not only decreased LDL (bad cholesterol), but increased the levels of HDL (good cholesterol). In order to make the transition as smooth as possible, Pfizer would have to have sufficient production capacity on line in 2011. Pfizer began production of torcetrapib as early as 2005 at a new $90 million plant in Loughberg, Ireland. This was reportedly a state of the art facility.

Unfortunately, things did not go as planned for Pfizer. After investing $800 million in the development of torcetrapib, Pfizer informed the United States Food and Drug Administration that it was suspending Phase 3 clinical trials. It was at that point that Pfizer discovered serious side effects with the drug. They had to make the painful decision to discontinue their efforts after investing nearly a billion dollars. To Pfizer's credit, it is rare that a drug must be pulled from the market as late as Phase 3. Still, Pfizer's failure to accurately forecast the outcomes of the trials resulted in serious losses.

Source: Bala et al. (2011). "Competition, Capacity, and 'Evergreening'". Unpublished manuscript. Indian School of Business.

11. A forecasting method used to predict can opener sales applies the following set of weights to the last five periods of data: .1, .1, .2, .2, .4 (with .4 being applied to the most recent observation). Observed values of can opener sales are

Period:	1	2	3	4	5	6	7	8
Observation:	18	22	26	33	14	28	30	52

Determine the following:

a. The one-step-ahead forecast for period 9.

b. The one-step-ahead forecast that was made for period 6.

12. A simple forecasting method for weekly sales of flash drives used by a local computer dealer is to form the average of the two most recent sales figures. Suppose sales for the drives for the past 12 weeks were

Week:	1	2	3	4	5	6	7	8	9	10	11	12
Sales:	86	75	72	83	132	65	110	90	67	92	98	73

a. Determine the one-step-ahead forecasts made for periods 3 through 12 using this method.

b. Determine the forecast errors for these periods.

c. Compute the MAD, the MSE, and the MAPE based on the forecast errors computed in part (*b*).

13. Two forecasting methods have been used to evaluate the same economic time series. The results are

Forecast from Method 1	Forecast from Method 2	Realized Value of the Series
223	210	256
289	320	340
430	390	375
134	112	110
190	150	225
550	490	525

Compare the effectiveness of these methods by computing the MSE, the MAD, and the MAPE. Do each of the measures of forecasting accuracy indicate that the same forecasting technique is best? If not, why?

14. What does the term *biased* mean in reference to a particular forecasting technique?
15. What is the estimate of the standard deviation of forecast error obtained from the data in Problem 12?

2.7 METHODS FOR FORECASTING STATIONARY SERIES

In this section we will discuss two popular techniques, moving averages and exponential smoothing, for forecasting stationary time series. A stationary time series is one in which each observation can be represented by a constant plus a random fluctuation. In symbols,

$$D_t = \mu + \epsilon_t,$$

where μ is an unknown constant corresponding to the mean of the series and ϵ_t is a random error with mean zero and variance σ^2.

The methods we consider in this section are more precisely known as single or simple exponential smoothing and single or simple moving averages. In addition, single moving averages also include weighted moving averages, which we do not discuss. For convenience, we will not use the modifiers single and simple in what follows. The meaning of the terms will be clear from the context.

Moving Averages

A simple but popular forecasting method is the method of moving averages. A moving average of order N is simply the arithmetic average of the most recent N observations. For the time being we restrict attention to one-step-ahead forecasts. Then F_t, the forecast made in period $t-1$ for period t, is given by

$$F_t = (1/N) \sum_{i=t-N}^{t-1} D_i = (1/N)(D_{t-1} + D_{t-2} + \cdots + D_{t-N}).$$

In words, this says that the mean of the N most recent observations is used as the forecast for the next period. We will use the notation MA(N) for N-period moving averages.

Example *2.2* Quarterly data for the failures of certain aircraft engines at a local military base during the last two years are 200, 250, 175, 186, 225, 285, 305, 190. Both three-quarter and six-quarter moving averages are used to forecast the numbers of engine failures. Determine the one-step-ahead

forecasts for periods 4 through 8 using three-period moving averages, and the one-step-ahead forecasts for periods 7 and 8 using six-period moving averages.

Solution

The three-period moving-average forecast for period 4 is obtained by averaging the first three data points.

$$F_4 = (1/3)(200 + 250 + 175) = 208.$$

The three-period moving-average forecast for period 5 is

$$F_5 = (1/3)(250 + 175 + 186) = 204.$$

The six-period moving-average forecast for period 7 is

$$F_7 = (1/6)(200 + 250 + 175 + 186 + 225 + 285) = 220.$$

Other forecasts are computed in a similar fashion. Arranging the forecasts and the associated forecast errors in a spreadsheet, we obtain

Quarter	Engine Failures	MA(3)	Error	MA(6)	Error
1	200				
2	250				
3	175				
4	186	208.33	22.33		
5	225	203.67	−21.33		
6	285	195.33	−89.67		
7	305	232.00	−73.00	220.17	−84.83
8	190	271.67	81.67	237.67	47.67

Cell Formulas

Cell	Formula	Copied to
C4	=1/3*SUM(B2:B4)	C5:C8
E7	=1/6*SUM(B2:B7)	E8
D5	=C5-B5	D6:D8
F8	=E8-B8	F9

An interesting question is, how does one obtain multiple-step-ahead forecasts? For example, suppose in Example 2.2 that we are interested in using MA(3) in period 3 to forecast for period 6. Because the moving-average method is based on the assumption that the demand series is stationary, the forecast made in period 3 for *any* future period will be the same. That is, the multiple-step-ahead and the one-step-ahead forecasts are identical (although the one-step-ahead forecast will generally be more accurate). Hence, the MA(3) forecast made in period 3 for period 6 is 208. In fact, the MA(3) forecast made in period 3 for any period beyond period 3 is 208 as well.

An apparent disadvantage of the moving-average technique is that one must recompute the average of the last N observations each time a new demand observation becomes available. For large N this could be tedious. However, recalculation of the full N-period average is not necessary every period, since

$$\begin{aligned} F_{t+1} &= (1/N)\sum_{i=t-N+1}^{t} D_i = (1/N)\left[D_t + \sum_{i=t-N}^{t-1} D_i - D_{t-N}\right] \\ &= F_t + (1/N)[D_t - D_{t-N}] \end{aligned}$$

This means that for one-step-ahead forecasting, we need only compute the difference between the most recent demand and the demand N periods old in order to update the forecast. However, we still need to keep track of all N past observations. Why?

Moving Average Lags behind the Trend

Consider a demand process in which there is a definite trend. For example, suppose that the observed demand is 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24. Consider the one-step-ahead MA(3) and MA(6) forecasts for this series.

Period	Demand	MA(3)	MA(6)
1	2		
2	4		
3	6		
4	8	4	
5	10	6	
6	12	8	
7	14	10	7
8	16	12	9
9	18	14	11
10	20	16	13
11	22	18	15
12	24	20	17

The demand and the forecasts for the respective periods are pictured in Figure 2–4. Notice that both the MA(3) and the MA(6) forecasts lag behind the trend. Furthermore, MA(6) has a greater lag. This implies that the use of simple moving averages is not an appropriate forecasting method when there is a trend in the series.

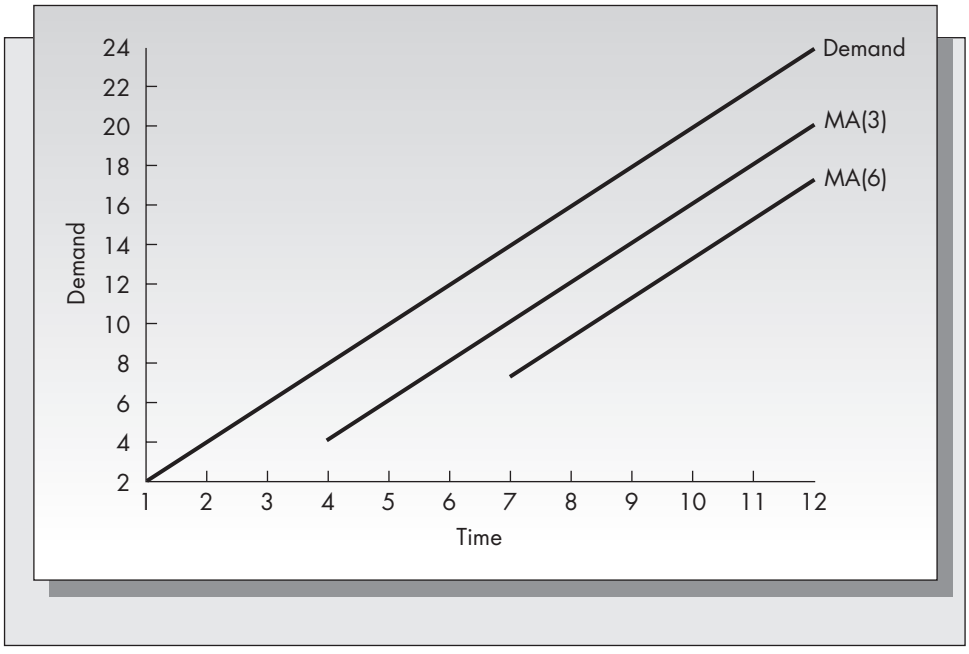


FIGURE 2–4
Moving-average forecasts lag behind a trend

Problems on Moving Averages

Problems 16 through 21 are based on the following data. Observations of the demand for a certain part stocked at a parts supply depot during the calendar year 2013 were

Month	Demand	Month	Demand
January	89	July	223
February	57	August	286
March	144	September	212
April	221	October	275
May	177	November	188
June	280	December	312

16. Determine the one-step-ahead forecasts for the demand for January 2014 using 3-, 6-, and 12-month moving averages.
17. Using a four-month moving average, determine the one-step-ahead forecasts for July through December 2013.
18. Using a four-month moving average, determine the two-step-ahead forecast for July through December 2013. (Hint: The two-step-ahead forecast for July is based on the observed demands in February through May.)
19. Compute the MAD for the forecasts obtained in Problems 17 and 18. Which method gave better results? Based on forecasting theory, which method should have given better results?
20. Compute the one-step-ahead three-month and six-month moving-average forecasts for July through December. What effect does increasing N from 3 to 6 have on the forecasts?
21. What would an MA(1) forecasting method mean? Compare the accuracy of MA(1) and MA(4) forecasts for July through December 2013.

Exponential Smoothing

Another very popular forecasting method for stationary time series is exponential smoothing. The current forecast is the weighted average of the last forecast and the current value of demand. That is,

$$\text{New forecast} = \alpha(\text{Current observation of demand}) + (1 - \alpha)(\text{Last forecast}).$$

In symbols,

$$F_t = \alpha D_{t-1} + (1 - \alpha)F_{t-1},$$

where $0 < \alpha \leq 1$ is the smoothing constant, which determines the relative weight placed on the current observation of demand. Interpret $(1 - \alpha)$ as the weight placed on past observations of demand. By a simple rearrangement of terms, the exponential smoothing equation for F_t can be written

$$\begin{aligned} F_t &= F_{t-1} - \alpha(F_{t-1} - D_{t-1}) \\ &= F_{t-1} - \alpha e_{t-1}. \end{aligned}$$

Written this way, we see that exponential smoothing can be interpreted as follows: the forecast in any period t is the forecast in period $t - 1$ minus some fraction of the observed forecast error in period $t - 1$. Notice that if we forecast high in period $t - 1$, e_{t-1} is positive and the adjustment is to decrease the forecast. Similarly, if we forecast low in period $t - 1$, the error is negative, and the adjustment is to increase the current forecast.

As before, F_t is the one-step-ahead forecast for period t made in period $t - 1$. Notice that since

$$F_{t-1} = \alpha D_{t-2} + (1 - \alpha)F_{t-2},$$

we can substitute above to obtain

$$F_t = \alpha D_{t-1} + \alpha(1 - \alpha)D_{t-2} + (1 - \alpha)^2 F_{t-2}.$$

We can now substitute for F_{t-2} in the same fashion. If we continue in this way, we obtain the infinite expansion for F_t,

$$F_t = \sum_{i=0}^{\infty} \alpha(1 - \alpha)^i D_{t-i-1} = \sum_{i=0}^{\infty} a_i D_{t-i-1},$$

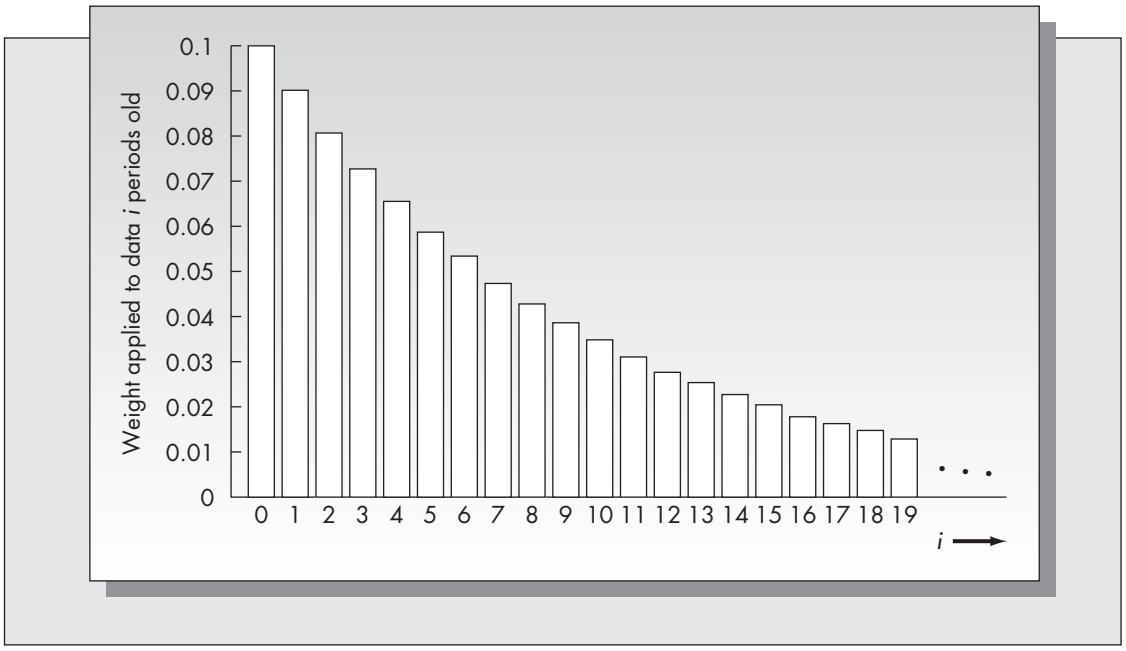
where the weights are $a_0 > a_1 > a_2 > \cdots > a_i = \alpha(1 - \alpha)^i$, and

$$\sum_{i=0}^{\infty} a_i = \sum_{i=0}^{\infty} \alpha(1 - \alpha)^i = \alpha \sum_{i=0}^{\infty} (1 - \alpha)^i = \alpha \times 1/[1 - (1 - \alpha)] = 1.$$

Hence, exponential smoothing applies a declining set of weights to all past data. The weights are graphed as a function of i in Figure 2–5.

In fact, we could fit the continuous exponential curve $g(i) = \alpha \exp(-\alpha i)$ to these weights, which is why the method is called exponential smoothing. The smoothing constant α plays essentially the same role here as the value of N does in moving

FIGURE 2–5
Weights in exponential smoothing



averages. If α is large, more weight is placed on the current observation of demand and less weight on past observations, which results in forecasts that will react quickly to changes in the demand pattern but may have much greater variation from period to period. If α is small, then more weight is placed on past data and the forecasts are more stable.

When using an automatic forecasting technique to predict demand for a production application, stable forecasts (that is, forecasts that do not vary a great deal from period to period) are very desirable. Demand forecasts are used as the starting point for production planning and scheduling. Substantial revision in these forecasts can wreak havoc with employee work schedules, component bills of materials, and external purchase orders. For this reason, a value of α between .1 and .2 is generally recommended for production applications. (See, for example, Brown, 1962.)

Multiple-step-ahead forecasts are handled the same way for simple exponential smoothing as for moving averages; that is, the one-step-ahead and the multiple-step-ahead forecasts are the same.

Example *2.3*

Consider Example 2.2 in which moving averages were used to predict aircraft engine failures. The observed numbers of failures over a two-year period were 200, 250, 175, 186, 225, 285, 305, 190. We will now forecast using exponential smoothing. In order to get the method started, let us assume that the forecast for period 1 was 200. Suppose that $\alpha = .1$. The one-step-ahead forecast for period 2 is

$$F_2 = \alpha D_1 + (1 - \alpha)F_1 = (.1)(200) + (.9)(200) = 200.$$

Similarly,

$$F_3 = \alpha D_2 + (1 - \alpha)F_2 = (.1)(250) + (.9)(200) = 205.$$

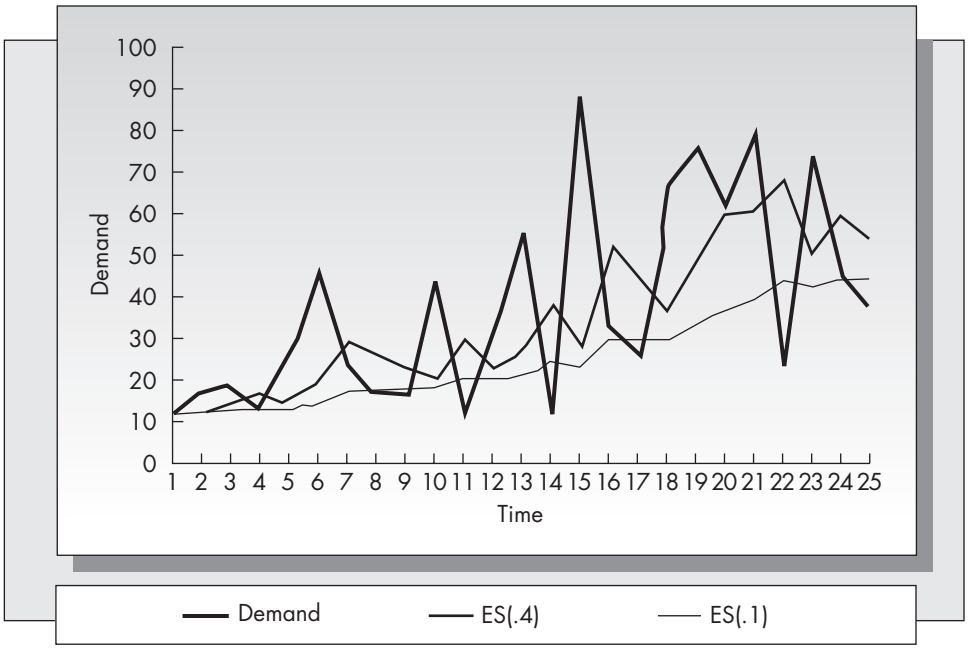
Other one-step-ahead forecasts are computed in the same fashion. The observed numbers of failures and the one-step-ahead forecasts for each quarter are the following:

Quarter	Failures	Forecast
1	200	200 (by assumption)
2	250	200
3	175	205
4	186	202
5	225	201
6	285	203
7	305	211
8	190	220

Notice the effect of the smoothing constant. Although the original series shows high variance, the forecasts are quite stable. Repeat the calculations with a value of $\alpha = .4$. There will be much greater variation in the forecasts.

Because exponential smoothing requires that at each stage we have the previous forecast, it is not obvious how to get the method started. We could assume that the initial forecast is equal to the initial value of demand, as we did in Example 2.3. However, this approach has a serious drawback. Exponential smoothing puts substantial weight on past observations, so the initial value of demand will have an unreasonably large effect on early forecasts. This problem can be overcome by

FIGURE 2–6
Exponential smoothing for different values of alpha



allowing the process to evolve for a reasonable number of periods (say 10 or more) and using the arithmetic average of the demand during those periods as the initial forecast.

In order to appreciate the effect of the smoothing constant, we have graphed a particularly erratic series in Figure 2–6, along with the resulting forecasts using values of $\alpha = .1$ and $\alpha = .4$. Notice that for $\alpha = .1$ the predicted value of demand results in a relatively smooth pattern, whereas for $\alpha = .4$, the predicted value exhibits significantly greater variation. Although smoothing with the larger value of α does a better job of tracking the series, the stability afforded by a smaller smoothing constant is very desirable for planning purposes.

Example *2.3* *(continued)*

Consider again the problem of forecasting aircraft engine failures. Suppose that we were interested in comparing the performance of MA(3) with the exponential smoothing forecasts obtained earlier (ES(.1)). The first period for which we have a forecast using MA(3) is period 4, so we will make the comparison for periods 4 through 8 only.

Quarter	Failures	Forecast	
1	200	200.00	
2	250	200.00	(by assumption)
3	175	205.00	
4	186	202.00	
5	225	200.40	
6	285	202.86	
7	305	211.07	
8	190	220.47	

Cell Formulas

Cell	Formula	Copied to
C3	=SUM(B$12*B2,(1-B$12)*C2)	C4:C9

The arithmetic average of the absolute errors, the MAD, is 57.6 for the three-period moving average and 49.2 for exponential smoothing. The respective values of the MSE are 4,215.6 and 3,458.4. Based on this comparison only, one might conclude that exponential smoothing is a superior method for this series. This is not necessarily true, however.

In Example 2.3 we compared exponential smoothing with $\alpha = .1$ and a moving average with $N = 3$. How do we know that those parameter settings are consistent? The MA(3) forecasts exhibit much greater variability than the ES(.1) forecasts do, suggesting that $\alpha = .1$ and $N = 3$ are not consistent values of these parameters.

Determining consistent values of α and N can be done in two ways. One is to equate the average age of the data used in making the forecast. A moving-average forecast consists of equal weights of $1/N$ applied to the last N observations. Multiplying the weight placed on each observation by its "age," we get the average age of data for moving averages as

$$\begin{aligned}\text{Average age} &= (1/N)(1 + 2 + 3 + \cdots + N) = (1/N)(N)(N + 1)/2 \\ &= (N + 1)/2.\end{aligned}$$

For exponential smoothing, the weight applied to data i periods old is $\alpha(1 - \alpha)^{i-1}$. Assume that we have an infinite history of past observations of demand. Hence, the average age of data in an exponential smoothing forecast is

$$\text{Average age} = \sum_{i=1}^{\infty} i\alpha(1 - \alpha)^{i-1} = 1/\alpha.$$

We omit the details of this calculation.

Equating the average age of data for the two methods, we obtain

$$\frac{N + 1}{2} = \frac{1}{\alpha},$$

which is equivalent to

$$\alpha = 2/(N + 1) \qquad \text{or} \qquad N = \frac{2 - \alpha}{\alpha}.$$

Hence, we see that we would have needed a value of $N = 19$ for $\alpha = .1$ or a value of $\alpha = .5$ for $N = 3$ in order for the methods to be consistent in the sense of average age of data.

In Appendix 2–A of this chapter, we derive the mean and variance of the forecast error for both moving averages and exponential smoothing in terms of the variance of each individual observation, assuming that the underlying demand process is stationary. We show that both methods are unbiased; that is, the expected value of the forecast error is zero. Furthermore, by equating the expressions for the variances of the forecast error, one obtains the same relationship between α and N as by equating the average age of data. This means that if both exponential smoothing and moving averages are used to predict the same stationary demand pattern, forecast errors are normally distributed, and $\alpha = 2/(N + 1)$, then both methods will have exactly the same distribution of forecast errors. (However, this does *not* mean that the forecasts obtained by the two methods are the same.)

Multiple-Step-Ahead Forecasts

Thus far, we have talked only about one-step-ahead forecasts. That is, we have assumed that a forecast in period t is for the demand in period $t + 1$. However, there are cases where we are interested in making a forecast for more than one step ahead. For example, a retailer planning for the Christmas season might need to make a forecast for December sales in June in order to have enough time to prepare. Since the underlying model assumed for both moving averages and exponential smoothing is stationary (i.e., not changing in time), the one-step-ahead and multiple-step-ahead forecasts for

moving averages and exponential smoothing are the same. That is, a forecast made in June for July sales is the same as a forecast made in June for December sales. (In the case of the retailer, the assumption of stationarity would probably be wrong, since December sales would likely be greater than a typical month's sales. That would suggest that these methods would *not* be appropriate in this case.)

Comparison of Exponential Smoothing and Moving Averages

There are several similarities and several differences between exponential smoothing and moving averages.

Similarities

1. Both methods are derived with the assumption that the underlying demand process is stationary (that is, can be represented by a constant plus a random fluctuation with zero mean). However, we should keep in mind that although the methods are appropriate for stationary time series, we don't necessarily believe that the series are stationary forever. By adjusting the values of N and α we can make the two methods more or less responsive to shifts in the underlying pattern of the data.

2. Both methods depend on the specification of a single parameter. For moving averages the parameter is N, the number of periods in the moving average, and for exponential smoothing the parameter is α, the smoothing constant. Small values of N or large values of α result in forecasts that put greater weight on current data, and large values of N and small values of α put greater weight on past data. Small N and large α may be more responsive to changes in the demand process, but will result in forecast errors with higher variance.

3. Both methods will lag behind a trend if one exists.

4. When $\alpha = 2/(N + 1)$, both methods have the same distribution of forecast error. This means that they should have roughly the same level of accuracy, but it does *not* mean that they will give the same forecasts.

Differences

1. The exponential smoothing forecast is a weighted average of *all* past data points (as long as the smoothing constant is strictly less than 1). The moving-average forecast is a weighted average of only the last N periods of data. This can be an important advantage for moving averages. An outlier (an observation that is not representative of the sample population) is washed out of the moving-average forecast after N periods, but remains forever in the exponential smoothing forecast.

2. In order to use moving averages, one must save all N past data points. In order to use exponential smoothing, one need only save the last forecast. This is the most significant advantage of the exponential smoothing method and one reason for its popularity in practice. In order to appreciate the consequence of this difference, consider a system in which the demand for 300,000 inventory items is forecasted each month using a 12-month moving average. The forecasting module alone requires saving $300{,}000 \times 12 = 3{,}600{,}000$ pieces of information. If exponential smoothing were used, only 300,000 pieces of information need to be saved. This issue is less important today than it has been, as the cost of information storage has decreased enormously in recent years. However, it is still easier to manage a system that requires less data. It is primarily for this reason that exponential smoothing appears to be more popular than moving averages for production-planning applications.

Problems for Section 2.7

22. Handy, Inc., produces a solar-powered electronic calculator that has experienced the following monthly sales history for the first four months of the year, in thousands of units:

January	23.3	March	30.3
February	72.3	April	15.5

    a. If the forecast for January was 25, determine the one-step-ahead forecasts for February through May using exponential smoothing with a smoothing constant of $\alpha = .15$.
    b. Repeat the calculation in part (*a*) for a value of $\alpha = .40$. What difference in the forecasts do you observe?
    c. Compute the MSEs for the forecasts you obtained in parts (*a*) and (*b*) for February through April. Which value of α gave more accurate forecasts, based on the MSE?
23. Compare and contrast exponential smoothing when α is small (near zero) and when α is large (near 1).
24. Observed weekly sales of ball peen hammers at the town hardware store over an eight-week period have been 14, 9, 30, 22, 34, 12, 19, 23.
    a. Suppose that three-week moving averages are used to forecast sales. Determine the one-step-ahead forecasts for weeks 4 through 8.
    b. Suppose that exponential smoothing is used with a smoothing constant of $\alpha = .15$. Find the exponential smoothing forecasts for weeks 4 through 8. [To get the method started, use the same forecast for week 4 as you used in part (*a*).]
    c. Based on the MAD, which method did better?
    d. What is the exponential smoothing forecast made at the end of week 6 for the sales in week 12?
25. Determine the following:
    a. The value of α consistent with $N = 6$ in moving averages.
    b. The value of N consistent with $\alpha = .05$.
    c. The value of α that results in a variance of forecast error, σ_e^2, 10 percent higher than the variance of each observation, σ^2 (refer to the formulas derived in Appendix 2–A).
26. Referring to the data in Problem 22, what is the exponential smoothing forecast made at the end of March for the sales in July? Assume $\alpha = .15$.
27. For the data for Problems 16 through 21, use the arithmetic average of the first six months of data as a baseline to initialize the exponential smoothing.
    a. Determine the one-step-ahead exponential smoothing forecasts for August through December, assuming $\alpha = .20$.
    b. Compare the accuracy of the forecasts obtained in part (*a*) with the one-step-ahead six-month moving-average forecasts determined in Problem 20.
    c. Comment on the reasons for the result you obtained in part (*b*).

Snapshot Application

SPORT OBERMEYER SLASHES COSTS WITH IMPROVED FORECASTING[1]

Sport Obermeyer is a leading supplier in the U.S. fashion ski apparel market. The firm, was founded in 1950 by engineer/ski instructor Klaus Obermeyer. Virtually all the firm's offerings are redesigned annually to incorporate changes in style, fabrics, and colors. For more than 50 years, the firm was able to successfully meet demands by producing during the summer months after receiving firm orders from customers.

During the 1990s, things changed and problems developed. First, volumes increased. There was insufficient capacity among suppliers to produce the required volume in the summer. Second, the firm developed a complex global supply chain strategy (see Section 6.10) to reduce costs. A parka sold in the United States might be sewn in China from fabrics and parts from Japan, South Korea, and Germany. Together these changes lengthened the production lead time, thus requiring the firm to commit to production before orders were placed by customers.

The firm undertook several "quick response" initiatives to reduce lead times. These included encouraging some customers to place orders earlier, locating raw materials near the Far East production facility, and instituting an air freight system to expedite delivery from the Far East to its Denver distribution center. Even with these changes in place, the problem of stockouts and markdowns due to oversupply were not solved. The company still had to commit about half the production based on forecasts. In the fashion industry, there is often no statistical history on which to base forecasts, and forecast errors can be huge. Products that outsell original forecasts by a factor of 2 or undersell original forecasts by a factor of 10 are common.

Sport Obermeyer needed some help with forecasting to avoid expensive miscalculations. The customary procedure was to base the forecasts on a consensus of members of the buying committee. The problem with consensus forecasting is that the dominant personalities in a group carry more weight. A forecast obtained in this way might represent only the opinion of one person. To overcome this problem, the research team (Fisher et al., 1994) recommended that members of the committee supply *individual* forecasts.

The dispersion among individual forecasts turned out to be a reliable indicator of forecast accuracy. When committee members' forecasts were close, forecasts were more accurate. This provided a mechanism for signaling the products whose sales were likely to be poorly forecast. This did not solve the problem of poorly forecast items, but it allowed the firm to commit first to production of items whose forecasts were likely to be accurate. By the time production had to begin on the problem items, information on early sales patterns would be available.

The team noticed that retailers were remarkably similar. That meant that even if only the first 20 percent of orders for a product were in, that information could dramatically improve forecasts. Production plans for these "trouble" items could now be committed with greater confidence. In this way, the firm could separate products into two categories: reactive and nonreactive. The nonreactive items are those for which the forecast is likely to be accurate. These are produced early in the season. The reactive items are those whose forecasts are updated later in the season from early sales figures. The firm's experience was that stockout and markdown costs were reduced from 10.2 percent of sales to 1.8 percent of sales on items that could be produced reactively. Sport Obermeyer was able to produce 30 percent of its season's volume reactively and experienced a cost reduction of about 5 percent of sales.

What are the lessons here? One is that even in cases where there is no statistical history, statistical methodology can be successfully applied to improve forecasting accuracy. Another is not to assume that things should be done a certain way. Sport Obermeyer assumed that consensus forecasting was the best approach. In fact, by requiring the buying committee to reach a consensus, valuable information was being ignored. The *differences* among individual forecasts proved to be important.

[1]This application is based on the work of a team from the Wharton School and the Harvard Business School. The results are reported in Fisher et al. (1994).

2.8 TREND-BASED METHODS

Both exponential smoothing and moving-average forecasts will lag behind a trend if one exists. We will consider two forecasting methods that specifically account for a trend in the data: regression analysis and Holt's method. Regression analysis is a method that fits a straight line to a set of data. Holt's method is a type of double exponential smoothing that allows for simultaneous smoothing on the series and on the trend.

Regression Analysis

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n paired data points for the two variables X and Y. Assume that y_i is the observed value of Y when x_i is the observed value of X. Refer to Y as the dependent variable and X as the independent variable. We believe that a relationship exists between X and Y that can be represented by the straight line

$$\hat{Y} = a + bX.$$

Interpret $\hat{Y}$ as the predicted value of Y. The goal is to find the values of a and b so that the line $\hat{Y} = a + bX$ gives the best fit of the data. The values of a and b are chosen so that the sum of the squared distances between the regression line and the data points is minimized (see Figure 2–7). In Appendix 2–B we derive the optimal values of a and b in terms of the given data.

When applying regression analysis to the forecasting problem, the independent variable often corresponds to time and the dependent variable to the series to be forecast. Assume that $D_1, D_2, \ldots, D_n$ are the values of the demand at times $1, 2, \ldots, n$. Then it is shown in Appendix 2–B that the optimal values of a and b are given by

$$b = \frac{S_{xy}}{S_{xx}}$$

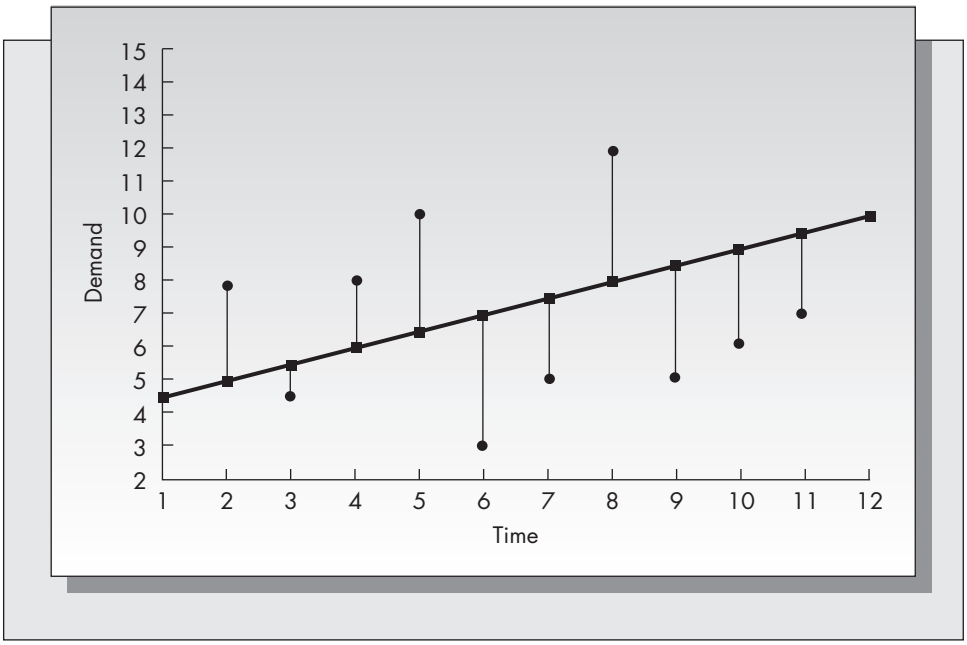


FIGURE 2–7
An example of a regression line

and

$$a = \overline{D} - b(n + 1)/2,$$

where

$$S_{xy} = n\sum_{i=1}^{n} iD_i - \frac{n(n + 1)}{2}\sum_{i=1}^{n} D_i,$$

$$S_{xx} = \frac{n^2(n + 1)(2n + 1)}{6} - \frac{n^2(n + 1)^2}{4},$$

and $\overline{D}$ is the arithmetic average of the observed demands during periods 1, 2, . . . , n.

Example *2.4*

We will apply regression analysis to the problem, treated in Examples 2.2 and 2.3, of predicting aircraft engine failures. Recall that the demand for aircraft engines during the last eight quarters was 200, 250, 175, 186, 225, 285, 305, 190. Suppose that we use the first five periods as a baseline in order to estimate our regression parameters. Then

$$\begin{aligned} S_{xy} &= 5[200 + (250)(2) + (175)(3) + (186)(4) + (225)(5)] \\ &\quad - [(5)(6)/2][200 + 250 + 175 + 186 + 225] \\ &= -70, \end{aligned}$$

$$S_{xx} = (25)(6)(11)/6 - (25)(36)/4 = 50.$$

Then

$$b = S_{xy}/S_{xx} = -70/50 = -7/5,$$

$$a = 207.2 - (-7/5)(3) = 211.4.$$

It follows that the regression equation based on five periods of data is

$$\hat{D}_t = 211.4 - (7/5)t.$$

$\hat{D}_t$ is the predicted value of demand at time t. We would use this regression equation to forecast from period 5 to any period beyond period 5. For example, the forecast made in period 5 for period 8 would be obtained by substituting $t = 8$ in the regression equation just given, which would result in the forecast $211.4 - (7/5)(8) = 200.2$. Note that if we were interested in forecasting in period 7 for period 8, then this regression equation would not be appropriate. We would have to repeat the entire calculation using the data from periods 1 through 7. In fact, one of the serious drawbacks of using regression for forecasting is that updating forecasts as new data become available is very cumbersome. (Note that Excel includes single and multiple linear regression capabilities.)

Problems for Section 2.8

28. Shoreline Park in Mountain View, California, has kept close tabs on the number of patrons using the park since its opening in January 1993. For the first six months of 2013, the following figures were recorded:

Month	Number of Patrons	Month	Number of Patrons
January	133	April	640
February	183	May	1,876
March	285	June	2,550

a. Draw a graph of these six data points. Assume that January = period 1, February = period 2, and so on. Using a ruler, "eyeball" the best straight-line fit of the data. Estimate the slope and intercept from your graph.

b. Compute the exact values of the intercept a and the slope b from the regression equations.

c. What are the forecasts obtained for July through December 2013 from the regression equation determined in part (*b*)?

d. Comment on the results you obtained in part (*c*). Specifically, how confident would you be about the accuracy of the forecasts that you obtained?

29. The Mountain View Department of Parks and Recreation must project the total use of Shoreline Park for calendar year 2014.

a. Determine the forecast for the total number of people using the park in 2014 based on the regression equation.

b. Determine the forecast for the same quantity using a six-month moving average.

c. Draw a graph of the most likely shape of the curve describing the park's usage by month during the calendar year, and predict the same quantity using your graph. Is your prediction closer to the answer you obtained for part (*a*) or part (*b*)? Discuss your results.

Double Exponential Smoothing Using Holt's Method

Holt's method is a type of double exponential smoothing designed to track time series with linear trend. The method requires the specification of two smoothing constants, α and β, and uses two smoothing equations: one for the value of the series (the intercept) and one for the trend (the slope). The equations are

$$S_t = \alpha D_t + (1 - \alpha)(S_{t-1} + G_{t-1}),$$
$$G_t = \beta(S_t - S_{t-1}) + (1 - \beta)G_{t-1}.$$

Interpret S_t as the value of the intercept at time t and G_t as the value of the slope at time t. The first equation is very similar to that used for simple exponential smoothing. When the most current observation of demand, D_t, becomes available, it is averaged with the prior forecast of the current demand, which is the previous intercept, S_{t-1}, plus 1 times the previous slope, G_{t-1}. The second equation can be explained as follows. Our new estimate of the intercept, S_t, causes us to revise our estimate of the slope to $S_t - S_{t-1}$. This value is then averaged with the previous estimate of the slope, G_{t-1}. The smoothing constants may be the same, but for most applications more stability is given to the slope estimate (implying $\beta \leq \alpha$).

The τ-step-ahead forecast made in period t, which is denoted by $F_{t,\,t+\tau}$, is given by

$$F_{t,\,t+\tau} = S_t + \tau G_t.$$

Example *2.5*

Let us apply Holt's method to the problem of developing one-step-ahead forecasts for the aircraft engine failure data. Recall that the original series was 200, 250, 175, 186, 225, 285, 305, 190. Assume that both α and β are equal to .1. In order to get the method started, we need estimates of both the intercept and the slope at time zero. Suppose that these are $S_0 = 200$ and $G_0 = 10$. Then we obtain

$$S_1 = (.1)(200) + (.9)(200 + 10) = 209.0$$
$$G \quad (.1)(209 \quad 200) + (.9)(10) \quad 9.9$$
$$S_2 = (.1)(250) + (.9)(209 + 9.9) = 222.0$$
$$G_2 = (.1)(222 - 209) + (.9)(9.9) = 10.2$$
$$S_3 = (.1)(175) + (.9)(222 + 10.2) = 226.5$$
$$G_3 = (.1)(226.5 - 222) + (.9)(10.2) = 9.6$$

and so on.

Comparing the one-step-ahead forecasts to the actual numbers of failures for periods 4 through 8, we obtain the following:

Period	Actual	S	G	Forecast	|Error|
1	200	200.00	10.00	200.00	0.00
2	250	209.00	9.90	218.90	31.10
3	175	222.01	10.21	232.22	57.22
4	186	226.50	9.64	236.14	50.14
5	225	231.12	9.14	240.26	15.26
6	285	238.74	8.98	247.72	37.28
7	305	251.45	9.36	260.81	44.19
8	190	265.23	9.80	275.02	85.02

$\alpha = 0.1$
$\beta = 0.1$
So = 200
Go = 10

Cell Formulas

Cell	Formula	Copied to
C3	SUM(B$12*B2,(1-B$12)*(C2+D2))	C4:C9
D3	SUM(B$13*(C3-C2),(1-B$13)*(D2))	D4:D9
E3	C3+D3	E4:E9
F3	ABS(B3-E3)	F4:F9

Averaging the numbers in the final column, we obtain a MAD of 46.4. Notice that this is lower than that for simple exponential smoothing or moving averages. Holt's method does better for this series because it is explicitly designed to track the trend in the data, whereas simple exponential smoothing and moving averages are not. Note that the forecasts in the given table are one-step-ahead forecasts. Suppose you needed to forecast the demand in period 2 for period 5. This forecast is $F_{2,5} = S_2 + (3)G_2 = 222 + (3)(10.2) = 252.6$.

The initialization problem also arises in getting Holt's method started. The best approach is to establish some set of initial periods as a baseline and use regression analysis to determine estimates of the slope and intercept using the baseline data.

Both Holt's method and regression are designed to handle series that exhibit trend. However, with Holt's method it is far easier to update forecasts as new observations become available.

More Problems for Section 2.8

30. For the data in Problem 28, use the results of the regression equation to estimate the slope and intercept of the series at the end of June. Use these numbers as the initial values of slope and intercept required in Holt's method. Assume that $\alpha = .15$, $\beta = .10$ for all calculations.
    a. Suppose that the actual number of visitors using the park in July was 2,150 and the number in August was 2,660. Use Holt's method to update the estimates of the slope and intercept based on these observations.
    b. What are the one-step-ahead and two-step-ahead forecasts that Holt's method gives for the number of park visitors in September and October?
    c. What is the forecast made at the end of July for the number of park attendees in December?
31. Because of serious flooding, the park was closed for most of December 1993. During that time only 53 people visited. Comment on the effect this observation would have on predictions of future use of the park. If you were in charge of forecasting the park's usage, how would you deal with this data point?

32. Discuss some of the problems that could arise when using either regression analysis or Holt's method for obtaining multiple-step-ahead forecasts.

2.9 METHODS FOR SEASONAL SERIES

This section considers forecasting methods for seasonal problems. A seasonal series is one that has a pattern that repeats every N periods for some value of N (which is at least 3). A typical seasonal series is pictured in Figure 2–8.

We refer to the number of periods before the pattern begins to repeat as the length of the season (N in the picture). Note that this is different from the popular usage of the word *season* as a time of year. In order to use a seasonal model, one must be able to specify the length of the season.

There are several ways to represent seasonality. The most common is to assume that there exists a set of multipliers c_t, for $1 \leq t \leq N$, with the property that $\sum c_t = N$. The multiplier c_t represents the average amount that the demand in the tth period of the season is above or below the overall average. For example, if $c_3 = 1.25$ and $c_5 = .60$, then, on average, the demand in the third period of the season is 25 percent above the average demand and the demand in the fifth period of the season is 40 percent below the average demand. These multipliers are known as seasonal factors.

Seasonal Factors for Stationary Series

In this part of the section we present a simple method of computing seasonal factors for a time series with seasonal variation and no trend. In the next part of this section we consider a method that is likely to be more accurate when there is also trend. Both methods require a minimum of two seasons of data.

The method is as follows:

1. Compute the sample mean of all the data.
2. Divide each observation by the sample mean. This gives seasonal factors for each period of observed data.
3. Average the factors for like periods within each season. That is, average all the factors corresponding to the

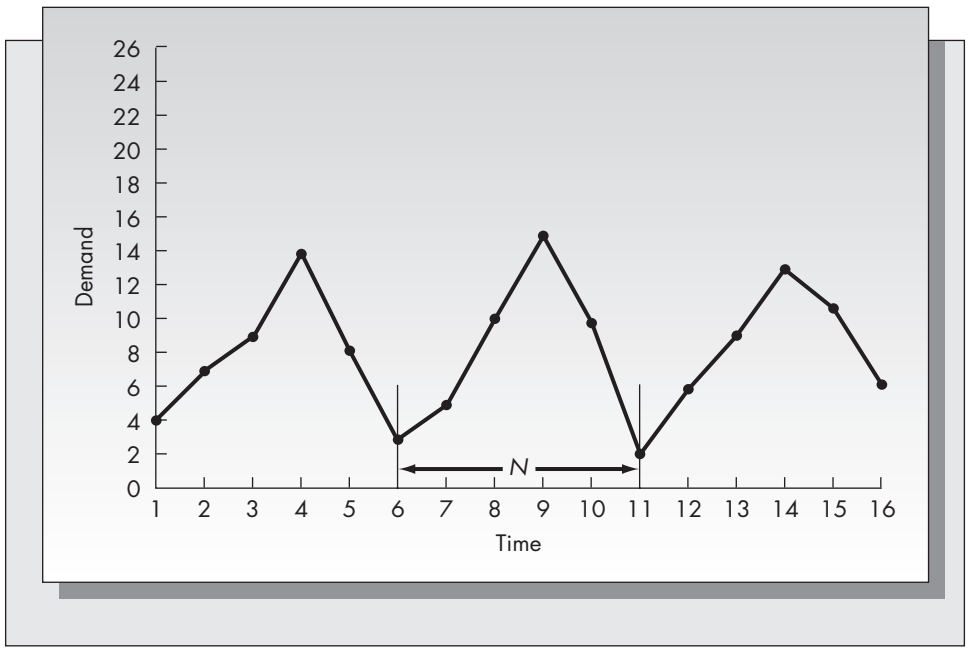


FIGURE 2–8
A seasonal demand series

second period of a season, and so on. The resulting averages are the N seasonal factors. They will always add to exactly N.

Example *2.6*

The County Transportation Department needs to determine usage factors by day of the week for a popular toll bridge connecting different parts of the city. In the current study, they are considering only working days. Suppose that the numbers of cars using the bridge each working day over the past four weeks were (in thousands of cars)

	Week 1	Week 2	Week 3	Week 4
Monday	16.2	17.3	14.6	16.1
Tuesday	12.2	11.5	13.1	11.8
Wednesday	14.2	15.0	13.0	12.9
Thursday	17.3	17.6	16.9	16.6
Friday	22.5	23.5	21.9	24.3

Find the seasonal factors corresponding to daily usage of the bridge.

Solution

To solve this problem we

1. Compute the arithmetic average of all of the observations (20 in this case).
2. Divide each observation by the average computed in step 1. This will give 20 factors.
3. Average factors corresponding to the same period of each season. That is, average all factors for Mondays, all factors for Tuesdays, and so on. This will give five seasonal factors: one for each day of the week. Note that these factors will sum to five.
4. Forecasts for the numbers of cars using the bridge by day of the week are obtained by multiplying the seasonal factors computed in step 3 by the average value computed in step 1.

These steps are summarized in the spreadsheet below.

	Week 1	Week 2	Week 3	Week 4
Monday	16.20	17.30	14.60	16.10
Tuesday	12.20	11.50	13.10	11.80
Wednesday	14.20	15.00	13.00	12.90
Thursday	17.30	17.60	16.90	16.60
Friday	22.50	23.50	21.90	24.30

Step 1: Compute the overall average of all of the observations

Average = 16.43 Formula = AVERAGE(B2:E6)

Step 2: Divide each observation by the Mean

	Week 1	Week 2	Week 3	Week 4
Monday	0.99	1.05	0.89	0.98
Tuesday	0.74	0.70	0.80	0.72
Wednesday	0.86	0.91	0.79	0.79
Thursday	1.05	1.07	1.03	1.01
Friday	1.37	1.43	1.33	1.48

Step 3: Average Factors corresponding to the same day of the week

	Seasonal Factor
Monday	0.98
Tuesday	0.74
Wednesday	0.84
Thursday	1.04
Friday	1.40

Note that these factors sum to exactly five.

Step 4: Build the forecasts by multiplying the mean, 16.425, by the appropriate factor

	Forecast
Monday	16.05
Tuesday	12.15
Wednesday	13.78
Thursday	17.10
Friday	23.05

Cell Formulas

Cell	Formula	Copied to
B10	AVERAGE(B2:E6)	
B14	B2/$B$10	B15:B18
C14	C2/$C$10	C15:C18
(Similar formulas and copies for cells D14 and E14)		
B22	AVERAGE(B14:E14)	B23:B26
B30	$B$10*B22	B31:B34

Note: Example 2.6

Determining the Deseasonalized Series

In cases where there is both seasonal variation and trend, a useful technique is to form the deseasonalized series by removing the seasonal variation from the original series. To illustrate this, consider the following simple example which consists of two seasons of data.

Example *2.7*

Period	Demand
1	10
2	20
3	26
4	17
5	12
6	23
7	30
8	22

Following the steps described above, the reader should satisfy himself or herself that we obtain the following four seasonal factors in this case:

0.550
1.075
1.400
0.975

To obtain the deseasonalized series, one simply divides each observation by the appropriate seasonal factor. For this example, that's 10/.550, 20/1.075, etc. In this case one obtains

Period	Demand	Deseasonalized Demand
1	10	18.182
2	20	18.605
3	26	18.571
4	17	17.436
5	12	21.818
6	23	21.395
7	30	21.429
8	22	22.564

Notice that the deseasonalized demand shows a clear trend. To forecast the deseasonalized series one could use any of the trend based methods discussed earlier in this chapter. Suppose that we fit a simple linear regression to this data where time is the independent variable as described in Section 2.8. Doing so one obtains the regression fit of this data as $y_t = 16.91 + 0.686t$. To forecast, one first applies the regression to forecast the deseasonalized series, and then re-seasonalizes by multiplying by the appropriate factor. For example, if one wishes to forecast for periods 9 through 12, the regression equation gives the following forecasts for the deseasonalized series: 23.08, 23.77, 24.46, 25.14. The final forecasts are obtained by multiplying by the appropriate seasonal factors, giving the final forecasts for periods 9 through 12 as 12.70, 25.55, 34.24, and 25.51.

Problems for Section 2.9

33. Sales of walking shorts at Hugo's Department Store in downtown Rolla appear to exhibit a seasonal pattern. The proprietor of the store, Wally Hugo, has kept careful records of the sales of several of his popular items, including walking shorts. During the past two years the monthly sales of the shorts have been

	Year 1	Year 2		Year 1	Year 2
Jan.	12	16	July	112	130
Feb.	18	14	Aug.	90	83
March	36	46	Sep.	66	52
April	53	48	Oct.	45	49
May	79	88	Nov.	23	14
June	134	160	Dec.	21	26

Assuming no trend in shorts sales over the two years, obtain estimates for the monthly seasonal factors.

34. A popular brand of tennis shoe has had the following demand history by quarters over a three-year period.

Snapshot Application

NATE SILVER PERFECTLY FORECASTS 2012 PRESIDENTIAL ELECTION

In a stunning victory, President Barack Obama won re-election in 2012 with 332 electoral votes versus Governor Romney's 206. This was in stark contrast to the Romney victory confidently predicted by Republican pundits just days before the election. Perhaps this was a result of a Gallup poll weeks before the election that showed Romney with a lead in the popular vote. However, American presidents are not elected by popular vote. They are elected by the Electoral College, which vote on a state by state basis. Even though several past Presidents have won the election while losing the popular vote, this was not the case in 2012.

A statistician named Nate Silver predicted the outcome of the election exactly. He based his forecasts on careful analysis of state by state (and even county by county) polls. His methodology had previously proven very successful in baseball. He also accurately predicted the outcome of the 2008 presidential election in which he was correct in 49 out of 50 states. But in 2012 he was correct in all fifty states, predicting not only an Obama victory with probability exceeding 90 percent, but predicting the number of electoral votes for each candidate exactly.

Silver's striking successes points to the importance in using sound analytics for forecasting. His is not an isolated case, however. Reports abound about such methods being successfully applied in many other contexts. (Note that a group from Princeton University, applying similar methods, was also able to accurately predict the outcome of the 2012 election.)

Year 1	Demand	Year 2	Demand	Year 3	Demand
1	12	1	16	1	14
2	25	2	32	2	45
3	76	3	71	3	84
4	52	4	62	4	47

a. Determine the seasonal factors for each quarter.

b. Based on the result of part (*a*), determine the deseasonalized demand series.

c. Predict the demand for each quarter of Year 4 for the deseasonalized series from a six-quarter moving average.

d. Using the results from parts (*a*) and (*c*), predict the demand for the shoes for each quarter of Year 4.

Winters's Method for Seasonal Problems

The moving-average method just described can be used to predict a seasonal series with or without a trend. However, as new data become available, the method requires that all seasonal factors be recalculated from scratch. Winters's method is a type of triple exponential smoothing, and this has the important advantage of being easy to update as new data become available.

We assume a model of the form

$$D_t = (\mu + G_t)c_t + \epsilon_t.$$

Interpret μ as the base signal or intercept at time $t = 0$ excluding seasonality, G_t as the trend or slope component, c_t as the multiplicative seasonal component in period t, and finally ϵ_t as the error term. Because the seasonal factor multiplies both the base level and the trend term, we are assuming that the underlying series has a form similar to that pictured in Figure 2–10.

Again assume that the length of the season is exactly N periods and that the seasonal factors are the same each season and have the property that $\Sigma\, c_t = N$. Three exponential

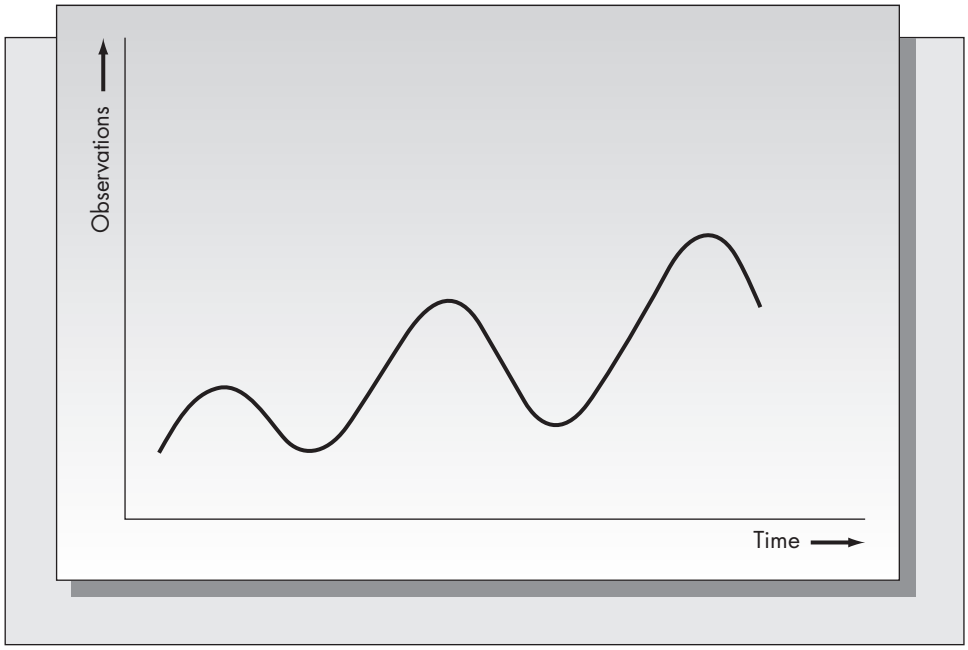


FIGURE 2–10
Seasonal series with increasing trend

smoothing equations are used each period to update estimates of deseasonalized series, the seasonal factors, and the trend. These equations may have different smoothing constants, which we will label α, β, and γ.

1. *The series.* The current level of the deseasonalized series, S_t, is given by

$$S_t = \alpha(D_t/c_{t-N}) + (1 - \alpha)(S_{t-1} + G_{t-1}).$$

Notice what this equation does. By dividing by the appropriate seasonal factor, we are deseasonalizing the newest demand observation. This is then averaged with the current forecast for the deseasonalized series, as in Holt's method.

2. *The trend.* The trend is updated in a fashion similar to Holt's method.

$$G_t = \beta[S_t - S_{t-1}] + (1 - \beta)G_{t-1}.$$

3. *The seasonal factors.*

$$c_t = \gamma(D_t/S_t) + (1 - \gamma)c_{t-N}.$$

The ratio of the most recent demand observation over the current estimate of the deseasonalized demand gives the current estimate of the seasonal factor. This is then averaged with the previous best estimate of the seasonal factor, c_{t-N}. Each time that a seasonal factor is updated, it is necessary to norm the most recent N factors to add to N.

Finally, the forecast made in period t for any future period $t + \tau$ is given by

$$F_{t,\,t+\tau} = (S_t + \tau G_t)c_{t+\tau-N}.$$

Note that this forecasting equation assumes that $t \leq N$. If $N < \tau \leq 2N$, the appropriate seasonal factor would be $c_{t+\tau-2N}$; if $2N < \tau \leq 3N$, the appropriate seasonal factor would be $c_{t+\tau-3N}$; and so on.

Initialization Procedure

In order to get the method started, we need to obtain initial estimates for the series, the slope, and the seasonal factors. Winters suggests that a minimum of two seasons of data be available for initialization. Let us assume that exactly two seasons of data are available; that is, $2N$ data points. Suppose that the current period is $t = 0$, so that the past observations are labeled $D_{-2N+1}, D_{-2N+2}, \ldots, D_0$.

1. Calculate the sample means for the two separate seasons of data.

$$V_1 = \frac{1}{N} \sum_{j=-2N+1}^{-N} D_j$$

$$V_2 = \frac{1}{N} \sum_{j=-N+1}^{0} D_j$$

2. Define $G_0 = (V_2 - V_1)/N$ as the initial slope estimate. If $m > 2$ seasons of data are available for initialization, then compute $V_1, \ldots, V_m$ as in step 1 and define $G_0 = (V_m - V_1)/[(m-1)N]$. If we locate V_1 at the center of the first season of data [at period $(-3N + 1)/2$] and V_2 at the center of the second season of data [at period $(-N + 1)/2$], then G_0 is simply the slope of the line connecting V_1 and V_2 (refer to Figure 2–11).

3. Set $S_0 = V_2 + G_0[(N - 1)/2]$. This estimates the value of the series at time $t = 0$. Note that S_0 is the value assumed by the line connecting V_1 and V_2 at $t = 0$ (see Figure 2–11).

4. *a.* The initial seasonal factors are computed for each period in which data are available and then averaged to obtain one set of seasonal factors. The initial seasonal factors are obtained by dividing each of the initial observations by the corresponding

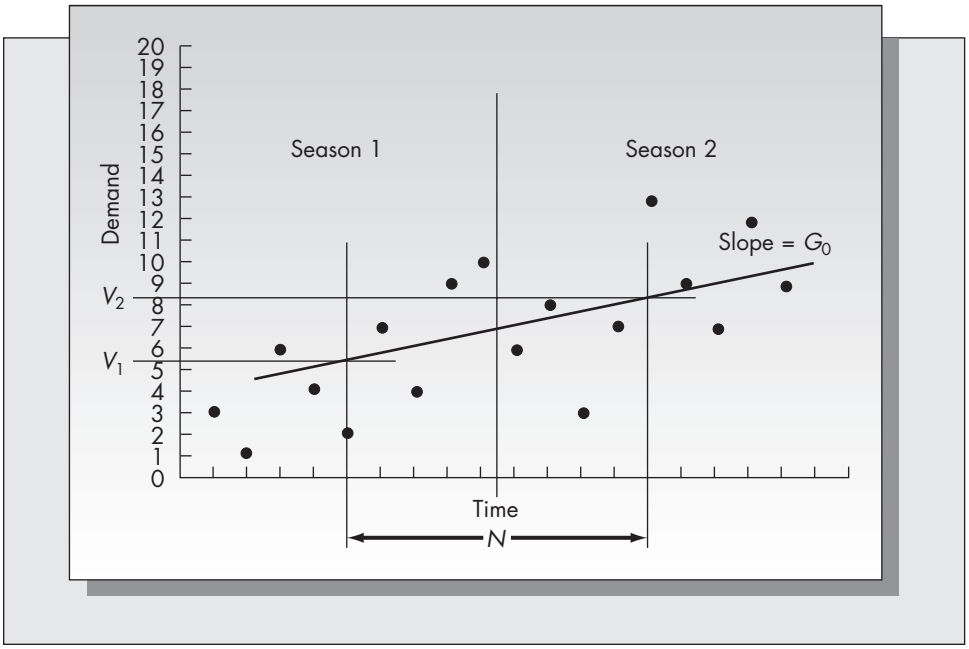


FIGURE 2–11
Initialization for Winters's method

point along the line connecting V_1 and V_2. This can be done graphically or by using the following formula:

$$c_t = \frac{D_t}{V_i - [(N+1)/2 - j]G_0} \qquad \text{for } -2N+1 \le t \le 0,$$

where $i = 1$ for the first season, $i = 2$ for the second season, and j is the period of the season. That is, $j = 1$ for $t = -2N + 1$ and $t = -N + 1$; $j = 2$ for $t = -2N + 2$ and $t = -N + 2$; and so on.

b. Average the seasonal factors. Assuming exactly two seasons of initial data, we obtain

$$c_{-N+1} = \frac{c_{-2N+1} + c_{-N+1}}{2}, \ldots, c_0 = \frac{c_{-N} + c_0}{2}.$$

c. Normalize the seasonal factors.

$$c_j = \left[\frac{c_j}{\sum_{i=0}^{-N+1} c_i} \right] \cdot N \qquad \text{for } -N+1 \le j \le 0.$$

This initialization procedure just discussed is the one suggested by Winters. It is not the only means of initializing the system. Seasonal factors could be determined by the method of moving averages discussed in the first part of this section. Another alternative would be to fit a linear regression to the baseline data and use the resulting slope and intercept values, as was done in Holt's method, to obtain S_0 and G_0. The seasonal factors would be obtained by dividing each demand observation in the baseline period by the corresponding value on the regression line, averaging like periods, and norming. The actual values of the initial estimates of the intercept, the slope, and the seasonal factors will be similar no matter which initialization scheme is employed.

Example *2.8*

Assume that the initial data set is the same as that of Example 2.7, in which centered moving averages were used to find the seasonal factors. Recall that we have two seasons of data: 10, 20, 26, 17, 12, 23, 30, 22. Then

$$V_1 = (10 + 20 + 26 + 17)/4 = 18.25,$$
$$V_2 = (12 + 23 + 30 + 22)/4 = 21.75,$$
$$G_0 = (21.75 - 18.25)/4 = 0.875,$$
$$S_0 = 21.75 + (0.875)(1.5) = 23.06.$$

The initial seasonal factors are computed as follows:

$$c_{-7} = \frac{10}{18.25 - (5/2 - 1)(0.875)} = 0.5904,$$
$$c_{-6} = \frac{20}{18.25 - (5/2 - 2)(0.875)} = 1.123.$$

The other factors are computed in a similar fashion. They are

$$c_{-5} = 1.391, \quad c_{-4} = 0.869, \quad c_{-3} = 0.5872,$$
$$c_{-2} = 1.079, \quad c_{-1} = 1.352, \quad c_0 = 0.9539.$$

We then average c_{-7} and c_{-3}, c_{-6} and c_{-2}, and so on, to obtain the four seasonal factors:

$$c_{-3} = 0.5888, \qquad c_{-2} = 1.1010, \qquad c_{-1} = 1.3720, \qquad c_0 = 0.9115.$$

Finally, norming the factors to ensure that the sum is 4 results in

$$c_{-3} = 0.5900, \qquad c_{-2} = 1.1100, \qquad c_{-1} = 1.3800, \qquad c_0 = 0.9200.$$

Notice how closely these factors agree with those obtained from the moving-average method.

Suppose that we wish to forecast the following year's demand at time $t = 0$. The forecasting equation is

$$F_{t,t+\tau} = (S_t + \tau G_t)c_{t+\tau-N},$$

which results in

$$\begin{aligned}
F_{0,1} &= (S_0 + G_0)c_{-3} = (23.06 + 0.875)(0.59) = 14.12,\\
F_{0,2} &= (S_0 + 2G_0)c_{-2} = [23.06 + (2)(0.875)](1.11) = 27.54,\\
F_{0,3} &= 35.44,\\
F_{0,4} &= 24.38.
\end{aligned}$$

Now, suppose that at time $t = 1$ we observe a demand of $D_1 = 16$. We now need to update our equations. Assume that $\alpha = .2$, $\beta = .1$, and $\gamma = .1$. Then

$$\begin{aligned}
S_1 &= \alpha(D_1/c_{-3}) + (1 - \alpha)(S_0 + G_0)\\
&= (0.2)(16/.59) + (0.8)(23.06 + 0.875) = 24.57,\\
G_1 &= \beta(S_1 - S_0) + (1 - \beta)G_0\\
&= (0.1)(24.57 - 23.06) + (0.9)(0.875) = 0.9385,\\
c_1 &= \gamma(D_1/S_1) + (1 - \gamma)c_{-3}\\
&= (0.1)(16/24.57) + (0.9)(0.59) = 0.5961.
\end{aligned}$$

At this point, we would renorm c_{-2}, c_{-1}, c_0, and the new value of c_1 to add to 4. Because the sum is 4.0061, it is close enough (the norming would result in rounding c_1 down to .59 once again).

Forecasting from period 1, we obtain

$$\begin{aligned}
F_{1,2} &= (S_1 + G_1)c_{-2} = (24.57 + 0.9385)(1.11) = 28.3144,\\
F_{1,3} &= (S_1 + 2G_1)c_{-1} = [24.57 + (2)(0.9385)](1.38) = 36.4969,
\end{aligned}$$

and so on.

Now suppose that we have observed one full year of demand, given by $D_1 = 16$, $D_2 = 33$, $D_3 = 34$, and $D_4 = 26$. Each time a new observation becomes available, the intercept, slope, and most current seasonal factor estimates are updated. One obtains

$$\begin{aligned}
S_2 &= 26.35, \qquad S_3 = 26.83, \qquad S_4 = 27.89,\\
G_2 &= 1.0227, \qquad G_3 = 0.9678, \qquad G_4 = 0.9770,\\
c_2 &= 1.124, \qquad c_3 = 1.369, \qquad c_4 = 0.9212.
\end{aligned}$$

As c_1, c_2, c_3, and c_4 sum to 4.01, normalization is not necessary. Suppose that we were interested in the forecast made in period 4 for period 10. The forecasting equation is

$$F_{t,t+\tau} = (S_t + \tau G_t)c_{t+\tau-2N},$$

which results in

$$F_{4,10} = (S_4 + 6G_4)c_2 = [27.89 + 6(0.9770)](1.124) = 37.94.$$

An important consideration is the choice of the smoothing constants α, β, and γ to be used in Winters's method. The issues here are the same as those discussed for simple exponential smoothing and Holt's method. Large values of the smoothing constants will result in more responsive but less stable forecasts. One method for setting α, β, and γ is to experiment with various values of the parameters that retrospectively give the best fit of previous forecasts to the observed history of the series. Because one must test many combinations of the three constants, the calculations are tedious. Furthermore, there is no guarantee that the best values of the smoothing constants based on past data will be the best values for future forecasts. The most conservative approach is to guarantee stable forecasts by choosing the smoothing constants to be between 0.1 and 0.2.

More Problems for Section 2.9

35. Consider the data for Problem 34.
    a. Using the data from Years 2 and 3, determine initial values of the intercept, slope, and seasonal factors for Winters's method.
    b. Assume that the observed demand for the first quarter of Year 4 was 18. Using $\alpha = .2$, $\beta = .15$, and $\gamma = .10$, update the estimates of the series, the slope, and the seasonal factors.
    c. What are the forecasts made at the end of the first quarter of Year 4 for the remaining three quarters of Year 4?
36. Suppose the observed quarterly demand for Year 4 was 18, 51, 86, 66. Compare the accuracy of the forecasts obtained for the last three quarters of Year 4 in Problems 34(*d*) and 35(*c*) by computing both the MAD and the MSE.
37. Determine updated estimates of the slope, the intercept, and the seasonal factors for the end of Year 4 based on the observations given in Problem 36. Using these updated estimates, determine the forecasts that Winters's method gives for all of Year 6 made at the end of Year 4. Use the values of the smoothing constants given in Problem 35(*b*).

2.10 BOX-JENKINS MODELS

The forecasting models introduced in this section are significantly more sophisticated than those previously discussed in this chapter. The goal is to present the basic concepts of Box-Jenkins analysis so that the reader can appreciate the power of these methods. However, an in-depth coverage is beyond the scope of this book. The methods are named for two well-known statisticians, George E. Box and Gwilym M. Jenkins formerly from the University of Wisconsin and University of Lancaster, respectively. The approach they developed is based on exploiting the autocorrelation structure of a time series. While Box-Jenkins methods are based on statistical relationships in a time series, much of the basic theory goes back to the famous book by Norbert Wiener (1949), and before.

Box-Jenkins models are also known as ARIMA models. ARIMA is an acronym for autoregressive integrated moving average. The autocorrelation function plays a central role in the development of these models, and is the feature that distinguishes ARIMA models from the other methods discussed in this chapter. As we have assumed throughout this chapter, denote the time series of interest as $D_1, D_2, \ldots$. We will assume initially that the series is stationary. That is, $E(D_i) = \mu$ and $\text{Var}(D_i) = \sigma^2$ for all $i = 1, 2, \ldots$. Practically speaking, *stationarity* means that there is no growth or

decline in the series, and variation remains relatively constant. It is important to note that stationarity does not imply independence. Hence, it is possible that values of D_i and D_j are dependent random variables when $i \neq j$ even though their marginal density functions are the same. It is this dependence we wish to exploit. (Note: A more precise way to characterize stationarity is that the joint distribution of $D_t, D_{t+1}, \ldots, D_{t+k}$ is the same as the joint distribution of $D_{t+m}, D_{t+m+1}, \ldots, D_{t+m+k}$ at any time t and pair of positive integers m and k.)

The assumption of stationarity implies that the marginal distributions of any two observations separated by the same time interval are the same. That is, D_t and D_{t+1} have the same joint distribution as D_{t+m} and D_{t+m+1} for any $m \geq 1$. This implies that the covariance of D_t and D_{t+1} is exactly the same as the covariance of D_{t+m} and D_{t+m+1}. Hence, the covariance of any two observations depends only on the number of periods separating them. In this context, the covariance is also known as the autocovariance, since we are comparing two values of the same series separated by a fixed lag.

Let Cov (D_{t+m}, D_{t+m+k}) be the covariance of D_{t+m} and D_{t+m+k} given by

$$\text{Cov}(D_{t+m}, D_{t+m+k}) = E(D_{t+m}D_{t+m+k}) - E(D_{t+m})E(D_{t+m+k}) \quad \text{for any integer } k \geq 1.$$

The correlation coefficient of these two random variables is given by

$$\rho_k = \frac{\text{Cov}(D_{t+m}, D_{t+m+k})}{\sqrt{\text{Var}(D_{t+m})}\sqrt{\text{Var}(D_{t+m+k})}}.$$

This is often referred to as the autocorrelation coefficient of lag k, since it refers to the correlation between all values of the series separated by k periods. These autocorrelation coefficients are typically computed for several values of k. It is these autocorrelation coefficients that will play the key role in building ARIMA models.

The autocorrelation coefficients are estimated from a history of the series. In order to guarantee reliable estimators, Box and Jenkins (1970) suggest that one have at least 72 data points of past history of the series. Hence, these models are only meaningful when one has a substantial and reliable history of the series being studied.

Estimating the Autocorrelation Function

Let $D_1, D_2, \ldots, D_n$ be a history of observations of a time series. The autocorrelation coefficient of lag k is estimated from the following formula:

$$r_k = \frac{\sum_{t=k+1}^{n} (D_t - \overline{D})(D_{t-k} - \overline{D})}{\sum_{t=1}^{n} (D_t - \overline{D})^2},$$

where $\overline{D}$ is the sample mean (that is, the average) of the observed values of the series. Refer to the r_k as sample autocorrelation coefficients. This calculation is typically done for 10 or 15 values of k. For most of the time series discussed earlier in the chapter, one identifies the appropriate patterns by just looking at a graph of the data. This is not the case here, however.

Example *2.9*

If observations are completely random (i.e., form a white noise process), then we would expect that there would be no significant autocorrelations among the observed values of the series. To test this, we generated a time series using the random number generator built into Excel. This series appears in Table 2–1. Each value is 100 times the RAND function. The reader can check that the sample autocorrelations for lags of 1 to 10 periods for these 36 observations are

TABLE 2–1
Time Series with 36 Values Generated by a Random Number Generator (White Noise Series)

Period	Value	Period	Value	Period	Value
1	42	13	47	25	88
2	93	14	52	26	73
3	17	15	28	27	60
4	5	16	58	28	56
5	38	17	41	29	49
6	2	18	47	30	51
7	67	19	48	31	59
8	66	20	50	32	80
9	11	21	81	33	40
10	65	22	93	34	60
11	88	23	45	35	20
12	91	24	24	36	35

$$
\begin{aligned}
r_1 &= \ \ 0.098 \\
r_2 &= -0.118 \\
r_3 &= \ \ 0.018 \\
r_4 &= -0.080 \\
r_5 &= \ \ 0.0752 \\
r_6 &= \ \ 0.006 \\
r_7 &= -0.270 \\
r_8 &= -0.207 \\
r_9 &= \ \ 0.117 \\
r_{10} &= \ \ 0.136
\end{aligned}
$$

If we had an infinite number of observations and a perfect white noise series, we would expect that all of the autocorrelations would be zero. However, since we only have a finite series, there will be statistical variation resulting in nonzero values of the autocorrelations. The question is whether these values are significantly different from zero. (The data from Table 2–1 appear in Figure 2–12 and the autocorrelations in Figure 2–13.)

FIGURE 2–12
Plot of data in Table 2–1

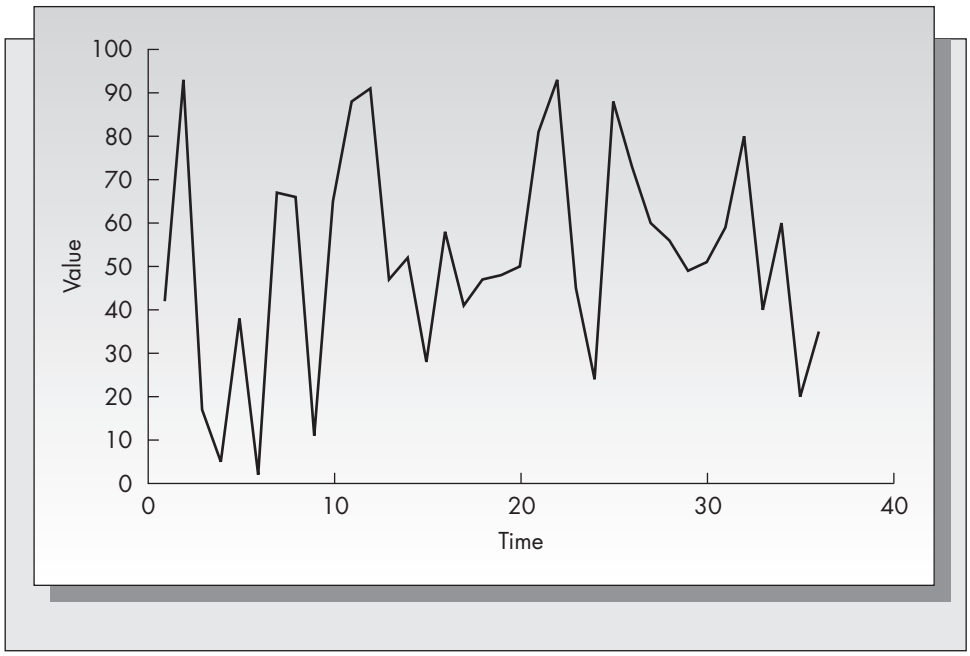
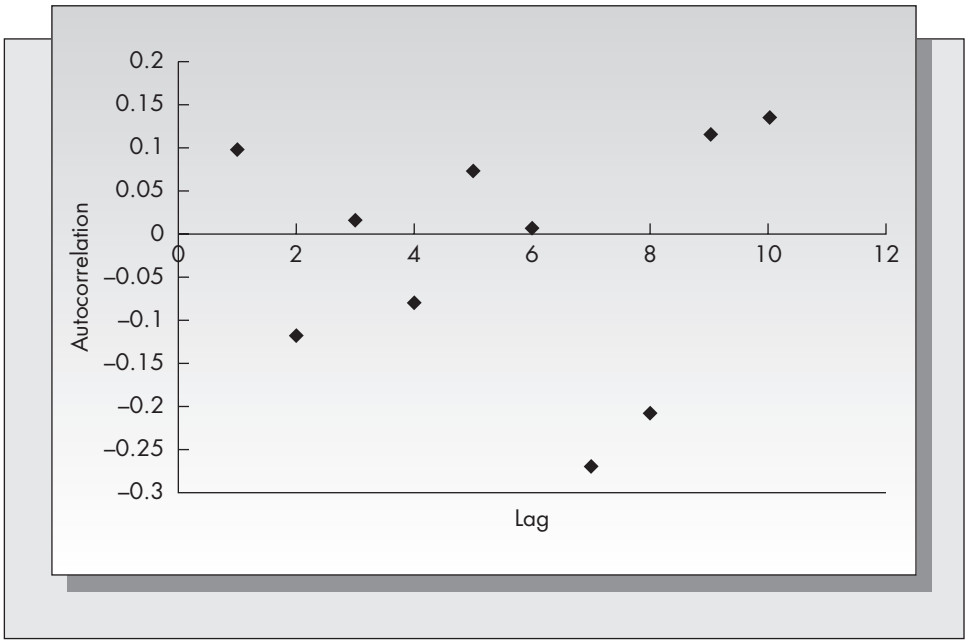


FIGURE 2–13
Plot of autocorrelations of series in Figure 2–12



Several statistical tests have been proposed to answer this question. One is the Box-Pierce Q statistic, computed from the formula

$$Q = n\sum_{k=1}^{h} r_k^2,$$

where h is the maximum length of the lag being considered, and n is the number of observations in the series. Under the null hypothesis that the series is white noise, the Q statistic has the chi-square distribution with $(h - m)$ degrees of freedom, where m is the number of parameters in the model that has been fitted to the data. This test can be applied to any set of data not necessarily fitted to a specific model by setting $m = 0$.

Applying the formula for the Q statistic to the preceding autocorrelations, we obtain a value of $Q = 6.62$. Comparing this to the critical values of the chi-square statistic in Table 2–2 with 10 degrees of freedom (df), we see that this value is substantially smaller than any value in the table (for example, the value for a right tail probability of .1 at 10 degrees of freedom is 15.99). Hence, we could not reject the null hypothesis that the data form a white noise process, and we conclude that the autocorrelations are not significant.

TABLE 2–2
Partial Table of Critical Values of the Chi-Square Statistic

	Tail Value Probability		
df	**0.1**	**0.05**	**0.01**
1	2.70	3.84	6.63
2	4.60	5.99	9.21
3	6.25	7.81	11.34
4	7.77	9.48	13.27
5	9.23	11.07	15.08
6	10.64	12.59	16.81
7	12.01	14.06	18.47
8	13.36	15.50	20.09
9	14.68	16.91	21.66
10	15.99	18.30	23.20

Of course, the real interest is in those cases where the autocorrelations are significant. The basic idea behind the method is to compare the graph of the autocorrelation function (known as the correlogram) to those of known processes to identify the appropriate model. (Note: In addition to considering autocorrelations, most texts on Box-Jenkins analysis also recommend examining the partial autocorrelations. We will not discuss partial autocorrelations here, but the reader should be aware that these also provide information about the underlying structure of the process.)

The Autoregressive Process

The autoregressive model is

$$D_t = a_0 + a_1 D_{t-1} + a_2 D_{t-2} + \cdots + a_p D_{t-p} + \epsilon_t,$$

where $a_0, a_1, \ldots, a_p$ are the linear regression coefficients and ϵ_t is the error term (generally assumed to be normal with mean 0 and variance σ^2 as earlier in the chapter). The reader familiar with linear regression will recognize this equation as being very similar to the standard regression equation with D_t playing the role of the dependent variable and $D_{t-1}, D_{t-2}, \ldots, D_{t-p}$ playing the role of the independent variables. Hence, the autoregressive model regresses the value of the series at time t on the values of the series at times $t - 1, t - 2, \ldots, t - p$. Note, however, that there is a fundamental difference between an autoregressive model and a simple linear regression, since in this case it is likely that the variables are correlated. We will use the notation AR(p) to represent this model.

Consider a basic AR(1) model,

$$D_t = a_0 + a_1 D_{t-1} + \epsilon_t.$$

In order for the process to be stable, we require $|a_1| < 1$. If $a_1 > 0$, it means that successive values of the series are positively correlated—that is, large values tend to be followed by large values, and small values tend to be followed by small values. This means that the series will be relatively smooth. If $a_1 < 0$, then the opposite is true, so the series will appear much spikier. The difference is illustrated in a realization of two AR(1) processes in Figure 2–14. (Figure 2–14 was generated in Excel using the built-in RAND function and the normal variate generator given in Problem 5.34 with $a_0 = 10$ and $\sigma = 30$.)

Of course, it is unlikely one can recognize an AR(1) process by simply examining a graph of the raw data. Rather, one would examine the autocorrelation function. It is easy to show that the autocorrelation function for an AR(1) process (see Nelson, 1973, page 39, for example) is

$$\rho_j = {a_1}^j.$$

The autocorrelation functions for the two cases illustrated in Figure 2–14 are given in Figure 2–15. If the sample autocorrelation function of a series has a pattern resembling one of those in Figure 2–15, it would suggest that an AR(1) process is appropriate. The theoretical autocorrelation functions for higher-order AR processes can be more complex. [In the case of AR(2), the patterns are either similar to one of the two pictured in Figure 2–15 or follow a damped sine wave.] In practice, one would rarely include more than one or two AR terms in the model. Determining the autocorrelation structure for higher-order AR processes is not difficult. One must solve a series of linear equations known as the Yule-Walker equations. We will not elaborate further here, but refer the interested reader to Box and Jenkins (1970).

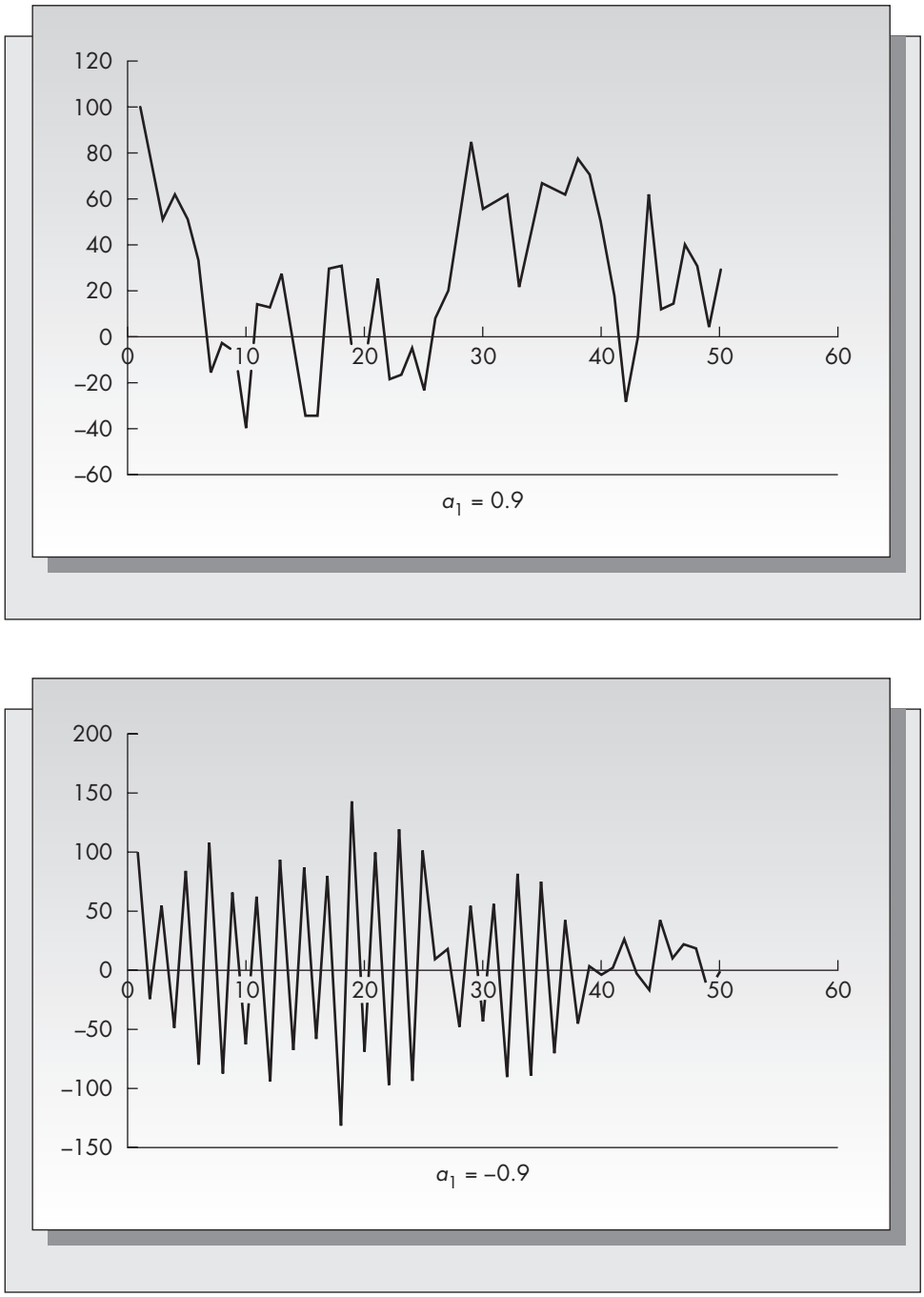
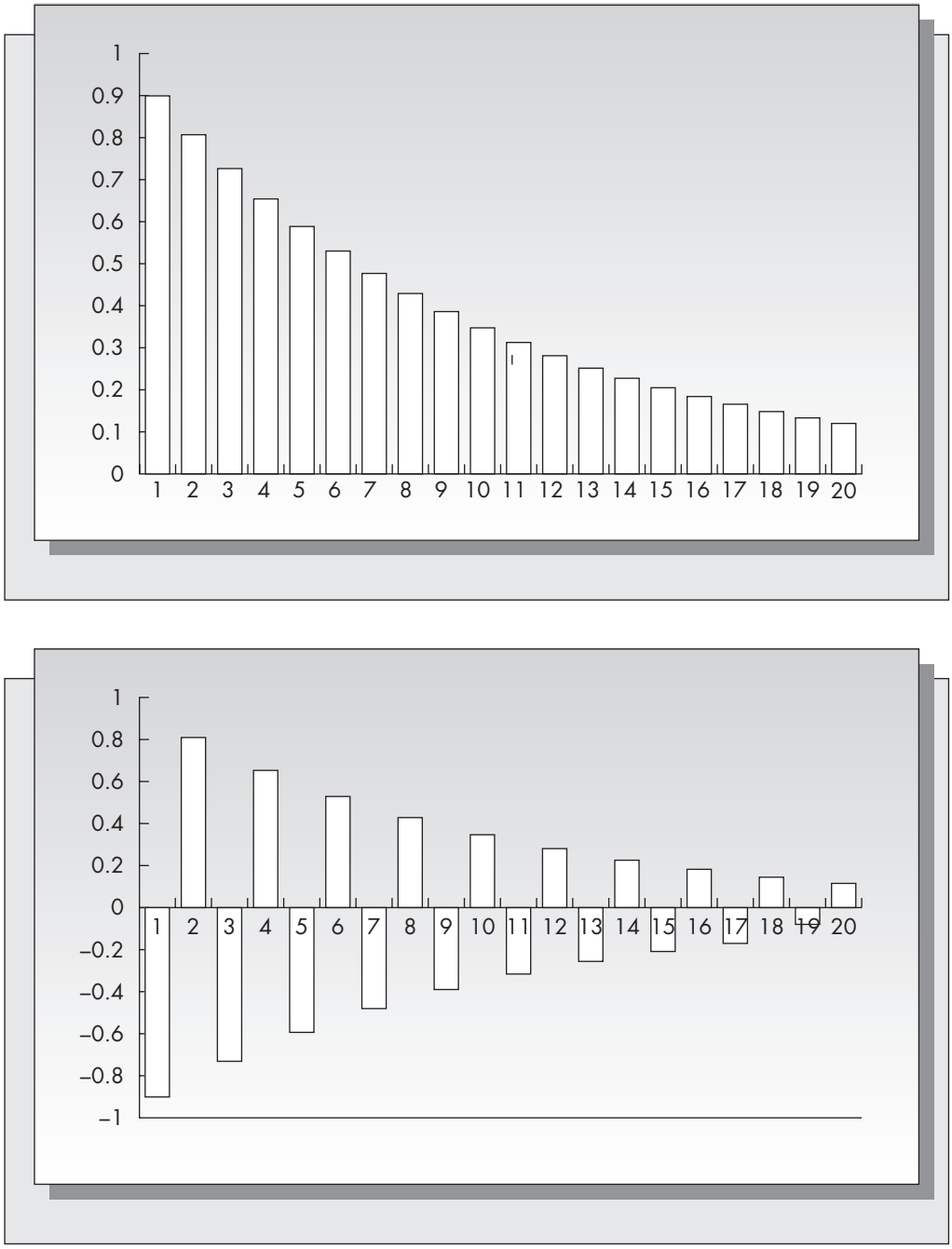


FIGURE 2–14
Realizations of an AR(1) process with $a_1 = 0.9$ and $a_1 = -0.9$, respectively.

The Moving-Average Process

The moving-average process provides another means of describing a stationary stochastic process used to model time series. The term *moving average* as used here should not be confused with the moving average discussed earlier in this chapter in Section 2.7. In that case, the moving average was an average of past values of the series. In this case, the moving average is a weighted average of past forecast errors.

FIGURE 2–15
Theoretical autocorrelation functions for AR(1) processes pictured in Figure 2–14.



The general moving-average process has the form

$$D_t = b_0 - b_1\epsilon_{t-1} - b_2\epsilon_{t-2} - \cdots - b_q\epsilon_{t-q} + \epsilon_t.$$

[The weights $b_1, b_2, \ldots, b_q$ are shown with negative signs by convention.] We will denote this model as MA(q). The intuition behind the moving-average process is not

as straightforward as the autoregressive process, but the two are related. Consider the first-order AR(1) process, $D_t = a_0 + a_1 D_{t-1} + \epsilon_t$. By back-substituting for D_{t-1}, $D_{t-2}, \ldots$, we see that the AR(1) process can also be written as

$$D_t = a_0 \sum_{i=0}^{\infty} a_1^i + a_1 \epsilon_{t-1} + a_1^2 \epsilon_{t-2} + \cdots + \epsilon_t,$$

which is easily recognized as an MA(∞) process.

The reader will better understand the power of MA processes when we examine the autocorrelation function. Consider first the simplest MA process, namely, an MA(1) process. In this case,

$$D_t = b_0 - b_1 \epsilon_{t-1} + \epsilon_t.$$

The autocorrelation structure for this case is very simple:

$$\rho_1 = \frac{-b_1}{1 + b_1^2},$$

$$\rho_2 = \rho_3 = \cdots = 0.$$

Hence, an MA(1) process has only one significant autocorrelation at lag 1. In general, MA(1) processes tend to have spiky patterns independent of the sign of b_1, because successive errors are assumed to be uncorrelated. Figure 2–16 shows realizations of an MA(1) process with b_1 equal to -0.9 and 0.9, respectively. Finding the autocorrelation structure for higher-order MA processes is a challenging mathematical problem, requiring the solution of a collection of nonlinear equations. Again, we refer the interested reader to Box and Jenkins (1970). The main characteristic identifying an MA(q) process is that only the first q autocorrelations are nonzero.

Mixtures: ARMA Models

The real power in Box-Jenkins methodology comes in being able to mix both AR and MA terms. Any model that contains one or more AR terms and one or more MA terms is known as an ARMA model, for autoregressive moving average. An ARMA(p, q) model contains p autoregressive terms and q moving-average terms. For example, one would write an ARMA(1,1) model as

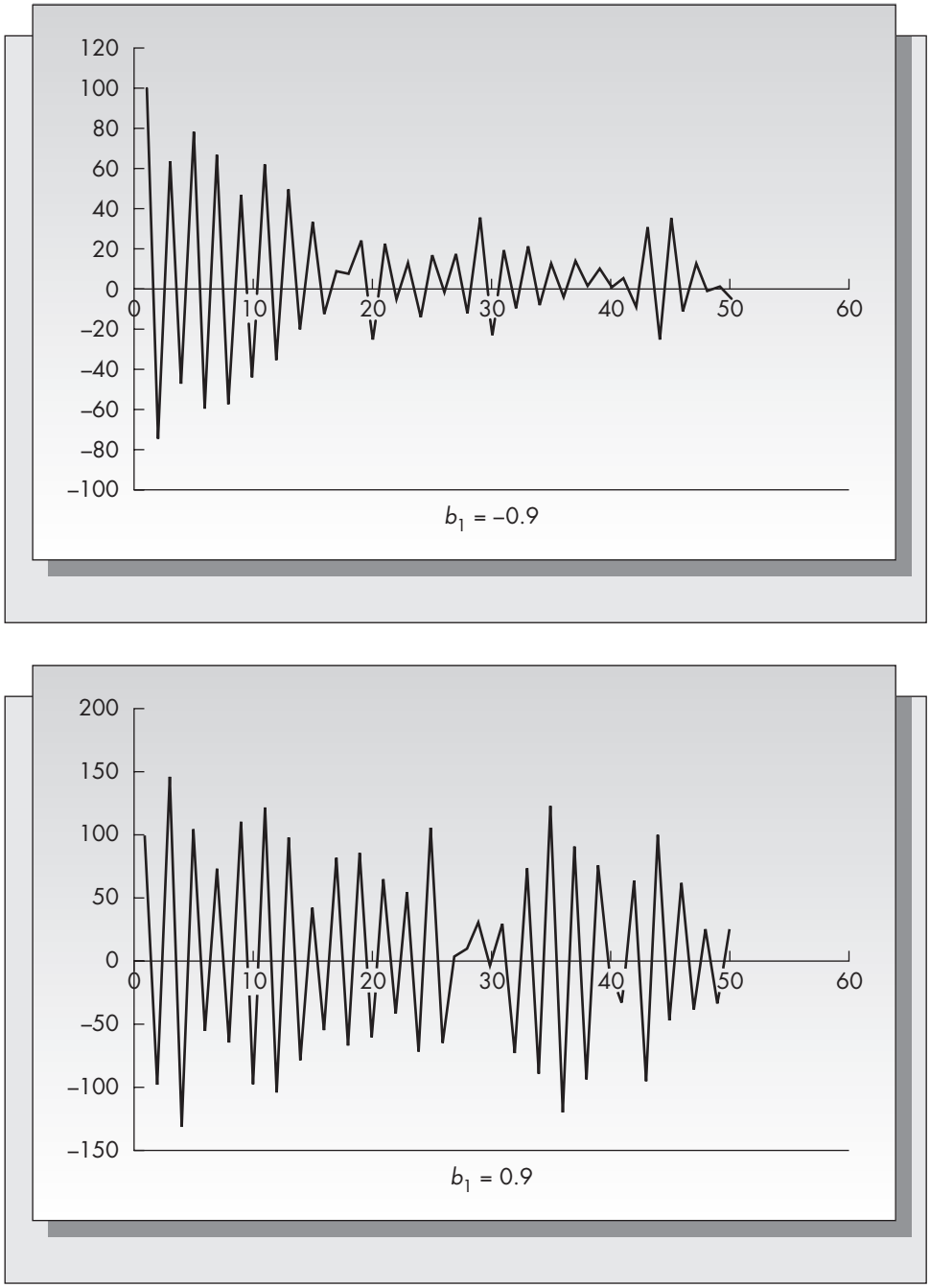
$$D_t = c + a_1 D_{t-1} - b_1 \epsilon_{t-1} + \epsilon_t.$$

The ARMA(1,1) model is quite powerful, and can describe many real processes accurately. It requires identification of the two parameters a_1 and b_1. The ARMA(1, 1) process is equivalent to an MA(∞) process and also equivalent to an AR(∞) process, thus showing the power that can be achieved with a parsimonious model. The autocorrelation function of the ARMA(1,1) process has characteristics of both the MA(1) and AR(1) processes. The autocorrelation at lag 1 is determined primarily by the MA(1) term, while autocorrelations at lags greater than 1 are determined by the AR(1) term.

ARIMA Models

Thus far we have assumed that the underlying stochastic process generating the time series is stationary. However, in many real problems, patterns such as trend or seasonality are present, which imply nonstationarity. The Box-Jenkins approach would be of limited utility if it were unable to address such situations. Fortunately, there is a simple technique for converting many nonstationary processes into stationary processes.

FIGURE 2–16
Realizations of an MA(1) process with $b_1 = -0.9$ and $b_1 = 0.9$, respectively



Consider first a process with a linear trend, such as the one pictured in Figure 2–7. Simple methods for dealing with linear trends were discussed in Section 2.8 of this chapter. How can a process with a linear trend be converted to one with no trend? The answer turns out to be surprisingly simple. Suppose our original process D_t has a linear trend. Consider the new process U_t given by

$$U_t = D_t - D_{t-1}.$$

The process U_t tracks the slope of the original process. If the original process had a linear trend, then the slope should be relatively constant, implying that U_t would be stationary. In the same way, if the original process increased or decreased according to a quadratic function, differencing the first difference process (forming a second difference) will result in a stationary process. Differencing is the discrete analogue of a derivative. Going from the process U_t back to the original process D_t requires summing values of U_t, which is the discrete analogue of integration. For that reason, when differencing is introduced, we use the term *integration* to describe it. An ARMA model based on data derived from differencing is denoted ARIMA, which stands for autoregressive integrated moving average.

A common notation is $U_t = \nabla D_t$. If two levels of differencing were required to achieve stationarity, then one would need a double summation to retrieve the original series. In the case of two levels of differencing,

$$\nabla^2 D_t = D_t - D_{t-1} - (D_{t-1} - D_{t-2}) = D_t - 2D_{t-1} + D_{t-2}.$$

Differencing can also be used to remove seasonality from a time series. Suppose that a seasonal pattern repeats every 12 months. Then defining

$$U_t = \nabla^{12} D_t = D_t - D_{t-12}$$

would result in a process with no seasonality.

An ARIMA process has three constants associated with it: p for the number of autoregressive terms, d for the order of differencing, and q for the number of moving-average terms. The general ARIMA process would be denoted ARIMA(p, d, q). Thus, for example, ARMA(1,1) can also be denoted ARIMA(1,0,1). While these parameters can be any nonnegative integers, it is very rare that any of the values of p, d, or q would exceed 2. Thus, virtually all the ARIMA models one finds in practice correspond to values of 0, 1, or 2 for the parameters p, d, and q. While this might seem limiting, these few cases cover an enormous range of practical forecasting scenarios.

It is important to note that observations are lost when differencing. For example, if one uses a single level of differencing ($U_t = D_t - D_{t-1}$), the first difference process, U_t, will have one less observation than the original series. Similarly, each level of differencing will reduce the sample size by 1. Seasonal differencing effectively reduces the data set by the length of the season. Another way to represent the differencing operation is via the backshift operator. That is,

$$BD_t = D_{t-1},$$

which means that one would represent the first difference process as

$$U_t = D_t - D_{t-1} = D_t - BD_t = (1 - B)D_t.$$

The backshift operator can also be used to simplify the notation for AR, MA, and ARMA models. Consider the simple AR(1) process given by $D_t = a_0 + a_1 D_{t-1} + \epsilon_t$. Writing this in the form $D_t - a_1 D_{t-1} = a_0 + \epsilon_t$ reduces the process to the alternate notation $(1 - a_1 B)D_t = a_0 + \epsilon_t$. Similarly, the reader should check that the MA(1) model using the backshift operator is $D_t = b_0 + (1 - b_1 B)e_t$.

Using ARIMA Models for Forecasting

Given an ARIMA model, how does one go about using it to provide forecasts of future values of the series? The approach is similar to that discussed earlier in this chapter for the simpler methods. For example, a forecast based on a simple AR(p) model is

a weighted average of past p observations of the series. A forecast based on an MA(q) model is a weighted average of the past q forecast errors. And finally, one must take into account the level of differencing and any transformations made on the original data.

As an example, consider an ARIMA (1, 1, 1) model (that is, a model having one level of differencing, one moving-average term, and one autoregressive term). This can be represented using the backshift notation as

$$(1 - B)(1 + a_1B)D_t = c + (1 - b_1B)\epsilon_t,$$

or as

$$(1 + a_1B)\nabla D_t = c + (1 - b_1B)\epsilon_t.$$

Writing out the model without backshift notation we have

$$D_t = c + (1 + a_1)D_{t-1} - a_1D_{t-2} + \epsilon_t - b_1\epsilon_{t-1}.$$

Let us suppose this model has been fitted to a time series with the result that $c = 15$, $a_1 = 0.24$, and $b_1 = 0.70$. Suppose that the last five values of the time series used to fit this model were 31.68, 29.10, 43.15, 56.74, and 62.44 based on a total of 76 observations. The first period in which one can forecast is period 77. The one-step-ahead forecast for period 77 made at the end of period 76 is

$$\hat{D}_{77} = 15 + (1 + 0.24)D_{76} - 0.24D_{75} - 0.70\epsilon_{76}.$$

$\hat{D}_{77}$ is the conditional expected value of D_{77} having observed the demand in periods 1, 2, . . . , 76. Since this is the first forecast made for this series, there is no observed previous value of the error, and the final term drops out. Hence, the one-step-ahead forecast made in period 76 for the demand in period 77 is $15 + (1.24)(62.44) - (0.24)(56.74) = 78.81$. Now, suppose we observed a value of 70 for the series in period 77. That means that the forecast error observed in period 77 is $\epsilon_{77} = 78.81 - 70 = 8.81$. The one-step-ahead forecast for period 78 made in period 77 would be $15 + (1.24)(70) - 0.24(62.44) - 0.70(8.81) = 86.25$.

When using an ARIMA model for multiple-step-ahead forecasts, the operative rule is to use the forecasts for the unobserved demand and use zero for the unobserved errors. Thus in the preceding example, a two-step-ahead forecast made at the end of period 76 for demand in period 78 would be based on the assumption that the observed demand in period 77 was the one-step-ahead forecast, 86.25. The observed forecast error for period 77 would be assumed to be zero.

Summary of the Steps Required for Building ARIMA Models

There are four major steps required for building Box-Jenkins forecasting models.

1. *Data transformations.* The Box-Jenkins methodology is predicated on starting with a stationary time series. To be certain that the series is indeed stationary, several preliminary steps might be required. We know that differencing eliminates trend and seasonality. However, if the mean of the series is relatively fixed, it still may be the case that the variance is not constant, thus possibly requiring a transformation of the data (for example, stock market data are often transformed by the logarithm).

2. *Model identification.* This step refers to determining exactly which ARIMA model seems to be most appropriate. Proper model identification is both art and science. It is difficult, if not impossible, to identify the appropriate model by only examining the series itself. It is far more effective to study the sample autocorrelations and

partial autocorrelations to discern patterns that match those of known processes. In some cases, the autocorrelation structure will point to a simple AR or MA process, but it is more common that some mixture of these terms would be required to get the best fit. However, one must not add terms willy-nilly. The operative concept is parsimony—that is, use the most economical model that adequately describes the data.

3. *Parameter estimation.* Once the appropriate model has been identified, the optimal values of the model parameters (i.e., $a_0, a_1, \ldots, a_p$ and $b_0, b_1, \ldots, b_q$) must be determined. Typically, this is done via either least squares fitting methods or the method of maximum likelihood. In either case, this step is done by a computer program.

4. *Forecasting.* Once the model has been identified and the optimal parameter values determined, the model provides forecasts of future values of the series. Box-Jenkins models are most effective in providing one-step-ahead forecasts, but can also provide multiple-step-ahead forecasts as well.

5. *Evaluation.* The pattern of residuals (forecast errors) can provide useful information regarding the quality of the model. The residuals should form a white noise (i.e., random) process with zero mean. Residuals should be normally distributed as well. When there are patterns in the residuals, it suggests that the forecasting model can be improved.

Case Study. *Using Box-Jenkins Methodology to Predict Monthly International Airline Passenger Totals*

This study is based on data that appeared originally in Brown (1962), but was analyzed using ARIMA methods in Box and Jenkins (1970). It illustrates the basic steps in transforming data and building ARIMA models. The data represent the monthly international airline sales from the period January 1949 to December 1960. The raw data appear in Table 2–3 and are pictured in Figure 2–17. From the figure, it is clear that there are several nonstationarities in this data. First, there is clearly an increasing linear trend. Second, there is seasonality, with a pattern repeating yearly. Third, there is increasing variance over time. In cases where the mean and variance increase at a comparable rate (which would occur if the series is

TABLE 2–3
International Airline Passengers: Monthly Totals (Thousands of Passengers), January 1949–December 1960*

Source: From Box and Jenkins (1970), p. 531.

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	550	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

*144 observations.

FIGURE 2–17
International airline passengers (thousands)

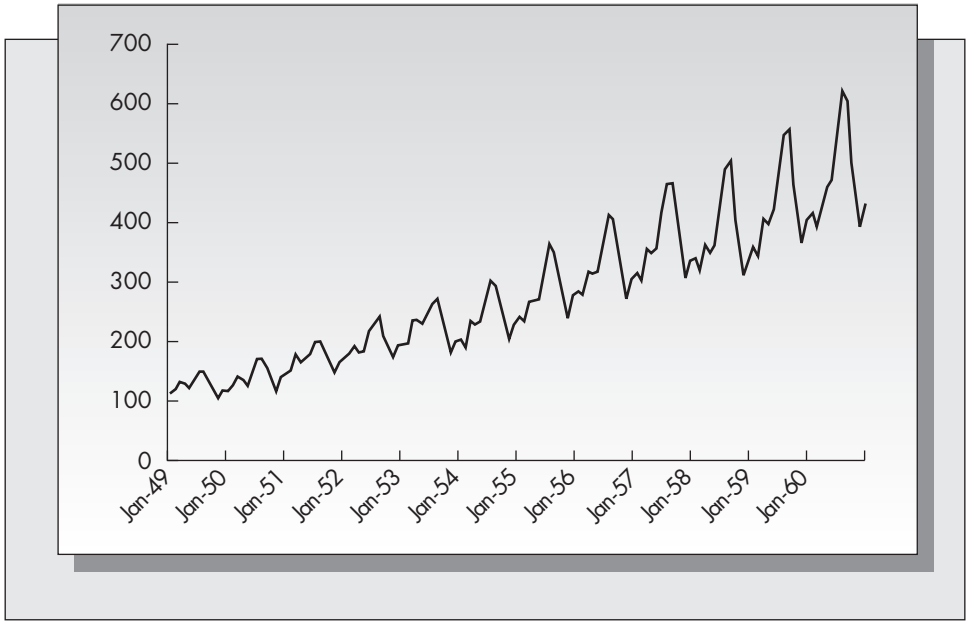
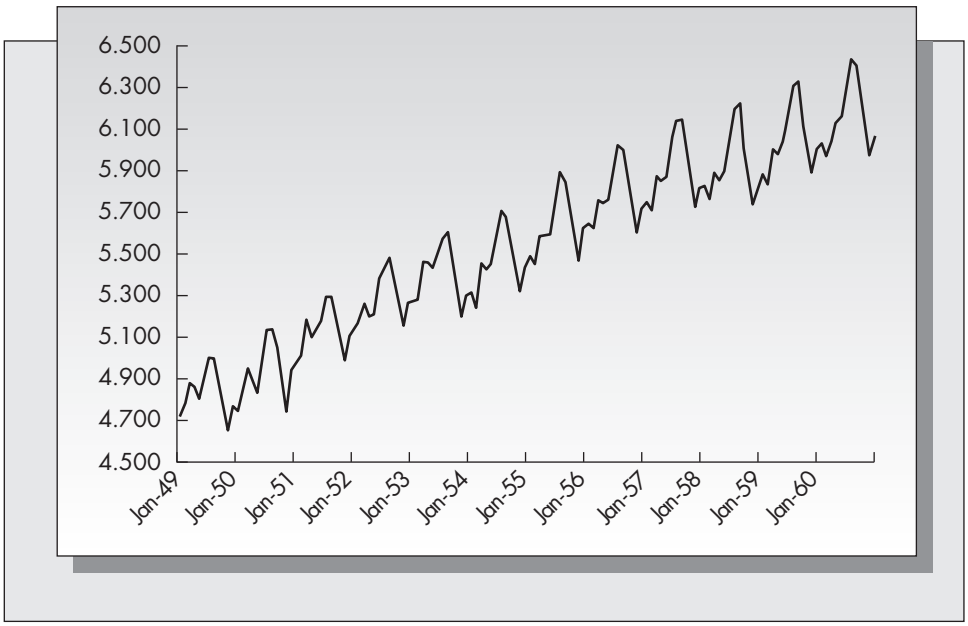


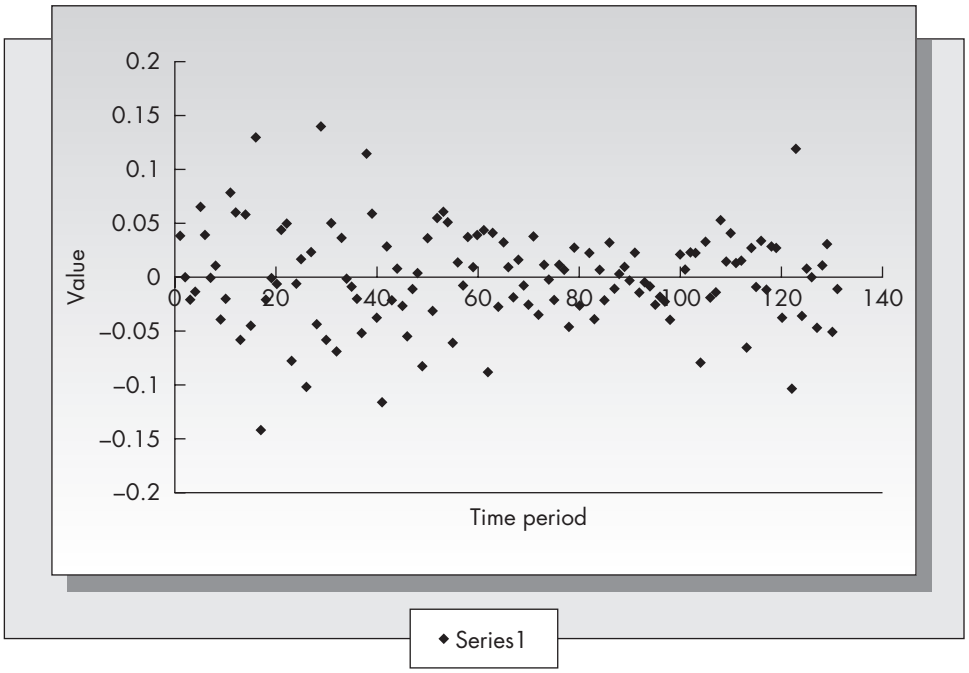
FIGURE 2–18
Natural log of international airline passengers



increasing by a fixed percentage), a logarithmic transformation will usually eliminate the nonstationarity due to increasing variance. (Changing variance is known as heteroscedasticity, and constant variance is known as homoscedasticity.) Applying a natural log transformation to the data yields a homoscedastic series as shown in Figure 2–18.

Next, we need to apply two levels of differencing to eliminate both trend and seasonality. The trend is eliminated by applying a single level of differencing and the seasonality by applying 12 periods of differencing. After these three transformations are applied to the original data, the resulting data appear in Figure 2–19. Note now the transformed data

FIGURE 2–19
Natural log of airline data with single and seasonal differencing



appear to form a random white noise process centered at zero showing neither trend nor seasonality.

It is to this set of data that we wish to fit an ARMA model. To do so, we determine the sample autocorrelations. While this can be accomplished with many of the software programs available for forecasting (including general statistical packages, such as SAS), we have done the calculations directly in Excel using the formulas for sample autocorrelations given earlier. The autocorrelations for lags of 1 to 12 periods for the series pictured in Figure 2–19 appear in Table 2–4.

Although not explicitly discussed in this section, when both seasonal differencing and period-to-period differencing are applied simultaneously, one must determine an ARMA model for each level of differencing. That is, one would like to find an ARMA model corresponding to the single level of differencing and to the seasonal level of differencing. Thus, when examining the autocorrelation function, we look for patterns at lags of 1 period and patterns at lags of 12 periods. From Table 2–4 it is clear that there are significant

TABLE 2–4
Autocorrelations for the Transformed Airline Data Pictured in Figure 2–19 (after taking logarithms and two levels of differencing)

Lag	Autocorrelation	Lag	Autocorrelation	Lag	Autocorrelation
1	−0.34	13	0.15	25	−0.10
2	0.11	14	−0.06	26	0.05
3	−0.20	15	0.15	27	−0.03
4	0.02	16	−0.14	28	0.05
5	0.06	17	0.07	29	−0.02
6	0.03	18	0.02	30	−0.05
7	−0.06	19	−0.01	31	−0.05
8	0.00	20	−0.12	32	0.20
9	0.18	21	0.04	33	−0.12
10	−0.08	22	−0.09	34	0.08
11	0.06	23	0.22	35	−0.15
12	−0.39	24	−0.02	36	−0.01

autocorrelations at lags of exactly 1 and 12 periods. This suggests that MA(1) models are appropriate for both differencing levels.

Note that if we let z_t represent the log-transformed series, we would denote the series pictured in Figure 2–19 as $(1 - B)(1 - B^{12})\, z_t$ or as $\nabla\nabla^{12} z_t$ to indicate that both first-order and 12th-order differencing were applied. Since we are assuming an MA(1) model for both the first difference process and the 12th-order difference process, the model we wish to fit can be denoted $\nabla\nabla^{12} z_t = c + (1 - b_1 B)(1 - b_2 B^{12})\epsilon_t$, where the parameters b_1 and b_2 are to be determined based on one of several fitting criteria. The exact values we obtain for the parameters will depend on the optimization method we use, but generally all methods will yield similar values. Least squares is probably the most common method used, but maximum likelihood and Bayesian methods have also been suggested. Using XLSTAT, a program that contains an ARIMA forecasting module and is embedded in Excel, we obtain the parameter values $b_1 = 0.333$ and $b_2 = 0.544$. The value of the constant c is small enough to be ignored. [These values differ slightly from those reported in Box and Jenkins (1970), since the search algorithm used by XLSTAT differs from the one used by Box and Jenkins.]

When forecasting using this model, it is convenient to write it out explicitly in the form

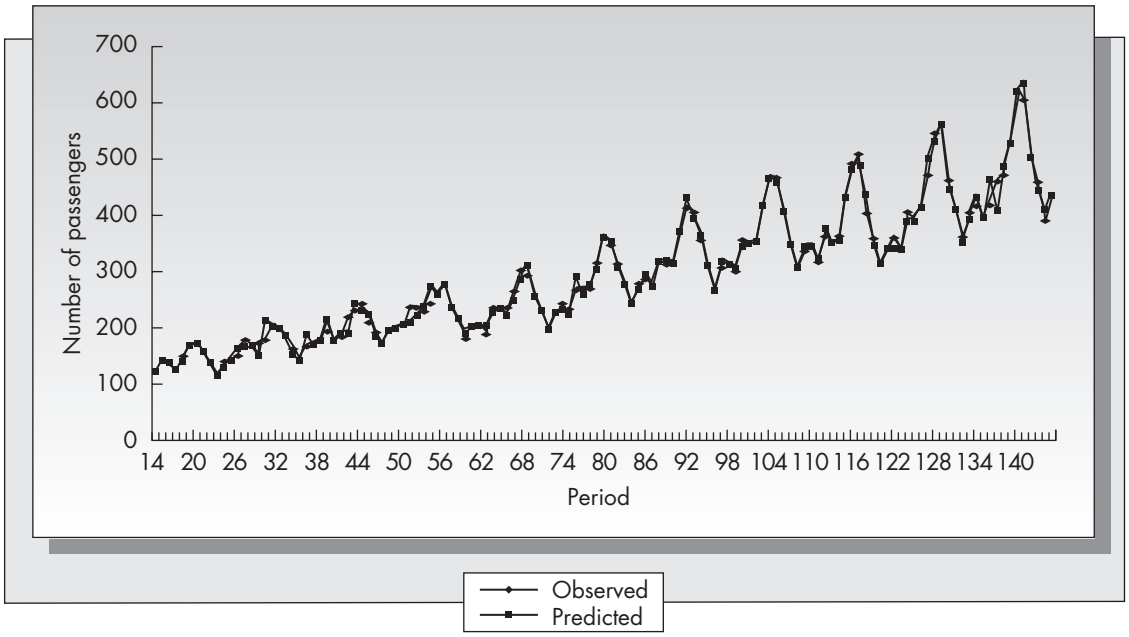
$$z_t - z_{t-1} - (z_{t-12} - z_{t-13}) = \epsilon_t - b_1\epsilon_{t-1} - b_2(\epsilon_{t-12} - b_1\epsilon_{t-13}).$$

Substituting for the parameter values and rearranging terms, the forecasting equation we obtain for the log series is

$$z_t = z_{t-1} + z_{t-12} - z_{t-13} + \epsilon_t - 0.333\epsilon_{t-1} - 0.544\epsilon_{t-12} + 0.181\epsilon_{t-13}.$$

To forecast the original series, D_t, we apply the antilog to z_t, namely, $D_t = \exp(z_t)$. Because of the two levels of differencing, period 14 is the first period for which we can determine a forecast. In Figure 2–20, we show the original series starting at period 14 and the one-step-ahead forecast using the preceding ARIMA model. Note how closely the ARIMA forecast tracks the original series.

FIGURE 2–20
Observed versus predicted number of airline sales



Snapshot Application

A SIMPLE ARIMA MODEL PREDICTS THE PERFORMANCE OF THE U.S. ECONOMY

In years past, a very complex and very large regression model known as the FRB-MIT-PENN (FMP) model (for Federal Reserve Bank—Massachusetts Institute of Technology—University of Pennsylvania) was used to predict several basic measures of the U.S. economy. This model required massive amounts of data and past history. Nelson (1972) employed the ARIMA methodology outlined in this section to obtain predictors for many of the same fundamental measures of the U.S. economy considered in the FMP model. These include gross national product, consumer's expenditures on durable goods, nonfarm inventory investment, and several others. Perhaps the most interesting case is the prediction of the gross national product. The ARIMA model he obtained is surprisingly simple:

$$z_t = z_{t-1} + 0.615(z_{t-1} - z_{t-2}) + 2.76 + \epsilon_t,$$

which is easily seen to be an AR(1) model with one level of differencing. What is most impressive and surprising is that the forecast errors obtained from this and Nelson's other ARIMA models had lower forecast errors in predicting future values of these measures than did the complex FMP model. Again, this points to the power of these methods in providing accurate forecasts in a variety of scenarios.

Box-Jenkins Modeling—A Critique

The preceding case study highlights the power of ARIMA models. However, one must be aware that there are only a limited number of situations in which one would or could use them. An important requirement is that one must have a substantial history of time series in question. Typical recommendations vary between 50 and 100 observations of past history, and that is for a nonseasonal series. When seasonality is present, the requirement is more severe. In some sense, each season of data is comparable to a single observation.

In the operations context, one would be hard-pressed to find many applications where that much history is available. Style goods, for example, are typically managed with a very small amount of data, if any; for example, recall the Sport Obermeyer Snapshot Application in this chapter. Even when enough data are available, it may not make sense to invest the amount of time and energy required to develop an ARIMA model. Consider forecasting usage rates in a Wal-Mart store, for example, where one must manage tens of thousands of SKUs. ARIMA models are most useful for forecasting economic series with substantial history (such as GNP) or continuous processes (such as chemical processes or stock market prices). In the latter case, one can conceivably choose a small enough time bucket to generate any number of observations.

Another shortcoming of ARIMA models is that they are not easily updated. If the problem parameters change, one must redo the analysis to find a new model, or at least rerun the software to find new optimal values of the parameters.

Even with all of these shortcomings, Box-Jenkins methods are very powerful. In the right situation, they can result in much more accurate short-term forecasts than any of the other methods discussed in this chapter.

Problems for Section 2.10

38. Consider the white noise process data in Table 2–1. Enter this data into Excel and generate the sample autocorrelations for lags of 1 to 10 periods, using the formula for the sample autocorrelations (r_k). Alternatively, if you have access to an ARIMA computing module, generate these autocorrelations using your software.

39. Use the random number generator in Excel and the normal variate generator given in Problem 5.34 to generate a sample path for an AR(1) process such as the ones shown in Figure 2–14 for $a_1 = 0.7$. Compute the sample autocorrelation function for the process you generated, and check that the correlogram has the same structure as shown in Figure 2–15a.
40. Use the random number generator in Excel and the normal variate generator given in Problem 5.34 to generate a sample path for an MA(1) process such as the ones shown in Figure 2–16 for $b_1 = 0.7$. Compute the sample autocorrelation function for the process you generated, and check that the correlogram shows only one significant autocorrelation at a lag of one period. What does the value of this autocorrelation appear to be?
41. Use the random number generator in Excel and the normal variate generator given in Problem 5.34 to generate a sample path for an ARMA(1,1) process. Use $a_1 = 0.8$ and $b_1 = -0.6$. Compute the sample autocorrelation function for this process.
42. Consider an ARIMA(2,1,1) process. Write out the resulting model using
    a. Backshift notation with the backshift operator B.
    b. Backshift notation with the operator ∇.
    c. No backshift notation.
43. Consider an ARIMA(0,2,2) process. Write out the resulting model using
    a. Backshift notation with the backshift operator B.
    b. Backshift notation with the operator ∇.
    c. No backshift notation.
44. Using back-substitution, show that an ARMA(1, 1) model may be written as either an AR(∞) or an MA(∞) model.
45. Consider the seasonal time series pictured in Figure 2–8. What level of differencing would be required to make this series stationary?
46. The U.S. Federal Reserve in St. Louis stores a host of economic data on its Web site at http://research.stlouisfed.org/fred2/categories/106. Download the following time series from this Web site:
    a. U.S. GNP.
    b. Annual Federal Funds rate.
    c. Consumer price index.

    In each case, use a minimum of 50 data points over a period not including the most recent 25 years. Graph the data points, and determine first the level of differencing that appears to be required. Once you have obtained a stationary process, use the methods outlined in this section to arrive at an appropriate ARIMA model for each case. Compare the most recent 25 years of observations with the predictions obtained from your model using both one-step-ahead and two-step-ahead forecasts.

2.11 PRACTICAL CONSIDERATIONS

Model Identification and Monitoring

Determining the proper model depends both on the characteristics of the history of observations and on the context in which the forecasts are required. When historical data are available, they should be examined carefully in order to determine if obvious

patterns exist, such as trend or seasonal fluctuations. Usually, these patterns can be spotted by graphing the data. Statistical tests, such as significance of regression, can be used to verify the existence of a trend, for example. Identifying complex relationships requires more sophisticated methods. The *sample autocorrelation function* can reveal intricate relationships that simple graphical methods cannot as we saw in Section 2.10.

Once a model has been chosen, forecasts should be monitored regularly to see if the model is appropriate or if some unforeseen change has occurred in the series. As we indicated, a forecasting method should not be biased. That is, the expected value of the forecast error should be zero. In addition to the methods mentioned in Section 2.6, one means of monitoring the bias is the *tracking signal* developed by Trigg (1964). Following earlier notation, let e_t be the observed error in period t and $|e_t|$ the absolute value of the observed error. The smoothed values of the error and the absolute error are given by

$$E_t = \beta e_t + (1 - \beta)E_{t-1},$$
$$M_t = \beta|e_t| + (1 - \beta)M_{t-1}.$$

The tracking signal is the ratio

$$T_t = \left|\frac{E_t}{M_t}\right|.$$

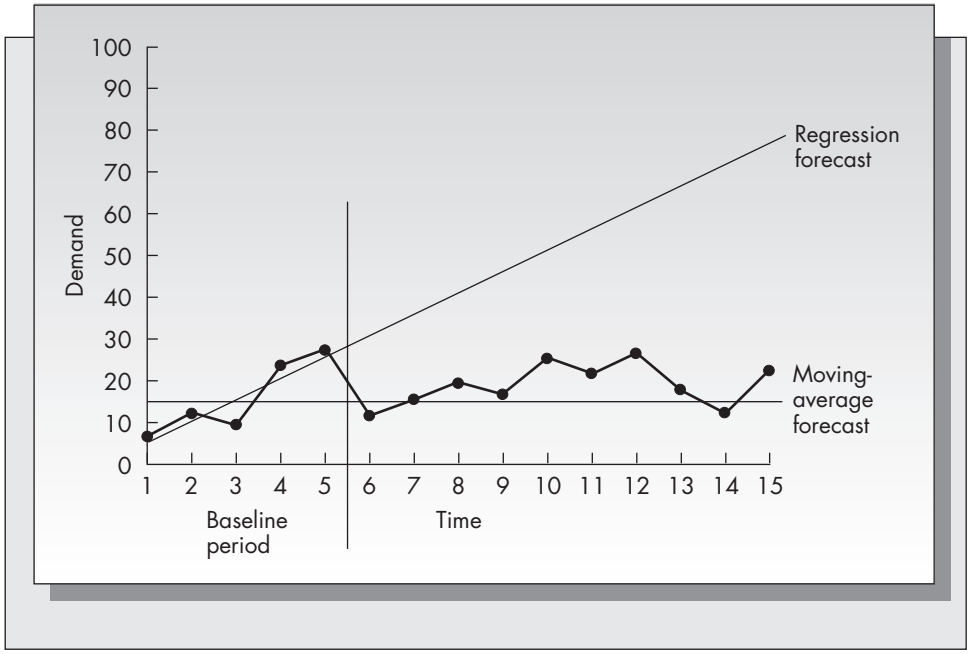
If forecasts are unbiased, the smoothed error E_t should be small compared to the smoothed absolute error M_t. Hence, a large value of the tracking signal indicates biased forecasts, which suggest that the forecasting model is inappropriate. The value of T_t that signals a significant bias depends on the smoothing constant β. For example, Trigg (1964) claims that a value of T_t exceeding 0.51 indicates nonrandom errors for a β of .1. The tracking signal also can be used directly as a variable smoothing constant. This is considered in Problem 55.

Simple versus Complex Time Series Methods

The literature on forecasting is voluminous. In this chapter we have touched only on a number of fairly simple techniques. The reader is undoubtedly asking himself or herself, do these methods actually work? The results from the literature suggest that the simplest methods are often as accurate as sophisticated ones. Armstrong (1984) reviews 25 years of forecasting case studies with the goal of ascertaining whether or not sophisticated methods work better. In comparing the results of 39 case studies, he found that in 20 cases sophisticated methods performed about as well as simple ones, in 11 cases they outperformed simple methods, and in 7 cases they performed significantly worse.

A more sophisticated forecasting method is one that requires the estimation of a larger number of parameters from the data. Trouble can arise when these parameters are estimated incorrectly. To give some idea of the nature of this problem, consider a comparison of simple moving averages and regression analysis for the following series: 7, 12, 9, 23, 27. Suppose that we are interested in forecasting at the end of period 5 for the demand in period 15 (that is, we require $F_{5,15}$). The five-period moving-average forecast made at the end of period 5 is 15.6, and this would be the forecast for period 15. The least squares fit of the data is $\hat{D}_t = 0.3 + 5.1t$. Substituting $t = 15$, we obtain the regression forecast of 76.8. In Figure 2–21 we picture the realization of the demand through period 15. Notice what has happened. The apparent trend that existed in the first five periods was extrapolated to period 15 by the regression equation. However, there really was no significant trend in this particular

FIGURE 2–21
The difficulty with long-term forecasts



case. The more complex model gave *significantly* poorer results for the long-term forecast.

There is some evidence that the arithmetic average of forecasts obtained from different methods is more accurate than a single method (see Makridakis and Winkler, 1983). This is perhaps because often a single method is unable to capture the underlying signal in the data and different models capture different aspects of the signal. (See a discussion of this phenomenon following Armstrong, 1984.)

What do these observations tell us about the application of forecasting techniques to production planning? At the aggregate level of planning, forecast accuracy is extremely important and multiple-step-ahead forecasts play an integral role in the planning of workforce and production levels. For that reason, blind reliance on time series methods is not advised at this level. At a lower level in the system, such as routine inventory management for spare parts, the use of simple time series methods such as moving averages or exponential smoothing makes a great deal of sense. At the individual item level, short-term forecasts for a large number of items are required, and monitoring the forecast for each item is impractical at best. The risk of severe errors is minimized if simple methods are used.

2.12 OVERVIEW OF ADVANCED TOPICS IN FORECASTING

Simulation as a Forecasting Tool

Computer simulation is a powerful technique for tackling complex problems. A computer simulation is a description of a problem reduced to a computer program. The program is designed to re-create the key aspects of the dynamics of a real situation. When a problem is too complex to model mathematically, simulation is a popular alternative. By rerunning the program under different starting conditions and/or different scenarios, one can, by a kind of trial-and-error process, discover the best strategy for managing a system.

Simulation is a common tool for modeling manufacturing planning problems such as complex material flow problems in the plant. It is less commonly used as a forecasting tool. Compaq Computer, a successful producer of personal computers based in Houston, Texas, has experimented with a powerful simulation-based forecasting tool for assisting with the process of new product introductions (McWilliams, 1995). The program recommends the optimal timing and pricing of new product introductions by incorporating forecasts of component availability and price changes, fluctuating demand for a given feature or price, and the impact of rival models.

Using this tool, Compaq decided to delay announcement of several new Pentium-based models in late 1994. This strategy "went against everything the company believed." Compaq's basic strategy had always been to be a technology leader, but its forecasting tool suggested that corporate customers were not quite ready to switch to Pentium-based machines at the end of 1994. The strategy proved to be very profitable for Compaq, which subsequently posted record earnings.

Forecasting Demand in the Presence of Lost Sales

Retailers rely heavily on forecasting. Basic items (items that don't change appreciably from season to season, such as men's dress shirts) generally have substantial sales history, arguing for the use of time series methods to forecast demand. However, there is an important difference between what is observed and what one wants to forecast. The goal is to forecast *demand,* but one only observes *sales*. What's the difference? Suppose a customer wants to buy a blouse in a certain size and color and finds it's not available on the shelf? What will she do? Perhaps she will place a special order with a salesperson, but, more likely, she will just leave the store and try to find the product somewhere else. This is known as a lost sale. The difficulty is that most retailers have no way to track lost sales. Thus, they observe sales but need to estimate demand.

As an example, consider an item that is restocked to 10 units at the beginning of each week. Suppose that over the past 15 weeks the sales history for the item was 7, 5, 10, 10, 8, 3, 6, 10, 10, 9, 5, 0, 10, 10, 4. Consider those weeks in which sales were 10 units. What were the demands in those weeks? The answer is that we don't know. We only know that it was *at least* 10. If you computed the sample mean and sample variance of these numbers, they would underestimate the true mean and variance of demand.

How does one go about forecasting demand in this situation? In the parlance of classical statistics, this is known as a censored sample. That means that we know the values of demand for only a portion of the sample. For the other portion of the sample, we know only a lower bound on the demand. Special statistical methods that incorporate censoring give significantly improved estimates of the population mean and variance in this case. These methods can be embedded into sequential forecasting schemes, such as exponential smoothing, to provide significantly improved forecasts.

Nahmias (1994) considered the problem of forecasting in the presence of lost sales when the true demand distribution was normal. He compared the method of maximum likelihood for censored samples and a new method, either of which could be incorporated into exponential smoothing routines. He also showed that both of these methods would result in substantially improved forecasts of both the mean and the variation of the demand.

To see how dramatic this difference can be, consider a situation in which the true weekly demand for a product is a normal random variable with mean 100 and standard deviation 30. Suppose that items are stocked up to 110 units at the start of each week. Exponential smoothing is used to obtain two sets of forecasts: the first accounts for lost sales (includes censoring) and the second does not (does not including censoring). Figures 2–22 and 2–23 show the estimators for the mean and the standard

FIGURE 2–22
Tracking the mean when lost sales are present

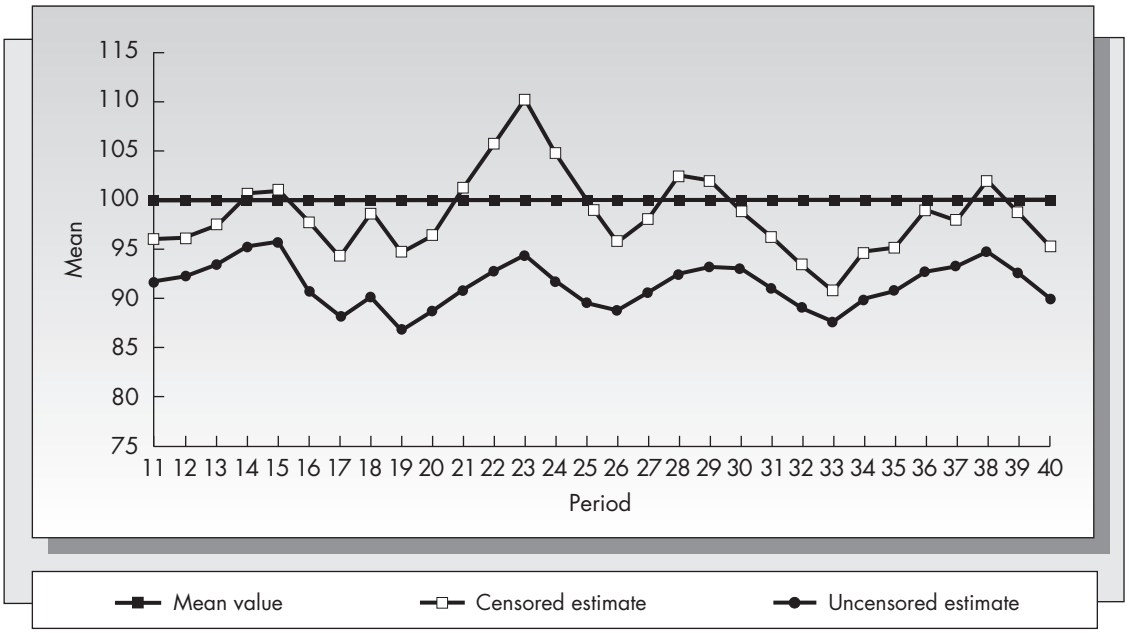
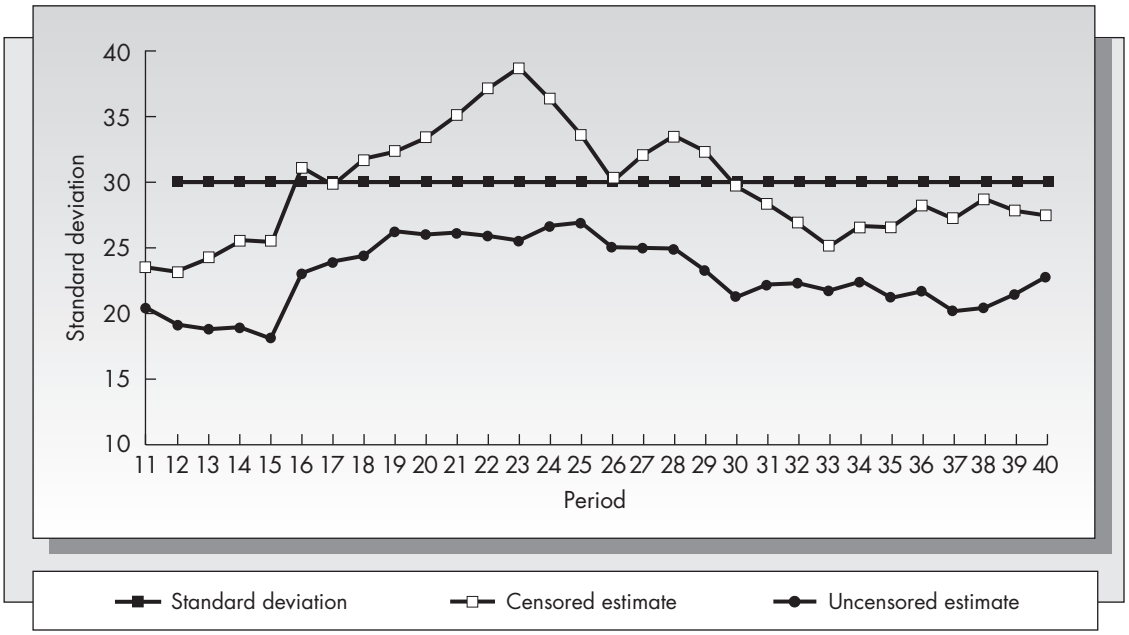


FIGURE 2–23
Tracking the standard deviation when lost sales are present



deviation, respectively, with and without censoring. Notice the severe low bias when lost sales are ignored in both cases. That means that by not correctly accounting for the difference between sales and demand, one underestimates both the mean and the variance of the demand. Since both the mean and the variance of demand are inputs for determining optimal stocking levels, these levels could be severely underestimated.

2.13 LINKING FORECASTING AND INVENTORY MANAGEMENT

Inventory control under demand uncertainty will be treated in detail in Chapter 5. In practice, inventory management and demand forecasting are closely linked. The forecasting method could be any one of the methods discussed in this chapter. One of the inputs required for inventory control models is the distribution of the demand over a period or over an order replenishment lead time. Is there a link between the distribution of forecast errors and the distribution of demand? The answer is that there is such a link. The forecast error distribution plays a key role in the correct application of inventory models in practice.

The first issue is to decide on the appropriate form of the distribution of demand. Most commercial systems assume that the demand distribution is normal. That means we need only estimate the mean μ and the standard deviation σ to specify the entire distribution. (Statistical methods, known as goodness-of-fit techniques, can be applied to test the accuracy of the normality assumption.) Whether or not the inventory system is linked to a forecasting system, we must have a history of observations of demand to obtain statistical estimates of the mean and variance. (When no history of demand exists, personal judgment must be substituted for statistical estimation. Methods for aggregating subjective judgment are discussed in Section 2.3.)

In the beginning of Chapter 5, we discuss how one would estimate the mean and the variance of demand directly from a history of demand observations. In practice, however, we don't generally use a long history of observations, because we believe that the underlying demand distribution does not remain constant indefinitely. That is why we adjust the N for moving averages or the α for exponential smoothing to balance stability and responsiveness. If that is the case, what is the appropriate variance estimator one should use?

In Appendix 2–A, we show that for simple moving averages of order N, the variance of forecast error, σ_e^2, is given by

$$\sigma_e^2 = \sigma^2 \left(\frac{N+1}{N} \right)$$

and for exponential smoothing the variance of forecast error is

$$\sigma_e^2 = \sigma^2 \left(\frac{2}{2-\alpha} \right).$$

Notice that in both cases, the value of σ_e^2 *exceeds* the value of σ^2. Also notice that as N gets large and as α gets small, the values of σ_e and σ grow close. These cases occur when one uses the entire history of demand observations to make a forecast. In Chapter 5 one of the inputs needed to determine safety stocks for inventory is the distribution of demand. The problem is, if we have estimators for both σ_e and σ, which should be used as the standard deviation estimator for setting safety stocks?

The obvious answer is that we should use the estimator for σ, since this represents the standard deviation of demand. The correct answer, however, is that we should use σ_e. The reason is that the process of forecasting introduces sampling error into the estimation process, and this sampling error is accounted for in the value of σ_e. The forecasting error variance is higher than the demand variance because the forecast is based on only a limited portion of the demand history.

We also can provide an intuitive explanation. If a forecast is used to estimate the mean demand, we keep safety stocks in order to protect against the error in this forecast.

Hence, the distribution of forecast errors is more relevant than the distribution of demands. This is an important practical point that has been the source of much confusion in the literature. We will return to this issue in Chapter 5 and discuss its relevance in the context of safety stock calculations.

Most inventory control systems use the method suggested by R. G. Brown (1959 and 1962) to estimate the value of σ_e. (In fact, it appears that Brown was the first to recognize the importance of the distribution of forecast errors in inventory management applications.) The method requires estimating the MAD of forecast errors using exponential smoothing. This is accomplished using the smoothing equation

$$\text{MAD}_t = \alpha|F_t - D_t| + (1 - \alpha)\text{MAD}_{t-1}$$

The MAD is converted to an estimate of the standard deviation of forecast error by multiplying by 1.25. That is, the estimator for σ_e obtained at time t is

$$\hat{\sigma}_e = 1.25\ \text{MAD}_t.$$

A small value of α, generally between 0.1 and 0.2, is used to ensure stability in the MAD estimator. This approach to estimating the MAD works for any of the forecasting methods discussed in this chapter. Safety stocks are then computed using this estimator for σ_e.

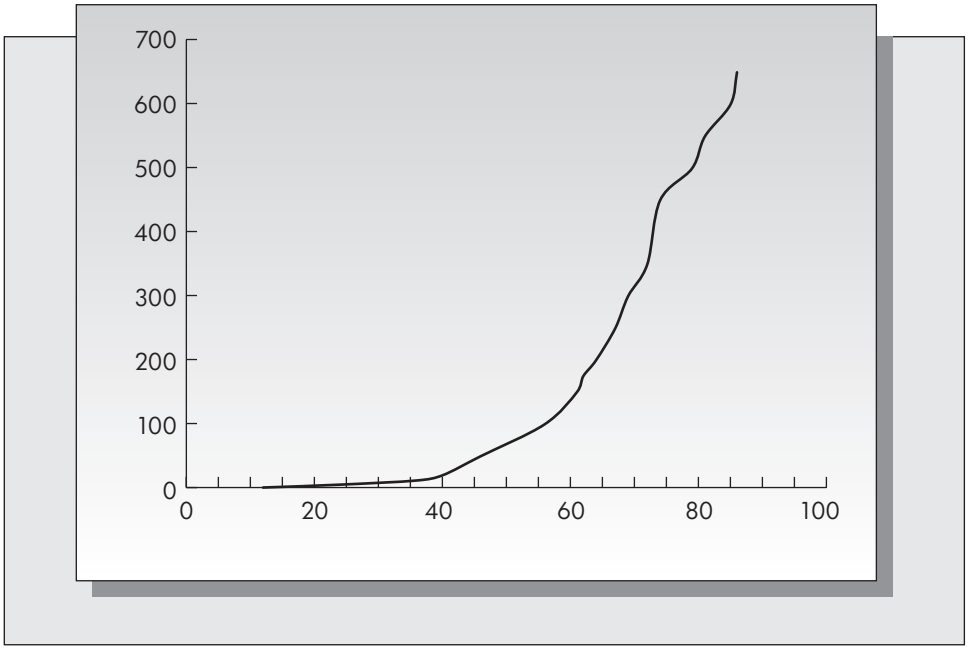
Case Study. Predicting the Growth of Facebook

Facebook, founded by Mark Zuckerberg in 2004, is the quintessential social network and one of the true phenomena of the early 21st century. While Myspace, another social network, was founded a year earlier than Facebook, it never enjoyed the kind of success experienced by Facebook. Consider the problem of forecasting the number of Facebook users in 2013 based on data up until 2011. The following chart tracks the numbers of Facebook users from the year of its founding in 2004 to early 2011:

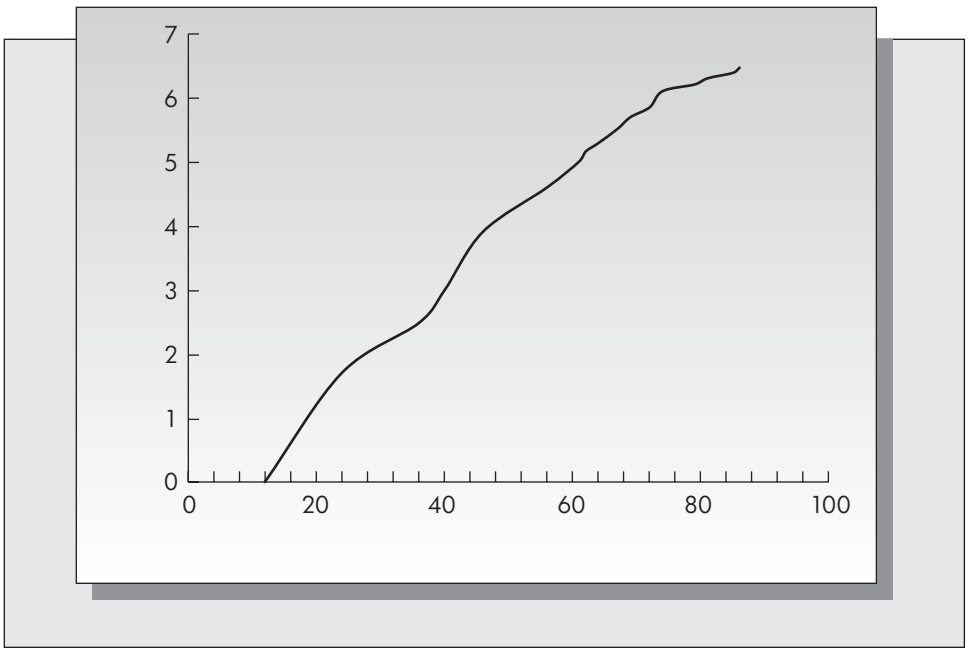
Date	Users (in millions)	Month Number
12/04	1	12
12/05	5.5	24
12/06	12	36
4/07	20	40
10/07	50	46
08/08	100	56
01/09	150	61
02/09	175	62
04/09	200	64
07/09	250	67
09/09	300	69
12/09	350	72
02/10	450	74
07/10	500	79
09/10	550	81
01/11	600	85
02/11	650	86

Note that the dates are not evenly spaced. One way to graph this data, which will also be useful in analyzing it, is to convert the dates to numbers of months elapsed from an arbitrary starting point. If we define month 1 as January 2004, then the dates in column 1 translate to the number of months elapsed in column 3.

Treating numbers of months as the independent variable, the graph of numbers of users versus months elapsed appears in the graph below:



The graph seems to show exponential growth. To test this hypothesis, we consider graphing the natural logarithm of the numbers of users versus elapsed numbers of months. Doing so results in the following graph:



The fact that the graph of the natural logs is almost linear means that the assumption of exponential growth is an accurate one. Fitting a simple linear regression to this curve results in the following:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.987306
R Square	0.974772
Adjusted R Square	0.973091
Standard Error	0.307178
Observations	17

ANOVA

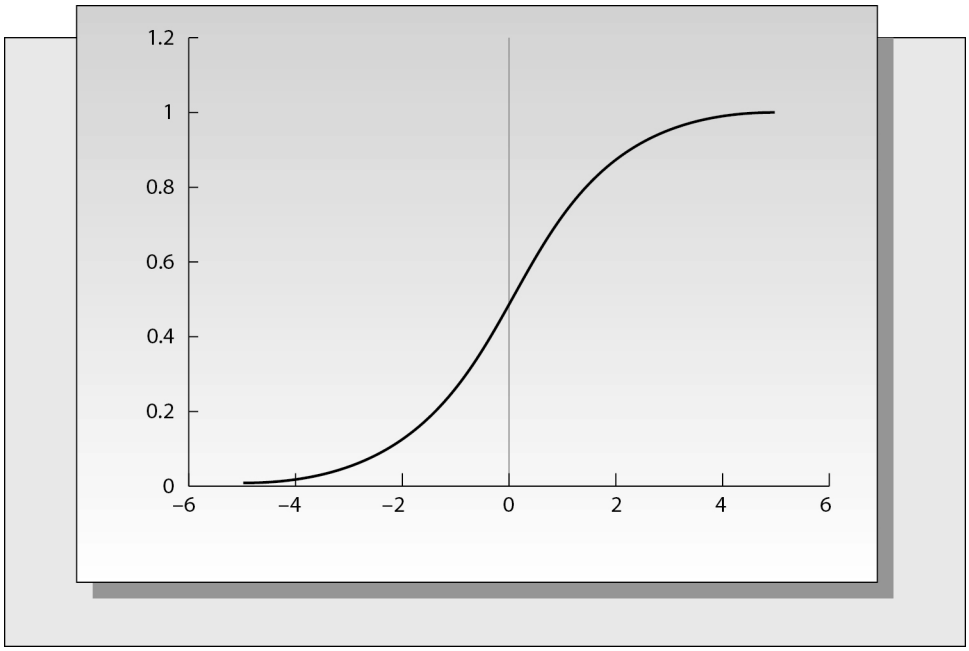
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	54.68893	54.68893	579.5877578	2.11649E-13
Residual	15	1.415375	0.094358		
Total	16	56.10431			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	−0.4425	0.22592	−1.95867	0.069013361	−0.924037859	0.039035397	−0.924037859	0.039035397
X Variable 1	0.086085	0.00357	24.07463	2.11649E-13	0.078463009	0.093706009	0.078463009	0.093706009

This Excel output shows that the logs follow a straight line relationship very closely. In fact, it's rare to see an R square value of more than .97, so this fit is extremely strong. Let's see what numbers of users would be projected by this model.

The regression model is $\bar{y} = a + bx$ where $\bar{y}$ represents the logarithm of the number of users, a is the intercept and b is the slope. The independent variable, x, is the number of months elapsed since January 2004. The regression output indicates that the least squares estimators are a = -.4425 and b = .086085. Let's consider how this model would be used for forecasting numbers of Facebook users. Consider the forecast for January 2012. This corresponds to month 97. Setting $x = 97$ gives $\bar{y} = 7.90785$. Since $\bar{y} = ln\ (y)$, it follows that $y = \exp(\bar{y}) = \exp(7.90785) = 2718.53$ millions of users. That is, if Facebook's growth continues at its current rate, there should be approximately 2.7 billion users as soon as January 2012. This seems unlikely, but possible. Let's consider the prediction for January 2014. This month corresponds to $x = 121$, which results in $\bar{y} = 9.34$ and a predicted number of users of more than 11 billion! Obviously, this prediction is absurd, as it is almost twice as many people as there are on earth.

The conclusion is that exponential growth cannot continue indefinitely if there are finite resources involved (in this case, the total number of computer users on the Earth). So the question remains, how should one forecast the growth of Facebook users? Clearly the model of continued exponential growth is unreasonable. What is likely is that the observed exponential growth is only the first phase of a more complex model. But what would be an appropriate model in this case? The answer is that at this point we're not sure, but based on past experience with company growth curves, it is reasonable to postulate a logistic curve. The mathematical form of the standard logistic curve is $P(t) = \dfrac{1}{1 + e^{-t}}$. This results in the curve pictured below.



The goal is to try to provide the best fit of this curve to the original Facebook data. Since we're not sure where the point of inflection occurs, our estimate at this stage is probably not going to be very accurate. However, we can do a lot better than simply extrapolating from an exponential growth curve. In order to fit a logistic curve to a set of data, we need to know a few things. First, where in the data set is the inflection point? Second, what is likely to be the value of the asymptote (which is one for the standard curve)? And finally, what is the appropriate time scale?

Consider estimating the asymptote. This is probably not going to be very accurate since we don't know where the inflection point is. From the original graph, it does appear that the inflection point occurs somewhere near month 80. The number of users in month 80 is approximately 525 million. This would imply a value of the asymptote of approximately (2)(525 million) = 1.05 billion. This would be our best guess of where the total numbers of Facebook users would top out in the coming years.

We also need to transform the time scale. If we assume in our original curve that the first half of the logistic curve corresponds to months 20 to 80, this would be the same as 5 to 0 in the standard logistic curve above. The reader should satisfy him or herself that this would correspond to the transformation of the time scale following the equation $u = 12*t + 80$. (Check the values of u and t at −5, 0, and 5 to verify this transformation.) Furthermore, if the asymptote is K rather than 1, we scale the function by multiplying by K. Applying these two transformations, the appropriate logistic curve to fit our original data would be

$$P(u) = \frac{K}{1 + \exp(-(u - 80)/12))}$$

where u corresponds to number of elapsed months, and $K = 1.05 \times 10^9$. Consider the forecast for January 2012 based on this model. Substituting $u = 97$ gives a forecast of $P(u) = 0.845$ billion users. This certainly seems reasonable. According to the logistic model, we would expect that the number of users will be close to the postulated maximum of 1.05 billion by month 140, which corresponds to August of 2015.

In late 2013 Facebook reported that there were 1.158 billion users. If true, this implies that our model predicted low. However, sources claimed that approximately 5 percent of those may have been duplicate accounts, while 1.3 percent may have been accounts that were improperly classified by users, and 0.9 percent may have been fake accounts (Cohen, 2013). That means the true value is likely to be closer to 1.07 billion, not far from the model's asymptote.

2.14 HISTORICAL NOTES AND ADDITIONAL TOPICS

Forecasting is a rich area of research. Its importance in business applications cannot be overstated. The simple time series methods discussed in this chapter have their roots in basic statistics and probability theory. The method of exponential smoothing is generally credited to R. G. Brown (1959 and 1962), who worked as a consultant to A. D. Little at the time. Although not grounded in basic theory, exponential smoothing is probably one of the most popular forecasting methods used today. Brown was also the first to recognize the importance of the forecast error distribution and its implications for inventory management. However, interest in using statistical methods for forecasting goes back to the turn of the century or before. (See, for example, Yule, 1926.)

Not discussed in this chapter is forecasting of time series using spectral analysis and state space methods. These methods are highly sophisticated and require substantial structure to exist in the data. They often rely on the use of the autocorrelation function, and in that sense are similar conceptually to Box-Jenkins methods discussed briefly in Section 2.10. The groundbreaking work in this area is due to Norbert Wiener (1949) and Rudolph Kalman (1960). However, there have been few applications of these methods to forecasting economic time series. Most applications have been in the area of signal processing in electrical engineering. [Davenport and Root (1958) provide a good summary of the fundamental concepts in this area.]

The *Kalman filter* is a type of exponential smoothing technique in which the value of the smoothing constant changes with time and is chosen in some sort of optimal fashion. The idea of adjusting the value of the smoothing constant based on some measure of prior performance of the model has been used in a number of ad hoc ways. A typical approach is the one suggested by Trigg and Leach (1967), which requires the calculation of the tracking signal. The tracking signal is then used as the value of the smoothing constant for the next forecast. The idea is that when the tracking signal is large, it suggests that the time series has undergone a shift; a larger value of the smoothing constant should be more responsive to a sudden shift in the underlying signal. Other methods have also been suggested. We have intentionally not included an explicit discussion of adaptive response rate methods for one reason: there is little evidence that they work in the context of predicting economic time series. Most studies that compare the effectiveness of different forecasting methods for many different series show no advantage for adaptive response rate models. (See, for example, Armstrong, 1984.) The method of Trigg and Leach is discussed in more detail in Problem 55.

With the sixth edition, we have included a reasonably self-contained discussion of Box-Jenkins ARIMA models. The basic ideas behind these methods, such as the autocorrelation structure of a process, go back many years. However, Box and Jenkins (1970) were the first to put these ideas together into a comprehensive step-by-step approach for building ARIMA models for short-term forecasting. Readers may find

their book daunting, as they focus on issues of mathematical concern. For the reader seeking a more comprehensive coverage of ARIMA models at a level consistent with ours, a good starting point is the text by Makridakis, Wheelright, and Hyndman (1998).

2.15 Summary

This chapter provided an introduction to a number of the more popular time series forecasting techniques as well as a brief discussion of other methods, including the Delphi method and causal models. A *moving-average* forecast is obtained by computing the arithmetic average of the N most recent observations of demand. An *exponential smoothing* forecast is obtained by computing the weighted average of the current observation of demand and the most recent forecast of that demand. The weight applied to the current observation is α and the weight applied to the last forecast (that is, the past observations) is $1 - \alpha$. Small N and large α result in responsive forecasts, and large N and small α result in stable forecasts. Although the two methods have similar properties, exponential smoothing is generally preferred because it requires storing only the previous forecast, whereas moving averages requires storing the last N demand observations.

When there is a trend in the series, both moving averages and exponential smoothing lag behind the trend. We discussed two time series techniques that are designed to track the trend. One is *regression analysis,* which uses least squares to fit a straight line to the data, and the other is *Holt's method,* which is a type of double exponential smoothing. Holt's method has the advantage that forecasts are easier to update as new demand observations become available.

We also discussed techniques for seasonal series. We employed *classical decomposition* to show how simple moving averages could be used to estimate seasonal factors and obtain the deseasonalized series when there is a trend and showed how seasonal factors could be estimated quickly when there is no trend. The extension of Holt's method to deal with seasonal problems, called *Winters's method,* is a type of triple exponential smoothing technique.

Section 2.12 provided a brief overview of advanced methods that are beyond the scope of this text. The final section discussed the relationship between forecasting and inventory control. The key point, which will be elaborated on in Chapter 5, is that the standard deviation of forecast error is the appropriate measure of variation for computing safety stocks.

Additional Problems on Forecasting

47. John Kittle, an independent insurance agent, uses a five-year moving average to forecast the number of claims made in a single year for one of the large insurance companies he sells for. He has just discovered that a clerk in his employ incorrectly entered the number of claims made four years ago as 1,400 when it should have been 1,200.

    a. What adjustment should Mr. Kittle make in next year's forecast to take into account the corrected value of the number of claims four years ago?

    b. Suppose that Mr. Kittle used simple exponential smoothing with $\alpha = .2$ instead of moving averages to determine his forecast. What adjustment is now required in next year's forecast? (Note that you do not need to know the value of the forecast for next year in order to solve this problem.)

48. A method of estimating the MAD discussed in Section 2.13 recomputes it each time a new demand is observed according to the following formula:

$$\text{MAD}_t = \alpha|e_t| + (1 - \alpha)\,\text{MAD}_{t-1}.$$

Consider the one-step-ahead forecasts for aircraft engine failures for quarters 2 through 8 obtained in Example 2.3. Assume an initial value of the MAD = 50 in period 1. Using the same value of α, what values of the MAD does this method give for periods 2 through 8? Discuss the advantages and disadvantages of this approach vis-à-vis direct computation of the MAD.

49. Herman Hahn is attempting to set up an integrated forecasting and inventory control system for his hardware store, Hahn's Hardware. When Herman indicates that outdoor lights are a seasonal item on the computer, he is prompted by the program to input the seasonal factors by quarter.

Unfortunately, Herman has not kept any historical data, but he estimates that first-quarter demand for the lights is about 30 percent below average, the second-quarter demand about 20 percent below average, third-quarter demand about average, and fourth-quarter demand about 50 percent above average. What should he input for the seasonal factors?

50. Irwin Richards, a publisher of business textbooks, publishes in the areas of management, marketing, accounting, production, finance, and economics. The president of the firm is interested in getting a relative measure of the sizes of books in the various fields. Over the past three years, the average numbers of pages of books published in each area were

	Year 1	Year 2	Year 3
Management	835	956	774
Marketing	620	540	575
Accounting	440	490	525
Production	695	680	624
Finance	380	425	410
Economics	1,220	1,040	1,312

Using the quick and dirty methods discussed in Section 2.9, find multiplicative factors for each area giving the percentage above or below the mean number of pages.

51. Over a two-year period, the Topper Company sold the following numbers of lawn mowers:

Month:	1	2	3	4	5	6	7	8	9	10	11	12
Sales:	238	220	195	245	345	380	270	220	280	120	110	85
Month:	13	14	15	16	17	18	19	20	21	22	23	24
Sales:	135	145	185	219	240	420	520	410	380	320	290	240

a. In column A input the numbers 1 to 24 representing the months and in column B the observed monthly sales. Compute the three-month moving-average forecast and place this in the third column. Be sure to align your forecast with the period *for* which you are forecasting (the average of sales for months 1, 2, and 3 should be placed in row 4; the average of sales for months 2, 3, and 4 in row 5; and so on.) In the fourth column, compute the forecast error for each month in which you have obtained a forecast.

b. In columns 5, 6, and 7 compute the absolute error, the squared error, and the absolute percentage error. Using these results, find the MAD, MSE, and MAPE for the MA(3) forecasts for months 4 through 24.

c. Repeat parts (*a*) and (*b*) for six-month moving averages. (These calculations should appear in columns 8 through 12.) Which method, MA(3) or MA(6), was more accurate for these data?

52. Repeat the calculations in Problem 51 using simple exponential smoothing, and allow the smoothing constant α to be a variable. That is, the smoothing constant should be a cell location. By experimenting with different values of α, determine the value that appears to minimize the

    a. MAD

    b. MSE

    c. MAPE

    Assume that the forecast for month 1 is 225.

53. Baby It's You, a maker of baby foods, has found a high correlation between the aggregate company sales (in $100,000) and the number of births nationally the preceding year. Suppose that the sales and the birth figures during the past eight years are

	Year							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Sales (in $100,000)	6.1	6.4	8.3	8.8	5.1	9.2	7.3	12.5
U.S. births (in millions)	2.9	3.4	3.5	3.1	3.8	2.8	4.2	3.7

    a. Assuming that U.S. births represent the independent variable and sales the dependent variable, determine a regression equation for predicting sales based on births. Use years 2 through 8 as your baseline. (Hint: You will require the general regression formulas appearing in Appendix 2–B to solve this problem.)

    b. Suppose that births are forecasted to be 3.3 million in year 9. What forecast for sales revenue in year 10 do you obtain using the results of part (*a*)?

    c. Suppose that simple exponential smoothing with $\alpha = .15$ is used to predict the number of births. Use the average of years 1 to 4 as your initial forecast for period 5, and determine an exponentially smoothed forecast for U.S. births in year 9.

    d. Combine the results in parts (*a*), (*b*), and (*c*) to obtain a forecast for the sum of total aggregate sales in years 9 and 10.

54. Hy and Murray are planning to set up an ice cream stand in Shoreline Park, described in Problem 28. After six months of operation, the observed sales of ice cream (in dollars) and the number of park attendees are

	Month					
	1	**2**	**3**	**4**	**5**	**6**
Ice cream sales	325	335	172	645	770	950
Park attendees	880	976	440	1,823	1,885	2,436

a. Determine a regression equation treating ice cream sales as the dependent variable and time as the independent variable. Based on this regression equation, what should the dollar sales of ice cream be in two years (month 30)? How confident are you about this forecast? Explain your answer.

b. Determine a regression equation treating ice cream sales as the dependent variable and park attendees as the independent variable. (Hint: You will require the general regression equations in Appendix 2–B in order to solve this part.)

c. Suppose that park attendance is expected to follow a logistic curve (see Figure 2–24).

The park department expects the attendance to peak out at about 6,000 attendees per month. Plot the data of park attendees by month and "eyeball" a logistics curve fit of the data using 6,000 as your maximum value. Based on your curve and the regression equation determined in part (*b*), predict ice cream sales for months 12 through 18.

55. A suggested method for determining the "right" value of the smoothing constant α in exponential smoothing is to retrospectively determine the α value that results in the minimum forecast error for some set of historical data. Comment on the appropriateness of this method and some of the potential problems that could result.

56. Lakeroad, a manufacturer of hard disks for personal computers, was founded in 1981 and has sold the following numbers of disks:

Year	Number Sold (in 000s)	Year	Number Sold (in 000s)
1981	0.2	1985	34.5
1982	4.3	1986	68.2
1983	8.8	1987	85.0
1984	18.6	1988	58.0

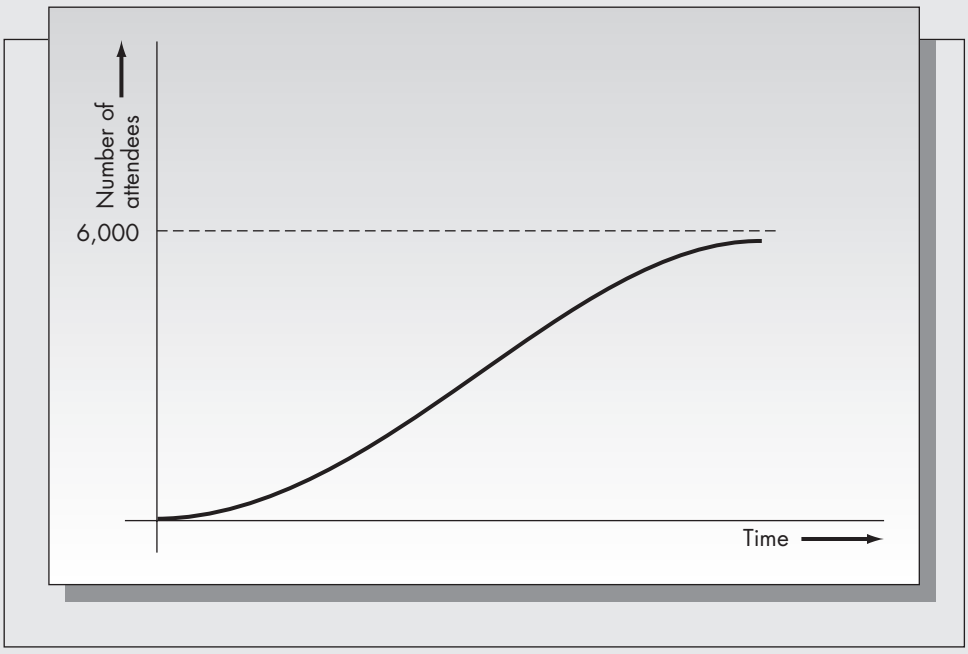


FIGURE 2–24
Logistic curve (for Problem 52)

a. Suppose the firm uses Holt's method for forecasting sales. Assume $S_0 = 0$ and $G_0 = 8$. Using $\alpha = .2$ and $\beta = .2$, find one-step-ahead forecasts for 1982 through 1989 and compute the MAD and MSE for the forecasts during this period. What is the sales forecast for the year 2000 made at the end of 1988? Based on the results of 1988, why might this forecast be very inaccurate?

b. By experimenting with various values of α and β, determine the values of the smoothing constants that appear to give the most accurate forecasts.

57. Trigg and Leach (1967) suggest the following adaptive response-rate exponential smoothing method. Along with smoothing the original series, also smooth the error e_t and the absolute error $|e_t|$ according to the equations

$$E_t = \beta e_t + (1 - \beta)E_{t-1},$$
$$M_t = \beta|e_t| + (1 - \beta)M_{t-1}$$

and define the smoothing constant to be used in forecasting the series in period t as

$$\alpha_t = \left|\frac{E_t}{M_t}\right|.$$

The forecast made in period t for period $t + 1$ is obtained by the usual exponential smoothing equation, using α_t as the smoothing constant. That is,

$$F_{t+1} = \alpha_t D_t + (1 - \alpha_t)F_t.$$

The idea behind the approach is that when E_t is close in magnitude to M_t, it suggests that the forecasts are biased. In that case, a larger value of the smoothing constant results that makes the exponential smoothing more responsive to sudden changes in the series.

a. Apply the Trigg-Leach method to the data in Problem 22. Using the MAD, compare the accuracy of these forecasts with the simple exponential smoothing forecasts obtained in Problem 22. Assume that $E_1 = e_1$ (the observed error in January) and $M_1 = |e_1|$. Use $\beta = .1$.

b. For what types of time series will the Trigg-Leach adaptive response rate give more accurate forecasts, and under what circumstances will it give less accurate forecasts? Comment on the advisability of using such a method for situations in which forecasts are not closely monitored.

58. The owner of a small brewery in Milwaukee, Wisconsin, is using Winters's method to forecast his quarterly beer sales. He has been using smoothing constants of $\alpha = .2$, $\beta = .2$, and $\gamma = .2$. He has currently obtained the following values of the various slope, intercept, and seasonal factors: $S_{10} = 120$, $G_{10} = 14$, $c_{10} = 1.2$, $c_9 = 1.1$, $c_8 = .8$, and $c_7 = .9$.

a. Determine the forecast for beer sales in quarter 11.

b. Suppose that the actual sales turn out to be 128 in quarter 11. Find S_{11} and G_{11}, and find the updated values of the seasonal factors. Also determine the forecast made at the end of quarter 11 for quarter 13.

59. The U.S. gross national product (GNP) in billions of dollars during the period 1964 to 1984 was as follows:

Year	GNP	Year	GNP
1964	649.8	1975	1,598.4
1965	705.1	1976	1,782.8
1966	772.0	1977	1,990.5
1967	816.4	1978	2,249.7
1968	892.7	1979	2,508.2
1969	963.9	1980	2,732.0
1970	1,015.5	1981	3,052.6
1971	1,102.7	1982	3,166.0
1972	1,212.8	1983	3,401.6
1973	1,359.3	1984	3,774.7
1974	1,472.8		

Source: *Economic Report of the President,* February 1986.

a. Use Holt's method to predict the GNP. Determine a regression fit of the data for the period 1964 to 1974 to estimate the initial values of the slope and intercept. (Hint: If you are doing the regression by hand, transform the years by subtracting 1963 from each value to make the calculations less cumbersome.) Using Holt's method, determine forecasts for 1975 to 1984. Assume that $\alpha = .2$ and $\beta = .1$. Compute the MAD and the MSE of the one-step-ahead forecasts for the period 1975 to 1984.

b. Determine the percentage increase in GNP from 1964 to 1984 and graph the resulting series. Use a six-month moving average and simple exponential smoothing with $\alpha = .2$ to obtain one-step-ahead forecasts of this series for the period 1975 to 1984. (Use the arithmetic average of the observations from 1964 to 1974 to initialize the exponential smoothing.)

In both cases (i.e., MA and ES forecasts), convert your forecasts of the percentage increase for the following year to a forecast of the GNP itself and compute the MAD and the MSE of the resulting forecasts. Compare the accuracy of these methods with that of part (*a*).

c. Discuss the problem of predicting GNP. What methods other than the ones used in parts (*a*) and (*b*) might give better predictions of this series?

Appendix **2–A**

Forecast Errors for Moving Averages and Exponential Smoothing

The forecast error e_t is the difference between the forecast for period t and the actual demand for that period. In this appendix we will derive the distribution of the forecast error for both moving averages and exponential smoothing.

The demand is assumed to be generated by the process

$$D_t = \mu + \epsilon_t,$$

where ϵ_t is normal with mean zero and variance σ^2.

CASE 1. MOVING AVERAGES

Consider first the case in which forecasts are generated by moving averages. Then the forecast error is $e_t = F_t - D_t$, where F_t is given by

$$F_t = \frac{1}{N} \sum_{i=t-N}^{t-1} D_i.$$

It follows that

$$E(F_t - D_t) = (1/N) \sum_{i=t-N}^{t-1} E(D_i) - E(D_t) = (1/N)(N\mu) - \mu = 0.$$

This proves that when demand is stationary, moving-average forecasts are unbiased.

Also,

$$\begin{aligned} \text{Var}(F_t - D_t) &= \text{Var}(F_t) + \text{Var}(D_t) \\ &= (1/N^2) \sum_{i=t-N}^{t-1} \text{Var}(D_i) + \text{Var}(D_t) \\ &= (1/N^2)(N\sigma^2) + \sigma^2 \\ &= \sigma^2(1 + 1/N) = \sigma^2[(N + 1)/N]. \end{aligned}$$

It follows that the standard deviation of the forecast error, σ_e, is

$$\sigma_e = \sigma \sqrt{\frac{N + 1}{N}}.$$

This is the standard deviation of the forecast error for simple moving averages in terms of the standard deviation of each observation.

Having derived the mean and the variance of the forecast error, we still need to specify the *form* of the forecast error distribution. By assumption, the values of D_t form a sequence of independent, identically distributed, normal random variables. Since F_t is a linear combination of $D_{t-1}, D_{t-2}, \ldots, D_{t-N}$, it follows that F_t is normally distributed and independent of D_t. It now follows that e_t is normal as well. Hence, the distribution of e_t is completely specified by its mean and variance.

As the expected value of the forecast error is zero, we say the method is unbiased. Notice that this is a result of the assumption that the demand process is stationary. Consider the variance of the forecast error. The value of N that minimizes σ_e is $N = +\infty$. This means that the variance is minimized if the forecast is the average of all the past data. However, our intuition tells us that we can do better if we use more recent data to make our forecast. The discrepancy arises because we really do not believe our assumption that the demand process is stationary for all time. A smaller value of N will allow the moving-average method to react more quickly to unforeseen changes in the demand process.

CASE 2. EXPONENTIAL SMOOTHING

Now consider the case in which forecasts are generated by exponential smoothing. In this case F_t may be represented by the weighted infinite sum of past values of demand.

$$F_t = \alpha D_{t-1} + \alpha(1-\alpha)D_{t-2} + \alpha(1-\alpha)^2 D_{t-3} + \cdots,$$
$$E(F_t) = \mu[\alpha + \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \cdots] = \mu.$$

Notice that this means that $E(e_t) = 0$, so that both exponential smoothing and moving averages are unbiased forecasting methods when the underlying demand process is a constant plus a random term.

$$\text{Var}(F_t) = \alpha^2\sigma^2 + (1-\alpha)^2\alpha^2\sigma^2 + \cdots$$
$$= \sigma^2\alpha^2 \sum_{n=0}^{\infty} (1-\alpha)^{2n}.$$

It can be shown that

$$\sum_{n=0}^{\infty} (1-\alpha)^{2n} = \frac{1}{1-(1-\alpha)^2}$$

so that

$$\text{Var}(F_t) = \frac{\sigma^2\alpha^2}{1-(1-\alpha)^2} = \frac{\sigma^2\alpha}{2-\alpha}.$$

Since

$$\text{Var}(e_t) = \text{Var}(F_t) + \text{Var}(D_t),$$
$$\text{Var}(e_t) = \sigma^2[\alpha/(2-\alpha) + 1] = \sigma^2[2/(2-\alpha)],$$

or

$$\sigma_e = \sigma\sqrt{\frac{2}{2-\alpha}}.$$

This is the standard deviation of the forecast error for simple exponential smoothing in terms of the standard deviation of each observation. The distribution of the forecast error for exponential smoothing is normal for essentially the same reasons as stated above for moving averages.

Notice that if we equate the variances of the forecast error for exponential smoothing and moving averages, we obtain

$$2/(2-\alpha) = (N+1)/N$$

or $\alpha = 2/(N+1)$, which is exactly the same result as we obtained by equating the average age of data for the two methods.

Appendix **2–B**

Derivation of the Equations for the Slope and Intercept for Regression Analysis

In this appendix we derive the equations for the optimal values of a and b for the regression model. Assume that the data are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, and the regression model to be fitted is $Y = a + bX$. Define

$$g(a, b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.$$

Interpret $g(a, b)$ as the sum of the squares of the distances from the line $a + bx$ to the data points y_i. The object of the analysis is to choose a and b to minimize $g(a, b)$. This is accomplished where

$$\frac{\partial g}{\partial a} = \frac{\partial g}{\partial b} = 0.$$

That is,

$$\frac{\partial g}{\partial a} = -\sum_{i=1}^{n} 2[y_i - (a + bx_i)] = 0,$$

$$\frac{\partial g}{\partial b} = -\sum_{i=1}^{n} 2x_i[y_i - (a + bx_i)] = 0,$$

which results in the two equations

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{1}$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{2}$$

These are two linear equations in the unknowns a and b. Multiplying Equation (1) by Σx_i and Equation (2) by n gives

$$an\sum_{i=1}^{n} x_i + b\left(\sum_{i=1}^{n} x_i\right)^2 = \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right), \tag{3}$$

$$an\sum_{i=1}^{n} x_i + bn\sum_{i=1}^{n} x_i^2 = n\sum_{i=1}^{n} x_i y_i. \tag{4}$$

Subtracting Equation (3) from Equation (4) results in

$$b\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] = n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right). \tag{5}$$

Define $S_{xy} = n\,\Sigma\, x_i y_i - (\Sigma\, x_i)(\Sigma\, y_i)$ and $S_{xx} = n\Sigma\, x_i^2 - (\Sigma x_i)^2$. It follows that Equation (5) may be written $bS_{xx} = S_{xy}$, which gives

$$b = \frac{S_{xy}}{S_{xx}}. \tag{6}$$

From Equation (1) we have

$$an = \sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i$$

or

$$a = \bar{y} - b\bar{x}, \tag{7}$$

where $\bar{y} = (1/n)\Sigma\, y_i$ and $\bar{x} = (1/n)\Sigma\, x_i$.

These formulas can be specialized to the forecasting problem when the independent variable is assumed to be time. In that case, the data are of the form $(1, D_1), (2, D_2), \ldots, (n, D_n)$, and the forecasting equation is of the form $\hat{D}_t = a + bt$. The various formulas can be simplified as follows:

$$\sum x_i = 1 + 2 + 3 + \cdots + n = \frac{n(n + 1)}{2},$$

$$\sum x_i^2 = 1 + 4 + 9 + \cdots + n^2 = \frac{n(n + 1)(2n + 1)}{6}.$$

Hence, we can write

$$S_{xy} = n \sum_{i=1}^{n} iD_i - n(n+1)/2 \sum_{i=1}^{n} D_i,$$

$$S_{xx} = \frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4}.$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$a = \overline{D} - \frac{b(n+1)}{2}$$

Appendix **2–C**

Glossary of Notation for Chapter 2

a = Estimate of the intercept in regression analysis.

α = Smoothing constant used for single exponential smoothing. One of the smoothing constants used for Holt's method or one of the smoothing constants used for Winters's method.

b = Estimate of the slope in regression analysis.

β = Second smoothing constant used for either Holt's method or Winters's method.

c_t = Seasonal factor for the tth period of a season.

γ = Third smoothing constant used for Winters's method.

D_t = Demand in period t. Refers to the series whose values are to be forecasted.

$e_t = F_t - D_t$ = (Observed) forecasting error in period t.

ϵ_t = Random variable representing the random component of the demand.

F_t = One-step-ahead forecast made in period $t-1$ for demand in period t.

$F_{t,t+\tau}$ = τ-step-ahead forecast made in period t for the demand in period $t+\tau$.

G_t = Smoothed value of the slope for Holt's and Winters's methods.

μ = Mean of the demand process.

MAD = Mean absolute deviation = $(1/n)\sum_{i=1}^{n} |e_i|$.

MAPE = Mean absolute percentage error = $(1/n)\sum_{i=1}^{n} |e_i/D_i| \times 100$.

MSE = Mean squared error = $(1/n)\sum_{i=1}^{n} e_i^2$.

S_t = Smoothed value of the series (intercept) for Holt's and Winters's methods.

σ^2 = Variance of the demand process.

T_t = Value of the tracking signal in period t (refer to Problem 57).

Bibliography

Armstrong, J. S. "Forecasting by Extrapolation: Conclusions from Twenty-Five Years of Research." *Interfaces* 14 (1984), pp. 52–66.

Box, G. E. P., and G. M. Jenkins. *Time Series Analysis, Forecasting, and Control*. San Francisco: Holden Day, 1970.

Brown, R. G. *Statistical Forecasting for Inventory Control*. New York: McGraw-Hill, 1959.

Brown, R. G. *Smoothing, Forecasting, and Prediction of Discrete Time Series.* Englewood Cliffs, NJ: Prentice Hall, 1962.

Cohen, D. "Facebook Has 1.15B Monthly Active Users, But How Many Are Invalid?" *AllFacebook* (www.allfacebook.com), accessed September 3, 2013.

Davenport, W. B., and W. L. Root. *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill, 1958.

Draper, N. R., and H. Smith. *Applied Regression Analysis*. New York: John Wiley & Sons, 1968.

Fisher, M. L.; J. H. Hammond; W. R. Obermeyer; and A. Raman, "Making Supply Meet Demand in an Uncertain World." *Harvard Business Review*, May–June 1994, pp. 221–40.

Helmer, O., and N. Rescher. "On the Epistemology of the Inexact Sciences." *Management Science* 6 (1959), pp. 25–52.

Kalman, R. E. "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering*, Ser. D 82, 1960, pp. 35–44.

Makridakis, S.; S. C. Wheelwright; and R. J. Hyndman. *Forecasting: Methods and Applications*. 3rd ed. New York: John Wiley & Sons, 1998.

Makridakis, S., and R. L. Winkler. "Averages of Forecasts." *Management Science* 29 (1983), pp. 987–96.

McWilliams, G. "At Compaq, a Desktop Crystal Ball." *Business Week*, March 20, 1995, pp. 96–97.

Nahmias, S. "Demand Estimation in Lost Sales Inventory Systems." *Naval Research Logistics* 41 (1994), pp. 739–57.

Nelson, C. R. "The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy." *The American Economic Review* 62, no. 5 (December 1972), pp. 902–917.

Nelson, C. R. *Applied Time Series Analysis for Managerial Forecasting*. San Francisco: Holden Day, 1973.

Trigg, D. W. "Monitoring a Forecasting System." *Operational Research Quarterly* 15 (1964), pp. 271–74.

Trigg, D. W., and A. G. Leach. "Exponential Smoothing with Adaptive Response Rate." *Operational Research Quarterly* 18 (1967), pp. 53–59.

Wiener, N. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA: MIT Press, 1949.

Yule, G. U. "Why Do We Sometimes Get Nonsense Correlations between Time Series? A Study of Sampling and the Nature of Time Series." *Journal of the Royal Statistical Society* 89 (1926), pp. 1–64.

Chapter **Three**

Sales and Operations Planning

"In preparing for battle I have always found that plans are useless, but planning is indispensable."

—Dwight D. Eisenhower

Chapter Overview

Purpose

To present the process by which companies go from technical forecasts to aggregate level sales and operations plans.

Key Points

1. *The sales and operations planning (S&OP) process.* This chapter could also be called Macro Planning, since the purpose of an S&OP process is to develop top-down sales and operations plans for the entire firm, or for some subset of the firm such as a product line or particular plant. The key goals of the process are to (a) make aggregate level plans that all divisions as well as suppliers can work to; (b) resolve the inherent tensions between sales and operations divisions; and (c) anticipate and escalate strategic challenges in matching supply with demand for the firm.
2. *Key Performance Indicators.* Inherent in any S&OP process are a set of metrics, or key performance indicators (KPIs), that the firm uses to judge the performance of the different divisions. Effective KPIs measure important factors, are relatively easy to compute, and are actionable, in the sense that those being measured by it can also effect its change. Operational KPIs may be efficiency or effectiveness focused and must be aligned with the strategic goals of the firm.
3. *The role of uncertainty.* It is important to explicitly recognize the role of uncertainty in planning. For a firm producing a wide range of products, or providing a service, uncertainty can be particularly problematic in planning, and management techniques must be adjusted accordingly. Furthermore, different types of uncertainty require different types of management responses. The S&OP process needs to explicitly acknowledge possible major sources of uncertainty and plan for them appropriately.

4. *Costs in aggregate operations plans.* The following are key costs to be considered in developing aggregate operations plans.
   - *Smoothing costs.* The cost of changing production and/or workforce levels.
   - *Holding costs.* The opportunity cost of dollars invested in inventory.
   - *Shortage costs.* The costs associated with back-ordered or lost demand.
   - *Labor costs.* These include direct labor costs on regular time, overtime, subcontracting costs, and idle time costs.
5. *Solving aggregate planning problems.* Approximate solutions to aggregate planning problems can be found graphically and exact solutions via linear programming. A level plan has constant production or workforce levels over the planning horizon, while a chase strategy keeps zero inventory and minimizes holding and shortage costs. A linear programming formulation assumes that all costs are linear and typically does not take into account management policy, such as avoiding hiring and firing as much as possible.
6. *Disaggregating plans.* While aggregate planning is useful for providing approximate solutions for macro planning at the firm level, the question is whether these aggregate plans provide any guidance for planning at the lower levels of the firm. A disaggregation scheme is a mean of taking an aggregate plan and breaking it down to get more detailed plans at lower levels of the firm.

As we go through life, we make both micro and macro decisions. Micro decisions might be what to eat for breakfast, what route to take to work, what auto service to use, or which movie to rent. Macro decisions are the kind that change the course of one's life: where to live, what to major in, which job to take, whom to marry. A company also must make both micro and macro decisions every day. Some macro decisions are highly strategic in nature, such as process technology or market entry choices, and were discussed in Chapter 1. Other macro decisions are more tactical, such as planning companywide workforce and production levels or setting sales target. In this chapter we explore tactical decisions made at the macro level in the context of a process known as **sales and operations planning** or **S&OP** for short.

S&OP begins with demand forecasts and turns them into targets for both sales and operations plans; techniques for demand forecasting were presented in Chapter 2. However, a firm will not want to just use such forecasts blindly. They may want to produce more than is forecast if stock-outs are unacceptable or if they are planning a major promotion; they may want to produce less than is forecast if overstocks are costly or they see the product winding down in its lifecycle. Such decisions must be made at a high strategic level and must involve both the sales and the operations staff.

One of the key goals in S&OP is to resolve the fundamental tension between sales and operations divisions in an organization. Sales divisions are typically measured on revenue; they want the product 100% available with as many different varieties as possible. Meanwhile, operations divisions are typically measured on cost; they want to keep both capacity and inventory costs low, which means limiting overproduction and product varieties. The best way to resolve these inherent tensions is with a formal S&OP process that gets the heads of sales and operations together in a room with other high level executives. This chapter explores what such processes look like.

Core to the S&OP process is a review of divisional **Key Performance Indicators** (KPIs). Section 3.2 reviews key challenges in KPI selection and key types of operational KPIs. One the fundamental challenges in KPI selection is in making sure the

KPI is both sufficiently high-level to be well aligned with the corporate strategy while being sufficiently low-level so that it provides a meaningful guide to behavior for those who are being evaluated by it.

The S&OP process would be a lot simpler if demand were certain; however, as noted in Chapter 2, it is not and therefore forecasts are generally wrong. In fact, there are a range of uncertainties that must be considered in the planning process from **known unknowns** to **unknown unknowns**, terms which will be formally defined in Section 3.3. Part of an S&OP process must include decisions around how to plan for and mitigate risk or uncertainty.

An important part of S&OP is **aggregate planning**, which might also be called macro production planning. It addresses the problem of deciding how many employees the firm should retain and, for a manufacturing firm, the quantity and the mix of products to be produced. Macro planning is not limited to manufacturing firms. Service organizations must determine employee staffing needs as well. For example, airlines must plan staffing levels for flight attendants and pilots, and hospitals must plan staffing levels for nurses. Macro planning strategies are a fundamental part of the firm's overall business strategy. Some firms operate on the philosophy that costs can be controlled only by making frequent changes in the size and/or composition of the workforce. The aerospace industry in California in the 1970s adopted this strategy. As government contracts shifted from one producer to another, so did the technical workforce. Other firms have a reputation for retaining employees, even in bad times. Traditionally, IBM and AT&T were two well-known examples.

Aggregate planning methodology is designed to translate demand forecasts into a blueprint for planning staffing and production levels for the firm over a predetermined planning horizon. Aggregate planning methodology is not limited to top-level planning. Although generally considered to be a macro planning tool for determining overall workforce and production levels, large companies may find aggregate planning useful at the plant level as well. Production planning may be viewed as a hierarchical process in which purchasing, production, and staffing decisions must be made at several levels in the firm. Aggregate planning methods may be applied at almost any level, although the concept is one of managing groups of items rather than single items.

This chapter outlines the S&OP process. Core to the process are KPIs, thus this chapter briefly reviews principles involved in setting KPIs. The chapter also describes the types of uncertainty that must be planned for in the process and how they are best addressed. Core to the operations component of any S&OP process is determining an aggregate production plan for capacity and inventory. This chapter reviews several techniques for determining aggregate production plans. Some of these are heuristic (i.e., approximate) and some are optimal. We hope to convey to the reader an understanding of the issues involved in S&OP, a knowledge of the basic tools available for providing production plans, and an appreciation of the difficulties associated with such planning in the real world.

3.1 The S&OP Process

The S&OP process is designed to produce a plan that all divisions within in the organization, as well as suppliers to the organization, can work to. The process is also sometimes referred to as sales, inventory, and operations planning (SIOP) to emphasize the important role that inventory can play as a buffer between sales planning and operations planning.

In any organization, demand is a function of sales effort and pricing and supply is a function of operations effort and capacity. Therefore, in order to best balance supply with demand a strategic approach must be applied. As defined by Grimson and Pyke (2007), "S&OP is a business process that links the corporate strategic plan to daily operations plans and enables companies to balance demand and supply for their products."

One of the key words in Grimson and Pike's definition of S&OP is that it is a "business process." That is, it a consistent set of steps that a company follows on a regular basis. Each business will apply its process slightly differently but some common elements include the following:

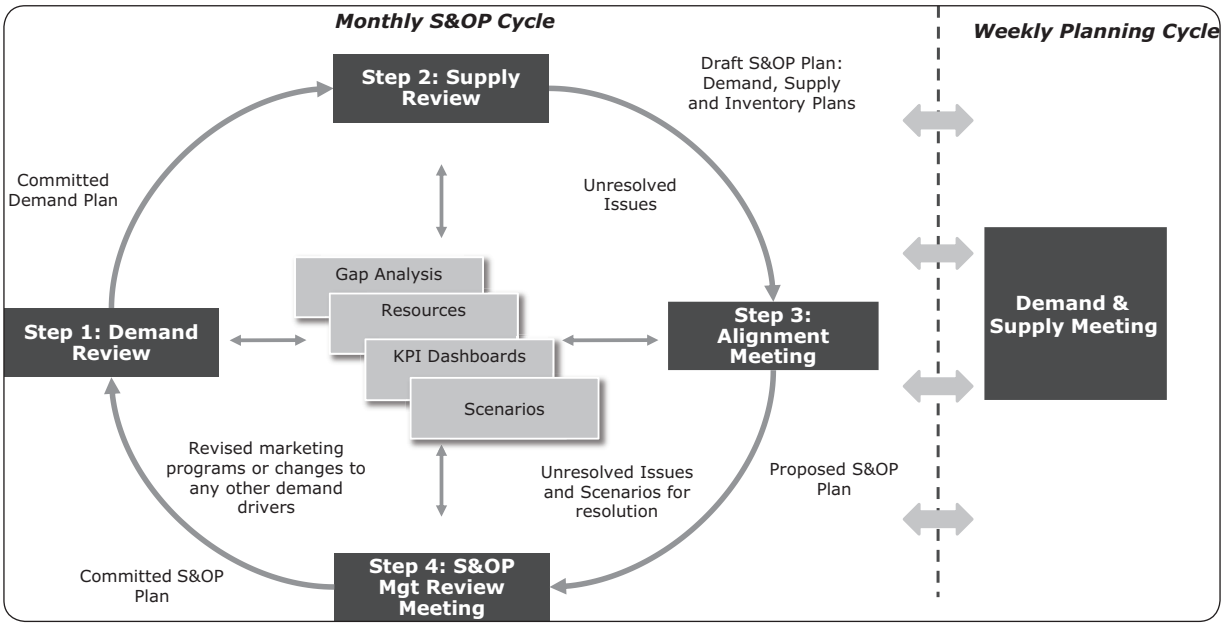
1. *Strategically focused.* As discussed above, finding the appropriate balance for supply and demand requires knowledge of the company's strategic plan. Typically c-level executives are involved in S&OP planning to ensure that the company's strategy is appropriately represented.
2. *Cross-functional team.* Because the S&OP process must balance the interests of different parts of the organization, the team performing the process must be both cross-functional and balanced.
3. *Aggregated.* It is typically not possible to forecast at the individual stock keeping unit (SKU) level. Thus S&OP processes typically work with aggregated units of production, such as product families. One natural unit of aggregation is sales dollars, but of course there is not a one-to-one mapping from sales dollars to units produced.
4. *Time fences.* The agreed sales and operations plan will typically be only fixed for a relatively short period (e.g., a week or month) and forecasts for future time periods will typically be assumed to be flexible within a given range. Oftentimes fences are used to show when the level of certainty changes and can be referred to as frozen (i.e., fixed), slushy (somewhat fixed), and liquid (highly flexible) time periods.

There are a number of key inputs required for the S&OP process as follows:

1. *Technical demand forecast*. Advanced forecasting techniques such as those discussed in Chapter 2 are used to understand raw demand.
2. *Sales plans*. The Marketing and Sales divisions will use the technical forecasts and their own promotion and marketing plans to provide a sales forecast.
3. *Operations plans*. Operations will produce a production plan that includes plans for capacity, supply, and inventory.
4. *Innovation plans*. The planned role out of new products and future product innovation must be considered in the context of sales and operations plans.
5. *Financial plans.* Any public company must announce earnings forecasts, typically with a breakdown of how such earnings are to be achieved. Because negative market surprises can affect a company's ability to raise capital, the sales and operations plans must also consider the financial forecasts. Further, in many smaller companies cash flow constraints must also be considered in the planning process.

The basic S&OP process must iterate between these five key inputs above to resolve tensions and unforeseen constraints (e.g., a supplier who cannot deliver until six weeks from today) to finalize the sales and operations plans for the company. It is then left to the sales and operations divisions of the company to execute on these plans, such as disaggregating sales dollars into actual production units.

FIGURE 3–1
S&OP Overview: Source: Smits and English (2010)



One sample iterative process is shown in Figure 3–1, which has been produced for Heineken by Smits and English (2010). Notice how the outcome of the S&OP meeting is a committed S&OP plan for the following month. Key uncertainties are shown as "scenarios" in the above.

Typically the chief executive officer (CEO) leads the S&OP meeting and makes the key decisions with respect to trade-offs (Sheldon, 2006). A well run meeting is focused on decision making, rather than information dissemination, and all participants are expected to have studied the materials that are prepared ahead of time by the different divisions. The vice president of sales and/or marketing will be present and is the process owner for the demand plan. The vice president of operations and/or supply chain management will also be present and is the process owner for the operations plan. The chief financial officer (CFO) will attend and prepares the financial plan, which measures performance for the master business plan. The master scheduler prepares documents and is the process owner for weekly schedules (which are not typically part of the S&OP process). In a larger organization this will take place at the divisional level, although the CEO may well attend at least semi-regularly divisional S&OP meetings.

A standard agenda for a monthly S&OP meeting is the following:

- Review of last 30 days—Accuracy of financial, demand, and operations plans (typically by product family)
- Review of 30- to 60-day plan expectations—Risks for each plan

Snapshot Application

Heineken International was founded in 1864 by Gerard Adriaan Heineken in Amsterdam and has a global network of distributors and 125 breweries in more than 70 countries. Heineken's corporate strategy is "to be a leading brewer in each of the markets in which it operates and to have the world's most prominent brand portfolio (with Heineken as leading premium beer)." Supply chain planning within the beer supply chain is complicated by the fact that most beers, including Heineken, are produced in batches. In fact, Heineken has a highly prescribed process for its brewing that all producers must follow to ensure global consistency in taste. It also has a robust S&OP process, which it implements in all of its subsidiaries. The key mantra for this process, pictured in Figure 3–1, is "one single plan." This process is described as follows:

> The global beer market is a dynamic environment with changing markets, strong competition, and changing customer preferences. Heineken realized that a Global Sales and Operations Planning (S&OP) program would become a key enabler in supporting aggressive expansion targets, and would become necessary to support a retail globalization landscape, which is applying increasing pressures on costs and service. Heineken's S&OP process integrates finance, marketing, sales, and supply chain departments with the objective of aligning the organization towards a synchronized operating plan globally. This program is supported by a very strong project management approach which has been designed to provide enough consistency across regions, yet provide enough flexibility to embrace and benefit from local cultural differences. (Smits & English, 2010)

Notice how this description matches with the general description of S&OP given above yet leaves some flexibility in terms of the actual running of meeting to allow for local customs to apply. The benefits Heineken has realized from their S&OP process are better cross-functional decisions; an enabler of growth; higher capacity utilization; lower supply chain costs; and reduced working capital (Rooijen, 2011).

- Review 90- to 120-day risks as required
- Review balance of 12-month horizon for exceptions (Sheldon, 2006)

Typically, an S&OP plan is considered fixed (with agreed windows of flexibility for later time periods) and, therefore, the operations division must plan for uncertainty within its own execution plan. Common strategies for dealing with uncertainty include the following:

1. *Buffering*: Maintaining excess resources (inventory, capacity, or time) to cover for fluctuations in supply or demand. Such buffering is an explicit component of planning under uncertainty, such as the inventory models considered in Chapter 5.
2. *Pooling*: Sharing buffers to cover multiple sources of variability (e.g., demand from different markets). This will be covered in more detail in Section 6.7.
3. *Contingency planning*: Establishing a preset course of action for an anticipated scenario. This may involve strategies such as bringing on a backup supplier or sourcing from the spot market. This is best done explicitly in the S&OP process.

Heineken Netherlands positions the S&OP process within their planning framework as given in Figure 3–2.

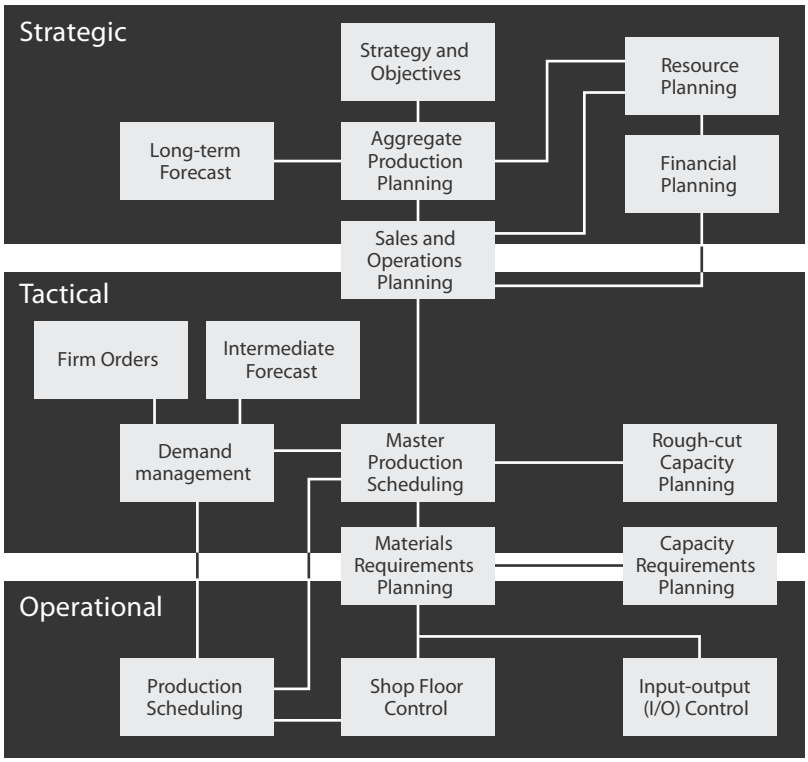


FIGURE 3–2
Planning and Control Framework for Heineken Netherlands. Source: Every Angle (2013)

Notice how in Figure 3–2, S&OP straddles the strategy and tactical portions of the planning framework. Many of the other planning topics, such as the master production schedule and materials requirements planning, will be covered in Chapter 8.

Problems for Section 3.1

1. Why don't the head of sales or the head of operations chair the S&OP meeting?
2. Why do you think it is important that the CFO attends the S&OP meeting?
3. Describe the likely key tensions between the sales and operations divisions of a small manufacturer that mostly produces to the local market. Now describe them for a global organization that sells in many regions and outsources most of its production to China.
4. Why is S&OP described as an iterative process?
5. What are the trade-offs between having long fixed or frozen periods versus short fixed or frozen periods as outputs of the S&OP process?
6. What are the trade-offs between using revenue dollars versus sales quantities as the units in the aggregate S&OP plan.

3.2 Key Performance Indicators

One of the quantitative inputs to the S&OP process will be the key performance indicators (KPIs) for the divisions and the organization. The CFO is present at the meeting to present key financial performance measurements such as net profit and return on investment. However, as discussed in Chapter 1, such financial measures tend to emphasize short-term goals. This issue is often counteracted by the **balanced score card approach**, which provides a more well-rounded performance measurement approach covering a wider scope of metrics than pure finance measures. The balanced score card approach was developed by Kaplan and Norton (1992) and represents performance measures designed to reflect strategy and communicate a vision to the organization. There are four standard perspectives for the approach, although many organizations have adapted these for their own use; these perspectives are (a) customer; (b) internal; (c) innovation and learning; and (d) financial. Figure 3–3 depicts these perspectives.

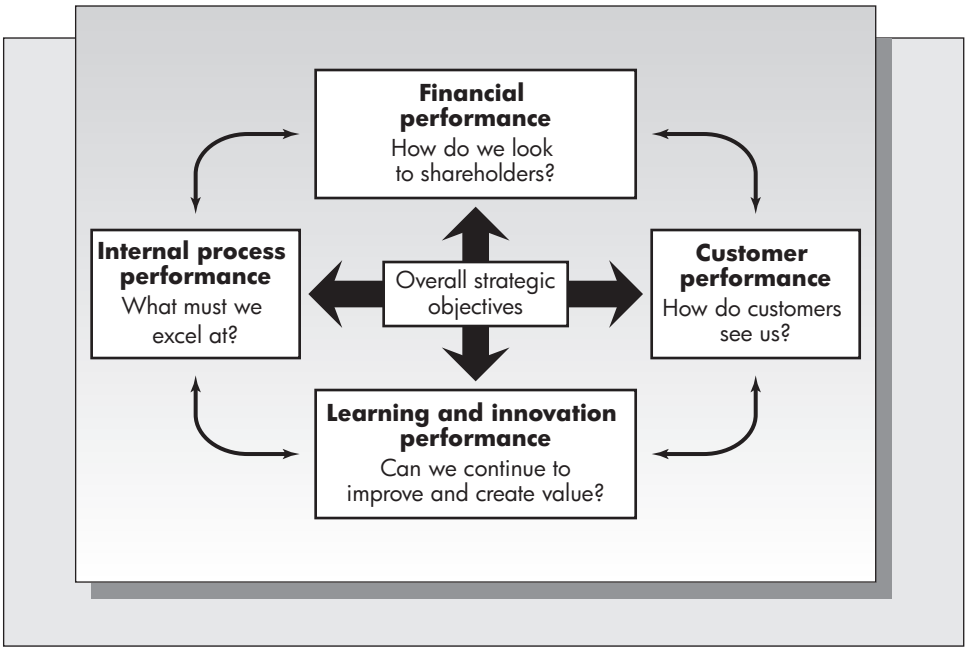


FIGURE 3–3
Perspectives for the Balanced Scorecard Approach; Source: Kaplan and Norton (1992)

Operational metrics typically sit within internal process measures. There are two key types of operational KPIs, namely **efficiency** related KPIs, which are a measure of resource utilization (e.g., time, cost, materials, etc.) and **effectiveness** related KPIs, which are a measure of how well the process meets or exceeds customer requirements (e.g., defects, complaints, satisfaction score, etc.). Which of these are most important will depend on the product-line strategy, with efficiency being the most important for products that compete on cost and effectiveness being the most important for high-margin products that compete on innovation or fashion dimensions.

Some evaluation criteria for the merits of a KPI (ABB, 2010) are: (a) *Importance*—Are you measuring the things that really matter? (b) *Ease*—Does the measurement 'flow' from the activity being monitored? (c) *Actionable*—Can the metric initiate

appropriate actions? The KPI must be well understood by employees if it is to be both measured accurately and to incentivize behavior correctly. One of the key challenges in choosing KPIs is in achieving organizational alignment. KPIs must be aligned to the strategic plan in order to monitor the essential performance results implied by the vision and mission statements and to monitor current strategic goals or objectives. It is then recommended that they be cascaded down to the business processes to maintain alignment. Finally, the KPIs may be assigned to business units once they have been associated with business processes. Going the opposite direction starting from business units is more likely to result in misaligned KPIs. In general, a high level KPI (e.g., monthly profit) is easier to align but if employees don't feel that they can control the KPI then they may disengage. This is the reason for the recommendation that KPIs be "actionable" in the above.

In the book *The Goal* by Goldratt and Cox (1992) it is stated that "The goal of a firm is to make money." Many of the issues described in the book arise from misaligned KPIs to this goal. Goldratt and Cox (1992) suggest that there are really only three operational KPIs, namely throughput, inventory, and operational expenses; however, their definitions of these terms are nonstandard. Throughput is defined as "the rate at which money is generated by the system through sales," inventory is defined as "all the money that the system has invested in purchasing things it intends to sell," and operating expenses are defined as "all the money that the system spends to turn inventory into throughput." The beauty of these metrics is that they indeed align incentives with the goal of making money, the challenge is that firms have found them difficult to implement in practice and they have not been widely adopted by the field of managerial accounting.

A final challenge in KPI use is gaming of the KPI by employees. If the KPI definition leaves room for interpretation then those being judged by it will typically interpret it in their favor. One of the most common operational KPIs is DIFOT, which stands for Delivery In-Full and On-Time. However, the definition of "on-time" needs to be spelled out carefully—does it relate to when the order is shipped or when it arrives at the customer? Of course, the customer only cares about the latter, but many firms measure the former as both easier to measure and more fully under operations control (i.e., more actionable). Clearly there are trade-offs when choosing between the two perspectives.

Problems for Section 3.2

7. In some S&OP processes all divisions are measured by the same set of KPIs. What are the advantages and disadvantages to this approach?
8. Give two efficiency-related KPIs and two effectiveness-related KPIs that you believe Wal-Mart is likely to use to evaluate its suppliers? Its in-store managers?
9. What types of organizations are likely to face the largest challenges in achieving alignment between operational KPIs and strategic priorities?
10. Evaluate the following metrics along the three dimensions of importance, ease, and actionablity: DIFOT, ROI, machine utilization, and number of customer complaints.
11. Do you think efficiency-related KPIs or effectiveness-related KPIs will be easier for employees to "game" in practice? Explain your answer.
12. Why do you think that the KPIs proposed by Goldratt and Cox (1992) have not been very widely implemented?

3.3 THE ROLE OF UNCERTAINTY

One of the key challenges in sales and operations planning is the effective management of uncertainty or risk. In a survey by IBM of 400 Supply Chain Executives (IBM, 2009) risk was identified as the second most important key challenge that comprise the Chief Supply Chain Officer agenda; it was identified by 60 percent of respondents as important "to a very great extent" or "to a significant extent" (the challenge that was ranked highest was supply chain visibility). Because S&OP is the highest level of tactical planning in the organization it must explicitly acknowledge major sources of uncertainty and explicitly work to manage and/or mitigate risk.

De Meyer et al. (2002) highlight four different types of uncertainty. While they particularly focus on project management, the categorization also applies to S&OP processes and operational risks.

1. *Variation*. Variation is anything that causes the system to depart from regular, predictable behavior that may be described by a probability distribution (see Appendix 5–A). Typical sources for variation include variability in customer orders, differential worker skill levels, and variability in quality or lead times.

2. *Foreseen uncertainty*. Foreseen uncertainty, or "known unknowns" as they are often called, are risks that can be foreseen and planned for. They are typically described by a probability of occurrence and have a larger impact than variation as described above. Typical sources include supply breakdowns, design changes, natural disasters, labor strikes, etc.

3. *Unforeseen uncertainty*. Unforeseen uncertainty, or "unknown unknowns," are risks that could not be predicted ahead of time or that are considered too unlikely to make contingency plans for. Typical sources include natural disasters of unusual scale or in regions where they do not typically occur (e.g., an earthquake on the eastern coast of the United States) or competitor innovations that were not anticipated. These are the so-called "black swan" events of Taleb (2007).

4. *Chaos*. Chaos is unforeseen uncertainty where the event not only affects the operations but also the fundamental goal of the project or company. For example, in the wake of the New Zealand city of Christchurch's 2010 earthquake, a number of local beverage manufacturers offered to switch to bottling water instead of regular products. Not only was the earthquake unforeseen because the fault rupture occurred along a previously unknown east-west fault line that is thought to not have moved for at least 16,000 years, but the goal for the beverage manufacturers shifted from profit from beverage sales to humanitarian aid.

Each of the above forms of uncertainty must be dealt with and planned for in different ways. However, all require a firm to have effective management and communication processes.

Variation is the type of uncertainty most commonly dealt with by operations analysts and managers and will be described in more detail in Section 5.1a. Because variation is considered routine and within an operations manager purview to anticipate and plan for, it is not usually much discussed during the S&OP process. Strategies used by operations to mitigate variation include holding inventory, deliberately maintaining excess capacity, or incurring deliberate delays in order fulfillment. Foreseen uncertainty is best dealt with specific contingency plans, risk mitigation processes, and clear ownership responsibility for processes. It is foreseen uncertainty that is best dealt with explicitly within the S&OP process. For example, if it is known that a supplier is currently struggling with labor issues, then it should be discussed at the S&OP meeting what plans should be put in place for the event that the supplier's workforce goes on strike.

Unforeseen uncertainty may be mitigated by generic contingency plans and good leadership while chaos requires strong leadership and crisis management processes. By definition, neither unforeseen uncertainty nor chaos may be explicitly planned for in the S&OP process. However, an organization with strong cross-functional ties, which good S&OP processes foster, is going to manage such occurrences better than a siloed organization with little cross-functional planning.

Problems for Section 3.3

13. Why does the S&OP process typically not explicitly recognize variation even though it is a fact of life for both operations and sales divisions?
14. Classify the following risks into variation, foreseen uncertainty, unforeseen uncertainty, and chaos:
    a. A hurricane on the U.S. East Coast floods a regional warehouse destroying a large amount of stock.
    b. A machine on the plant floor breaks down for an hour.
    c. Bad weather on the weekend causes an increase in demand for umbrellas.
    d. A cool summer causes a decrease in demand for air conditioners for that season.
    e. The excavation process for a new manufacturing plant in the U.S. Midwest uncovers an archaeological find of such significance that no building can take place on that site and a new site for the plant must be found.
    f. Competitors to the iPad launch smaller tablet computers before the iPad mini is ready to launch, thus negatively affecting demand for the iPad.
    g. The Second World War caused auto manufacturers to switch to producing military vehicles.
    h. A drug is found to have dangerous side effects following its launch.
    i. The transportation disruptions, including the grounding of all airplanes, following the attacks on September 11, 2001 severed many supply chains.
15. List two examples each of variation, foreseen uncertainty, and unforeseen uncertainty that you have personally experienced in your studies.
16. Give an example of chaos, either from your own experience or from others, within the educational domain.

3.4 AGGREGATE PLANNING OF CAPACITY

The operations division is responsible for determining an aggregate plan for capacity usage throughout the planning horizon. This plan uses the agreed upon set of sales forecasts that comes out of the S&OP meeting. Such a forecast is expressed in terms of aggregate production units or dollars. The operations division must then determine aggregate production quantities and the levels of resources required to achieve these production goals. In practice, this translates to finding the number of workers that should be employed and the number of aggregate units to be produced in each of the planning periods 1, 2, …, T. The objective of such aggregate planning is to balance the advantages of producing to meet demand as closely as possible against the disturbance caused by changing the levels of production and/or the workforce levels.

As just noted, aggregate planning methodology requires the assumption that demand is known with certainty. This is simultaneously a weakness and a strength

of the approach. It is a weakness because it does not provide any buffer against unanticipated forecast errors. However, most inventory models that allow for random demand require that the average demand be constant over time. Aggregate planning allows the manager to focus on the systematic changes that are generally not present in models that assume random demand. By assuming deterministic demand, the effects of seasonal fluctuations and business cycles can be incorporated into the planning function. As discussed in Section 3.1, variation caused by demand uncertainty may be buffered using inventory, capacity, or customer delays.

Costs in Aggregate Capacity Planning

As with most of the optimization problems considered in production management, the goal of the analysis is to choose the aggregate plan that minimizes cost. It is important to identify and measure those specific costs that are affected by the planning decision.

1. *Smoothing costs*. Smoothing costs are those costs that accrue as a result of changing the production levels from one period to the next. In the aggregate planning context, the most salient smoothing cost is the cost of changing the size of the workforce. Increasing the size of the workforce requires time and expense to advertise positions, interview prospective employees, and train new hires. Decreasing the size of the workforces means that workers must be laid off. Severance pay is thus one cost of decreasing the size of the workforce. However, there are other costs associated with firing workers that may be harder to measure.

Firing workers could have far-reaching consequences. Firms that hire and fire frequently develop a poor public image. This could adversely affect sales and discourage potential employees from joining the company. It may adversely affect employee morale. Furthermore, workers that are laid off might not simply wait around for business to pick up. Firing workers can have a detrimental effect on the future size of the labor force if those workers obtain employment in other industries. Finally, most companies are simply not at liberty to hire and fire at will. Labor agreements restrict the freedom of management to freely alter workforce levels.

Most of the models that we consider assume that the costs of increasing and decreasing the size of the workforce are linear functions of the number of employees that are hired or fired. That is, there is a constant dollar amount charged for each employee hired or fired. The assumption on linearity is probably reasonable up to a point. As the supply of labor becomes scarce, there may be additional costs required to hire more workers, and the costs of laying off workers may go up substantially if the number of workers laid off is too large. A typical cost function for changing the size of the workforce appears in Figure 3–4.

2. *Holding costs*. Holding costs are the costs that accrue as a result of having capital tied up in inventory. If the firm can decrease its inventory, the money saved could be invested elsewhere with a return that will vary with the industry and with the specific company. (A more complete discussion of holding costs is deferred to Chapter 4.) Holding costs are almost always assumed to be linear in the number of units being held at a particular point in time. We will assume for the purposes of the aggregate planning analysis that the holding cost is expressed in terms of dollars per unit held per planning period. We also will assume that holding costs are charged against the inventory remaining on hand at the *end* of the planning period. This assumption is made for

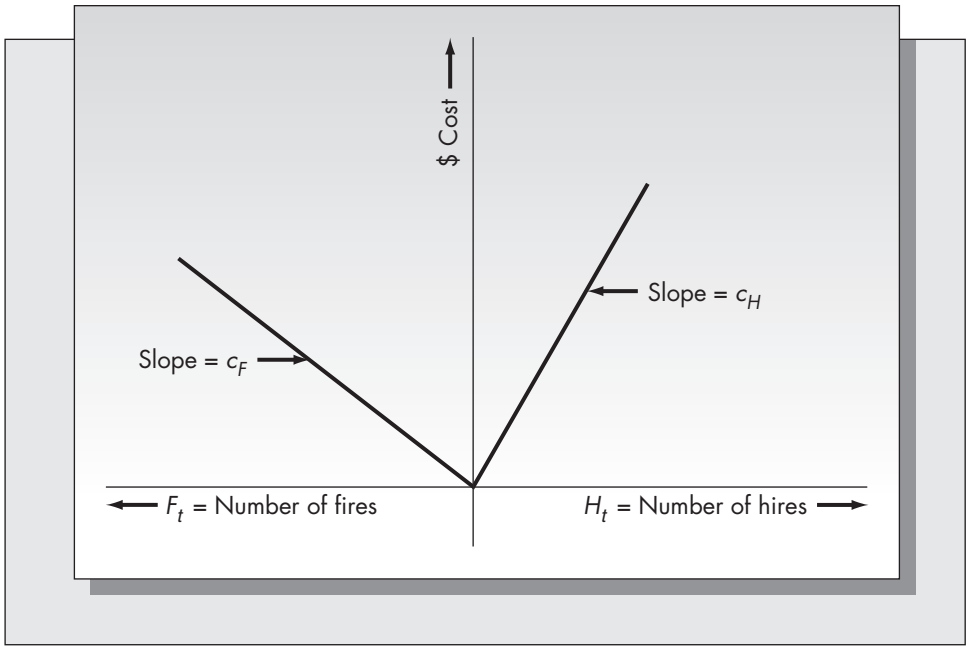


FIGURE 3–4
Cost of changing the size of the workforce

convenience only. Holding costs could be charged against starting inventory or average inventory as well.

3. *Shortage costs.* Holding costs are charged against the aggregate inventory as long as it is positive. In some situations it may be necessary to incur shortages, which are represented by a negative level of inventory. Shortages can occur when forecasted demand exceeds the capacity of the production facility or when demands are higher than anticipated. For the purposes of aggregate planning, it is generally assumed that excess demand is backlogged and filled in a future period. In a highly competitive situation, however, it is possible that excess demand is lost and the customer goes elsewhere. This case, which is known as lost sales, is more appropriate in the management of single items and is more common in a retail than in a manufacturing context.

As with holding costs, shortage costs are generally assumed to be linear. Convex functions also can accurately describe shortage costs, but linear functions seem to be the most common. Figure 3–5 shows a typical holding/shortage cost function.

4. *Regular time costs.* These costs involve the cost of producing one unit of output during regular working hours. Included in this category are the actual payroll costs of regular employees working on regular time, the direct and indirect costs of materials, and other manufacturing expenses. When all production is carried out on regular time, regular payroll costs become a "sunk cost," because the number of units produced must equal the number of units demanded over any planning horizon of sufficient length. If there is no overtime or worker idle time, regular payroll costs do not have to be included in the evaluation of different strategies.

5. *Overtime and subcontracting costs.* Overtime and subcontracting costs are the costs of production of units not produced on regular time. Overtime refers to production by regular-time employees beyond the normal workday, and subcontracting refers to the production of items by an outside supplier. Again, it is generally assumed that both of these costs are linear.

6. *Idle time costs.* The complete formulation of the aggregate planning problem also includes a cost for underutilization of the workforce, or idle time. In most contexts,

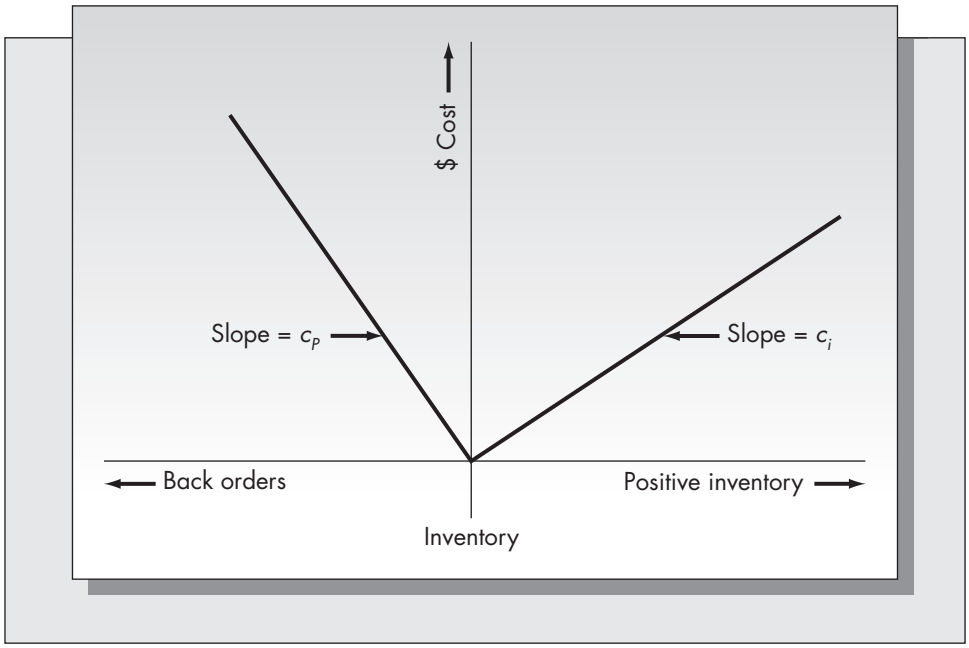


FIGURE 3–5
Holding and back-order costs

the idle time cost is zero, as the direct costs of idle time would be taken into account in labor costs and lower production levels. However, idle time could have other consequences for the firm. For example, if the aggregate units are input to another process, idle time on the line could result in higher costs to the subsequent process. In such cases, one would explicitly include a positive idle cost.

When planning is done at a relatively high level of the firm, the effects of intangible factors are more pronounced. Any solution to the aggregate planning problem obtained from a cost-based model must be considered carefully in the context of company policy. An optimal solution to a mathematical model might result in a policy that requires frequent hiring and firing of personnel. Such a policy may be infeasible because of prior contract agreements, or undesirable because of the potential negative effects on the firm's public image.

A Prototype Problem

We will illustrate the various techniques for solving aggregate planning problems with the following example.

Example *3.1*

Densepack is to plan workforce and production levels for the six-month period January to June. The firm produces a line of disk drives for mainframe computers that are plug compatible with several computers produced by major manufacturers. Forecast demands over the next six months for a particular line of drives produced in the Milpitas, California, plant are 1,280, 640, 900, 1,200, 2,000, and 1,400. There are currently (end of December) 300 workers employed in the Milpitas plant. Ending inventory in December is expected to be 500 units, and the firm would like to have 600 units on hand at the end of June.

There are several ways to incorporate the starting and the ending inventory constraints into the formulation. The most convenient is simply to modify the values of the predicted demand. Define net predicted demand in period 1 as the predicted demand minus initial inventory. If there is a minimum ending inventory constraint, then this amount should be added to the demand in period T. Minimum buffer inventories also can be handled by modifying the predicted demand. If there is a minimum buffer inventory in every period, this amount should be added to the first period's demand. If there is a minimum buffer inventory in only one period, this amount should be added to that period's demand and subtracted from the next period's demand. Actual ending inventories should be computed using the original demand pattern, however.

Returning to our example, we define the net predicted demand for January as 780 (1,280 − 500) and the net predicted demand for June as 2,000 (1,400 + 600). By considering net demand, we may make the simplifying assumption that starting and ending inventories are both zero. The net predicted demand and the net cumulative demand for the six months January to June are as follows:

Month	Net Predicted Demand	Net Cumulative Demand
January	780	780
February	640	1,420
March	900	2,320
April	1,200	3,520
May	2,000	5,520
June	2,000	7,520

The cumulative net demand is pictured in Figure 3–6. A production plan is the specification of the production levels for each month. If shortages are not permitted, then cumulative production must be at least as great as cumulative demand each period. In addition to the cumulative net demand, Figure 3–6 also shows one feasible production plan.

In order to illustrate the cost trade-offs of various production plans, we will assume in the example that there are only three costs to be considered: cost of hiring workers, cost of firing workers, and cost of holding inventory. Define

c_H = Cost of hiring one worker = \$500,

c_F = Cost of firing one worker = \$1,000,

c_I = Cost of holding one unit of inventory for one month = \$80.

We require a means of translating aggregate production in units to workforce levels. Because not all months have an equal number of working days, we will use a day as an indivisible unit of measure and define

K = Number of aggregate units produced by one worker in one day.

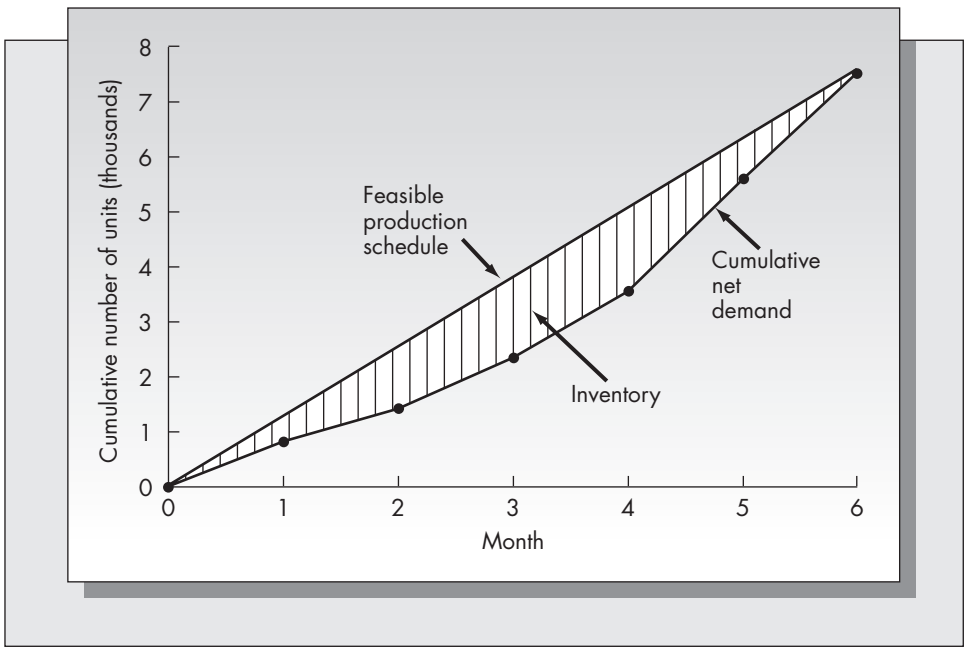


FIGURE 3–6
A feasible aggregate plan for Densepack

In the past, the plant manager observed that over 22 working days, with the workforce level constant at 76 workers, the firm produced 245 disk drives. That means that on average the production rate was $245/22 = 11.1364$ drives per day when there were 76 workers employed at the plant. It follows that one worker produced an average of $11.1364/76 = 0.14653$ drive in one day. Hence, $K = 0.14653$ for this example.

The final data needed to evaluate a plan for this example is the number of working days per month. In what follows we assume this to be 20, 24, 18, 26, 22, and 15 for January to June, respectively.

Chase, Level, and Mixed Strategies

Two extremes in capacity planning are the *zero inventory plan*, also known as a *chase strategy*, and the *constant workforce plan*, also known as a *level strategy*. Under the zero inventory plan the workforce is changed each month in order to produce enough units to most closely match the demand pattern. Capacity is adjusted up and down (i.e., workers are hired and fired) to achieve this matching. Under the constant workforce plan, capacity is kept constant during the planning period (i.e., no workers are hired or fired) and instead inventory is kept between periods; capacity is set to the minimum possible to ensure no shortages in any period.

For Example 3–1, a zero inventory plan hires at total of 755 workers, fires a total 145 workers, and achieves a total cost of $572,900 (calculations are not shown and are left as a reader's exercise). The best constant workforce plan sets capacity to 411 workers each month (hiring 111 at the beginning of January) with no further hiring or firing, has a total inventory cost of $524,960, and a total cost of $580,460 once the initial workers are considered (calculations again left to the reader). While the zero inventory plan has slightly lower costs it is unlikely to be practical because there may be constraints on the total capacity of the plant or on the maximum change that is possible from one month to the next.

The zero inventory plan and constant workforce strategies are pure strategies: they are designed to achieve one objective. They are useful in enhancing intuition and for ballpark calculations. However, with more flexibility, small modifications can result in dramatically lower costs. Such plans are described as *mixed strategies*. Optimal mixed strategies are typically found using linear programming formulations, which can incorporate a variety of additional practical constraints. Linear programming formulations are the subject of the next section.

Problems for Sections 3.4

17. A local machine shop employs 60 workers who have a variety of skills. The shop accepts one-time orders and also maintains a number of regular clients. Discuss some of the difficulties with using the aggregate planning methodology in this context.
18. A large manufacturer of household consumer goods is considering integrating an aggregate planning model into its manufacturing strategy. Two of the company vice presidents disagree strongly as to the value of the approach. What arguments might each of the vice presidents use to support his or her point of view?
19. Describe the following costs and discuss the difficulties that arise in attempting to measure them in a real operating environment.
    a. Smoothing costs
    b. Holding costs
    c. Payroll costs

Snapshot Application

HP ENTERPRISE SERVICES USES OPTIMIZATION FOR WORKFORCE PLANNING

The S&OP planning problem discussed in this chapter includes optimization of several of the firm's functions, including production planning and workforce planning. In practice, the workforce planning problem is much more difficult than the problem framed in this chapter, since it is often the case that both the supply and demand for workers is uncertain. This means that both demand and supply are likely to be random variables, thus making the problem potentially very complex. HP Enterprise Services (HPES) is a part of the HP Corporation that focuses on global business and technology. These services are provided by more than 100,000 employees located in over 60 countries. To provide an optimal solution to HPES' planning problem is probably not realistic due to the scale of the problem. Traditional operations research methods such as Markov Decision Process modeling or large scale mixed integer programming (two methods that have been proposed in the literature) quickly become computationally unwieldy.

A group of researchers (Santos, et. al. (2013)) suggested a two stage approach for solving this problem. The first stage (supply and demand consolidation) indicates those jobs for which an employee is fully qualified. For those employees that are partially qualified, they develop a transition table that provides scores which indicate the degree of qualification. Supply uncertainty in this context is primarily due to employee attrition, which can be 30% or more in some locations. Under ideal circumstances, the firm would like to be able to match job requirements with those employees that are 100% qualified. However, such a stringent rule results in poor demand fulfilment levels. Rather, the group designed a flexible matching scheme based concepts developed for the analytical hierarchy process, a tool that allows one to determine appropriate weights in a multi-criteria decision problem.

The second stage of the analysis is to build a mixed integer programming (MIP) module to allocate workers to jobs. This is done in stages: first, available employees are allocated to jobs. Second, employees that are currently committed to jobs, but will be freed up at some future time are allocated. If neither of these schemes covers the job requirements, then new hires are recommended. Finally, if positions are still unfilled, a "gap" is declared. As noted above, mixed integer optimization can be very demanding computationally, so the authors employ a heuristic for this phase. Gaps are quite common, so procedures are recommended to deal with them.

HP first implemented this approach at its facility in Bangalore, India in 2009. Resource utilization rates improved from approximately 75% to approximately 90% as a result of this planning tool. HP is now in the process of implementing this system on a worldwide scale.

Source: Santos, C., et al. "HP Enterprise Services Uses Optimization for Resource Planning". *Interfaces* 43 (2013), pp. 152–169.

20. Discuss the following statement: "Since we use a rolling production schedule, I really don't need to know the demand beyond next month."
21. St. Clair County Hospital is attempting to assess its needs for nurses over the coming four months (January to April). The need for nurses depends on both the numbers and the types of patients in the hospital. Based on a study conducted by consultants, the hospital has determined that the following ratios of nurses to patients are required:

Patient Type	Numbers of Nurses Required per Patient	Patient Forecasts			
		Jan.	Feb.	Mar.	Apr.
Major surgery	0.4	28	21	16	18
Minor surgery	0.1	12	25	45	32
Maternity	0.5	22	43	90	26
Critical care	0.6	75	45	60	30
Other	0.3	80	94	73	77

a. How many nurses should be working each month to most closely match patient forecasts?

b. Suppose the hospital does not want to change its policy of not increasing the nursing staff size by more than 10 percent in any month. Suggest a schedule of nurse staffing over the four months that meets this requirement and also meets the need for nurses each month.

22. Give an example of an environment where a chase strategy would be highly disruptive. Now give one where a level strategy would be highly disruptive.
23. Is a chase or level strategy more appropriate for aggregate planning in an air conditioning manufacturing plant where demand is highly seasonal and the workforce is relatively skilled? Explain your answer.

3.5 SOLVING AGGREGATE PLANNING PROBLEMS

Linear programming is a term used to describe a general class of optimization problems. The objective is to determine values of n nonnegative real variables in order to maximize or minimize a linear function of these variables that is subject to m linear constraints of these variables.[1] The primary advantage in formulating a problem as a linear program is that optimal solutions can be found very efficiently by the simplex method.

When all cost functions are linear, there is a linear programming formulation of the general aggregate planning problem. Because of the efficiency of commercial linear programming codes, this means that (essentially) optimal solutions can be obtained for very large problems.[2]

Cost Parameters and Given Information

The following values are assumed to be known:

c_H = Cost of hiring one worker,
c_F = Cost of firing one worker,
c_I = Cost of holding one unit of stock for one period,
c_R = Cost of producing one unit on regular time,
c_O = Incremental cost of producing one unit on overtime,
c_U = Idle cost per unit of production,
c_S = Cost to subcontract one unit of production,
n_t = Number of production days in period t,
K = Number of aggregate units produced by one worker in one day,
I_0 = Initial inventory on hand at the start of the planning horizon,
W_0 = Initial workforce at the start of the planning horizon,
D_t = Forecast of demand in period t.

The cost parameters also may be time dependent; that is, they may change with t. Time-dependent cost parameters could be useful for modeling changes in the costs of hiring or firing due, for example, to shortages in the labor pool, or changes in the costs of production and/or storage due to shortages in the supply of resources, or changes in interest rates.

[1] An overview of linear programming can be found in Supplement 1, which follows this chapter.

[2] The qualifier is included because rounding may give suboptimal solutions. There will be more about this point later.

Problem Variables

The following are the problem variables:

W_t = Workforce level in period t,
P_t = Production level in period t,
I_t = Inventory level in period t,
H_t = Number of workers hired in period t,
F_t = Number of workers fired in period t,
O_t = Overtime production in units,
U_t = Worker idle time in units ("undertime"),
S_t = Number of units subcontracted from outside.

The overtime and idle time variables are determined in the following way. The term Kn_t represents the number of units produced by one worker in period t, so that Kn_tW_t would be the number of units produced by the entire workforce in period t. However, we do not require that $Kn_tW_t = P_t$. If $P_t > Kn_tW_t$, then the number of units produced exceeds what the workforce can produce on regular time. This means that the difference is being produced on overtime, so that the number of units produced on overtime is exactly $O_t = P_t - Kn_tW_t$. If $P_t < Kn_tW_t$, then the workforce is producing less than it should be on regular time, which means that there is worker idle time. The idle time is measured in units of production rather than in time, and is given by $U_t = Kn_tW_t - P_t$.

Problem Constraints

Three sets of constraints are required for the linear programming formulation. They are included to ensure that conservation of labor and conservation of units are satisfied.

1. Conservation of workforce constraints.

$$\begin{array}{ccccccc} W_t & = & W_{t-1} & + & H_t & - & F_t \\ \text{Number of workers in } t & = & \text{Number of workers in } t-1 & + & \text{Number hired in t} & - & \text{Number fired in } t \end{array} \qquad \text{for } 1 \le t \le T.$$

2. Conservation of units constraints.

$$\begin{array}{ccccccccc} I_t & = & I_{t-1} & + & P_t & + & S_t & - & D_t \\ \text{Inventory in } t & = & \text{Inventory in } t-1 & + & \text{Number of units produced in } t & + & \text{Number of units subcontracted in } t & - & \text{Demand in } t \end{array} \qquad \text{for } 1 \le t \le T.$$

3. Constraints relating production levels to workforce levels.

$$\begin{array}{ccccccc} P_t & = & Kn_tW_t & + & O_t & - & U_t \\ \text{Number of units produced in } t & = & \text{Number of units produced by regular workforce in } t & + & \text{Number of units produced on overtime in } t & - & \text{Number of units of idle production in } t \end{array} \qquad \text{for } 1 \le t \le T.$$

In addition to these constraints, linear programming requires that all problem variables be nonnegative. These constraints and the nonnegativity constraints are the minimum that must be present in any formulation. Notice that (1), (2), and (3) constitute $3T$ constraints, rather than 3 constraints, where T is the length of the forecast horizon.

The formulation also requires specification of the initial inventory, I_0, and the initial workforce, W_0, and may include specification of the ending inventory in the final period, I_T.

The objective function includes all the costs defined earlier. The linear programming formulation is to choose values of the problem variables W_t, P_t, I_t, H_t, F_t, O_t, U_t, and S_t to

$$\text{Minimize} \sum_{t=1}^{T} (c_H H_t + c_F F_t + c_I I_t + c_R P_t + c_O O_t + c_U U_t + c_S S_t)$$

subject to

$$W_t = W_{t-1} + H_t - F_t \qquad \text{for } 1 \le t \le T \text{ (conservation of workforce),} \tag{A}$$

$$P_t = Kn_t W_t + O_t - U_t \qquad \text{for } 1 \le t \le T \text{ (production and workforce)} \tag{B}$$

$$I_t = I_{t-1} + P_t + S_t - D_t \qquad \text{for } 1 \le t \le T \text{ (inventory balance),} \tag{C}$$

$$H_t, F_t, I_t, O_t, U_t, S_t, W_t, P_t \ge 0 \text{ (nonnegativity),} \tag{D}$$

plus any additional constraints that define the values of starting inventory, starting workforce, ending inventory, or any other variables with values that are fixed in advance.

Rounding the Variables

In general, the optimal values of the problem variables will not be integers. However, fractional values for many of the variables do not make sense. These variables include the size of the workforce, the number of workers hired each period, and the number of workers fired each period, and also may include the number of units produced each period. (It is possible that fractional numbers of units could be produced in some applications.) One way to deal with this problem is to require in advance that some or all of the problem variables assume only integer values. Unfortunately, this makes the solution algorithm considerably more complex. The resulting problem, known as an integer linear programming problem, requires much more computational effort to solve than does ordinary linear programming. For a moderate-sized problem, solving the problem as an integer linear program is certainly a reasonable alternative.

If an integer programming code is unavailable or if the problem is simply too large to solve by integer programming, linear programming still provides a workable solution. However, after the linear programming solution is obtained, some of the problem variables must be rounded to integer values. Simply rounding off each variable to the closest integer may lead to an infeasible solution and/or one in which production and workforce levels are inconsistent. It is not obvious what is the best way to round the variables. We recommend the following conservative approach: round the values of the numbers of workers in each period t to W_t, the next larger integer. Once the values of W_t are determined, the values of the other variables, H_t, F_t, and P_t, can be found along with the cost of the resulting plan.

Conservative rounding will always result in a feasible solution, but will rarely give the optimal solution. The conservative solution generally can be improved by trial-and-error experimentation.

There is no guarantee that if a problem can be formulated as a linear program, the final solution makes sense in the context of the problem. In the aggregate planning problem, it does not make sense that there should be both overtime production and idle time in the same period, and it does not make sense that workers should be hired and fired in the same period. This means that either one or both of the variables O_t and U_t must be zero, and either one or both of the variables H_t and F_t must be zero for each t, $1 \leq t \leq T$. This requirement can be included explicitly in the problem formulation by adding the constraints

$$O_t U_t = 0 \qquad \text{for } 1 \leq t \leq T,$$
$$H_t F_t = 0 \qquad \text{for } 1 \leq t \leq T,$$

since if the product of two variables is zero it means that at least one must be zero. Unfortunately, these constraints are not linear, as they involve a product of problem variables. However, it turns out that it is not necessary to explicitly include these constraints, because the optimal solution to a linear programming problem always occurs at an extreme point of the feasible region. It can be shown that every extreme point solution automatically has this property. If this were not the case, the linear programming solution would be meaningless.

Extensions

Linear programming also can be used to solve somewhat more general versions of the aggregate planning problem. Uncertainty of demand can be accounted for indirectly by assuming that there is a minimum buffer inventory B_t each period. In that case we would include the constraints

$$I_t \geq B_t \qquad \text{for } 1 \leq t \leq T.$$

The constants B_t would have to be specified in advance. Upper bounds on the number of workers hired and the number of workers fired each period could be included in a similar way. Capacity constraints on the amount of production each period could easily be represented by the set of constraints:

$$P_t \leq C_t \qquad \text{for } 1 \leq t \leq T.$$

The linear programming formulation introduced in this section assumed that inventory levels would never go negative. However, in some cases it might be desirable or even necessary to allow demand to exceed supply, for example, if forecast demand exceeded production capacity over some set of planning periods. In order to treat backlogging of excess demand, the inventory level I_t must be expressed as the difference between two nonnegative variables, say I_t^+ and I_t^-, satisfying

$$I_t = I_t^+ - I_t^-,$$
$$I_t^+ \geq 0, \qquad I_t^- \geq 0.$$

The holding cost would now be charged against I_t^+ and the penalty cost for back orders (say c_P) against I_t^-. However, notice that for the solution to be sensible, it must be true that I_t^+ and I_t^- are not both positive in the same period t. As with the overtime and idle time and the hiring and firing variables, the properties of linear programming will guarantee that this holds without having to explicitly include the constraint $I_t^+ I_t^- = 0$ in the formulation.

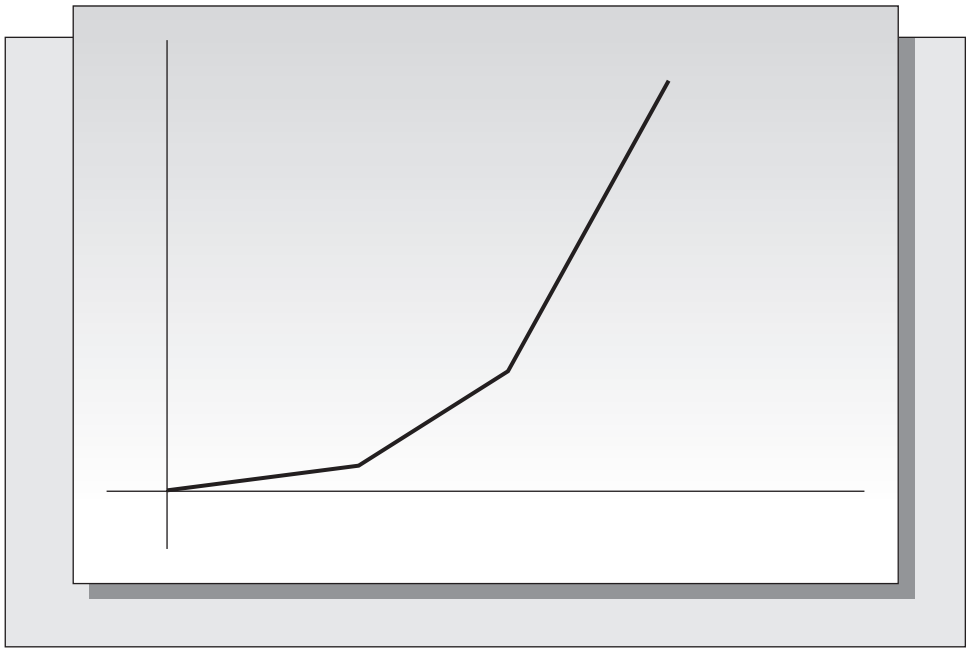


FIGURE 3–7
A convex piecewise-linear function

In the development of the linear programming model, we stated the requirement that all the cost functions must be linear. This is not strictly correct. Linear programming also can be used when the cost functions are *convex piecewise-linear functions*.

A convex function is one with an increasing slope. A piecewise-linear function is one that is composed of straight-line segments. Hence, a convex piecewise-linear function is a function composed of straight lines that have increasing slopes. A typical example is presented in Figure 3–7.

In practice, it is likely that some or all of the cost functions for aggregate planning are convex. For example, if Figure 3–7 represents the cost of hiring workers, then the marginal cost of hiring one additional worker increases with the number of workers that have already been hired. This is probably more accurate than assuming that the cost of hiring one additional worker is a constant independent of the number of workers previously hired. As more workers are hired, the available labor pool shrinks and more effort must be expended to hire the remaining available workers.

In order to see exactly how convex piecewise-linear functions would be incorporated into the linear programming formulation, we will consider a very simple case. Suppose that the cost of hiring new workers is represented by the function pictured in Figure 3–8. According to the figure, it costs c_{H1} to hire each worker until H^* workers are hired, and it costs c_{H2} for each worker hired beyond H^* workers, with $c_{H1} < c_{H2}$. The variable H_t, the number of workers hired in period t, must be expressed as the sum of two variables:

$$H_t = H_{1t} + H_{2t}.$$

Interpret H_{1t} as the number of workers hired up to H^* and H_{2t} as the number of workers hired beyond H^* in period t. The cost of hiring is now represented in the objective function as

$$\sum_{t=1}^{T} (c_{H1}H_{1t} + c_{H2}H_{2t}),$$

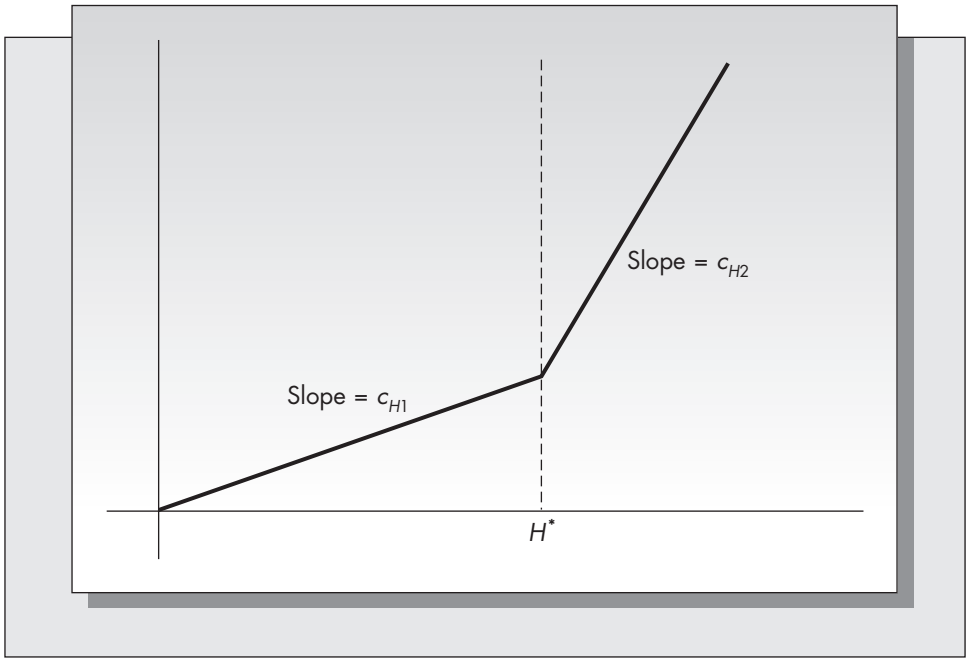


FIGURE 3–8
Convex piecewise-linear hiring cost function

and the additional constraints

$$\begin{aligned} H_t &= H_{1t} + H_{2t} \\ 0 &\leq H_{1t} \leq H^* \\ 0 &\leq H_{2t} \end{aligned}$$

must also be included.

In order for the final solution to make sense, it can never be the case that $H_{1t} < H^*$ and $H_{2t} > 0$ for some t. (Why?) However, because linear programming searches for the minimum cost solution, it will force H_{1t} to its maximum value before allowing H_{2t} to become positive, since $c_{H1} < c_{H2}$. This is the reason that the cost functions must be convex. This approach can easily be extended to more than two linear segments and to any of the other cost functions present in the objective function. The technique is known as separable convex programming and is discussed in greater detail in Hillier and Lieberman (1990).

We will demonstrate the use of linear programming by finding the optimal solution to the example presented in Section 3.4. As there is no subcontracting, overtime, or idle time allowed, and the cost coefficients are constant with respect to time, the objective function is simply

$$\text{Minimize} \left(500 \sum_{t=1}^{6} H_t + 1{,}000 \sum_{t=1}^{6} F_t + 80 \sum_{t=1}^{6} I_t \right).$$

The boundary conditions comprise the specifications of the initial inventory of 500 units, the initial workforce of 300 workers, and the ending inventory of 600 units. These are best handled by including a separate additional constraint for each boundary condition.

The constraints are obtained by substituting $t = 1, \ldots, 6$ into Equations (A), (B), and (C). The full set of constraints expressed in standard linear programming format (with all problem variables on the left-hand side and nonnegative constants on the right-hand side) is as follows:

$$\begin{aligned} W_1 - W_0 - H_1 + F_1 &= 0, \\ W_2 - W_1 - H_2 + F_2 &= 0, \\ W_3 - W_2 - H_3 + F_3 &= 0, \\ W_4 - W_3 - H_4 + F_4 &= 0, \\ W_5 - W_4 - H_5 + F_5 &= 0, \\ W_6 - W_5 - H_6 + F_6 &= 0; \end{aligned} \tag{A}$$

$$\begin{aligned} P_1 - I_1 + I_0 &= 1{,}280, \\ P_2 - I_2 + I_1 &= 640, \\ P_3 - I_3 + I_2 &= 900, \\ P_4 - I_4 + I_3 &= 1{,}200, \\ P_5 - I_5 + I_4 &= 2{,}000, \\ P_6 - I_6 + I_5 &= 1{,}400; \end{aligned} \tag{B}$$

$$\begin{aligned} P_1 - 2.931W_1 &= 0, \\ P_2 - 3.517W_2 &= 0, \\ P_3 - 2.638W_3 &= 0, \\ P_4 - 3.810W_4 &= 0, \\ P_5 - 3.224W_5 &= 0, \\ P_6 - 2.198W_6 &= 0; \end{aligned} \tag{C}$$

$$W_1, \ldots, W_6, P_1, \ldots, P_6, I_1, \ldots, I_6, F_1, \ldots, F_6, H_1, \ldots, H_6 \geq 0; \tag{D}$$

$$\begin{aligned} W_0 &= 300, \\ I_0 &= 500, \\ I_6 &= 600. \end{aligned} \tag{E}$$

The values in equations (C) come from multiplying the value of K, number of aggregate units per worker, found earlier by the number of working days in the month, which shows this formulation in Excel.

	A	B	C	D	E	F	G	H	I	J
1	**Cost of hiring**	$500								
2	**Cost of firing**	$1,000								
3	**Holding cost**	$80								
4	**K**	0.14653								
5	**Ending inv**	600								
6										
7	**Month**	**Hired**	**Fired**	**Inventory**	**Workers**	**Production**		**Worker**	**Inventory**	**Adjusted**
8		(H_t)	(F_t)	(I_t)	(W_t)	(P_t)		**balance**	**balance**	**Demand**
9	Start			500	300					
10	January	0.00	27.02	20.00	272.98	800.00		0	1280	1280
11	February	0.00	0.00	340.00	272.98	960.00		0	640	640
12	March	0.00	0.00	160.00	272.98	720.00		0	900	900
13	April	0.00	0.00	0.00	272.98	1040.00		0	1200	1200
14	May	464.80	0.00	378.38	737.78	2378.38		0	2000	2000
15	June	0.00	0.00	600.00	737.78	1621.62		0	1400	1400
16	Totals	464.80	27.02	1498.38						
17										
18	**Month**	**Days**	**Units per**	**Production**						
19		(n_t)	**worker**	**balance**						
20	January	20	2.931	0						
21	February	24	3.517	0						
22	March	18	2.638	0						
23	April	26	3.810	0						
24	May	22	3.224	0						
25	June	15	2.198	0						
26										
27	**Total cost**	$379,292.22								

Cell Formulas

Cell	Formula	Copied to
B4	=245/22/76	
H9	=E9-E8-B9+C9	H10:H14
I9	=F9-D9+D8	I10:I14
C20	=B20*$B$4	C21:C25
D20	=F10-C20*E10	D21:D25
B16	=SUM(B10:B15)	C16:D16
B27	=$B$1*$B$16+$B$2*$C$16+$B$3*$D$16	

The Solver window for this problem is as follows.

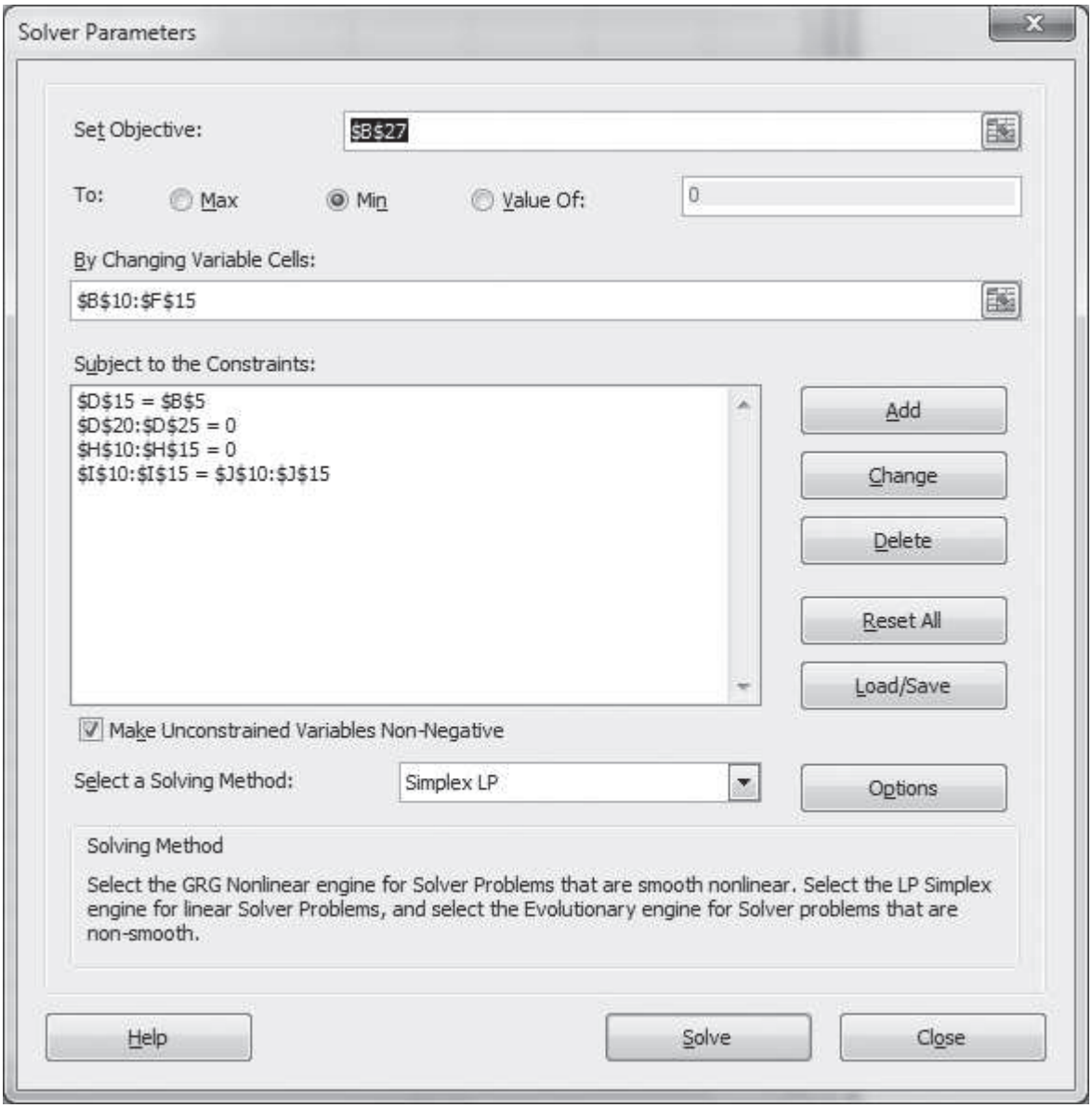


TABLE 3–1 Aggregate Plan for Densepack Obtained from Rounding the Linear Programming Solution

A	B	C	D	E	F	G	H	I
Month	Number of Workers	Number Hired	Number Fired	Number of Units per Worker	Number of Units Produced (B × E)	Cumulative Production	Cumulative Net Demand	Ending Inventory (G − H)
January	273		27	2.931	800	800	780	20
February	273			3.517	960	1,760	1,420	340
March	273			2.638	720	2,480	2,320	160
April	273			3.810	1,040	3,520	3,520	0
May	738	465		3.224	2,379	5,899	5,520	379
June	738			2.198	1,622	7,521	7,520	1
Totals		465	27					900

The value of the objective function at the optimal solution is 379,292.22,which is *considerably* less than that achieved with either the zero inventory plan or the constant workforce plan. However, this cost is based on fractional values of the variables. The actual cost will be slightly higher after rounding.

Following the rounding procedure recommended earlier, we will round all the values of W_t to the next higher integer. That gives $W_1 = \cdots = W_4 = 273$ and $W_5 = W_6 = 738$. This determines the values of the other problem variables. This means that the firm should fire 27 workers in January and hire 465 workers in May. The complete solution is given in Table 3–6.

Again, because column H in Table 3–1 corresponds to net demand, we add the 600 units of ending inventory in June, giving a total inventory of 900 + 600 = 1,500 units. Hence, the total cost of this plan is (500)(465) + (1,000)(27) + (80)(1,500) = $379,500, which represents a substantial savings over both the zero inventory plan and the constant workforce plan.

The results of the linear programming analysis suggest another plan that might be more suitable for the company. Because the optimal strategy is to decrease the workforce in January and build it back up again in May, a reasonable alternative might be to not fire the 27 workers in January and to hire fewer workers in May. In this case, the most efficient method for finding the correct number of workers to hire in May is to simply re-solve the linear program, but without the variables $F_1, \ldots, F_6$, as no firing of workers means that these variables are forced to zero. (If you wish to avoid reentering the problem into the computer, simply append the old formulation with the constraints $F_1 = 0, F_2 = 0, \ldots, F_6 = 0$.) The optimal number of workers to hire in May turns out to be 374 if no workers are fired, and the cost of the plan is approximately $386,120. This is only slightly more expensive than the optimal plan, and has the important advantage of not requiring the firing of any workers.

Problems for Section 3.5

24. Mr. Meadows Cookie Company makes a variety of chocolate chip cookies in the plant in Albion, Michigan. Based on orders received and forecasts of buying habits, it is estimated that the demand for the next four months is 850,

1,260, 510, and 980, expressed in thousands of cookies. During a 46-day period when there were 120 workers, the company produced 1.7 million cookies. Assume that the number of workdays over the four months are respectively 26, 24, 20, and 16. There are currently 100 workers employed, and there is no starting inventory of cookies.

a. What is the minimum constant workforce required to meet demand over the next four months?

b. Assume that $c_I = 10$ cents per cookie per month, $c_H = \$100$, and $c_F = \$200$. Evaluate the cost of the plan derived in part (a).

c. Formulate as a linear program. Be sure to define all variables and include the required constraints.

d. Solve for the optimal solution.

25. Harold Grey owns a small farm in the Salinas Valley that grows apricots. The apricots are dried on the premises and sold to a number of large supermarket chains. Based on past experience and committed contracts, he estimates that sales over the next five years in thousands of packages will be as follows:

Year	Forecasted Demand (thousands of packages)
1	300
2	120
3	200
4	110
5	135

Assume that each worker stays on the job for at least one year, and that Grey currently has three workers on the payroll. He estimates that he will have 20,000 packages on hand at the end of the current year. Assume that, on the average, each worker is paid $25,000 per year and is responsible for producing 30,000 packages. Inventory costs have been estimated to be 4 cents per package per year, and shortages are not allowed.

Based on the effort of interviewing and training new workers, Farmer Grey estimates that it costs $500 for each worker hired. Severance pay amounts to $1,000 per worker.

a. Assuming that shortages are not allowed, determine the minimum constant workforce that he will need over the next five years.

b. Evaluate the cost of the plan found in part (*a*).

c. Formulate this as a linear program.

d. Solve the problem and round-off the solution and determine the cost of the resulting plan.

26. A local semiconductor firm, Superchip, is planning its workforce and production levels over the next year. The firm makes a variety of microprocessors and uses sales dollars as its aggregate production measure. Based on orders received

and sales forecasts provided by the marketing department, the estimate of dollar sales for the next year by month is as follows:

Month	Production Days	Predicted Demand (in $10,000)
January	22	340
February	16	380
March	21	220
April	19	100
May	23	490
June	20	625
July	24	375
August	12	310
September	19	175
October	22	145
November	20	120
December	16	165

Inventory holding costs are based on a 25 percent annual interest charge. It is anticipated that there will be 675 workers on the payroll at the end of the current year and inventories will amount to $120,000. The firm would like to have at least $100,000 of inventory at the end of December next year. It is estimated that each worker accounts for an average of $60,000 of production per year (assume that one year consists of 250 working days). The cost of hiring a new worker is $200, and the cost of laying off a worker is $400.

a. Formulate this as a linear program.

b. Solve the problem. Round the variables in the resulting solution and determine the cost of the plan you obtain.

27. Consider Mr. Meadows Cookie Company, described in Problem 24. Suppose that the cost of hiring workers each period is $100 for each worker until 20 workers are hired, $400 for each worker when between 21 and 50 workers are hired, and $700 for each worker hired beyond 50.

    a. Write down the complete linear programming formulation of the revised problem.

    b. Solve the revised problem for the optimal solution. What difference does the new hiring cost function make in the solution?

28. Leather-All produces a line of handmade leather products. At the present time, the company is producing only belts, handbags, and attaché cases. The predicted demand for these three types of items over a six-month planning horizon is as follows:

Month	Number of Working Days	Belts	Handbags	Attaché Cases
1	22	2,500	1,250	240
2	20	2,800	680	380
3	19	2,000	1,625	110
4	24	3,400	745	75
5	21	3,000	835	126
6	17	1,600	375	45

The belts require an average of two hours to produce, the handbags three hours, and the attaché cases six hours. All the workers have the skill to work on any item. Leather-All has 46 employees who each has a share in the firm and cannot be fired. There are an additional 30 locals that are available and can be hired for short periods at higher cost. Regular employees earn $8.50 per hour on regular time and $14.00 per hour on overtime. Regular time comprises a seven-hour workday, and the regular employees will work as much overtime as is available. The additional workers are hired for $11.00 per hour and are kept on the payroll for at least one full month. Costs of hiring and firing are negligible.

Because of the competitive nature of the industry, Leather-All does not want to incur any demand back orders.

a. Using worker hours as an aggregate measure of production, convert the forecasted demands to demands in terms of aggregate units.

b. What would be the size of the workforce needed to satisfy the demand for the coming six months on regular time only? Would it be to the company's advantage to bring the permanent workforce up to this level? Why or why not?

c. Formulate the problem of optimizing Leather-All's hiring schedule as a linear program. Define all problem variables and include whatever constraints are necessary.

d. Solve the problem formulated in part (*a*) for the optimal solution. Round all the relevant variables and determine the cost of the resulting plan.

3.6 DISAGGREGATING PLANS

Aggregate

As we saw earlier in this chapter, aggregate planning may be done at several levels of the firm. Example 3.1 considered a single plant, and showed how one might define an aggregate unit to include the six different items produced at that plant. Aggregate planning might be done for a single plant, a product family, a group of families, or for the firm as a whole.

There are two views of production planning: bottom-up or top-down. The bottom-up approach means that one would start with individual item production plans. These plans could then be aggregated up the chain of products to produce aggregate plans. The top-down approach, which is the one treated in this chapter, is to start with an aggregate plan at a high level. These plans would then have to be "disaggregated" to produce detailed production plans at the plant and individual item levels.

It is not clear that disaggregation is an issue in all circumstances. If the aggregate plan is used only for macro planning purposes, and not for planning at the detail level, then one need not worry about disaggregation. However, if it is important that individual item production plans and aggregate plans be consistent, then it might be necessary to consider disaggregation schemes.

The disaggregation problem is similar to the classic problem of resource allocation. Consider how resources are allocated in a university, for example. A university receives revenues from tuition, gifts, interest on the endowment, and research grants. Costs include salaries, maintenance, and capital expenditures. Once an

annual budget is determined, each school (arts and science, engineering, business, law, etc.) and each budget center (staff, maintenance, buildings and grounds, etc.) would have to be allocated its share. Budget centers would have to allocate funds to each of their subgroups. For example, each school would allocate funds to individual departments, and departments would allocate funds to faculty and staff in that department.

In the manufacturing context, a disaggregation scheme is just a means of allocating aggregate units to individual items. Just as funds are allocated on several levels in the university, aggregate units might have to be disaggregated at several levels of the firm. This is the idea behind hierarchical production planning championed by several researchers at MIT and reported in detail in Bitran and Hax (1981) and Hax and Candea (1984).

We will discuss one possible approach to the disaggregation problem from Bitran and Hax (1981). Suppose that X^* represents the number of aggregate units of production for a particular planning period. Further, suppose that X^* represents an aggregation of n different items ($Y_1, Y_2, \ldots, Y_n$). The question is how to divide up (i.e., disaggregate) X^* among the n items. We know that holding costs are already included in the determination of X^*, so we need not include them again in the disaggregation scheme. Suppose that K_j represents the fixed cost of setting up for production of Y_j, and λ_j is the annual usage rate for item j. A reasonable optimization criterion in this context is to choose $Y_1, Y_2, \ldots, Y_n$ to minimize the average annual cost of setting up for production. As we will see in Chapter 4, the average annual setup cost for item j is $K_j \lambda_j / Y_j$. Hence, disaggregation requires solving the following mathematical programming problem:

$$\text{Minimise} \sum_{J=1}^{J} \frac{K_j \lambda_j}{Y_j}$$

subject to

$$\sum_{j=1}^{J} Y_j = X^*$$

and

$$a_j \leq Y_j \leq b_j \qquad \text{for } 1 \leq j \leq J.$$

The upper and lower bounds on Y_j account for possible side constraints on the production level for item j.

A number of feasibility issues need to be addressed before the family run sizes Y_j are further disaggregated into lots for individual items. The objective is to schedule the lots for individual items within a family so that they run out at the scheduled setup time for the family. In this way, items within the same family can be produced within the same production setup.

Snapshot Application

WELCH'S USES AGGREGATE PLANNING FOR PRODUCTION SCHEDULING

Welch's is a make-to-stock food manufacturer based in Concord, Massachusetts. They are probably best known for grape jelly and grape juices, but they produce a wide variety of processed foods. Allen and Schuster (1994) describe an aggregate planning model for Welch's primary production facility.

The characteristics of the production system for which their system was designed are

- Dynamic, uncertain demand, resulting in changing buffer stock requirements.
- Make-to-stock environment.
- Dedicated production lines.
- Production lines that each produces two or more families of products.
- Large setup times and setup costs for families, as opposed to low setup times and costs for individual items.

The two primary objectives of the production system as described by the authors are to smooth peak demands through time so as not to exceed production capacity and to allocate production requirements among the families to balance family holding and setup costs. The planning is done with a six-month time horizon for demand forecasting. The six-month period is divided into two portions: the next four weeks and the remaining five months. Detailed plans are developed for the near term, including regular and overtime production allocation.

The model has two primary components: a family planning model, which finds the optimal timing and sizing of family production runs, and a disaggregation planning model, which takes the results of family planning and determines lot sizes for individual items within families.

The authors also discuss several implementation issues specific to Welch's environment. Product run lengths must be tied to the existing eight-hour shift structure. To do so; they recommend that production run lengths be expressed as multiples of one-quarter shift (two hours).

The model was implemented on a personal computer. Computing times are very moderate. Solution techniques include a mixed integer mathematical program and a linear programming formulation with relaxation (that is, rounding of variables to integer values).

This case demonstrates that the concepts discussed in this chapter can be useful in a real production planning environment. Although the system described here is not based on any of the specific models discussed in this chapter, this application shows that aggregation and disaggregation are useful concepts. Hierarchical aggregation for production scheduling is a valuable planning tool.

The concept of disaggregating the aggregate plan along organizational lines in a fashion that is consistent with the aggregation scheme is an appealing one. Whether or not the methods discussed in this section provide a workable link between aggregate plans and detailed item schedules remains to be seen.

Another approach to the disaggregation problem has been explored by Chung and Krajewski (1984). They develop a mathematical programming formulation of the problem. Inputs to the program include aggregate plans for each product family. This includes setup time, setup status, total production level for the family, inventory level, workforce level, overtime, and regular time availability. The goal of the analysis is to specify lot sizes and timing of production runs for each individual item, consistent with the aggregate information for the product family. Although such a formulation provides a potential link between the aggregate plan and the master production schedule, the resulting mathematical program requires many inputs and can result in a very large mixed integer problem that could be very time-consuming to solve.

Problems for Section 3.6

29. What does "disaggregation of aggregate plans" mean?
30. Discuss the following quotation made by a production manager: "Aggregate planning is useless to me because the results have nothing to do with my master production schedule."

3.7 SALES AND OPERATION PLANNING ON A GLOBAL SCALE

Globalization of manufacturing operations is commonplace. Many major corporations are now classified as multinationals; manufacturing and distribution activities routinely cross international borders. With the globalization of both sources of production and markets, firms must rethink production planning strategies. One issue explored in this chapter was smoothing of production plans over time; costs of increasing or decreasing workforce levels (and, hence, production levels) play a major role in the optimization of any aggregate plan. When formulating global production strategies, other smoothing issues arise. Exchange rates, costs of direct labor, and tax structure are just some of the differences among countries that must be factored into a global strategy.

Why the increased interest in global operations? In short, cost and competitiveness. According to McGrath and Bequillard (1989):

> The benefits of a properly executed international manufacturing strategy can be very substantial. A well developed strategy can have a direct impact on the financial performance and ultimately be reflected in increased profitability. In the electronic industry, there are examples of companies attributing 5% to 15% reduction in cost of goods sold, 10% to 20% increase in sales, 50% to 150% improvement in asset utilization, and 30% to 100% increase in inventory turnover to their internationalization of manufacturing.

Cohen et al. (1989) outline some of the issues that a firm must consider when planning production levels on a worldwide basis. These include

- In order to achieve the kinds of economies of scale required to be competitive today, multinational plants and vendors must be managed as a global system.
- Duties and tariffs are based on material flows. Their impact must be factored into decisions regarding shipments of raw material, intermediate product, and finished product across national boundaries.
- Exchange rates fluctuate randomly and affect production costs and pricing decisions in countries where the product is produced and sold.
- Corporate tax rates vary widely from one country to another.
- Global sourcing must take into account longer lead times, lower unit costs, and access to new technologies.
- Strategies for market penetration, local content rules, and quotas constrain product flow across borders.
- Product designs may vary by national market.
- Centralized control of multinational enterprises creates difficulties for several reasons, and decentralized control requires coordination.
- Cultural, language, and skill differences can be significant.

Determining optimal globalized manufacturing strategies is clearly a daunting problem for any multinational firm. One can formulate and solve mathematical models similar to the linear programming formulations of the aggregate planning models presented in this chapter, but the results of these models must always be balanced against judgment and experience. Cohen et al. (1989) consider such a model. They assume multiple products, plants, markets, raw materials, vendors, vendor supply contract alternatives, time periods, and countries. Their formulation is a large-scale mixed integer, nonlinear program.

One issue not treated in their model is that of exchange rate fluctuations and their effect on both pricing and production planning. Pricing, in particular, is traditionally done by adding a markup to unit costs in the home market. This completely ignores the issue of exchange rate fluctuations and can lead to unreasonable prices in some countries. For example, this issue has arisen at Caterpillar Tractor (Caterpillar Tractor Company, 1985). In this case, dealers all over the world were billed in U.S. dollars based on U.S. production costs. When the dollar was strong relative to other currencies, retail prices charged to overseas customers were not competitive in local markets. Caterpillar found itself losing market share abroad as a result. In the early 1980s the firm switched to a locally competitive pricing strategy to counteract this problem.

The notion that manufacturing capacity can be used as a hedge against exchange rate fluctuations has been explored by several researchers. Kogut and Kulatilaka (1994), for example, develop a mathematical model for determining when it is optimal to switch production from one location to another. Since the cost of switching is assumed to be positive, there must be a sufficiently large difference in exchange rates before switching is recommended. As an example, they consider a situation where a firm can produce its product in either the United States or Germany. If production is currently being done in one location, the model provides a means of determining if it is economical to switch locations based on the relative strengths of the euro and dollar. While such models are in the early stages of development, they provide a means of rationalizing international production planning strategies. Similar issues have been explored by Huchzermeier and Cohen (1996) as well.

3.8 HISTORICAL NOTES

The aggregate planning problem was conceived in an important series of papers that appeared in the mid-1950s. The first, Holt, Modigliani, and Simon (1955), discussed the structure of the problem and introduced the quadratic cost approach, and the later study of Holt, Modigliani, and Muth (1956) concentrated on the computational aspects of the model. A complete description of the method and its application to production planning for a paint company is presented in Holt, Modigliani, Muth, and Simon (1960).

It should be recognized that the text by Holt et al. (1960) represents a landmark work in the application of quantitative methods to production planning problems. The authors developed a solution method that results in a set of formulas that are easy to implement and they actually undertook the implementation of the method. The work details the application of the approach to a large manufacturer of household paints in the Pittsburgh area. The analysis was implemented in the company but a subsequent visit to the firm indicated that serious problems arose when the linear decision rule was followed, primarily because of the firm's policy of not firing workers when the model indicated that they should be fired.

That production planning problems could be formulated as linear programs appears to have been known in the early 1950s. Bowman (1956) discussed the use of a transportation model for production planning. The particular linear programming formulation of the aggregate planning problem discussed in Section 3.5 is essentially the same as the one developed by Hansmann and Hess (1960). Other linear programming formulations of the production planning problem generally involve multiple products or more complex cost structures (see, for example, Newson, 1975a and 1975b).

More recent work on the aggregate planning problem has focused on aggregation and disaggregation issues (Axsater, 1981; Bitran and Hax, 1981; and Zoller, 1971), the incorporation of learning curves into linear decision rules (Ebert, 1976), extensions to allow for multiple products (Bergstrom and Smith, 1970), and inclusion of marketing and/or financial variables (Damon and Schramm, 1972, and Leitch, 1974).

Taubert (1968) considers a technique he refers to as the search decision rule. The method requires developing a computer simulation model of the system and searching the response surface using standard search techniques to obtain a (not necessarily optimal) solution. Taubert's approach, which was described in detail in Buffa and Taubert (1972), gives results that are comparable to those of Holt, Modigliani, Muth, and Simon (1960) for the case of the paint company.

Kamien and Li (1990) developed a mathematical model to examine the effects of subcontracting on aggregate planning decisions. The authors show that under certain circumstances it is preferred to producing in-house, and provides an additional means of smoothing production and workforce levels.

3.9 Summary

Modern firms take a much less siloed approach to planning than firms of the past. They have come to realize that large benefits can accrue from a collaborative approach among the different divisions including operations, sales and marketing, and finance. A well designed S&OP process can manage the inherent tensions between these divisions and make tradeoffs from a strategic perspective. The key output from an operational perspective of this process is a fixed sales plan that the operations division can plan to.

A common component of the S&OP process will be reviewing divisional KPIs. Because KPIs have such a large impact on employee incentives they need to be carefully chosen. The primary challenge in KPI selection is to ensure that the KPI is *aligned* with the strategic imperatives of the firm while still remaining *actionable* as far as the person being measured by it is concerned. It is also important to try to mitigate opportunities for gaming of KPIs by employees.

The key output from the S&OP process is a fixed forecast and therefore planning for routine uncertainty in these numbers is typically left to the operations division. While risk pooling (see Chapter 6) can be used to mitigate some of the uncertainty, there will always be natural variation that must be buffered. Such buffering can take the form of inventory, spare capacity, or time in the form of customer lead times. This sort of buffering typically takes place outside the S&OP process and will be discussed in later chapters in this text.

At a higher level than routine variation are more major risks that are discussed within the S&OP process. Such *known unknowns* should have contingency plans and/or mitigation strategies associated with them. They may be demand-side risks, such as new product introductions by competitors or demand shocks caused by extreme weather, or they may be supply-side risks such as supplier failure or major quality issues. An effective S&OP can anticipate such risks and set strategies for dealing with them that are in line with the company strategy. This may include the explicit relaxing of certain KPIs that become less relevant when working under exceptional circumstances.

Determining optimal production levels for all products produced by a large firm can be an enormous undertaking. *Aggregate planning* addresses this problem by assuming that individual items can be grouped together. However, finding an effective aggregating scheme can be difficult and often revenue dollars are used for simplicity. One particular aggregating scheme that has been suggested is *items*, *families*, and *types*. Items (or stock keeping units, SKUs), represent the finest level of detail, are identified by separate part numbers, bar codes, and/or radio frequency ID (RFID) tags when appropriate. Families are groups of items that share a common manufacturing setup, and types are natural groups of families. This particular aggregation scheme is fairly general but there is no guarantee that it will work in every application.

As mentioned above, once the strategies for dealing with uncertainty have been determined, most firms assume *determinist demand* in coming up with an aggregate production plan. Indeed, fixed demand forecasts over a specified planning horizon are required input. This assumption is not made for the sake of realism, but to allow the analysis to focus on the changes in the demand that are systematic rather than random. The goal of the analysis is to determine for each period the number of workers that should be employed, the production that should occur, and the inventory that should be carried over for each period.

The objective of an aggregate production plan is to minimize costs of production, payroll, holding, and changing the size of the workforce or capacity. The costs of making changes are generally referred to as *smoothing costs*. The aggregate planning models discussed in this chapter assume that all the costs are *linear functions*. This assumption is probably a reasonable approximation for most real systems within a given reasonable range of values. It is unlikely that the primary problem with applying a linear programming formulation to a real situation will be that the shape of the cost function is incorrect; it is more likely that the primary difficulty will be in correctly estimating the costs and other required input information.

Aggregate production plans will be of little use to the firm if they cannot be coordinated with detailed item schedules (i.e., the master production schedule). The problem of *disaggregating aggregate plans* is a difficult one, but one that must be addressed if the aggregate plans are to have value to the firm. There have been some mathematical programming formulations of the disaggregating problem suggested in the literature but these disaggregation schemes have yet to be proven in practice.

Additional Problems on Aggregate Planning

31. An aggregate planning model is being considered for the following applications. Suggest an aggregate measure of production and discuss the difficulties of applying aggregate planning in each case.
    a. Planning for the size of the faculty in a university.
    b. Determining workforce requirements for a travel agency.
    c. Planning workforce and production levels in a fish-processing plant that produces canned sardines, anchovies, kippers, and smoked oysters.

32. A local firm manufactures children's toys. The projected demand over the next four months for one particular model of toy robot is

Month	Workdays	Forecasted Demand (in aggregate units)
July	23	3,825
August	16	7,245
September	20	2,770
October	22	4,440

Assume that a normal workday is eight hours. Hiring costs are $350 per worker and firing costs (including severance pay) are $850 per worker. Holding costs are $4.00 per aggregate unit held per month. Assume that it requires an average of 1 hour and 40 minutes for one worker to assemble one toy. Shortages are not permitted. Assume that the ending inventory for June was 600 of these toys and the manager wishes to have at least 800 units on hand at the end of October. Assume that the current workforce level is 35 workers. Find the optimal plan by formulating as a linear program.

33. The Paris Paint Company is in the process of planning labor force requirements and production levels for the next four quarters. The marketing department has provided production with the following forecasts of demand for Paris Paint over the next year:

Quarter	Demand Forecast (in thousands of gallons)
1	380
2	630
3	220
4	160

Assume that there are currently 280 employees with the company. Employees are hired for at least one full quarter. Hiring costs amount to $1,200 per employee and firing costs are $2,500 per employee. Inventory costs are $1 per gallon per quarter. It is estimated that one worker produces 1,000 gallons of paint each quarter.

Assume that Paris currently has 80,000 gallons of paint in inventory and would like to end the year with an inventory of at least 20,000 gallons.

a. Determine the minimum constant workforce plan for Paris Paint and the cost of the plan. Assume that stock-outs are not allowed.

b. If Paris were able to back-order excess demand at a cost of $2 per gallon per quarter, determine a minimum constant workforce plan that holds less inventory than the plan you found in part (*a*), but incurs stock-outs in quarter 2. Determine the cost of the new plan.

c. Formulate this as a linear program. Assume that stock-outs are not allowed.

d. Solve the linear program. Round the variables and determine the cost of the resulting plan.

34. Consider the problem of Paris Paint presented in Problem 33. Suppose that the plant has the capacity to employ a maximum of 370 workers. Suppose that regular-time employee costs are $12.50 per hour. Assume seven-hour days, five-day weeks, and four-week months. Overtime is paid on a time-and-a-half basis. Subcontracting is available at a cost of $7 per gallon of paint produced. Overtime is limited to three hours per employee per day, and no more than 100,000 gallons can be subcontracted in any quarter.

a. Formulate as a linear program.

b. Solve the linear program. Round the variables and determine the cost of the resulting plan.

35. The Mr. Meadows Cookie Company can obtain accurate forecasts for 12 months based on firm orders. These forecasts and the number of workdays per month are as follows:

Month	Demand Forecast (in thousands of cookies)	Workdays
1	850	26
2	1,260	24
3	510	20
4	980	18
5	770	22
6	850	23
7	1,050	14
8	1,550	21
9	1,350	23
10	1,000	24
11	970	21
12	680	13

During a 46-day period when there were 120 workers, the firm produced 1,700,000 cookies. Assume that there are 100 workers employed at the beginning of month 1 and zero starting inventory.

a. Find the minimum constant workforce needed to meet monthly demand.

b. Assume c_I = \$0.10 per cookie per month, c_H = \$100, and c_F = \$200. Add columns that give the cumulative on-hand inventory and inventory cost. What is the total cost of the constant workforce plan?

c. Solve for the optimal plan using linear programming. Compare your solution to *b*.

36. The Yeasty Brewing Company produces a popular local beer known as Iron Stomach. Beer sales are somewhat seasonal, and Yeasty is planning its production and workforce levels on March 31 for the next six months. The demand forecasts are as follows:

Month	Production Days	Forecasted Demand (in hundreds of cases)
April	11	85
May	22	93
June	20	122
July	23	176
August	16	140
September	20	63

As of March 31, Yeasty had 86 workers on the payroll. Over a period of 26 working days when there were 100 workers on the payroll, Yeasty produced 12,000 cases of beer. The cost to hire each worker is \$125 and the cost of laying off each worker is \$300. Holding costs amount to 75 cents per case per month.

As of March 31, Yeasty expects to have 4,500 cases of beer in stock, and it wants to maintain a minimum buffer inventory of 1,000 cases each month. It plans to start October with 3,000 cases on hand.

a. Based on this information, find the minimum constant workforce plan for Yeasty over the six months, and determine hiring, firing, and holding costs associated with that plan.

b. Suppose that it takes one month to train a new worker. How will that affect your solution?

c. Suppose that the maximum number of workers that the company can expect to be able to hire in one month is 10. How will that affect your solution to part (*a*)?

d. Formulate the problem levels as a linear program. [You may ignore the conditions in parts (*b*) and (*c*).]

e. Solve the resulting linear program. Round the appropriate variables and determine the cost of your solution.

f. Suppose Yeasty does not wish to fire any workers. What is the optimal plan subject to this constraint?

37. A local canning company sells canned vegetables to a supermarket chain in the Minneapolis area. A typical case of canned vegetables requires an average of 0.2 day of labor to produce. The aggregate inventory on hand at the end of June is 800 cases. The demand for the vegetables can be accurately predicted for about 18 months based on orders received by the firm. The predicted demands for the next 18 months are as follows:

Month	Forecasted Demand (hundreds of cases)	Workdays	Month	Forecasted Demand (hundreds of cases)	Workdays
July	23	21	April	29	20
August	28	14	May	33	22
September	42	20	June	31	21
October	26	23	July	20	18
November	29	18	August	16	14
December	58	10	September	33	20
January	19	20	October	35	23
February	17	14	November	28	18
March	25	20	December	50	10

The firm currently has 25 workers. The cost of hiring and training a new worker is $1,000, and the cost to lay off one worker is $1,500. The firm estimates a cost of $2.80 to store a case of vegetables for a month. They would like to have 1,500 cases in inventory at the end of the 18-month planning horizon.

a. Develop a spreadsheet to find the minimum constant workforce aggregate plan and determine the total cost of that plan.

b. Develop a spreadsheet to find a plan that hires and fires workers monthly in order to minimize inventory costs. Determine the total cost of that plan as well.

Appendix **3–A**

Glossary of Notation for Chapter 3

α = Smoothing constant for production and demand used in Bowman's model.

β = Smoothing constant for inventory used in Bowman's model.

c_F = Cost of firing one worker.

c_H = Cost of hiring one worker.

c_I = Cost of holding one unit of stock for one period.

c_O = Cost of producing one unit on overtime.

c_P = Penalty cost for back orders.

c_R = Cost of producing one unit on regular time.

c_S = Cost to subcontract one unit of production.

c_U = Idle cost per unit of production.

D_t = Forecast of demand in period t.

F_t = Number of workers fired in period t.

H_t = Number of workers hired in period t.

I_t = Inventory level in period t.

K = Number of aggregate units produced by one worker in one day.

λ_j = Annual demand for family j (refer to Section 3.9).

n_t = Number of production days in period t.

O_t = Overtime production in units.

P_t = Production level in period t.

S_t = Number of units subcontracted from outside.

T = Number of periods in the planning horizon.

U_t = Worker idle time in units ("undertime").

W_t = Workforce level in period t.

Bibliography

ABB Group. "Key Performance Indicators: Identifying and using key metrics for performance." PowerPoint accessed from http://www.abb.com/Search.aspx?q=pr&abbcontext=products&num=10&filetype=mspowerpoint&filter=0&start=10

Allen, S. J., and E. W. Schuster. "Practical Production Scheduling with Capacity Constraints and Dynamic Demand: Family Planning and Disaggregation." *Production and Inventory Management Journal* 35 (1994), pp. 15–20.

Axsater, S. "Aggregation of Product Data for Hierarchical Production Planning." *Operations Research* 29 (1981), pp. 744–56.

Bergstron, G. L., and B. E. Smith. "Multi-Item Production Planning—An Extension of the HMMS Rules." *Management Science* 16 (1970), pp. 100–103.

Bitran, G. R., and A. Hax. "Disaggregation and Resource Allocation Using Convex Knapsack Problems with Bounded Variables." *Management Science* 27 (1981), pp. 431–41.

Buffa, E. S., and W. H. Taubert. *Production-Inventory Systems: Planning and Control.* Rev. ed. New York: McGraw-Hill/Irwin, 1972.

Chopra, S. and P. Meindl. *Supply Chain Management: Strategy, Planning & Operation*. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2006.

Chung, C., and L. J. Krajewski. "Planning Horizons for Master Production Scheduling." *Journal of Operations Management* (1984), pp. 389–406.

De Meyer, A., C. H. Loch, and M. T. Pich. "Managing Project Uncertainty: From Variation to Chaos." *MIT Sloan Management Review* (2002).

Damon, W. W., and R. Schramm. "A Simultaneous Decision Model for Production, Marketing, and Finance." *Management Science* 19 (1972), pp. 16 –72.

Ebert, R. J. "Aggregate Planning with Learning Curve Productivity." *Management Science* 23 (1976), pp. 171–82.

Erenguc, S., and S. Tufekci. "A Transportation Type Aggregate Production Model with Bounds on Inventory and Backordering." *European Journal of Operations Research* 35 (1988), pp. 414–25.

EveryAngle Software.Accessed from http://www.everyangle.com/downloads/customer-cases/en/customercase_heineken.pdf

Goldratt, E. M. and J. Cox. *The Goal: A Process of Ongoing Improvement.* 2nd Revised Edition. Great Barrington, MA: North River Press, 1992.

Grimson, J. A., and D. F. Pyke. "Sales and operations planning: an exploratory study and framework." *International Journal of Logistics Management* 18 (2007), pp.322–346.

Hansmann, F., and S. W. Hess. "A Linear Programming Approach to Production and Employment Scheduling." *Management Technology* 1 (1960), pp. 46–51.

Hax, A. C., and D. Candea. *Production and Inventory Management.* Englewood Cliffs, NJ: Prentice Hall, 1984.

Hax, A. C., and H. C. Meal., "Hierarchical Integration of Production Planning and Scheduling." In *TIMS Studies in Management Science.* Volume 1, *Logistics*, ed. M. Geisler. New York: Elsevier, 1975.

Hiller, F. S., and G. J. Lieberman. *Introduction to Operations Research.* 5th ed. San Francisco: Holden Day, 1990.

Holt, C. C., F. Modigliani, and J. F. Muth. "Derivation of a Linear Decision Rule for Production and Employment." *Management Science* 2 (1956), pp. 159–77.

Holt, C. C., F. Modigliani, J. F. Muth, and H. A. Simon. *Planning Production, Inventories, and Workforce*. Englewood Cliffs, NJ: Prentice Hall, 1960.

Holt, C. C., F. Modigliani; and H. A. Simon. "A Linear Decision Rule for Employment and Production Scheduling." *Management Science* 2 (1955), pp. 1–30.

Hopp, W. J., and M. L. Spearman. *Factory Physics.* (1996). Boston, MA: McGraw Hill.

IBM. "IBM Global Chief Supply Chain Officer Study." Accessed from http://www-935.ibm.com/services/us/gbs/bus/html/gbs-csco-study.html

Jordan, W. C., and S. C. Graves. "Principles on the benefits of manufacturing process flexibility." *Management Science* 41 (1995), pp. 577–594.

Kaplan, R. S., and D. P. Norton. "The Balanced Scorecard: Measures that Drive Performance." *Harvard Business Review* (1992), pp. 71–80.

Kamien, M. I., and L. Li. "Subcontracting, Coordination, and Flexibility, and Production Smoothing Aggregate Planning." *Management Science* 36 (1990), pp. 1352–63.

Leitch, R. A. "Marketing Strategy and Optimal Production Schedule." *Management Science* 20 (1974), pp. 903–11.

Newson, E. F. P. "Multi-Item Lost Size Scheduling by Heuristic, Part 1: With Fixed Resources." *Management Science* 21 (1975a), pp. 1186–93.

Newson, E. F. P. "Multi-Item Lost Size Scheduling by Heuristic, Part 2: With Fixed Resources." *Management Science* 21 (1975b), pp. 1194–1205.

Rooijen, H. "Connecting our supply chain to our customers. Accessed from http://www.heinekeninternational.com/content/live/files%202011/investors/6.%20Henk%20van%20Rooijen.pdf

Santos, C., et al. "HP Enterprise Services Uses Optimization for Resource Planning." *Interfaces* 43 (2013), pp. 152–169.

Sheldon, D. H. *World Class Sales & Operations Planning.* Ft. Lauderdale, FL: J Ross Publishing and APICS, 2006.

Smits, J. and M. English. "The Journey to Worldclass S&OP at Heineken." Accessed from http://supply-chain.org/node/17283

Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable.* New York: Random House, 2007.

Taubert, W. H. "A Search Decision Rule for the Aggregate Scheduling Problem." *Management Science* 14 (1968), pp. B343–53.

Vollmann, T. E., W. L. Berry, and D. C. Whybark. *Manufacturing, Planning, and Control Systems.* 3rd ed. New York: McGraw-Hill/Irwin, 1992.

Zoller, K. "Optimal Disaggregation of Aggregate Production Plans." *Management Science* 17 (1971), pp. B53–49.

Supplement **One**

Linear Programming

S1.1 INTRODUCTION

Linear programming is a mathematical technique for solving a broad class of optimization problems. These problems require maximizing or minimizing a linear function of n real variables subject to m constraints. One can formulate and solve a large number of real problems with linear programming. A partial list includes

1. Scheduling of personnel.
2. Several varieties of blending problems including cattle feed, sausages, ice cream, and steel making.
3. Inventory control and production planning.
4. Distribution and logistics problems.
5. Assignment problems.

Problems with thousands of variables and thousands of constraints are easily solvable on computers today. Linear programming was developed to solve logistics problems during World War II. George Dantzig, a mathematician employed by the RAND Corporation at the time, developed a solution procedure he labeled the Simplex Method. That the method turned out to be so efficient for solving large problems quickly was a surprise even to its developer. That fact, coupled with the simultaneous development of the electronic computer, established linear programming as an important tool in logistics management. The success of linear programming in industry spawned the development of the disciplines of operations research and management science. The Simplex Method has withstood the test of time. Only in recent years has another method been developed that potentially could be more efficient than the Simplex Method for solving very large, specially structured, linear programs. This method, known as Karmarkar's Algorithm, is named for the Bell Labs mathematician who conceived it.

In Section S1.2 we consider a typical manufacturing problem that we formulate and solve using linear programming. Later we explore how one solves small problems (that is, having exactly two decision variables) graphically, and how one solves large problems using a computer.

S1.2 A PROTOTYPE LINEAR PROGRAMMING PROBLEM

Example *S1.1* Sidneyville manufactures household and commercial furnishings. The Office Division produces two desks, rolltop and regular. Sidneyville constructs the desks in its plant outside Medford, Oregon, from a selection of woods. The woods are cut to a uniform thickness of 1 inch. For this reason, one measures the wood in units of square feet. One rolltop desk requires 10 square feet of pine, 4 square feet of cedar, and 15 square feet of maple. One regular desk requires

20 square feet of pine, 16 square feet of cedar, and 10 square feet of maple. The desks yield respectively \$115 and \$90 profit per sale. At the current time the firm has available 200 square feet of pine, 128 square feet of cedar, and 220 square feet of maple. The firm has backlogged orders for both desks and would like to produce the number of rolltop and regular desks that would maximize profit. How many of each should it produce?

Solution

The first step in formulating a problem as a linear program is to identify the decision variables. In this case there are two decisions required: the number of rolltop desks to produce and the number of regular desks to produce. We must assign symbol names to each of these decision variables.

Let

$$x_1 = \text{Number of rolltop desks to be produced,}$$

$$x_2 = \text{Number of regular desks to be produced.}$$

Now that we have identified the decision variables, the next step is to identify the objective function and the constraints. The objective function is the quantity we wish to maximize or minimize. The objective is to maximize the profits, so the objective function equals the total profit when producing x_1 rolltop desks and x_2 regular desks. Each rolltop desk contributes \$115 to profit, so the total contribution to profit from all rolltop desks is $115x_1$. Similarly, the contribution to profit from all regular desks is $90x_2$. Hence, the total profit is $115x_1 + 90x_2$. This is known as the objective function.

The next step is to identify the constraints. The number of desks Sidneyville can produce is limited by the amount of wood available. The three types of wood constitute the critical resources. To obtain the constraints, we need to find expressions for the amount of each type of wood consumed by construction of x_1 rolltop desks and x_2 regular desks. Those expressions are then bounded by the amount of each type of wood available.

The number of square feet of pine used to make x_1 rolltop desks is $10x_1$. The number of square feet of pine used to make x_2 regular desks is $20x_2$. It follows that the total amount of pine consumed in square feet is $10x_1 + 20x_2$. This quantity cannot exceed the number of square feet of pine available, which is 200. Hence we obtain the first constraint:

$$10x_1 + 20x_2 \leq 200.$$

The other two constraints are similar. The second constraint is to ensure that the firm does not exceed the available supply of cedar. Each rolltop desk requires 4 square feet of cedar, so x_1 rolltop desks require $4x_1$ square feet of cedar. Each regular desk requires 16 square feet of cedar, so x_2 regular desks require $16x_2$ square feet of cedar. It follows that the constraint on the supply of cedar is

$$4x_1 + 16x_2 \leq 128.$$

In the same way, the final constraint ensuring that we do not exceed the supply of maple is

$$15x_1 + 10x_2 \leq 220.$$

Because we cannot produce a negative number of desks, we also include nonnegativity constraints:

$$x_1 \geq 0,$$

$$x_2 \geq 0.$$

We have now constructed the complete linear programming formulation of the Sidneyville problem. The goal is to find values of x_1 and x_2 to maximize $115x_1 + 90x_2$, subject to the constraints.

$$10x_1 + 20x_2 \leq 200,$$

$$4x_1 + 16x_2 \leq 128,$$

$$15x_1 + 10x_2 \leq 220,$$

$$x_1, x_2 \geq 0.$$

TABLE S1–1 Partial Computer Output for Sidneyville Problem

```
LP OPTIMUM FOUND AT STEP 2

OBJECTIVE FUNCTION VALUE

1) 1740.000000

VARIABLE        VALUE          REDUCED COST
   X1         12.000000          .000000
   X2          4.000000          .000000
```

We next consider how such a problem is solved. We will briefly outline the theory behind the solution technique known as the Simplex Method. However, as linear programming problems are almost never solved by hand any longer, you will not need to understand the mechanics of the Simplex Method in order to use linear programming. You will need to know how to formulate problems as linear programs, enter the formulations into the computer, recognize special problems, and analyze the computer's output.

A typical computer output is given in Table S1–1.

This output tells us that at the optimal solution, the value of x_1 is 12 and the value of x_2 is 4. That is, Sidneyville should produce exactly 12 rolltop desks and 4 regular desks. The value of the objective function at the optimal solution is \$1,740.00. That this is the (unique) optimal solution means the following: every other pair of values of x_1 and x_2 will result in either a lower profit, not meeting the constraints, or both.

Sidneyville's manager of production planning is very skeptical about mathematics. When presented with this solution, his response was, "There's only one problem with this production plan. We have a specialist make the rolltop portion of the rolltop desk. She can only do four desks a day and we want to be ready to ship out in two days. There is no way we can produce twelve of those desks in two days. I knew this math stuff was a lot of hooey!"

The manager was wrong. The trouble is not that the formulation is incorrect, but that it does not include all relevant constraints, as labor hours turned out to be a critical resource. The lesson here is that for the final solution to be meaningful, the model must include *all* relevant constraints.

S1.3 STATEMENT OF THE GENERAL PROBLEM

The Sidneyville problem is an example of a linear program in which there are two decision variables and three constraints. Linear programming problems may have any number of decision variables and any number of constraints. Suppose that there are n decision variables, labeled $x_1, x_2, \ldots, x_n$, subject to m resource constraints. Then we may write the problem of maximizing the objective subject to the constraints as

$$\begin{aligned}
&\text{Maximize } c_1x_1 + c_2x_2 + \cdots + c_nx_n,\\
&\text{subject to } a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \le b_1,\\
&\qquad\qquad a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \le b_2,\\
&\qquad\qquad\qquad\qquad \vdots\\
&\qquad\qquad a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \le b_m,\\
&\qquad\qquad\qquad x_1, x_2, \ldots, x_n \ge 0.
\end{aligned}$$

Interpret $c_1, c_2, \ldots, c_n$ as the profit coefficients per unit of output of the activities $x_1, x_2, \ldots, x_n$; a_{ij} as the amount of resource i consumed by one unit of activity j; and

b_1 as the amount of resource i available, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. We require that the constants $b_1, \ldots, b_m$ be nonnegative. This particular formulation includes problems in which we want to maximize profit subject to constraints on the available resources. However, linear programming can be used to solve a much larger variety of problems. Other possible formulations will be discussed later.

Definitions of Commonly Used Terms

1. *Objective function.* This is the quantity we wish to maximize or minimize. In the given formulation, the objective function is the term $c_1x_1 + c_2x_2 + \cdots + c_nx_n$. In business applications, one typically minimizes cost or maximizes profit. We use the abbreviations "min" for a minimization problem and "max" for a maximization problem.

2. *Constraints.* Each constraint is a linear inequality or equation, that is, a linear combination of the problem variables followed by a relational operator ($\leq$ or $=$ or $\geq$) followed by a nonnegative constant. Although the given formulation shows all $\leq$ constraints, $\geq$ and $=$ type constraints are also common. For example, suppose there is a contractual agreement that requires a minimum number of labor hours daily. This would result in a $\geq$ constraint.

3. *Right-hand side.* The right-hand side is the constant following the relational operator in a constraint. In the given constraint, the constants $b_1, b_2, \ldots, b_m$ are the right-hand sides. These constants are required to be nonnegative numbers. However, we do *not* require that the constants a_{ij} be nonnegative. This means that any constraint can be written with a nonnegative right-hand side by multiplying through by the constant -1 whenever the right-hand side is negative. Consider the following simple example. Suppose that when formulating a problem as a linear program we obtain the constraint

$$4x_1 - 2x_2 \leq -5.$$

Because the right-hand side is negative, this is not a legal constraint. However, if we multiply through by -1, this constraint becomes

$$-4x_1 + 2x_2 \geq 5,$$

which is acceptable.

4. *Feasible region.* The feasible region is the set of values of the decision variables, $x_1, x_2, \ldots, x_n$, that satisfy the constraints. Because each of the constraints is generated by a linear equation or linear inequality, the feasible region has a particular structure. The technical term for this structure is *convex polytope*. In two dimensions, a convex polytope is a convex set with boundaries that are straight lines. In three dimensions, the boundaries are formed by planes. A convex set is characterized as follows: pick any two points in the set and connect them with a straight line; the line lies entirely within the set.

5. *Extreme points.* Because of the structure of the feasible region, there will be a finite number of feasible points, with the property that they cannot be expressed as a linear combination of any other set of feasible points. These points are known as extreme points or corner points, and they play an important role in linear programming. The concept of extreme points will become clearer when we discuss graphical solutions.

6. *Feasible solution.* A feasible solution is one particular set of values of the decision variables that satisfies the constraints. A feasible solution is also one point in the feasible region. It may be an extreme point or an interior point.

7. *Optimal solution.* The optimal solution is the feasible solution that maximizes or minimizes the objective function. In some cases the optimal solution may not be unique. When this is the case, there will be an infinite number of optimal solutions.

Features of Linear Programs

Linear programming is a very powerful tool. Many real problems have been successfully formulated and solved using this technique. However, to use the method correctly, one must be aware of its limitations. Two important features of linear programs are linearity and continuity. Many problems that may appear to be solvable by linear programming fail one or both of these two crucial tests.

Linearity

Optimization problems can be formulated as linear programs only when (*a*) the objective can be expressed as a linear function of the decision variables and (*b*) all constraints can be expressed as linear functions of the decision variables.

Linearity implies that quantities change in fixed proportions. For example, if it costs \$10 to produce one unit, then it costs \$20 to produce two units, and \$100 to produce ten units. If one ounce of orange juice supplies 30 mg of vitamin C, then three ounces must supply 90 mg. Linearity must hold in the objective function and the constraints. In the objective function, this means that the profit or cost per unit must be the same independent of the number of units. In the constraints, linearity means that the amount of each resource consumed is the same per unit whether one produces a single unit or many units.

However, one often observes nonlinear relationships in the real world. Economies of scale in production mean that the marginal cost of producing a unit decreases as the number of units produced increases. When this occurs, the cost of production is a nonlinear function of the number of units produced. An example of scale economies is a fixed setup cost for production. The formula for the EOQ discussed in Chapter 4 says that the lot size increases as the square root of the demand rate. Hence, EOQ is a nonlinear function of the demand. When either the objective function or a constraint is a nonlinear function of the decision variables, the problem is a nonlinear programming problem and cannot be solved by linear programming.[1]

Continuity

This means that the decision variables should be continuous (that is, able to assume any nonnegative value) as opposed to discrete or integer valued. This can be a serious restriction. The solution to many problems makes sense only if the decision variables are integer valued. In particular, Example S1.1 is, strictly speaking, not a linear programming problem because the number of desks produced must be integer valued. (We were lucky that the optimal solution turned out to be integer valued in this case.) One might think that the optimal integer solution is equal to the continuous solution rounded off to the nearest integer. Unfortunately, this is not always the case. First, rounding may lead to infeasibility; that is, the rounded-off solution may lie outside the feasible region. Second, even if the rounded solution is feasible, it may not be optimal. It can happen that the optimal integer solution is in an entirely different portion of the feasible region than the rounded-off linear programming solution!

To give the reader some idea of the difficulties that can arise when the solution must be integer valued, consider Example S1.1. Suppose that the profit from selling rolltop desks was \$150 rather than \$115. Then the objective would be to maximize $150x_1 + 90x_2$ subject to the same set of constraints. The optimal linear programming solution is

$$x_1 = 14.666666\ldots.$$
$$x_2 = 0.$$

[1] In some circumstances, convex programming problems can be solved by linear programming by approximating the objective function with a piecewise-linear function. An example of this approach appears at the end of Section 3.5.

Rounding the solution to the nearest integer gives $x_1 = 15$ and $x_2 = 0$, which is infeasible. Substituting these values into the final constraint results in a requirement of 225 square feet of maple. However, only 220 square feet are available. Rounding x_1 down to 14 results in a feasible but suboptimal solution. At $x_1 = 14$ and $x_2 = 0$, there are 10 feet of maple still available, which is enough to make one regular desk. The optimal integer solution in this case is $x_1 = 14$ and $x_2 = 1$.

When the decision variables must be integer valued, we say that the problem is an integer linear programming problem. Finding optimal integer solutions to linear programs can be very time-consuming, even for modest-sized problems. However, Excel does offer an option for defining some or all of the problem variables as integer valued. Excel does a fine job solving small integer linear programming problems. For larger problems, one would use a computer program designed to solve integer programming problems. In some cases, and especially when the values of the variables are relatively large, careful rounding of the linear programming solution should give acceptable results.

S1.4 SOLVING LINEAR PROGRAMMING PROBLEMS GRAPHICALLY

Graphing Linear Inequalities

In this section we will show how to solve two-variable linear programming problems graphically. Although most real problems have more than two variables, understanding the procedure for solving two-variable linear programs will improve your grasp of the concepts underlying the Simplex Method.

The first step is to graph the linear inequalities represented by the constraints. A linear inequality corresponds to all points in the plane on one side of a straight line. There are thus two steps to graphing linear inequalities:

1. Draw the straight line representing the boundary of the region corresponding to the constraint expressed as an equation.
2. Determine which side of the line corresponds to the inequality.

To illustrate the method, consider the first constraint in Example S1.1.

$$10x_1 + 20x_2 \leq 200.$$

The boundary of the region represented by this inequality is the straight line

$$10x_1 + 20x_2 = 200.$$

The easiest way to graph a straight line is to determine the two intercepts. These are found by setting x_2 to zero and solving for x_1, and then setting x_1 to zero and solving for x_2. First setting x_2 to zero gives

$$10x_1 = 200 \qquad \text{or} \qquad x_1 = 20.$$

Similarly, setting x_1 to zero and solving for x_2 gives

$$20x_2 = 200 \qquad \text{or} \qquad x_2 = 10.$$

Hence, the line $10x_1 + 20x_2 = 200$ must pass through the points (20, 0) (the x_1 intercept) and (0, 10) (the x_2 intercept). A graph of this line is shown in Figure S1–1. Now that we have graphed the line defining the boundary of the half space, we must determine which side of the line corresponds to the inequality. To do so, we pick any point *not* on the line, substitute the values of x_1 and x_2, and see if the inequality is satisfied or

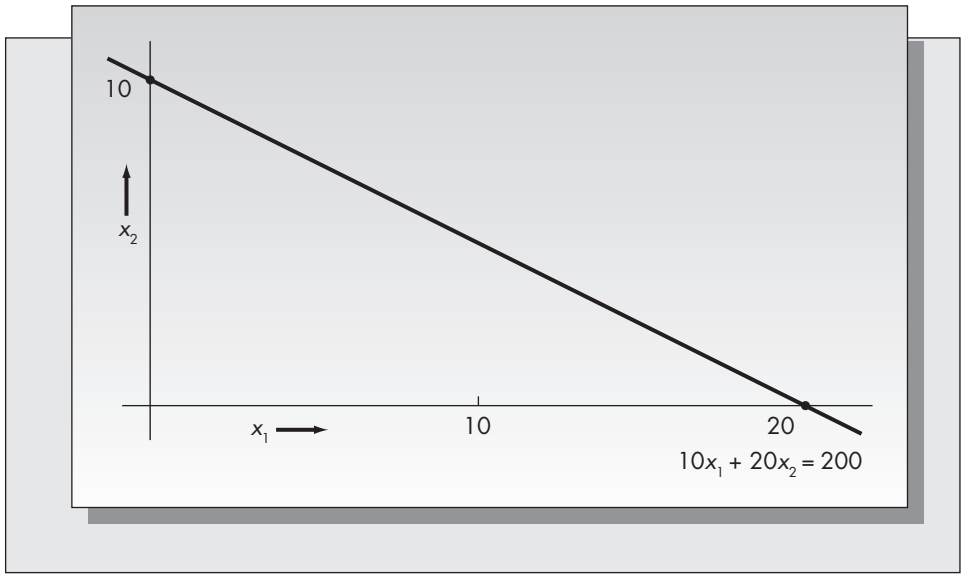


FIGURE S1–1
Graphing a constraint boundary

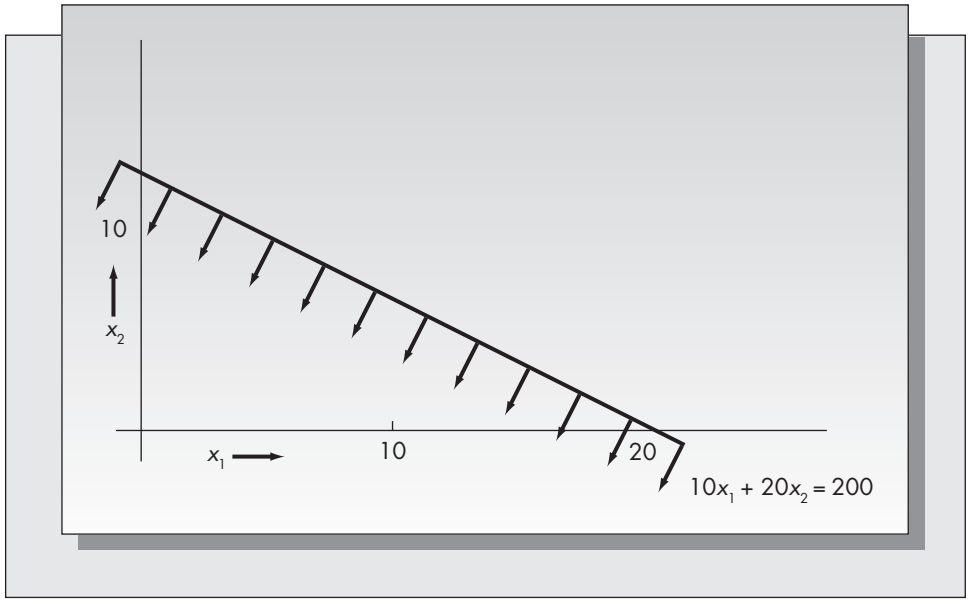


FIGURE S1–2
Half space representing the inequality $10x_1 + 20x_2 \leq 200$

not. If the inequality is satisfied, then that point belongs in the half space; if it is not, then that point does not belong in the half space. If the boundary line does not go through the origin [that is, the point (0, 0)], then the most straightforward approach is to use the origin as the point to be tested.

Substituting (0, 0) into the inequality, we obtain

$$(10)(0) + (20)(0) = 0 \leq 200.$$

Because 0 is less than 200, the test is satisfied and the origin lies within the region represented by the inequality. This means that the graph of the inequality includes all the points below the boundary line in Figure S1–1, as shown in Figure S1–2.

The origin test to determine which side of the line is appropriate only works when the boundary line itself does not go through the origin. If it does, then some other point

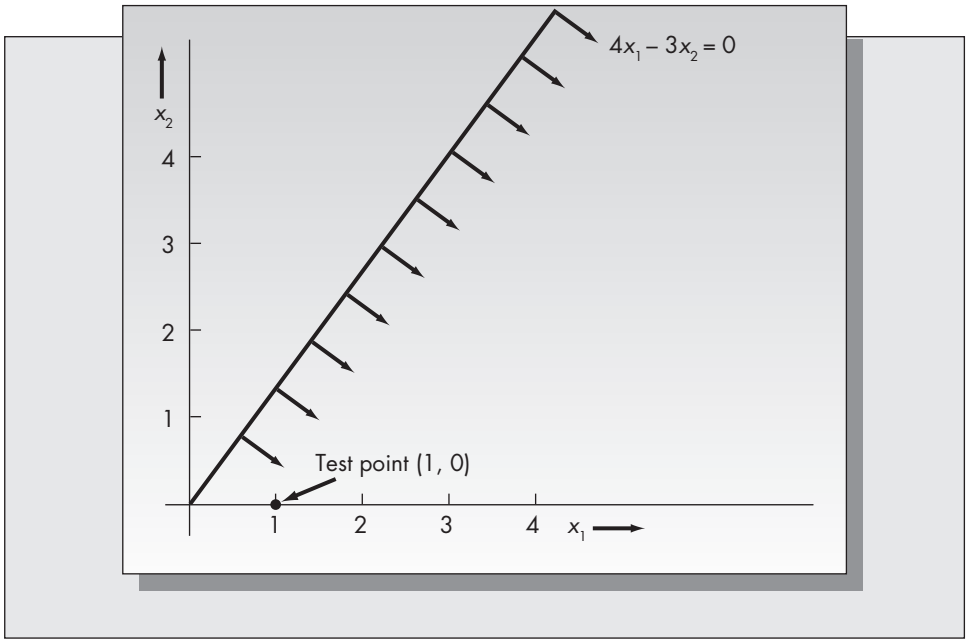


FIGURE S1–3
Graphing a constraint boundary that passes through the origin

not on the line must be used for the test. For example, consider the constraint

$$4x_1 - 3x_2 \geq 0.$$

When we try to graph the line $4x_1 - 3x_2 = 0$, we see that substituting $x_1 = 0$ gives $x_2 = 0$ and substituting $x_2 = 0$ gives $x_1 = 0$. This means that the line passes through the origin. In order to graph the line, we must determine another point that lies on it. Just pick any value of x_1 and solve for the corresponding value of x_2. For example, substituting $x_1 = 3$ gives $x_2 = 4$, meaning that the point (3, 4) lies on the line as well as the point (0, 0). The boundary line is graphed by connecting these points, as shown in Figure S1–3.

Next, we determine which side of the inequality corresponds to the region of interest. As noted above, the origin test does not work when the line passes through the origin. We pick any point *not* on the line to do the test. In this case, one point that does not lie on the line is $x_1 = 1$ and $x_2 = 0$. Substituting these values into the inequality gives

$$(4)(1) - 0 = 4 > 0.$$

The inequality is satisfied, so the point (1, 0) lies in the region. The inequality corresponds to the points below the line as pictured in Figure S1–3.

Graphing the Feasible Region

The graph of the feasible region is found by graphing the linear inequalities represented by the constraints and determining the region of intersection of the corresponding half spaces. We will determine a graph of the feasible region for the Sidneyville problem in this way.

We have graphed the feasible region corresponding to the first constraint in Figure S1–2. Consider the other two constraints:

$$4x_1 + 16x_2 \leq 128,$$
$$15x_1 + 10x_2 \leq 220.$$

The half spaces corresponding to these constraints are found in the same way. First we graph the straight lines corresponding to the region boundaries. In the first case

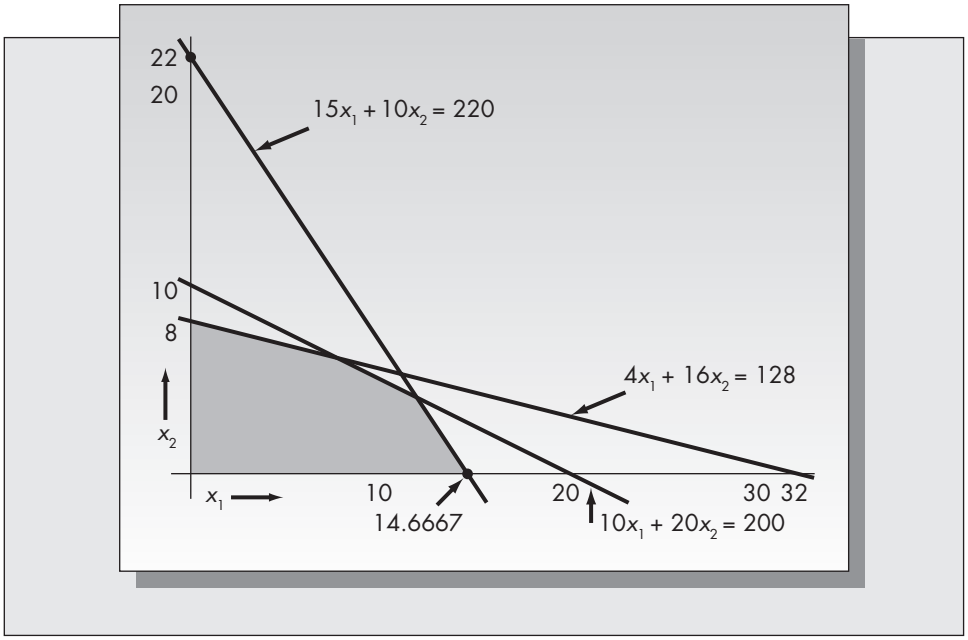


FIGURE S1–4
Feasible region for Sidneyville problem (Example S1.1)

the intercepts are $x_1 = 32$ and $x_2 = 8$, and in the second case they are $x_1 = 14.6667$ and $x_2 = 22$. Using the origin test, we see that the appropriate half spaces are the points lying below both lines. In addition, we must also include the nonnegativity constraints, $x_1 \geq 0$ and $x_2 \geq 0$. The resulting feasible region is pictured in Figure S1–4.

Finding the Optimal Solution

The feasible region pictured in Figure S1–4 has several interesting properties that are common to all linear programming problems. Pick any two feasible solutions (that is, points in the region) and connect these points with a straight line. The resulting line lies completely in the region, meaning that the feasible region is a convex set. The region boundaries are straight lines. These lines intersect at points known as the extreme points. There are a total of five extreme points in the feasible region pictured in Figure S1–4.

An important property of linear programs is that the optimal solution always occurs at an extreme point of the feasible region. For the Sidneyville problem, this means that among the infinite number of solutions (i.e., points) in the feasible region, the optimal solution will be one of only five points![2] This is true no matter what objective function we assume. This means that we can find the optimal solution to the Sidneyville problem by identifying the five extreme points, substituting the (x_1, x_2) coordinates of these points into the objective function, and determining the point that results in the maximum profit. We will consider this method first, even though there is a more efficient graphical solution procedure.

From Figure S1–4 we see that one of the extreme points is the origin (0, 0). Another is the x_2 intercept corresponding to the constraint $4x_1 + 16x_2 = 128$, which is (0, 8). A third is the x_1 intercept corresponding to the constraint $15x_1 + 10x_2 = 220$, which is (14.6667, 0). The other two extreme points correspond to the intersections of pairs of boundary lines. They are found by simultaneously solving the equations corresponding to the boundaries.

[2] It can happen that two extreme points are optimal, in which case all the points on the line connecting them are optimal as well.

First, we simultaneously solve the equations

$$4x_1 + 16x_2 = 128,$$
$$10x_1 + 20x_2 = 200.$$

These equations can be solved in several ways. Multiplying the first equation by 10 and the second by 4 gives

$$40x_1 + 160x_2 = 1{,}280,$$
$$40x_1 + 80x_2 = 800.$$

Subtracting the second equation from the first yields

$$80x_2 = 480$$
$$x_2 = 6.$$

The value of x_1 is found by substituting $x_2 = 6$ into either equation (both will yield the same value for x_1). Substituting into the first equation gives

$$4x_1 + (16)(6) = 128$$
$$4x_1 = 128 - 96 = 32$$
$$x_1 = 8.$$

Check that substituting $x_2 = 6$ into the equation $10x_1 + 20x_2 = 200$ gives the same result.

The last extreme point is found by solving

$$15x_1 + 10x_2 = 220,$$
$$10x_1 + 20x_2 = 200$$

simultaneously. We will not present the details of this calculation. The reader should be able to show by the same methods just used that the simultaneous solution in this case is

$$x_1 = 12,$$
$$x_2 = 4.$$

We have now identified all five extreme points. The next step is to substitute the corresponding values of x_1 and x_2 into the objective function and see which gives the largest profit. The objective function is $115x_1 + 90x_2$.

Extreme Point	Value of Objective Function
(0, 0)	(115)(0) + (90)(0) = 0
(0, 8)	(115)(0) + (90)(8) = 720
(14.666 . . . , 0)	(115)(14.666 . . .) + (90)(0) = 1,686.67
(8, 6)	(115)(8) + (90)(6) = 1,460
(12, 4)	(115)(12) + (90)(4) = 1,740

The maximum value of the objective function is 1,740 and is achieved at the point (12, 4). This agrees with the computer output in Table S1–1.

Hence, we have shown that one method of finding the optimal solution to a linear programming problem is to find all the extreme points, substitute their values into the objective function, and pick the one that gives the largest objective function value

for maximization problems or the smallest objective function value for minimization problems. Next we show how one can quickly identify the optimal extreme point graphically.

Identifying the Optimal Solution Directly by Graphical Means

One identifies the optimal solution directly in the following way. The objective function is a linear combination of the decision variables. In our example the objective function is $115x_1 + 90x_2$. Consider the family of straight lines defined by the equation

$$Z = 115x_1 + 90x_2.$$

As Z is varied, one generates a family of parallel lines. The variable Z is the profit obtained when producing x_1 rolltop desks and x_2 regular desks such that (x_1, x_2) lies on the line $Z = 115x_1 + 90x_2$. As an example, consider $Z = 4{,}140$. Figure S1–5 shows the line $4{,}140 = 115x_1 + 90x_2$. Notice that it lies completely outside the feasible region, meaning that no feasible combination of x_1 and x_2 results in a profit of \$4,140. Reducing Z to 2,070 drops the Z line closer to the feasible region, as also pictured in Figure S1–5.

The graphical method of identifying the optimal solution is to pick one value of Z, such as $Z = 3{,}000$, that takes us beyond the feasible region, place a ruler on the Z line, and move the ruler parallel to the Z line toward the feasible region. The extreme point that is hit first is the optimal solution. Once one graphically determines which extreme point is optimal, the coordinates of that point are found by solving the appropriate equations as shown above. This approach avoids having to identify all the extreme points of the feasible region.

One problem can arise. If we pick a small starting value of Z and move toward the feasible region from below, this method will identify a *different* extreme point. In Figure S1–5, suppose that we chose $Z = -3{,}000$. Then the Z line would be located under the feasible region. As we moved the Z line toward the feasible region, the first extreme point encountered is the origin. This means that the origin solves Example S1.1 with a

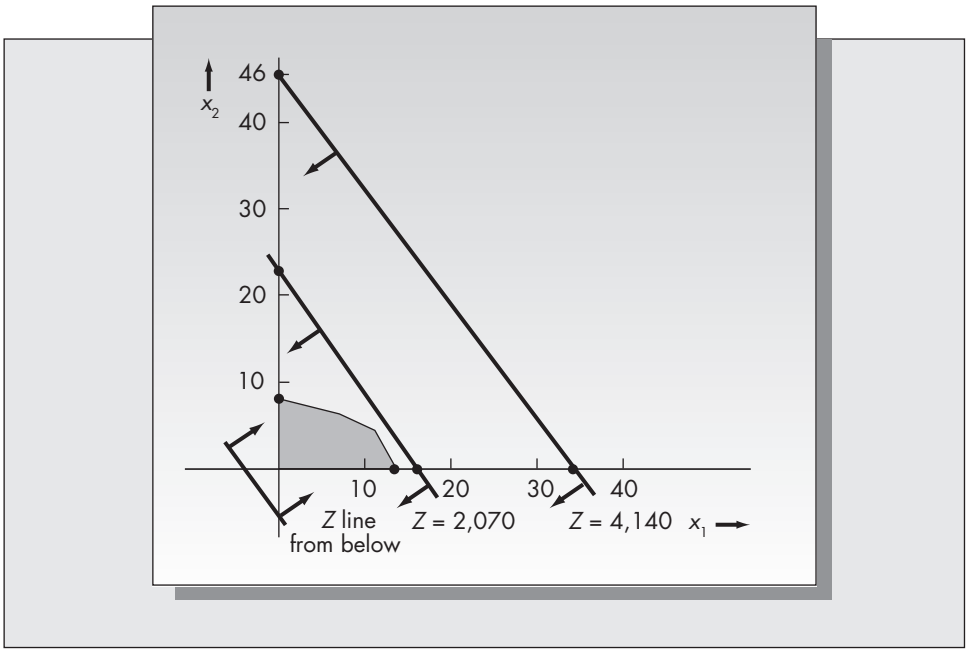


FIGURE S1–5
Approaching the feasible region with the Z line

minimization objective rather than a maximization objective (i.e., if your goal is to *minimize* profit, the best strategy is not to produce any desks).

Hence, this approach identifies two extreme points. One corresponds to the maximum solution and one to the minimum solution. In this case it is obvious which is which. When it is not obvious, the (x_1, x_2) values for each extreme point should be found and substituted into the objective function to be certain which is the minimum solution and which is the maximum solution.

S1.5 THE SIMPLEX METHOD: AN OVERVIEW

The Simplex Method is an algorithm that moves sequentially from one extreme point to another until it reaches the optimal solution. If the origin (that is, all problem variables set to zero) is a feasible solution, it will serve as the initial extreme point.[3] At each iteration, the method considers all adjacent extreme points (those that can be reached by moving along an edge), and moves to the one that gives the best improvement in the objective function. The algorithm continues to move from one adjacent extreme point to another, finally terminating when the optimal solution is reached.

For the problem of Example S1.1 the origin is feasible, so that is the initial feasible solution. The two extreme points adjacent to the origin are $(x_1, x_2) = (14.6667, 0)$ and $(x_1, x_2) = (0, 8)$. The largest improvement in the objective function is obtained by moving to the point (14.6667, 0). There are two extreme points adjacent to (14.6667, 0). They are (0, 0) and (12, 4). Clearly, a greater improvement is obtained by moving to (12, 4). At this point the method recognizes that the current solution is optimal, as a movement to another adjacent extreme point lowers the profit.

In the worst case, the Simplex Method could conceivably need to search all the extreme points of the feasible region before identifying the optimal solution. If this were common, the Simplex Method would not be a practical solution method for solving linear programming problems. Let us consider why.

Suppose that a linear program had 25 variables and 25 less than or equal to constraints. For each constraint, one adds a slack variable converting the problem to one with only equality constraints. This is known as standard form. Hence, in standard form we have a problem of 50 variables and 25 constraints. Each basic solution (extreme point) corresponds to setting 25 variables to zero and solving the resulting system of 25 linear equations in 25 unknowns. It follows that the number of such solutions (that is, extreme points) equals the number of combinations of 50 things taken 25 at a time. This turns out to be about 1.264×10^{14} (about 126 trillion). To give the reader some idea of the magnitude of this number, suppose that we had a computer program that could identify 100 extreme points every second. At that rate, it would take about 40,000 years to find all the extreme points for a problem of this size!

It is indeed fortunate that the Simplex Method rarely needs to search all the extreme points of the feasible region to discover the optimal solution. In fact, for a problem of this size, on average it would need to evaluate only about 25 extreme points.[4] Hence, the Simplex Method turned out to be a very efficient solution procedure.

[3] When the origin is not feasible, there are techniques available for determining an initial feasible solution to get the method started.

[4] The reason for this was understood only very recently. The proof requires very sophisticated mathematics.

We will not explore the mechanics of the method or additional theory underpinning its concepts. With today's easy access to computing and the wide availability of excellent software, it is unlikely that anyone with a real problem would solve it manually. There are many excellent texts that delve more deeply into the theoretical and computational aspects of linear programming. A good starting point for the interested reader is Hillier and Lieberman (1990). A more detailed treatment of the theory can be found in Hadley (1962).

S1.6 SOLVING LINEAR PROGRAMMING PROBLEMS WITH EXCEL

Excel has become the standard for spreadsheet programs. One of the useful features of Excel is that it comes bundled with an add-in called Solver that solves both linear and nonlinear programming problems. While the reliability of the nonlinear and the integer portions of the program are suspect, the linear programming part of Solver is excellent.

Because Solver is part of a spreadsheet program, problems are not entered algebraically (as they are with a system known as LINDO discussed in previous editions of this book). Consider Example S1.1. The algebraic representation is

$$\text{Maximize } 115x_1 + 90x_2,$$

subject to

$$\begin{aligned} 10x_1 + 20x_2 &\leq 200, \\ 4x_1 + 16x_2 &\leq 128, \\ 15x_1 + 10x_2 &\leq 220, \\ x_1, x_2 &\geq 0. \end{aligned}$$

It is convenient to write the problem in a matrix format before entering the information into the Excel spreadsheet. The matrix representation for this problem is

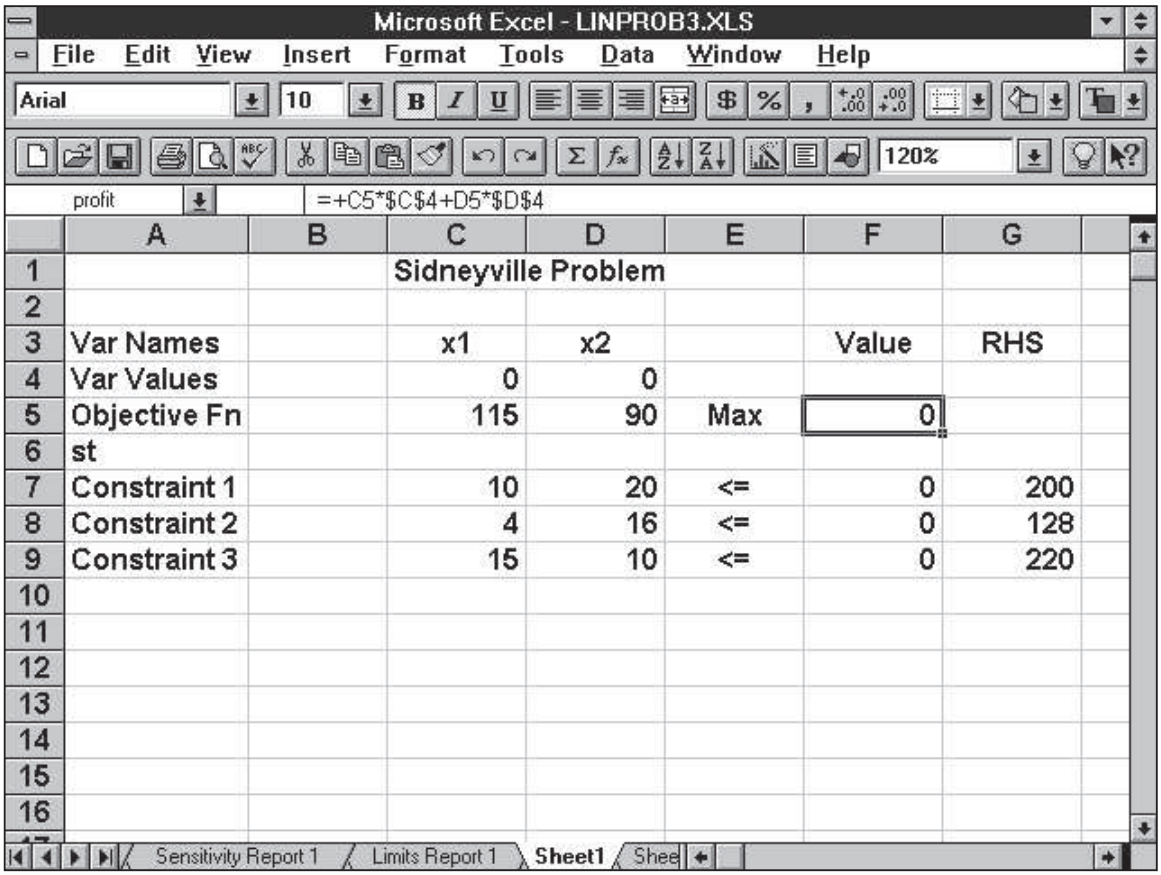
Variable names:	x_1	x_2	Operator	RHS
Objective function:	115	90	max	
subject to				
Constraint 1	10	20	$\leq$	200
Constraint 2	4	16	$\leq$	128
Constraint 3	15	10	$\leq$	220

The spreadsheet will look very much like this. The only differences are that one must specify the locations of the variable values (which I recommend be a row located directly under the variable names) and the algebraic formulas for the objective function and the constraints. These will be located in a column between "Operator" and "RHS." We will label this column "Value" in the spreadsheet.

Note that the column labeled "Operator" is not required. It is a convenience for the user to help keep track of the direction of the objective function and constraints. Excel requires this information, and it is entered manually when Solver is invoked.

At this point the spreadsheet should look similar to Figure S1–6. Notice the additional row labeled "Var Values" and the additional column labeled "Value." It is in the column labeled "Value" that one enters the algebraic form of the linear programming

FIGURE S1–6
Excel spreadsheet for Sidneyville problem



problem. The locations of the variable values are cells C4 and D4. The algebraic form of the objective function will be entered in cell F5 and the constraints in cells F7, F8, and F9. The formula for cell F5 is =C5*$C$4+D5*$D$4. The formula can be typed in or entered using the mouse to point and click cell locations.

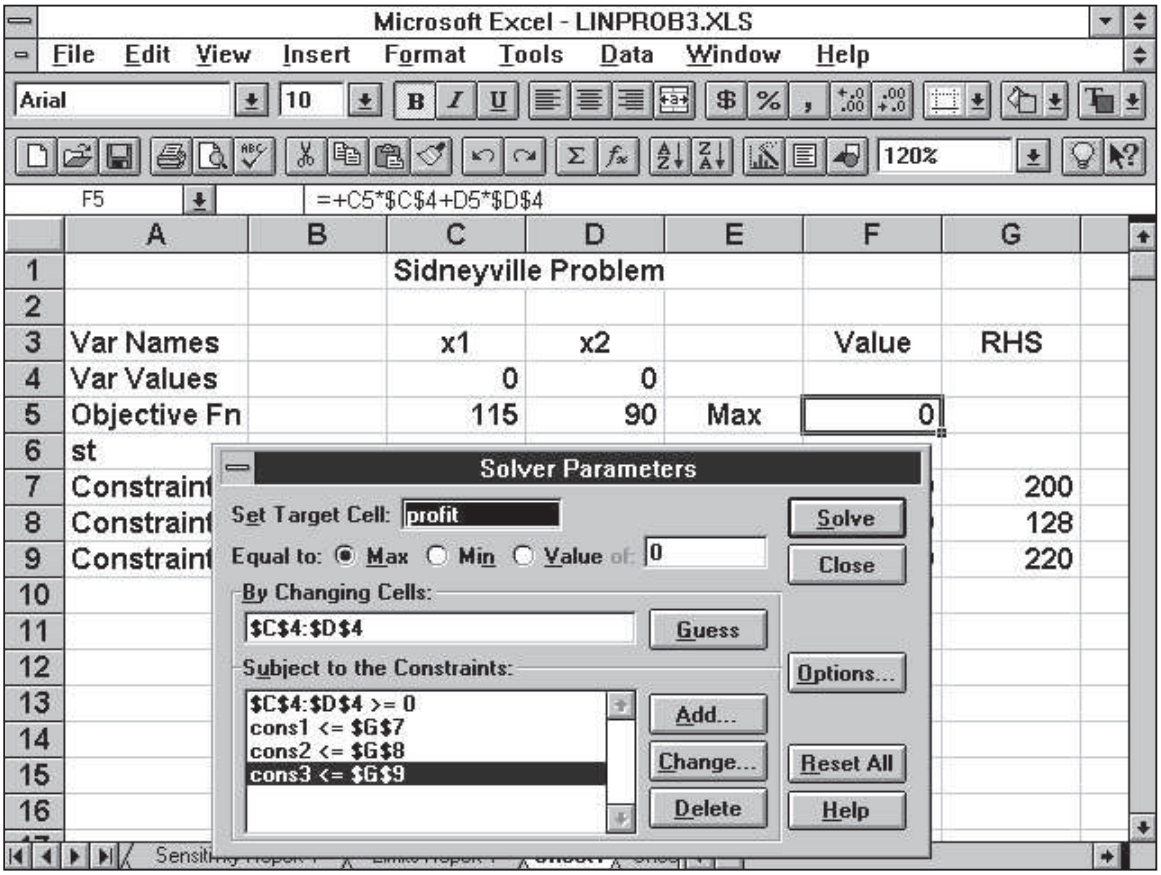
Notice that we have used absolute addressing for the variable values (C4 and D4). This allows us to copy the formula from cell F5 to cells F7, F8, and F9 without having to retype the algebraic form for the constraints. You may wish to assign name labels to these cells so that they can later be referred to by name rather than by cell location. (This is done most conveniently by invoking the formula bar, accessing the label area, and typing in a label of your choice. Note that the label "profit" appears just below the name of the current font. This was the name assigned to cell F5.)

Check that the algebraic form of the constraints is correct after copying cell F5 to cells F7, F8, and F9. For example, the formula corresponding to cell F7 should be: =C7*$C$4 + D7*$D$4.

Place 0s in the cells corresponding to the variable values (C4 and C5). After Solver has completed its search, the optimal values of these variables will appear in these cell locations.

The problem is now completely defined. Invoke Solver to obtain the solution. This is done by accessing Tools from the Menu Bar and choosing Solver. The Solver dialog

FIGURE S1–7
Spreadsheet with Solver dialog box



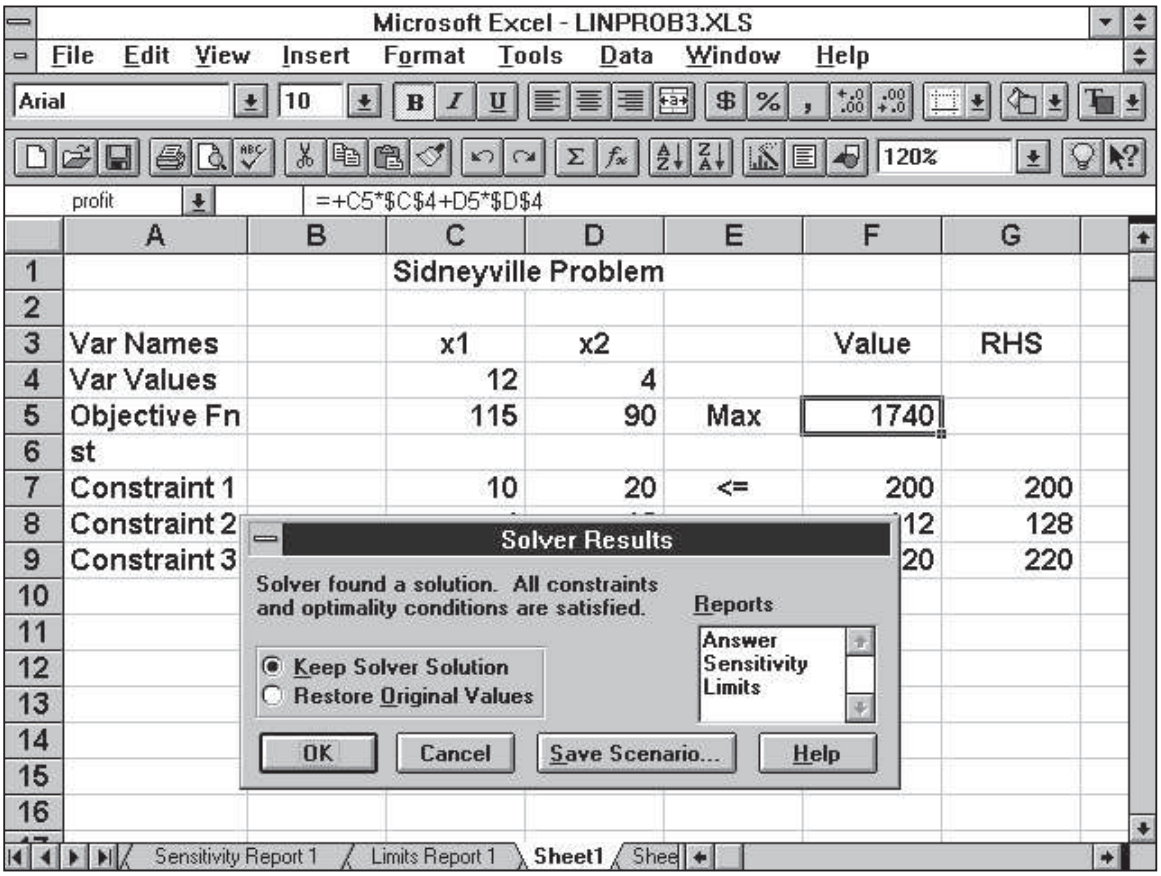
box will appear as in Figure S1–7. The first requirement is setting the location of the target cell. This corresponds to the cell location of the algebraic formula for the objective function—cell F5 in our spreadsheet. This can be accomplished by typing in F5 or clicking on the cell with the mouse. Notice that in this case I assigned the name "profit" to cell F5, so I simply type in "profit" for the target cell.

Next, specify if the problem is a min (minimum) or a max (maximum). Excel refers to the variables in the problem as "changing cells." You must tell Solver where to find these cells. In this spreadsheet they are C4 and D4. (These can be indicated by pointing and clicking with the mouse, manually typing in the cell locations, or entering preassigned cell names as we did for the objective function.)

Next, tell Excel where to find the algebraic definitions for the constraints. The constraints are entered into the system one at a time by clicking on the Add button. For each constraint you first tell the system where to find the algebraic form for the left-hand side of the constraint (F7, F8, and F9 in our case), the logical operator for the constraint ($\leq$, $=$, or $\geq$), and the location of the RHS value (G7, G8, and G9 in our case).

Because Solver is a general-purpose mathematical programming tool, two additional pieces of information must be included. Both of these can be done at the same time by clicking the Options key and choosing the two options "Assume linear model" and "Assume non-negative." In this way, the nonnegativity constraints do not need

FIGURE S1–8
The Excel spreadsheet displaying the optimal solution

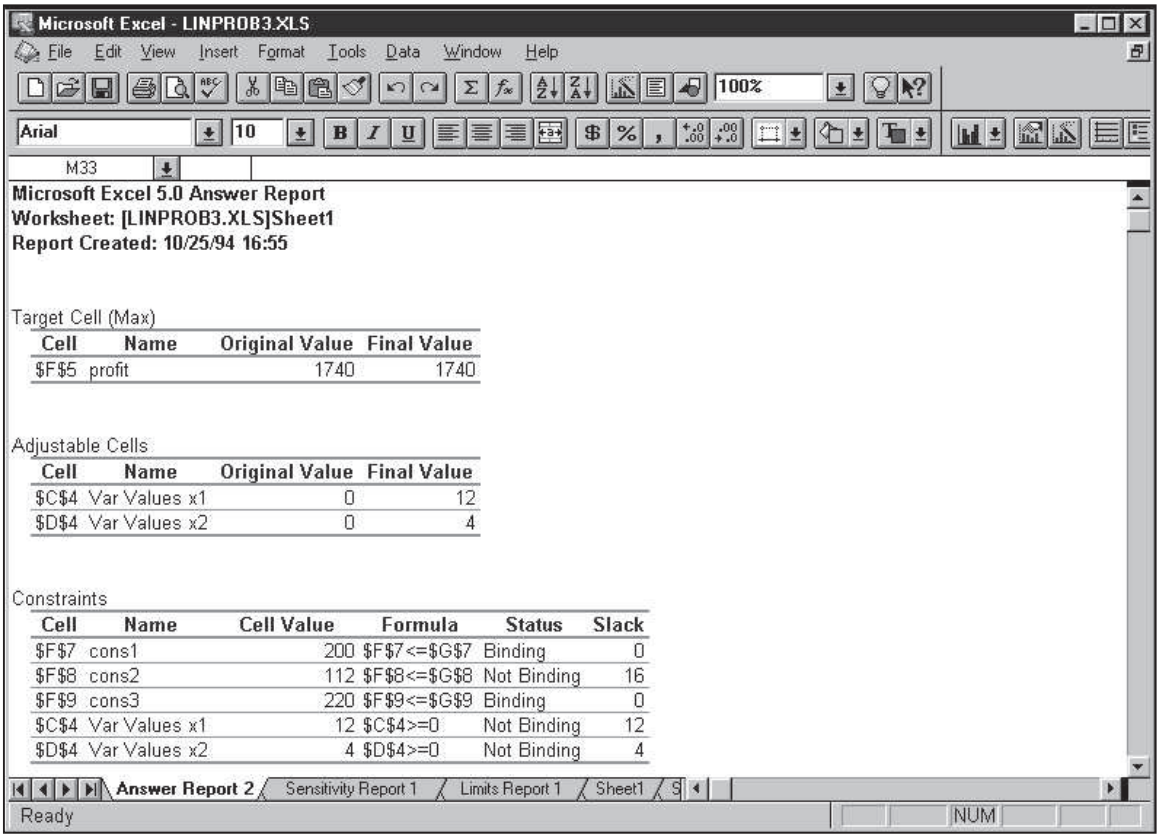


to be explicitly entered. (Note: The nonnegativity option was not available in Excel Version 5 or earlier. Nonnegativity constraints had to be entered into the problem explicitly.) Telling the program to assume a linear model ensures that the problem is solved by the Simplex algorithm rather than the gradient method used for nonlinear problems. If your output includes values of Lagrange variables, you'll know that you forgot to specify this option.

At this point your dialog box should look like Figure S1–7. Notice that I have named some of the cells and the names appear in the dialog box rather than the cell locations. Using named cells that have some meaning relative to your problem will be very helpful later when you obtain the solution and sensitivity reports. In the dialog box, we have named the objective cell (F5) "profit" and the cells corresponding to the constraints "cons1," "cons2," and "cons3."

Check that all information is correctly entered and that you have specified a linear problem. Now simply click the mouse on the Solve button and Excel will whirl away and quickly produce a solution. After solving, the resulting spreadsheet should look like the one in Figure S1–8. Notice that the values of the variables in cells C4 and D4 now reflect the optimal solution of 12 and 4, respectively. The value in cell F5 is the value of the optimal profit of $1,740. The values in cells C7, C8, and C9 are the values of the left-hand sides of the constraints.

FIGURE S1–9
Answer report for Sidneyville problem



Microsoft Excel - LINPROB3.XLS

File Edit View Insert Format Tools Data Window Help

Arial 10 100%

M33

Microsoft Excel 5.0 Answer Report
Worksheet: [LINPROB3.XLS]Sheet1
Report Created: 10/25/94 16:55

Target Cell (Max)

Cell	Name	Original Value	Final Value
$F$5	profit	1740	1740

Adjustable Cells

Cell	Name	Original Value	Final Value
$C$4	Var Values x1	0	12
$D$4	Var Values x2	0	4

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$F$7	cons1	200	$F$7<=$G$7	Binding	0
$F$8	cons2	112	$F$8<=$G$8	Not Binding	16
$F$9	cons3	220	$F$9<=$G$9	Binding	0
$C$4	Var Values x1	12	$C$4>=0	Not Binding	12
$D$4	Var Values x2	4	$D$4>=0	Not Binding	4

Answer Report 2 | Sensitivity Report 1 | Limits Report 1 | Sheet1 | S

Ready NUM

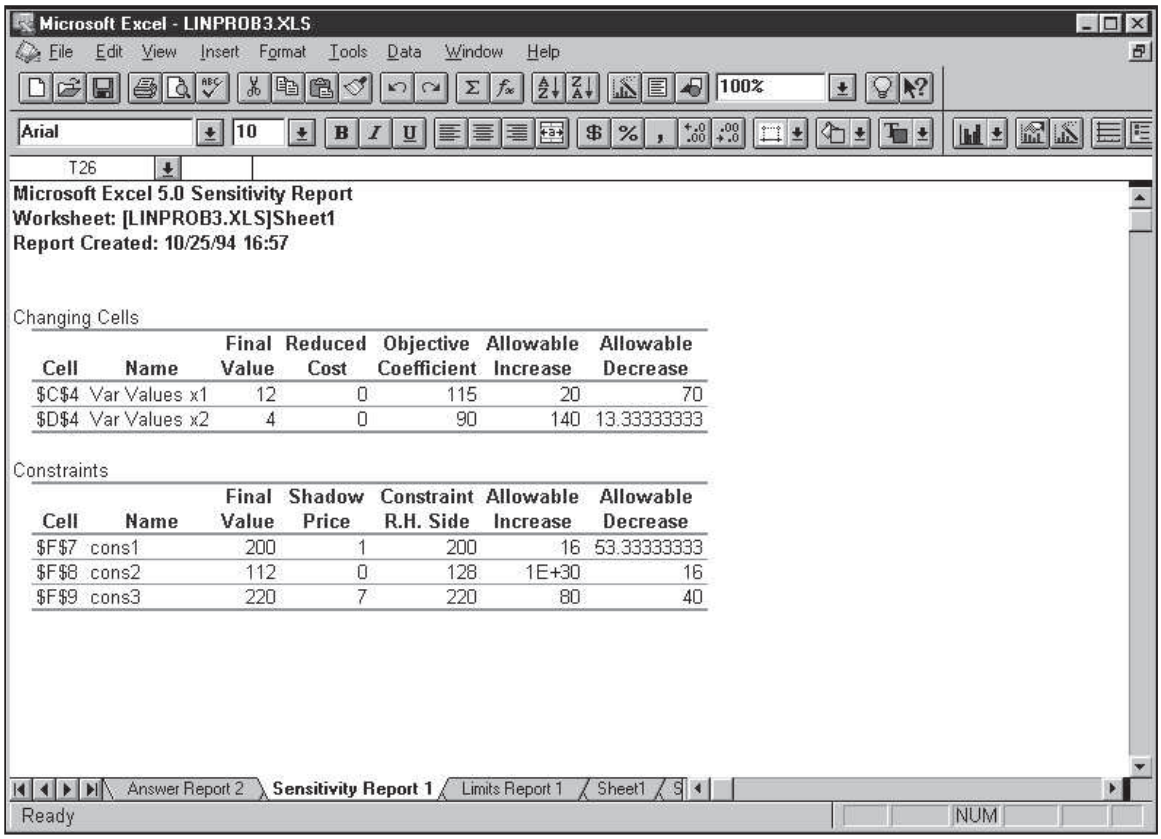
Although the optimal values of the variables and the optimal value of the objective function now appear in your spreadsheet, Excel has the option of printing several types of reports. The two that are relevant for linear programming are the Answer and Sensitivity reports. These reports appear on different sheets of the Excel workbook and are shown in Figures S1–9 and S1–10. Most of the information in the Answer report appears in the original spreadsheet. We also are told which constraints are binding. A variable that is labeled "not binding" is one that is nonzero.

Entering Large Problems Efficiently

The steps outlined for solving linear programming problems on a spreadsheet are fine for small problems. Excel has several features that allow more efficient entry of larger problems, however.

One such feature is the SUMPRODUCT function. SUMPRODUCT is a vector or array product, which multiplies the elements of one array times another array term by term and adds the results. This means that the algebraic form for the objective function and for the constraints can be entered with this function. Recall that for cell F5 we used the formula =C5*$C$4+D5*$D$4. This formula also could have been entered as =SUMPRODUCT($C$4:$D$4, C5:D5). While this may not appear to be much of an improvement here, it saves a lot of typing when entering large problems.

FIGURE S1–10
Sensitivity report for Sidneyville problem

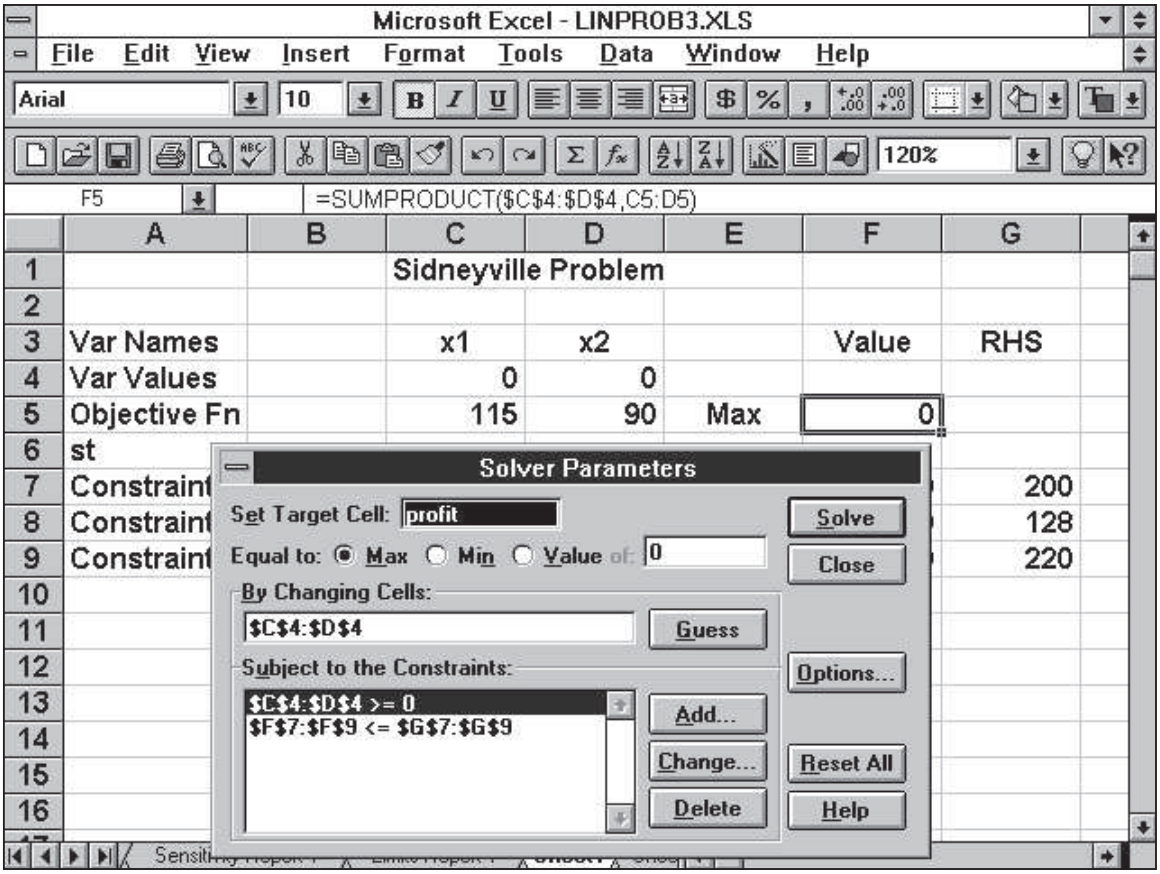


The second shortcut for entering large problems is to group constraints by type and enter all constraints in a group at one time. In the case of the Sidneyville problem, there are three $\leq$ constraints. These constraints could be entered with one command, similar to the single command for entering the nonnegativity constraints. The appropriate formulas for the constraints appear in cells F7, F8, and F9. The single command for entering these three constraints into Solver is: \$F\$7:\$F\$9 $\leq$ \$G\$7:\$G\$9. This can be typed in directly or entered by pointing to the appropriate cells in the spreadsheet. The Solver dialog box for the Sidneyville problem using this approach is shown in Figure S1–11.

Using SUMPRODUCT and entering constraints in groups can save a lot of time for large problems. However, the Solver dialog box is not as informative, since only cell locations and not cell names appear. For problems of fewer than 10 constraints total, entering the constraints one at a time is fine.

We noted earlier that an advantage of the spreadsheet for solving linear programming is that one can construct a general-purpose template. A template might have up to 10 variables and 10 constraints. Variable names could be x1 through x10 and constraint names const1 through const10. The SUMPRODUCT functions for the objective function and the constraints would be programmed in advance. One would then simply enter the coefficients of the problem to be solved and save as a new file name. (This is much faster than typing in the full algebraic representation for every new problem as one must do with other linear programming systems.)

FIGURE S1–11
Solver dialog box with efficient data entry



S1.7 INTERPRETING THE SENSITIVITY REPORT

Shadow Prices

An interesting issue is the value of additional resources in a linear programming problem. In Example S1.1, the resources are the three types of wood: pine, cedar, and maple. An increase in the level of any of these resources results in an increase in the value of the right-hand side of the appropriate constraint. The sensitivity report (Figure S1–10) gives information about the value to the objective function of additional resources. This information is contained in the values of the shadow prices for the constraints.

The shadow price is defined as the improvement in the objective function realized by adding one additional unit of a resource. For Example S1.1, the first constraint refers to the amount of pine needed. Since this constraint is binding at the optimal solution (since the final value and the constraint right-hand-side values are the same), it is likely that if we had additional pine we could increase revenue. The shadow price tells us just how beneficial. Since the shadow price for this constraint is $1, it means that for each additional unit square foot of pine, the objective (profit) increases by $1. (This will hold only within a certain range as discussed subsequently.) Consider the

second constraint (cedar). Since the final value and the right-hand side for this constraint are different (112 versus 120), there is slack in this constraint. That means we are not consuming the entire quantity of cedar at the optimal solution, and additional cedar will not improve the profit. This is borne out by the fact that the shadow price for this constraint is zero. The final shadow price of 7 indicates that every additional square foot of maple contributes \$7 to the profit.

Objective Function Coefficients and Right-Hand Sides

The shadow prices remain valid as long as the optimal basis does not change. (The basis is the set of positive variables in the final solution.) The columns Allowable Increase and Allowable Decrease indicate over what range the shadow prices remain valid. This means that we can determine the effect on the objective function of changes in constraint right-hand sides without re-solving the problem.

The first part of the sensitivity report (Figure S1–10) gives the values of the objective function coefficients for which the shadow prices remain valid. The current values of the profits are \$115 and \$90, respectively. The shadow prices are valid as long as the first objective function coefficient does not increase more than 20 or decrease more than 70 (i.e., the objective function coefficient for x_1 is between 45 and 135). Similarly, the allowable range for the objective function coefficient for x_2 is $76\frac{2}{3}$ to 230.

The second part of the sensitivity report reports the ranges on the right-hand sides of the constraints for which the shadow prices remain valid. Hence, the shadow price of \$1 for pine is valid for an increase of 16 or less and a decrease of $53\frac{1}{3}$ or less in the right-hand side of the first constraint. That is, the right-hand side of the first constraint could be any value between $146\frac{2}{3}$ and 216. The shadow price of 0 for cedar is valid for any increase (1E+30 should be interpreted as infinity) and a decrease of 16 or less, and the shadow price for maple is valid for an increase of 80 or less and a decrease of 40 or less.

If either the objective function coefficients or the right-hand sides increase or decrease beyond the allowable ranges, the shadow prices no longer remain valid, and the problem would have to be re-solved to determine the effect. Note that Excel uses the convention that a positive shadow price means an increase in the objective function per unit increase in the right-hand side of a constraint, and a negative shadow price means a decrease in the objective function per unit increase in the right-hand side, irrespective of whether the problem is a max or a min. (Other linear programming systems may have other conventions.) Also note that changes to objective function coefficients or right-hand sides can be only one at a time in these ranges. The rules for simultaneous changes in right-hand sides or objective function coefficients are much more complex and will not be discussed here.

Adding a New Variable

We can use the results of sensitivity analysis to determine whether it is profitable to add a new activity (variable) without re-solving the problem. For small problems, such as Example S1.1, simply inputting and solving the new problem on the computer is quick and easy. However, in real applications, the number of decision variables and constraints could be in the hundreds or even in the thousands. Reentering a problem of this magnitude is a major task, to be avoided whenever possible.

Suppose the firm is considering producing a third product, a vanity table, which would require the same woods used in making the desks. Each vanity table would contribute \$75 to profit, but each would require 8 square feet of pine, 6 square feet of cedar, and 10 square feet of maple. We could determine if it would be worth producing vanity tables in addition to the desks by solving the problem with three activities and comparing the values of the objective functions at the optimal solutions.

There is a faster way. The dual prices in Figure S1–10 tell us the value of each unit of resource at the current solution. The decrease in profit resulting from reducing the supply of pine is \$1 per square foot, which translates to \$8.00 for 8 square feet of pine. There is no cost for decreasing the supply of cedar. The cost of decreasing the supply of maple by 10 square feet is (10)(7) = \$70. Hence the total decrease in profit from the consumption of resources required to produce one vanity table is \$78. The contribution to profit is only \$75. We conclude that it is not optimal to produce vanity tables in addition to the desks with the current resources. Had we determined, however, that it was profitable to produce the vanity table, we would have had to re-solve the problem with three activities to find the optimal numbers of desks and vanity tables to produce.

Using Sensitivity Analysis

To cement your understanding of the information in Figure S1–10, consider the following questions:

Example *S1.1 (continued)*

a. Sidneyville's sales manager has renegotiated the contract for regular desks and now expects to make a profit of \$125 on each. He excitedly conveys this information to the firm's production manager, expecting that the optimal mix of rolltop and regular desks will change as a result. Does it?

b. Suppose that the new contract also has a higher profit for the rolltop desks. If the new profit for the rolltop desks is \$140, how will this change the optimal solution?

c. A logging company has offered to sell Sidneyville an additional 50 square feet of maple for \$5.00 per square foot. Based on the original objective function, would you recommend that it accept the offer?

d. Assuming that Sidneyville purchases the 50 square feet of maple, how is the optimal solution affected?

e. The firm is considering a pine desk that would require 25 square feet of pine and no other wood. What profit for pine desks would be required to make its production worthwhile, assuming current levels of resources and original profits on regular and rolltop desks?

f. During inspection, the quality department discovered that 50 square feet of pine had water damage and could not be used. Will it be optimal to produce both desks under these circumstances? Will the product mix change?

Solution

a. According to Figure S1–10, the allowable increase in the coefficient of the objective function for variable x_2, the regular desks, is 140. Because the increase to \$125 is still within the allowable range, the optimal mix of rolltop and regular desks will remain the same: namely, $x_1 = 12$ and $x_2 = 4$.

b. The allowable increase in the objective function for the rolltop desks (x_1) is 20, or to a maximum value of 135. As 140 is outside the allowable range, it is possible that the basis will change. However, the allowable ranges in Figure S1–10 are only valid if the profit for regular desks is \$90. The allowable ranges will change when the profit for regular desks is changed to \$125, even though the optimal solution does not. The output for part (*a*) (that is, with profits of \$115 and \$125) is

	OBJ COEFFICIENT RANGES		
VARIABLE	CURRENT COEF	ALLOWABLE INCREASE	ALLOWABLE DECREASE
X1	115.000000	72.500000	52.500000
X2	125.000000	105.000000	48.333330

This shows that the allowable increase in the coefficient for x_1 is now 72.5. Because 140 is within the allowable range, the solution for parts (*a*) and (*b*) will be the same, which is also the same as our original solution of $x_1 = 12$ and $x_2 = 4$.

c. Since the dual price for the third constraint corresponding to maple is 7, it is profitable to purchase the maple for $5.00 a square foot. The allowable increase of the right-hand side over which this dual price applies is 80, so it is worth purchasing the full 50 additional square feet.

d. Because the increase of 50 is within the allowable right-hand-side range, we know that the basis will not change. That is, it still will be optimal to produce both the rolltop and the regular desks. However, if the right-hand side changes, the values of the basic variables *will* change. We must re-solve the problem with the new right-hand-side value to determine the updated solution. The solution is

```
                        LP OPTIMUM FOUND AT STEP 1

OBJECTIVE FUNCTION VALUE

            1)          2090.000000

       VARIABLE        VALUE                REDUCED COST
       X1              17.000000                .000000
       X2               1.500000                .000000
```

To retain feasibility we round x_2 to 1. (This is *not* the optimal integer solution, however. The optimal integer solution is $x_1 = 18$ and $x_2 = 0$ with a profit of $2,070, which is obtained from Excel by identifying both x_1 and x_2 as integer variables. The suboptimal solution of $x_1 = 17$ and $x_2 = 1$ results in a profit of $2,045.)

e. The dual price for pine is $1 per square foot. As each desk consumes 25 square feet of pine, the profit for each pine desk must exceed $25 for pine desks to be profitable to produce.

f. The right-hand side of the first constraint can decrease as much as 53.333330 and the current basis will remain optimal. That means that a decrease of 50 square feet will not change the basis; it will still be profitable to produce both desks. However, the production quantities will decrease. We must re-solve the problem to determine the correct levels of the new quantities. They are

```
LP OPTIMUM FOUND AT STEP 2
 OBJECTIVE FUNCTION VALUE
        1)     1690.000000
  VARIABLE      VALUE                    REDUCED COST
        X1      14.500000                    .000000
        X2        .250000                    .000000
```

Again, we need to round these variables. Rounding both x_1 and x_2 down guarantees feasibility. If we produce 14 rolltop desks and 0 regular desks, we require 140 square feet of pine (150 are available), 56 square feet of cedar (128 are available), and 210 square feet of maple (220 are available). There does not appear to be enough wood to produce an additional desk of either type, so we leave the solution at $x_1 = 14$ and $x_2 = 0$. (This is the optimal integer solution.)

S1.8 RECOGNIZING SPECIAL PROBLEMS

Several problems can occur when solving linear programming problems. In this section we will discuss the causes of these problems and how one recognizes them when using Excel.

Unbounded Solutions

The feasible region of a linear program is not necessarily bounded. The feasible region for the Sidneyville problem pictured in Figure S1–4 is bounded. However, consider the following linear programming problem.

Example *S1.2*

Maximize

$$2x_1 + 3x_2$$

subject to

$$\begin{aligned} x_1 + 4x_2 &\geq 8, \\ x_1 + x_2 &\geq 5, \\ 2x_1 + x_2 &\geq 7, \\ x_1, x_2 &\geq 0. \end{aligned}$$

Figure S1–12 shows the feasible region. Notice that it is unbounded. Because we can make x_1 and x_2 as large as we like, there is no limit to the size of the objective function. When this occurs, the problem is unbounded and there is no optimal solution.

When this problem is inputted in Excel, Solver writes in very large values for the problem variables and displays a message that set target values do not converge. The Excel output for this problem appears in Figure S1–13.

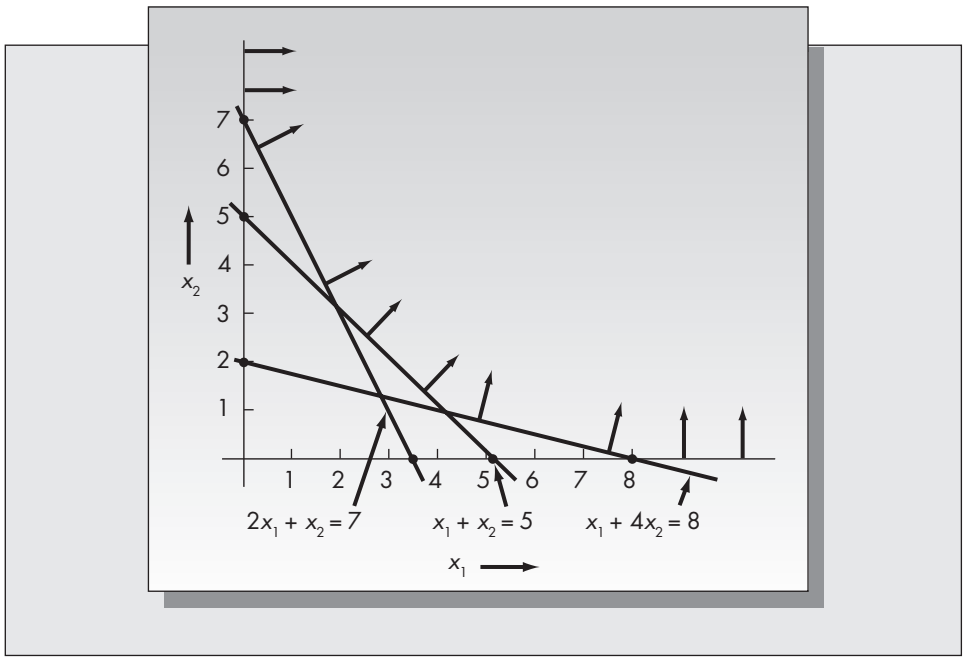
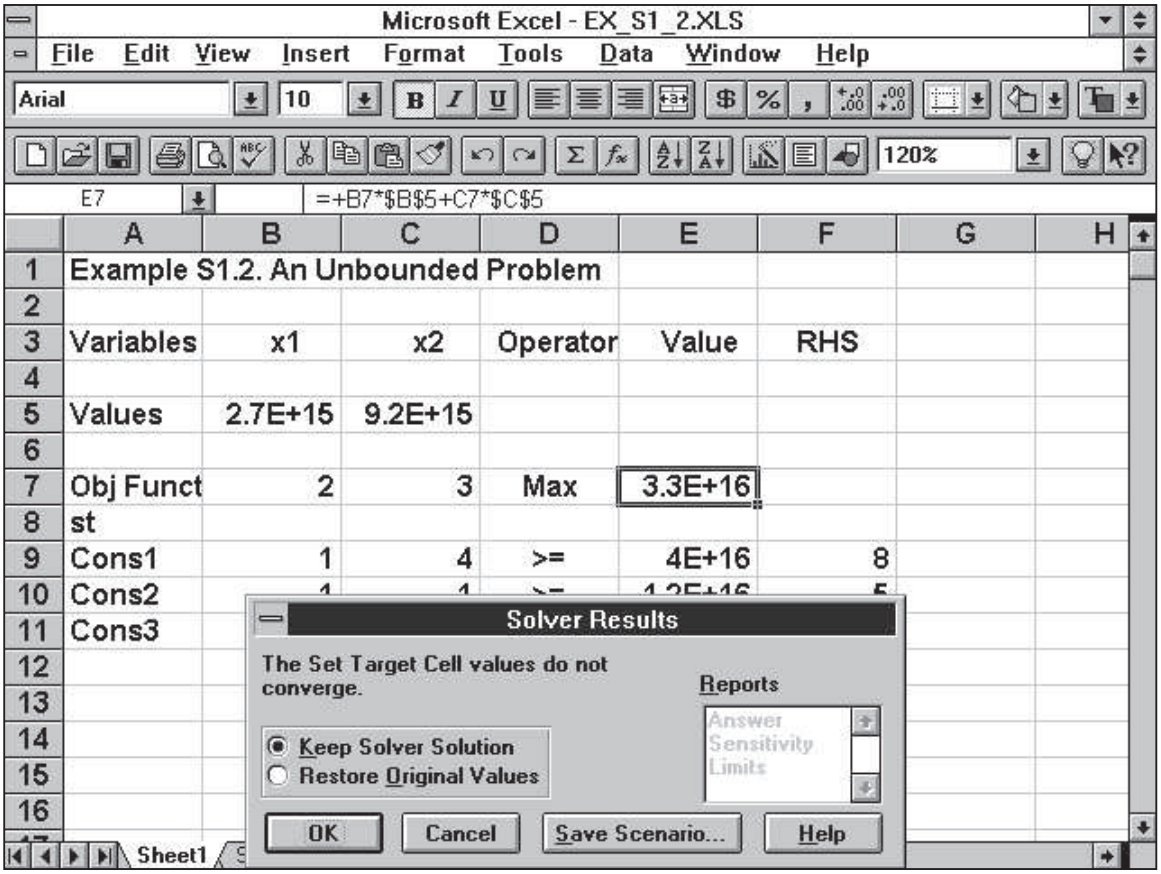


FIGURE S1–12
Feasible region for Example S1.2

FIGURE S1–13
Excel output for Example S1.2



Empty Feasible Region

It is possible for two or more constraints to be inconsistent. When that occurs, there will be no feasible solution. Consider the following problem.

Example *S1.3* Maximize

$$2x_1 + 3x_2$$

subject to

$$x_1 + 4x_2 \leq 8,$$
$$x_1 + x_2 \geq 10,$$
$$x_1, x_2 \geq 0.$$

The feasible region for this example appears in Figure S1–14. Notice that there is no intersection of the half spaces defined by the two constraints in the positive quadrant. In this case, the feasible region is empty and we say that the problem is infeasible. The Excel output appears in Figure S1–15. Note that the solution $x_1 = 8$ and $x_2 = 0$ shown is not feasible because it results in negative slack in the first constraint.

FIGURE S1–14
Feasible region for Example S1.3

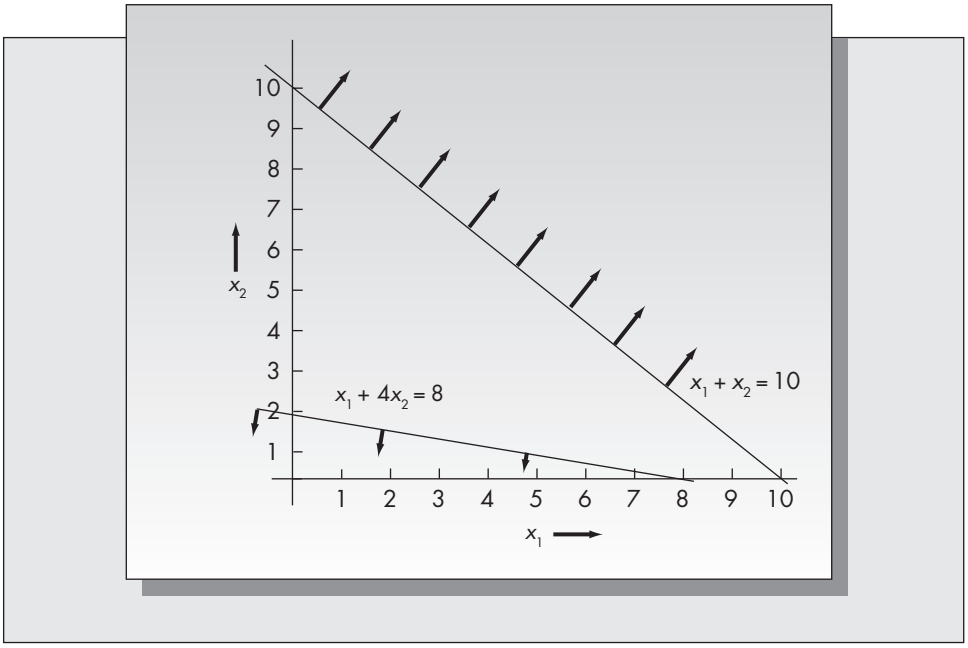
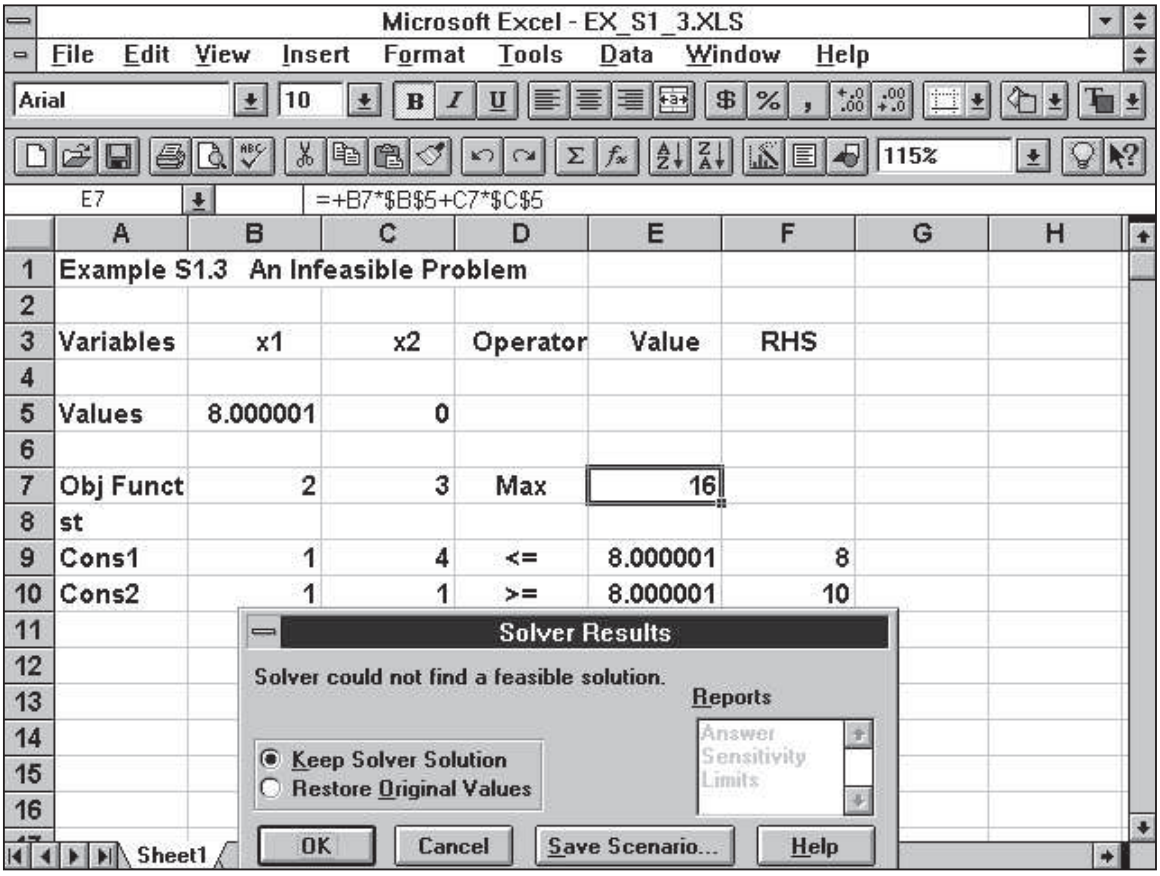


FIGURE S1–15
Excel output for Example S1.3



Degeneracy

In linear programming there are two types of variables: basic variables and nonbasic variables. The number of basic variables equals the number of constraints, and basic variables may be either original variables or slack or surplus variables. What defines basic variables? Consider any linear program in standard form. By including slack and surplus variables, all constraints are expressed as equations. In standard form, there always will be more variables than constraints. Suppose that after a linear programming problem has been expressed in standard form, there are $n + m$ variables and n constraints. A basic solution is found by setting m variables to zero and solving the resulting n equations in n unknowns. In most cases the values of the n basic variables will be positive. A degenerate solution occurs when one or more basic variables are zero at the optimal solution. In Excel, a basic variable is one with zero reduced cost. Degeneracy occurs when the value of a variable is zero and its reduced cost or shadow price is also zero.

Why are we interested in degeneracy? It is possible that if degenerate solutions occur, the Simplex Method will cycle through some set of solutions and never recognize the optimal solution. The phenomenon of cycling has never been observed in practice, and most computer programs have means of guaranteeing that it never occurs. The bottom line is that degeneracy is an issue about which we need not worry, but one of which we should be aware.

Multiple Optimal Solutions

The optimal solution to a linear program is not always unique. There are cases in which there are multiple optimal solutions. In Chapter 3 we saw that two-variable problems could be solved by graphical means by approaching the feasible region with the Z line. Assuming that we approach the feasible region from the correct side, the first feasible point with which the Z line comes into contact is the optimal solution.

However, suppose that the Z line is parallel to one of the constraints. In that case it does not contact a single point first, but an entire edge.

Example *S1.4*

Consider the feasible region pictured in Figure S1–12 corresponding to the constraints of Example S1.2. Suppose that the objective function is min $3x_1 + 3x_2$. Then the Z line has slope -1 and is parallel to the constraint boundary $x_1 + x_2 = 5$. As the Z line approaches the feasible region, it meets the edge defined by this constraint and both extreme points along this edge. This means that both extreme points *and* all points along the edge are optimal.

Unfortunately, Excel does not indicate that there are multiple optimal solutions to this problem. Our only clue is that the solution is degenerate; the surplus variable for the third constraint has both zero dual price and zero value. Our graphical solution tells us that both extreme points (2, 3) and (4, 1) are optimal, and so are all points along the edge connecting these extreme points. [The points along the edge can be written in the form

$$x_1 = \alpha(2) + (1 - \alpha)(4),$$

$$x_2 = \alpha(3) + (1 - \alpha)(1),$$

where α is a number between zero and one. This is known as a convex combination of these two extreme points.]

Redundant Constraints

It is possible for one or more constraints to be redundant. That means that these constraints can be eliminated from the formulation without affecting the solution. In simple two-variable problems, redundant constraints can be recognized graphically because they lie outside the feasible region. Excel does not recognize or signal if one or more constraints are redundant. Sometimes redundant constraints can cause degeneracy, but degeneracy can result when constraints are not redundant as well.

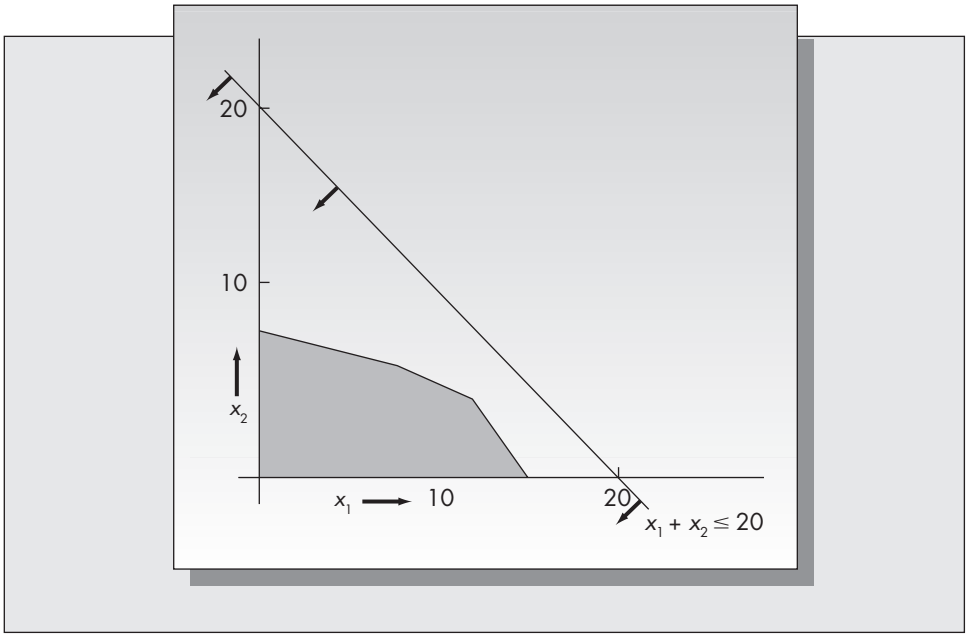


FIGURE S1–16
Feasible region for Example S1.5 showing a redundant constraint

Example *S1.5* Consider Example S1.1. Suppose that we add the following additional constraint:

$$x_1 + x_2 \leq 20.$$

Figure S1–16 shows the resulting feasible region. It is exactly the same as the feasible region pictured in Figure S1–4. The additional constraint has no effect, as it lies completely outside the feasible region defined by the first three constraints. The optimal solution will, of course, be exactly the same.

If we had originally formulated the problem with the four constraints

$$\begin{aligned} 10x_1 + 20x_2 &\leq 200, \\ 4x_1 + 16x_2 &\leq 128, \\ 15x_1 + 10x_2 &\leq 220, \\ x_1 + x_2 &\leq 20, \end{aligned}$$

and solved, the output indicates that the optimal solution is $x_1 = 12$ and $x_2 = 4$ as before, and gives us no clue that the last constraint is redundant. The only way to see that the final constraint is redundant is to graph the feasible region as we did in Figure S1–16. Graphing is possible, however, in two-variable problems only.

Does redundancy cause a problem? Not really. We would certainly like to be able to write our linear program as economically as possible, but if one or more constraints are redundant, the optimal solution is unaffected.

S1.9 THE APPLICATION OF LINEAR PROGRAMMING TO PRODUCTION AND OPERATIONS ANALYSIS

In Chapter 3 we showed how linear programming could be used to find optimal solutions (subject to rounding errors) for aggregate planning problems. Although this is the only explicit use of linear programming in this book, there have been successful linear

programming applications for many operations management problems.[5] Scheduling and distribution are perhaps two areas in which applications are most common.

Fisher et al. (1982) describe an application of linear programming to the problem of providing a coordinated vehicle scheduling and routing system for delivery of consumable products to customers of the Du Pont Company. The primary issue was to determine delivery routes (loops) for the trucks to the company's clients in various regions of the country. A refrigerated truck drives a weekly loop that includes several dozen customers. The largest region considered, Chicago, had 16 loops and several hundred cities, whereas the Houston region, the smallest, had 4 loops and less than 80 cities.

The basic mathematical formulation of the problem was a generalized assignment problem. (The assignment problem is discussed in Chapter 9. It is a linear programming problem in which the decision variables are restricted to be zeros and ones.) The mathematical formulation used in this study is the following:

1. Given data

$$d_{ik} = \text{Cost of including customer } i \text{ in loop } k,$$
$$a_i = \text{Demand from customer } i.$$

2. Problem variables

$$y_{ik} = \begin{cases} 1 & \text{if customer } i \text{ is assigned to loop } k, \\ 0 & \text{if customer } i \text{ is not assigned to loop } k. \end{cases}$$

3. Generalized assignment problem

$$\text{Min} \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ik}\, y_{ik}$$

subject to

$$\sum_{k=1}^{K} y_{ik} = 1, \qquad \text{for } i = 1, \ldots, n,$$
$$\sum_{i=1}^{n} a_i\, y_{ik} \le b_k, \qquad \text{for } k = 1, \ldots, K,$$
$$y_{ik} = 0 \text{ or } 1, \quad \text{for all } i \text{ and } k,$$

where K is the total number of loops in the region and n is the number of customers.

The implementation of this model was reported to have saved Du Pont over $200 million. A more complex mathematical model solving a similar problem for Air Products Corporation was reported by Bell et al. (1983). This study won the Institute of Management Sciences Practice Award in 1983.

Linear programming (or, more generally, mathematical programming) has been an important tool for logistics planning for a wide variety of operations management problems. Today, Microsoft bundles Solver, a general purpose mathematical programming Excel add-in, with its office software, thus making linear programming accessible to a much wider audience.

[5] In this section we interpret linear programming in the broad sense to include integer linear programming.

Bibliography

Bell, W. J.; L. M. Dalberto; M. L. Fisher; A. J. Greenfield; R. Jaikumar; P. Kedia; R. G. Mack; and P. J. Prvtzman. "Improving the Distribution of Industrial Gases with an On-Line Computerized Routing and Scheduling Optimizer." *Interfaces* 13 (1983), pp. 4–23.

Fisher, M.; A. J. Greenfield; R. Jaikumar; and J. T. Uster III. "A Computerized Vehicle Routing Application." *Interfaces* 12 (1982), pp. 42–52.

Hadley, G. *Linear Programming.* Reading, MA: Addison-Wesley, 1962.

Hillier, F. S., and G. J. Lieberman. *Introduction to Operations Research.* 5th ed. San Francisco: Holden Day, 1990.

Chapter **Four**

Inventory Control Subject to Known Demand

"We want to turn our inventory faster than our people."
—James Sinegal

Chapter Overview

Purpose

To consider methods for controlling individual item inventories when product demand is assumed to follow a known pattern (that is, demand forecast error is zero).

Key Points

1. *Classification of inventories*
   - *Raw materials.* These are resources required for production or processing.
   - *Components.* These could be raw materials or subassemblies that will later be included into a final product.
   - *Work-in-process (WIP).* These are inventories that are in the plant waiting for processing.
   - *Finished goods.* These are items that have completed the production process and are waiting to be shipped out.
2. *Why hold inventory?*
   - *Economies of scale.* It is probably cheaper to order or produce in large batches than in small batches.
   - *Uncertainties.* Demand uncertainty, lead time uncertainty, and supply uncertainty all provide reasons for holding inventory.
   - *Speculation.* Inventories may be held in anticipation of a rise in their value or cost.
   - *Transportation.* Refers to pipeline inventories that are in transit from one location to another.
   - *Smoothing.* As noted in Chapter 3, inventories provide a means of smoothing out an irregular demand pattern.
   - *Logistics.* System constraints that may require holding inventories.
   - *Control costs.* Holding inventory can lower the costs necessary to monitor a system. (For example, it may be less expensive to order yearly and hold the units than to order weekly and closely monitor orders and deliveries.)

3. *Characteristics of inventory systems*
   - *Patterns of demand.* The two patterns are (*a*) constant versus variable and (*b*) known versus uncertain.
   - *Replenishment lead times.* The time between placement of an order (or initiation of production) until the order arrives (or is completed).
   - *Review times.* The points in time that current inventory levels are checked.
   - *Treatment of excess demand.* When demand exceeds supply, excess demand may be either backlogged or lost.
4. *Relevant costs*
   - *Holding costs.* These include the opportunity cost of lost investment revenue, physical storage costs, insurance, breakage and pilferage, and obsolescence.
   - *Order costs.* These generally consist of two components: a fixed component and a variable component. The fixed component is incurred whenever a positive order is placed (or a production run is initiated), and the variable component is a unit cost paid for each unit ordered or produced.
   - *Penalty costs.* These are incurred when demand exceeds supply. In this case excess demand may be back-ordered (to be filled at a later time) or lost. Lost demand results in lost profit, and back orders require record keeping and in both cases, one risks losing customer goodwill.
5. *The basic EOQ model.* The EOQ model dates back to 1915 and forms the basis for all the inventory control models developed subsequently. It treats the basic trade-off between the fixed cost of ordering and the variable cost of holding. If h represents the holding cost per unit time and K the fixed cost of setup, then we show that the order quantity that minimizes costs per unit time is $Q = \sqrt{2K\lambda/h}$, where λ is the rate of demand. This formula is very robust for several reasons: (*a*) It is a very accurate approximation for the optimal order quantity when demand is uncertain (treated in Chapter 5), and (*b*) we show that deviations from the optimal Q generally result in modest cost errors. For example, a 25 percent error in Q results in an average annual holding and setup cost error of only 2.5 percent.
6. *The EOQ with finite production rate.* This is an extension of the basic EOQ model to take into account that when items are produced internally rather than ordered from an outside supplier, the rate of production is finite rather than infinite, as would be required in the simple EOQ model. We show that the optimal size of a production run now follows the formula $Q = \sqrt{2K\lambda/h'}$ where $h' = h(1 - \lambda/P)$ and P is the rate of production ($P > \lambda$). Note that since $h' < h$, the batch size when the production rate is taken into account exceeds the batch size obtained by the EOQ formula.
7. *Quantity discounts.* We consider two types of quantity discounts: all-units and incremental discounts. In the case of all-units discounts, the discount is applied to all the units in the order, while in the case of incremental discounts, the discount is applied to only the units above the break point. The all-units case is by far the most common in practice, but one does encounter incremental discounts in industry. In the case of all-units discounts, the optimization procedure requires searching for the lowest point on a broken annual cost curve. In the incremental discounts case, the annual cost curve is continuous, but has discontinuous derivatives.

8. *Resource-constrained multiple product systems.* Consider a retail store that orders many different items, but cannot exceed a fixed budget. If we optimize the order quantity of each item separately, then each item should be ordered according to its EOQ value. However, suppose doing so exceeds the budget. In this section, a model is developed that explicitly takes into account the budget constraint and adjusts the EOQ values accordingly. In most cases, the optimal solution subject to the budget constraint requires an iterative search of the Lagrange multiplier. However, when the condition $c_1/h_1 = c_2/h_2 = \cdots = c_n/h_n$ is met, the optimal order quantities are a simple scaling of the optimal EOQ values. Note that this problem is mathematically identical to one in which the constraint is on available space rather than available budget.
9. *EOQ models for production planning.* Suppose that n distinct products are produced on a single production line or machine. Assume we know the holding costs, order costs, demand rates, and production rates for each of the items. The goal is to determine the optimal sequence to produce the items, and the optimal batch size for each of the items to meet the demand and minimize costs. Note that simply setting a batch size for each item equal to its EOQ value (that is, optimal lot size with a finite production rate), is likely to be suboptimal since it is likely to result in stock-outs. The problem is handled by considering the optimal cycle time, T, where we assume we produce exactly one lot of each item each cycle. The optimal size of the production run for item j is simply $Q_j = \lambda_j T$, where T is the optimal cycle time. Finding T is nontrivial, however.

The current investment in inventories in the United States is enormous. In the third quarter of 2007, the total dollar investment was estimated to be $1.57 trillion.[1] Figure 4–1 shows investment in inventories broken down by sectors of the economy. The inventory models we will be discussing in this chapter and in Chapter 5 can be applied

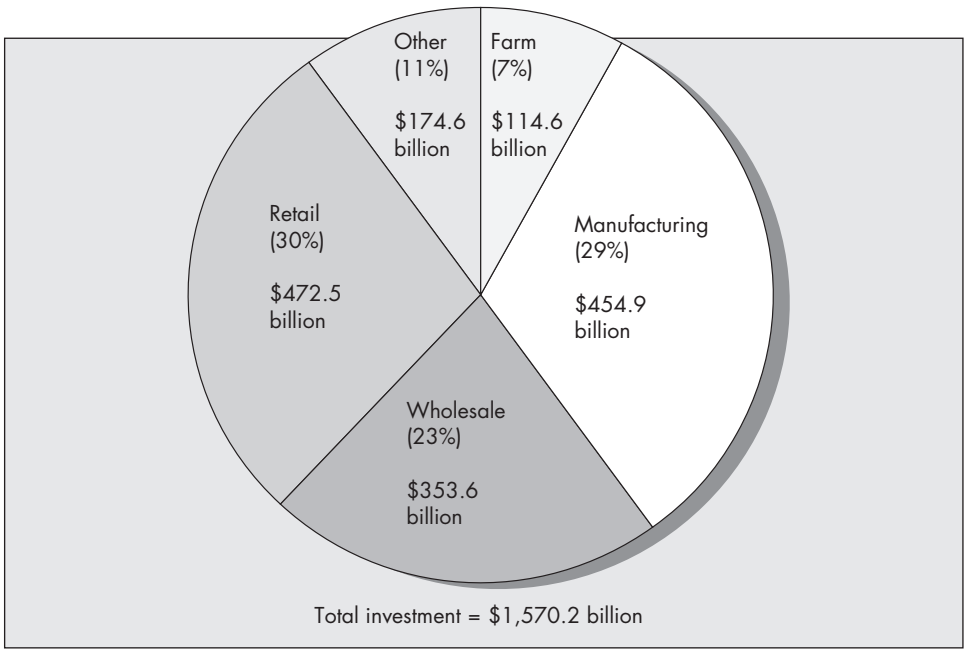


FIGURE 4–1
Breakdown of the total investment in inventories in the U.S. economy (2007)

[1] *Survey of Current Business* (July 2007).

to all the sectors of the economy shown in Figure 4–1, but are most applicable to the manufacturing, wholesale, and retail sectors, which compose approximately 82 percent of the total. The trillion-dollar investment in inventories accounts for between 20 and 25 percent of the total annual GNP. Clearly there is enormous potential for improving the efficiency of our economy by intelligently controlling inventories. Companies that use scientific inventory control methods have a significant competitive advantage in the marketplace.

A major portion of this text is devoted to presenting and analyzing several mathematical models that can assist with controlling the replenishment of inventories. Both Chapters 4 and 5 assume that the demand for the item is external to the system. In most cases, this means that the inventory is being acquired or produced to meet the needs of a customer. In a manufacturing environment, however, demands for certain parts are the result of production schedules for higher-level assemblies; the production-lot-sizing decisions at one level of the system result in the demand patterns at other levels. The interaction of components, subassemblies, and final products plays an important role in determining future demand. Systems of this type are referred to as *materials requirements planning* (MRP) systems or dependent demand systems. MRP is treated in detail in Chapter 8.

The fundamental problem of inventory management can be succinctly described by the two questions (1) When should an order be placed? and (2) How much should be ordered? The complexity of the resulting model depends upon the assumptions one makes about the various parameters of the system. The major distinction is between models that assume known demand (this chapter) and those that assume random demand (Chapter 5), although, as we will see, the form of the cost functions and the assumptions one makes about physical characteristics of the system also play an important role in determining the complexity of the resulting model.

In general, the models that we discuss can be used interchangeably to describe either replenishment from an outside vendor or internal production. This means that from the point of view of the model, inventory control and production planning are often synonymous. For example, the lot-sizing methods treated in Chapter 8 could just as well have been included in this chapter. The issue is not the label that is placed on a technique, but whether it is being correctly applied to the problem being addressed.

4.1 TYPES OF INVENTORIES

When we consider inventories in the context of manufacturing and distribution, there is a natural classification scheme suggested by the value added from manufacturing or processing. (This certainly is not the only means of categorizing inventories, but it is the most natural one for manufacturing applications.)

1. *Raw materials.* These are the resources required in the production or processing activity of the firm.
2. *Components.* Components correspond to items that have not yet reached completion in the production process. Components are sometimes referred to as subassemblies.
3. *Work-in-process.* Work-in-process (WIP) is inventory either waiting in the system for processing or being processed. Work-in-process inventories include component inventories and may include some raw materials inventories as well. The level of work-in-process inventory is often used as a measure of the efficiency of a production scheduling system. The just-in-time approach, discussed in detail in Chapter 8, is aimed at reducing WIP to a minimum.

4. *Finished goods.* Also known as end items, these are the final products of the production process. During production, value is added to the inventory at each level of the manufacturing operation, culminating with finished goods.

The appropriate label to place on inventory depends upon the context. For example, components for some operations might be the end products for others.

4.2 MOTIVATION FOR HOLDING INVENTORIES

1. *Economies of scale.* Consider a company that produces a line of similar items, such as air filters for automobiles. Each production run of a particular size of filter requires that the production line be reconfigured and the machines recalibrated. Because the company must invest substantial time and money in setting up to produce each filter size, enough filters should be produced at each setup to justify this cost. This means that it could be economical to produce a relatively large number of items in each production run and store them for future use. This allows the firm to amortize fixed setup costs over a larger number of units.[2]

2. *Uncertainties.* Uncertainty often plays a major role in motivating a firm to store inventories. Uncertainty of external demand is the most important. For example, a retailer stocks different items so that he or she can be responsive to consumer preferences. If a customer requests an item that is not available immediately, it is likely that the customer will go elsewhere. Worse, the customer may never return. Inventory provides a buffer against the uncertainty of demand.

Other uncertainties provide a motivation for holding inventories as well. One is the uncertainty of the lead time. Lead time is defined as the amount of time that elapses from the point that an order is placed until it arrives. In the production planning context, interpret the lead time as the time required to produce the item. Even when future demand can be predicted accurately, the company needs to hold buffer stocks to ensure a smooth flow of production or continued sales when replenishment lead times are uncertain.

A third significant source of uncertainty is the supply. The OPEC oil embargo of the late 1970s is an example of the chaos that can result when supply lines are threatened. Two industries that relied (and continue to rely) heavily on oil and gasoline are the electric utilities and the airlines. Firms in these and other industries risked having to curtail operations because of fuel shortages.

Additional uncertainties that could motivate a firm to store inventory include the uncertainty in the supply of labor, the price of resources, and the cost of capital.

3. *Speculation.* If the value of an item or natural resource is expected to increase, it may be more economical to purchase large quantities at current prices and store the items for future use than to pay the higher prices at a future date. In the early 1970s, for example, the Westinghouse Corporation sustained severe losses on its contracts to build nuclear plants for several electric utility companies because it guaranteed to supply the uranium necessary to operate the plants at a fixed price. Unfortunately for Westinghouse, the price of the uranium skyrocketed between the time the contracts were signed and the time the plants were built.

Other industries require large quantities of costly commodities that have experienced considerable fluctuation in price. For example, silver is required for the production of photographic film. By correctly anticipating a major price increase in

[2] This argument assumes that the setup cost is a fixed constant. In some circumstances it can be reduced, thus justifying smaller lot sizes. This forms the basis of the just-in-time philosophy discussed in detail in Chapter 8.

silver, a major producer of photographic film, such as Kodak, could purchase and store large quantities of silver in advance of the increase and realize substantial savings.

The speculative motive also can be a factor for a firm facing the possibility of a labor strike. The cost of production could increase significantly when there is a severe shortage of labor.

4. *Transportation.* In-transit or *pipeline* inventories exist because transportation times are positive. When transportation times are long, as is the case when transporting oil from the Middle East to the United States, the investment in pipeline inventories can be substantial. One of the disadvantages of producing overseas is the increased transportation time, and hence the increase in pipeline inventories. This factor has been instrumental in motivating some firms to establish production operations domestically.

5. *Smoothing.* Changes in the demand pattern for a product can be deterministic or random. Seasonality is an example of a deterministic variation, while unanticipated changes in economic conditions can result in random variation. Producing and storing inventory in anticipation of peak demand can help to alleviate the disruptions caused by changing production rates and workforce levels. Smoothing costs and planning for anticipated swings in the demand were considered in the aggregate planning models in Chapter 3.

6. *Logistics.* We use the term *logistics* to describe reasons for holding inventory different from those already outlined. Certain constraints can arise in the purchasing, production, or distribution of items that force the system to maintain inventory. One such case is an item that must be purchased in minimum quantities. Another is the logistics of manufacture; it is virtually impossible to reduce all inventories to zero and expect any continuity in a manufacturing process.

7. *Control costs.* An important issue, and one that often is overlooked, is the cost of maintaining the inventory control system. A system in which more inventory is carried does not require the same level of control as one in which inventory levels are kept to a bare minimum. It can be less costly to the firm in the long run to maintain large inventories of inexpensive items than to expend worker time to keep detailed records for these items. Even though control costs could be a major factor in determining the suitability of a particular technique or system, they are rarely factored into the types of inventory models we will be discussing.

4.3 CHARACTERISTICS OF INVENTORY SYSTEMS

1. *Demand.* The assumptions one makes about the pattern and characteristics of the demand often turn out to be the most significant in determining the complexity of the resulting control model.

   a. *Constant versus variable.* The simplest inventory models assume that the rate of demand is a constant. The economic order quantity (EOQ) model and its extensions are based on this assumption. Variable demand arises in a variety of contexts, including aggregate planning (Chapter 3) and materials requirements planning (Chapter 8).

   b. *Known versus random.* It is possible for demand to be constant in expectation but still be random. Synonyms for random are *uncertain* and *stochastic.* Virtually all stochastic demand models assume that the average demand rate is constant. Random demand models are generally both more realistic and more complex than their deterministic counterparts.

2. *Lead time.* If items are ordered from the outside, the lead time is defined as the amount of time that elapses from the instant that an order is placed until it arrives. If items are produced internally, however, then interpret lead time as the amount of time required to produce a batch of items. We will use the Greek letter τ to represent lead time, which is expressed in the same units of time as demand. That is, if demand is expressed in units per year, then lead time should be expressed in years.

3. *Review time.* In some systems the current level of inventory is known at all times. This is an accurate assumption when demand transactions are recorded as they occur. One example of a system in which inventory levels are known at all times is a modern supermarket with a visual scanning device at the checkout stand that is linked to a storewide inventory database. As an item is passed through the scanner, the transaction is recorded in the database, and the inventory level is decreased by one unit. We will refer to this case as *continuous review.* In the other case, referred to as *periodic review,* inventory levels are known only at discrete points in time. An example of periodic review is a small grocery store in which physical stock-taking is required to determine the current levels of on-hand inventory.

4. *Excess demand.* Another important distinguishing characteristic is how the system reacts to excess demand (that is, demand that cannot be filled immediately from stock). The two most common assumptions are that excess demand is either back-ordered (held over to be satisfied at a future time) or lost (generally satisfied from outside the system). Other possibilities include partial back-ordering (part of the demand is back-ordered and part of the demand is lost) or customer impatience (if the customer's order is not filled within a fixed amount of time, he or she cancels). The vast majority of inventory models, especially the ones that are used in practice, assume full back-ordering of excess demand.

5. *Changing inventory.* In some cases the inventory undergoes changes over time that may affect its utility. Some items have a limited shelf life, such as food, and others may become obsolete, such as automotive spare parts. Mathematical models that incorporate the effects of perishability or obsolescence are generally quite complex and beyond the scope of this text. A brief discussion can be found in Section 5.8.

4.4 RELEVANT COSTS

Because we are interested in optimizing the inventory system, we must determine an appropriate optimization or performance criterion. Virtually all inventory models use cost minimization as the optimization criterion. An alternative performance criterion might be profit maximization. However, cost minimization and profit maximization are essentially equivalent criteria for most inventory control problems. Although different systems have different characteristics, virtually all inventory costs can be placed into one of three categories: holding cost, order cost, or penalty cost. We discuss each in turn.

Holding Cost

The **holding cost,** also known as the carrying cost or the inventory cost, is the sum of all costs that are proportional to the amount of inventory physically on hand at any point in time. The components of the holding cost include a variety of seemingly unrelated items. Some of these are

- Cost of providing the physical space to store the items.
- Taxes and insurance.
- Breakage, spoilage, deterioration, and obsolescence.
- Opportunity cost of alternative investment.

The last item often turns out to be the most significant in computing holding costs for most applications. Inventory and cash are in some sense equivalent. Capital must be invested to either purchase or produce inventory, and decreasing inventory levels results in increased capital. This capital could be invested by the company either internally, in its own operation, or externally.

What is the interest rate that could be earned on this capital? You and I can place our money in a simple passbook account with an interest rate of 2 percent, or possibly a long-term certificate of deposit with a return of maybe 5 percent. We could earn somewhat more by investing in high-yield bond funds or buying short-term industrial paper or second deeds of trust.

In general, however, most companies must earn higher rates of return on their investments than do individuals in order to remain profitable. The value of the interest rate that corresponds to the opportunity cost of alternative investment is related to (but not the same as) a number of standard accounting measures, including the internal rate of return, the return on assets, and the hurdle rate (the minimum rate that would make an investment attractive to the firm). Finding the right interest rate for the opportunity cost of alternative investment is very difficult. Its value is estimated by the firm's accounting department and is usually an amalgam of the accounting measures listed earlier. For convenience, we will use the term *cost of capital* to refer to this component of the holding cost. We may think of the holding cost as an aggregated interest rate comprised of the four components we listed. For example,

$$\begin{aligned} 28\% &= \text{Cost of capital} \\ 2\% &= \text{Taxes and insurance} \\ 6\% &= \text{Cost of storage} \\ 1\% &= \text{Breakage and spoilage} \\ \hline 37\% &= \text{Total interest charge} \end{aligned}$$

This would be interpreted as follows: We would assess a charge of 37 cents for every dollar that we have invested in inventory during a one-year period. However, as we generally measure inventory in units rather than in dollars, it is convenient to express the holding cost in terms of dollars per unit per year rather than dollars per dollar per year. Let c be the dollar value of one unit of inventory, I be the annual interest rate, and h be the holding cost in terms of dollars per unit per year. Then we have the relationship

$$h = Ic.$$

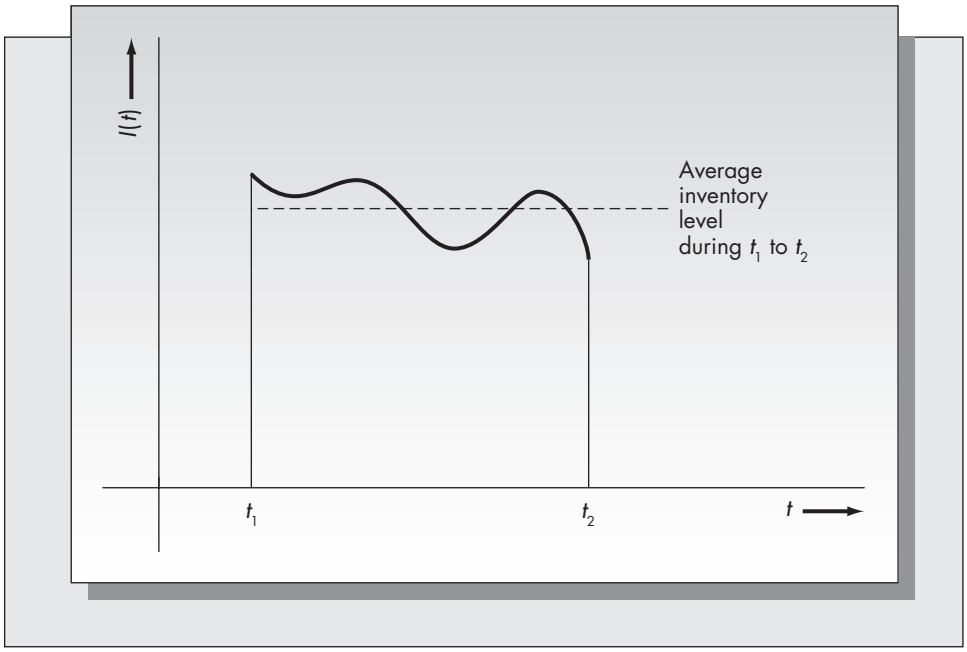
Hence, in this example, an item valued at \$180 would have an annual holding cost of $h = (0.37)(\$180) = \$66.60$. If we held 300 of these items for five years, the total holding cost over the five years would be

$$(5)(300)(66.60) = \$99{,}900.$$

This example raises an interesting question. Suppose that during the five-year period the inventory level did not stay fixed at 300 but varied on a continuous basis. We would expect inventory levels to change over time. Inventory levels decrease when items are used to satisfy demand and increase when units are produced or new orders arrive. How would the holding cost be computed in such a case? In particular, suppose the inventory level $I(t)$ during some interval (t_1, t_2) behaves as in Figure 4–2.

The holding cost incurred at any point in time is proportional to the inventory level at that point in time. In general, the total holding cost incurred from a time t_1 to a time t_2 is h multiplied by the area under the curve described by $I(t)$. The *average* inventory

FIGURE 4–2
Inventory as a function of time



level during the period (t_1, t_2) is the area under the curve divided by $t_2 - t_1$. For the cases considered in this chapter, simple geometry can be used to find the area under the inventory level curve. When $I(t)$ is described by a straight line, its average value is obvious. In cases such as that pictured in Figure 4–2, in which the curve of $I(t)$ is complex, the average inventory level would be determined by computing the integral of $I(t)$ over the interval (t_1, t_2) and dividing by $t_2 - t_1$.

Order Cost

The holding cost includes all those costs that are proportional to the amount of inventory on hand, whereas the **order cost** depends on the amount of inventory that is ordered or produced.

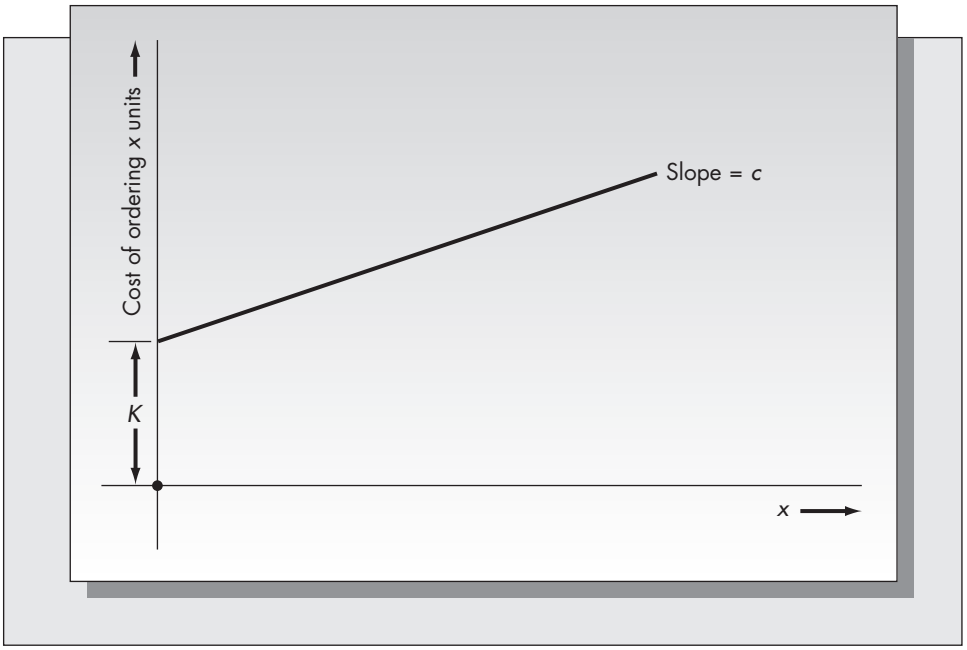
In most applications, the order cost has two components: a fixed and a variable component. The fixed cost, K, is incurred independent of the size of the order as long as it is not zero. The variable cost, c, is incurred on a per-unit basis. We also refer to K as the setup cost and c as the proportional order cost. Define $C(x)$ as the cost of ordering (or producing) x units. It follows that

$$C(x) = \begin{cases} 0 & \text{if } x = 0, \\ K + cx & \text{if } x > 0. \end{cases}$$

The order cost function is pictured in Figure 4–3.

When estimating the setup cost, one should include *only* those costs that are relevant to the current ordering decision. For example, the cost of maintaining the purchasing department of the company is *not* relevant to daily ordering decisions and should not be factored into the estimation of the setup cost. This is an overhead cost that is independent of the decision of whether or not an order should be placed. The appropriate costs comprising K would be the bookkeeping expense associated with the order, the fixed costs independent of the size of the order that might be required by the vendor, costs of order generation and receiving, and handling costs.

FIGURE 4–3
Order cost function



Penalty Cost

The **penalty cost,** also known as the shortage cost or the stock-out cost, is the cost of not having sufficient stock on hand to satisfy a demand *when it occurs*. This cost has a different interpretation depending on whether excess demand is back-ordered (orders that cannot be filled immediately are held on the books until the next shipment arrives) or lost (known as lost sales). In the back-order case, the penalty cost includes whatever bookkeeping and/or delay costs might be involved. In the lost-sales case, it includes the lost profit that would have been made from the sale. In either case, it would also include the *loss-of-goodwill* cost, which is a measure of customer satisfaction. Estimating the loss-of-goodwill component of the penalty cost can be very difficult in practice.

We use the symbol p to denote penalty cost and assume that p is charged on a per-unit basis. That is, each time a demand occurs that cannot be satisfied immediately, a cost p is incurred independent of how long it takes to eventually fill the demand. An alternative means of accounting for shortages is to charge the penalty cost on a per-unit-per-unit-time basis (as we did with the holding cost). This approach is appropriate if the time that a back order stays on the books is important, for example, if a back order results in stopping a production line because of the unavailability of a part. The models considered in this chapter assume that penalty costs are charged on a per-unit basis only. Penalty cost models are not considered in this chapter. Penalty costs are included in Chapter 5, but models incorporating a time-weighted penalty cost are not.

We present those inventory models that have had the greatest impact in the user community. Many of the techniques discussed in both this chapter and in Chapter 5 form the basis for commercial inventory control systems or in-house systems. In most cases, the models are simple enough that optimal operating policies can be calculated by hand, but they are often complex enough to capture the essential trade-offs in inventory management.

Problems for Sections 4.1–4.4

1. What are the two questions that inventory control addresses?
2. Discuss the cost penalties incurred by a firm that holds too much inventory and one that holds too little inventory.
3. ABC, Inc., produces a line of touring bicycles. Specifically, what are the four types of inventories (raw materials, components, work-in-process, and finished goods) that would arise in the production of this item?
4. I Carry rents trucks for moving and hauling. Each truck costs the company an average of $8,000, and the inventory of trucks varies monthly depending on the number that are rented out. During the first eight months of last year, I Carry had the following ending inventory of trucks on hand:

Month	Number of Trucks	Month	Number of Trucks
January	26	May	13
February	38	June	9
March	31	July	16
April	22	August	5

I Carry uses a 20 percent annual interest rate to represent the cost of capital. Yearly costs of storage amount to 3 percent of the value of each truck, and the cost of liability insurance is 2 percent.

   a. Determine the total handling cost incurred by I Carry during the period January to August. Assume for the purposes of your calculation that the holding cost incurred in a month is proportional to the inventory on hand at the end of the month.

   b. Assuming that these eight months are representative, estimate the average annual cost of holding trucks.

5. Stationery Supplies is considering installing an inventory control system in its store in Provo, Utah. The store carries about 1,400 different inventory items and has annual gross sales of about $80,000. The inventory control system would cost $12,500 to install and about $2,000 per year in additional supplies, time, and maintenance. If the savings to the store from the system can be represented as a fixed percentage of annual sales, what would that percentage have to be in order for the system to pay for itself in five years or less?
6. For Stationery Supplies, discussed in Problem 5, list and discuss all the uncertainties that would motivate the store to maintain inventories of its 1,400 items.
7. Stationery Supplies orders plastic erasers from a company in Nürnberg, Germany. It takes six weeks to ship the erasers from Germany to Utah. Stationery Supplies maintains a standing order of 200 erasers every six months (shipped on the first of January and the first of July).

   a. Assuming the ordering policy the store is using does not result in large buildups of inventory or long-term stock-outs, what is the annual demand for erasers?

   b. Draw a graph of the pipeline inventory (that is, the inventory ordered but not received) of the erasers during one year. What is the average pipeline inventory of erasers during the year?

c. Express the replenishment lead time in years and multiply the annual demand you obtained in part (*a*) by the lead time. What do you notice about the result that you obtain?

8. Penalty costs can be assessed only against the number of units of demand that cannot be satisfied, or against the number of units weighted by the amount of time that an order stays on the books. Consider the following history of supply and demand transactions for a particular part:

Month	Number of Items Received	Demand during Month
January	200	520
February	175	1,640
March	750	670
April	950	425
May	500	280
June	2,050	550

Assume that starting inventory at the beginning of January is 480 units.

a. Determine the ending inventory each month. Assume that excess demands are back-ordered.

b. Assume that each time a unit is demanded that cannot be supplied immediately, a one-time charge of $10 is made. Determine the stock-out cost incurred during the six months (1) if excess demand at the end of each month is lost, and (2) if excess demand at the end of each month is back-ordered.

c. Suppose that each stock-out costs $10 per unit per month that the demand remains unfilled. If demands are filled on a first-come, first-served basis, what is the total stock-out cost incurred during the six months using this type of cost criterion? (Assume that the demand occurs at the beginning of the month for purposes of your calculation.) Notice that you must assume that excess demands are back-ordered for this case to make any sense.

d. Discuss under what circumstances the cost criterion used in part (*b*) might be appropriate and under what circumstances the cost criterion used in part (*c*) might be appropriate.

9. HAL Ltd. produces a line of high-capacity disk drives for mainframe computers. The housings for the drives are produced in Hamilton, Ontario, and shipped to the main plant in Toronto. HAL uses the drive housings at a fairly steady rate of 720 per year. Suppose that the housings are shipped in trucks that can hold 40 housings at one time. It is estimated that the fixed cost of loading the housings onto the truck and unloading them on the other end is $300 for shipments of 120 or fewer housings (i.e., three or fewer truckloads). Each trip made by a single truck costs the company $160 in driver time, gasoline, oil, insurance, and wear and tear on the truck.

a. Compute the annual costs of transportation and loading and unloading the housings for the following policies: (1) shipping one truck per week, (2) shipping one full truckload as often as needed, and (3) shipping three full truckloads as often as needed.

b. For what reasons might the policy in part (*a*) with the highest annual cost be more desirable from a systems point of view than the policy having the lowest annual cost?

4.5 THE EOQ MODEL

The **EOQ model** (or *economic order quantity model*) is the simplest and most fundamental of all inventory models. It describes the important trade-off between fixed order costs and holding costs, and is the basis for the analysis of more complex systems.

The Basic Model

Assumptions:

1. The demand rate is known and is a constant λ units per unit time. (The unit of time may be days, weeks, months, etc. In what follows we assume that the default unit of time is a year. However, the analysis is valid for other time units as long as all relevant variables are expressed in the same units.)
2. Shortages are not permitted.
3. There is no order lead time. (This assumption will be relaxed.)
4. The costs include
   a. Setup cost at K per positive order placed.
   b. Proportional order cost at c per unit ordered.
   c. Holding cost at h per unit held per unit time.

Assume with no loss in generality that the on-hand inventory at time zero is zero. Shortages are not allowed, so we must place an order at time zero. Let Q be the size of the order. It follows that the on-hand inventory level increases instantaneously from zero to Q at time $t = 0$.

Consider the next time an order is to be placed. At this time, either the inventory is positive or it is again zero. A little reflection shows that we can reduce the holding costs by waiting until the inventory level drops to zero before ordering again. At the instant that on-hand inventory equals zero, the situation looks exactly the same as it did at time $t = 0$. If it was optimal to place an order for Q units at that time, then it is still optimal to order Q units. It follows that the function that describes the changes in stock levels over time is the familiar sawtooth pattern of Figure 4–4.

The objective is to choose Q to minimize the average cost per unit time. Unless otherwise stated, we will assume that a unit of time is a year, so that we minimize the average annual cost. Other units of time, such as days, weeks, or months, are also acceptable, as long as all time-related variables are expressed in the same units. One might think that the appropriate optimization criterion would be to minimize the *total* cost in a cycle. However, this ignores the fact that the cycle length itself is a function of Q and must be explicitly included in the formulation.

Next, we derive an expression for the average annual cost as a function of the lot size Q. In each cycle, the total fixed plus proportional order cost is $C(Q) = K + cQ$. In order to obtain the order cost per unit time, we divide by the cycle length T. As Q units are consumed each cycle at a rate λ, it follows that $T = Q/\lambda$. This result also can be obtained by noting that the slope of the inventory curve pictured in Figure 4–4, $-\lambda$, equals the ratio $-Q/T$.

Consider the holding cost. Because the inventory level decreases linearly from Q to 0 each cycle, the average inventory level during one order cycle is $Q/2$. Because all cycles are identical, the average inventory level over a time horizon composed

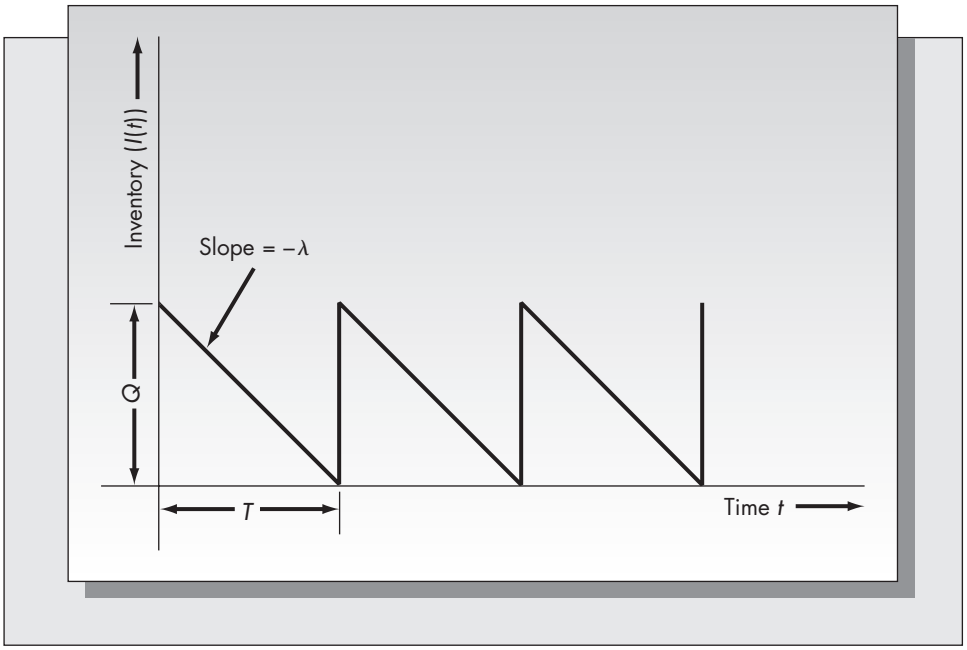


FIGURE 4–4
Inventory levels for the EOQ model

of many cycles is also $Q/2$. It follows that the average annual cost, say $G(Q)$, is given by

$$G(Q) = \frac{K + cQ}{T} + \frac{hQ}{2} = \frac{K + cQ}{Q/\lambda} + \frac{hQ}{2}$$
$$= \frac{K\lambda}{Q} + \lambda c + \frac{hQ}{2}.$$

The three terms composing $G(Q)$ are annual setup cost, annual purchase cost, and annual holding cost, respectively.

We now wish to find Q to minimize $G(Q)$. Consider the shape of the curve $G(Q)$. We have that

$$G'(Q) = -K\lambda/Q^2 + h/2$$

and

$$G''(Q) = 2K\lambda/Q^3 > 0 \qquad \text{for } Q > 0.$$

Since $G''(Q) > 0$, it follows that $G(Q)$ is a convex function of Q. Furthermore, since $G'(0) = -\infty$ and $G'(\infty) = h/2$, it follows that $G(Q)$ behaves as pictured in Figure 4–5.

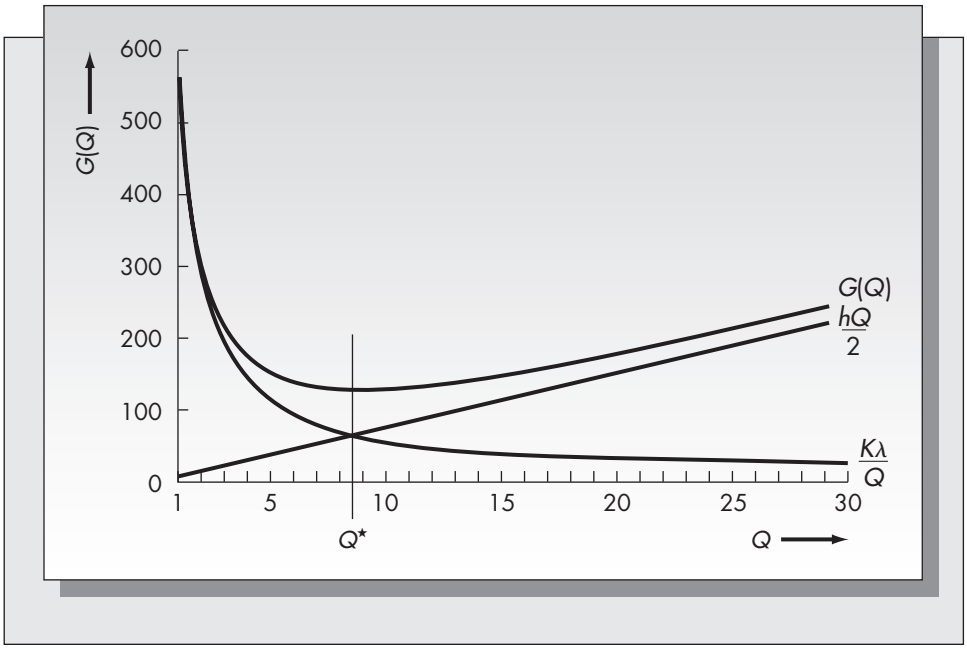
The optimal value of Q occurs where $G'(Q) = 0$. This is true when $Q^2 = 2K\lambda/h$, which gives

$$Q^* = \sqrt{\frac{2K\lambda}{h}}.$$

Q^* is known as the economic order quantity (EOQ). There are a number of interesting points to note:

1. In Figure 4–5, the curves corresponding to the fixed order cost component $K\lambda/Q$ and the holding cost component $hQ/2$ also are included. Notice that Q^* is the value of

FIGURE 4–5
The average annual cost function $G(Q)$



Q where the two curves intersect. (If you equate $hQ/2$ and $K\lambda/Q$ and solve for Q, you will obtain the EOQ formula.) In general, the minimum of the sum of two functions will *not* occur at the intersection of the two functions. It is an interesting coincidence that it does in this case.

2. Notice that the proportional order cost component, c, does not appear explicitly in the expression for Q^*. This is because the term λc appearing in the definition of $G(Q)$ is independent of Q. As all feasible policies must replenish inventory at the rate of demand, the proportional order cost incurred per unit time is λc independent of Q. Because λc is a constant, we generally ignore it when computing average costs. Notice that c *does* affect the value of Q^* indirectly, as h appears in the EOQ formula and $h = Ic$.

Example *4.1*

Number 2 pencils at the campus bookstore are sold at a fairly steady rate of 60 per week. The pencils cost the bookstore 2 cents each and sell for 15 cents each. It costs the bookstore $12 to initiate an order, and holding costs are based on an annual interest rate of 25 percent. Determine the optimal number of pencils for the bookstore to purchase and the time between placement of orders. What are the yearly holding and setup costs for this item?

Solution

First, we convert the demand to a yearly rate so that it is consistent with the interest charge, which is given on an annual basis. (Alternatively, we could have converted the annual interest rate to a weekly interest rate.) The annual demand rate is $\lambda = (60)(52) = 3{,}120$. The holding cost h is the product of the annual interest rate and the variable cost of the item. Hence, $h = (0.25)(0.02) = 0.005$. Substituting into the EOQ formula, we obtain

$$Q^* = \sqrt{\frac{2K\lambda}{h}} = \sqrt{\frac{(2)(12)(3{,}120)}{0.005}} = 3{,}870.$$

The cycle time is $T = Q/\lambda = 3{,}870/3{,}120 = 1.24$ years. The average annual holding cost is $h(Q/2) = 0.005(3{,}870/2) = \$9.675$. The average annual setup cost is $K\lambda/Q$, which is also $9.675.

Example 4.1 illustrates some of the problems that can arise when using simple models. The optimal solution calls for ordering almost 4,000 pencils every 15 months. Even though this value of Q minimizes the yearly holding and setup costs, it could be infeasible: the store may not have the space to store 4,000 pencils. Simple models cannot account for all the constraints present in a real problem. For that reason, every solution must be considered in context and modified, if necessary, to fit the application.

Notice also that the optimal solution did not depend on the selling price of 15 cents. Even if each pencil sold for $2, we would recommend the same order quantity, because the pencils are assumed to sell at a rate of 60 per week no matter what their price. This is, of course, a simplification of reality. It is reasonable to assume that the demand is relatively stable for a range of prices. Inventory models explicitly incorporate selling price in the formulation only when the selling price is included as part of the optimization.

Inclusion of Order Lead Time

One of the assumptions made in our derivation of the EOQ model was that there was no order lead time. We now relax that assumption. Suppose in Example 4.1 that the pencils had to be ordered four months in advance. If we were to place the order exactly four months before the end of the cycle, the order would arrive at exactly the same point in time as in the zero lead time case. The optimal timing of order placement for Example 4.1 is shown in Figure 4–6.

Rather than say that an order should be placed so far in advance of the end of a cycle, it is more convenient to indicate reordering in terms of the on-hand inventory. Define R, the reorder point, as the level of on-hand inventory at the instant an order should be placed. From Figure 4–6, we see that R is the product of the lead time and the demand rate ($R = \lambda\tau$). For the example, $R = (3{,}120)(0.3333) = 1{,}040$. Notice that we converted the lead time to years before multiplying. *Always express all relevant variables in the same units of time.*

Determining the reorder point is more difficult when the lead time exceeds a cycle. Consider an item with an EOQ of 25, a demand rate of 500 units per year, and a lead time of six weeks. The cycle time is $T = 25/500 = 0.05$ year, or 2.6 weeks. Forming

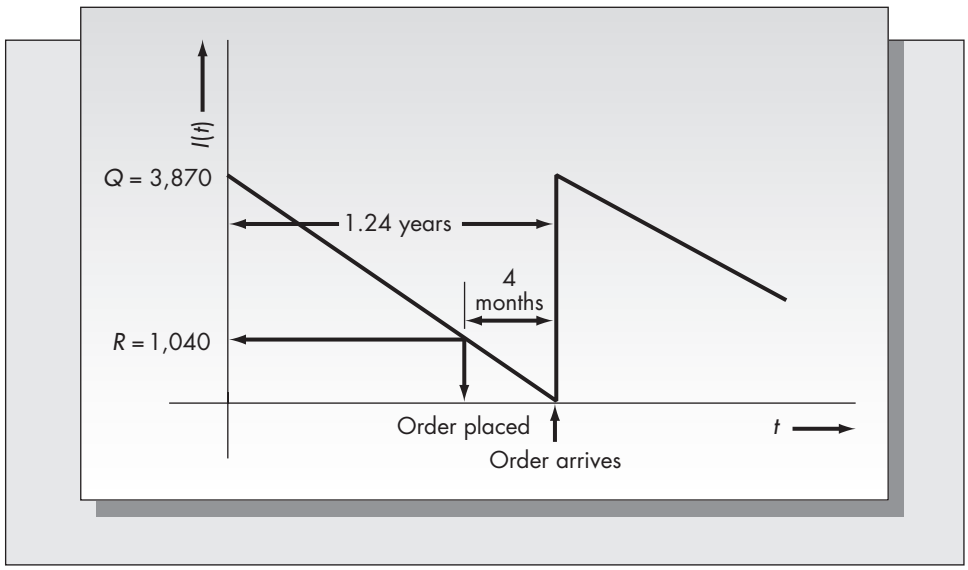


FIGURE 4–6
Reorder point calculation for Example 4.1

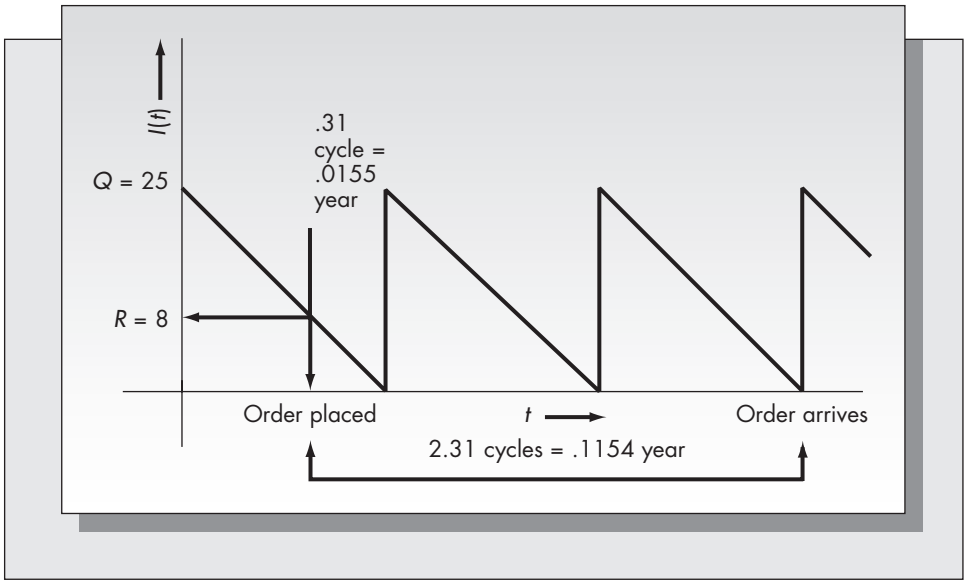


FIGURE 4–7
Reorder point calculation for lead times exceeding one cycle

the ratio of τ/T, we obtain 2.31. This means that there are exactly 2.31 cycles in the lead time. Every order must be placed 2.31 cycles in advance (see Figure 4–7).

Notice that for the purpose of computing the reorder point, this is exactly the same as placing the order *0.31 cycle in advance*. This is true because the level of on-hand inventory is the same whether we are at a point 2.31 or 0.31 cycle before the arrival of an order. In this case, 0.31 cycle is 0.0155 year, thus giving a reorder point of $R = (0.0155)(500) = 7.75 \approx 8$. In general, when $\tau > T$, use the following procedure:

a. Form the ratio τ/T.
b. Consider only the fractional remainder of the ratio. Multiply this fractional remainder by the cycle length to convert back to years.
c. Multiply the result of step (*b*) by the demand rate to obtain the reorder point.

Sensitivity

In this part we examine the issue of how sensitive the annual cost function is to errors in the calculation of Q. Consider Example 4.1. Suppose that the bookstore orders pencils in batches of 1,000, rather than 3,870 as the optimal solution indicates. What additional cost is it incurring by using a suboptimal solution? To answer the question, we consider the average annual cost function $G(Q)$. By substituting $Q = 1{,}000$, we can find the average annual cost for this lot size and compare it to the optimal cost to determine the magnitude of the penalty. We have

$$G(Q) = K\lambda/Q + hQ/2 = (12)(3{,}120)/1{,}000 + (0.005)(1{,}000)/2 = \$39.94,$$

which is considerably larger than the optimal cost of $19.35.

One can find the cost penalty for suboptimal solutions in this manner for any particular problem. However, it is more instructive and more convenient to obtain a universal solution to the sensitivity problem. We do so by deriving an expression for the ratio of the suboptimal cost over the optimal cost as a function of the ratio of the optimal and suboptimal order quantities. Let G^* be the average annual holding and

setup cost at the optimal solution. Then

$$G^* = K\lambda/Q^* + hQ^*/2 = \frac{K\lambda}{\sqrt{2K\lambda/h}} + \frac{h}{2}\sqrt{\frac{2K\lambda}{h}} = 2\sqrt{\frac{K\lambda h}{2}}$$
$$= \sqrt{2K\lambda h}.$$

It follows that for any Q,

$$\frac{G(Q)}{G^*} = \frac{K\lambda/Q + hQ/2}{\sqrt{2K\lambda h}}$$
$$= \frac{1}{2Q}\sqrt{\frac{2K\lambda}{h}} + \frac{Q}{2}\sqrt{\frac{h}{2K\lambda}}$$
$$= \frac{Q^*}{2Q} + \frac{Q}{2Q^*}$$
$$= \frac{1}{2}\left[\frac{Q^*}{Q} + \frac{Q}{Q^*}\right].$$

To see how one would use this result, consider using a suboptimal lot size in Example 4.1. The optimal solution was $Q^* = 3{,}870$, and we wished to evaluate the cost error of using $Q = 1{,}000$. Forming the ratio Q^*/Q gives 3.87. Hence, $G(Q)/G^* = (0.5)(3.87 + 1/3.87) = (0.5)(4.128) = 2.06$. This says that the average annual holding and setup cost with $Q = 1{,}000$ is 2.06 times the optimal average annual holding and setup cost.

In general, the cost function $G(Q)$ is relatively insensitive to errors in Q. For example, if Q is twice as large as it should be, Q/Q^* is 2 and G/G^* is 1.25. Hence, an error of 100 percent in the value of Q results in an error of only 25 percent in the annual holding and setup cost. Notice that you obtain the same result if Q is half Q^*, since $Q^*/Q = 2$. However, this does *not* imply that the average annual cost function is symmetric. In fact, suppose that the order quantity differed from the optimal by ΔQ units. A value of $Q = Q^* + \Delta Q$ would result in a *lower* average annual cost than a value of $Q = Q^* - \Delta Q$.

EOQ and JIT

Largely as a result of the success of Toyota's kanban system, a new philosophy is emerging about the role and importance of inventories in manufacturing environments. This philosophy, known as *just-in-time* (JIT), says that excess work-in-process inventories are not desirable, and inventories should be reduced to the bare essentials. (We discuss the just-in-time philosophy in more detail in Chapter 1 and the mechanics of kanban in Chapter 8.) EOQ is the result of traditional thinking about inventories and scale economies in economics. Are the EOQ and JIT approaches at odds with each other?

Proponents argue that an essential part of implementing JIT is reducing setup times, and hence setup costs. As setup costs decrease, traditional EOQ theory says that lot sizes should be reduced. In this sense, the two ways of thinking are compatible. However, there are times when they may not be. We believe that there is substantial value to the JIT approach that may not be incorporated easily into a mathematical model. Quality problems can be identified and rectified before inventories of defective parts accumulate. Plants can be more flexible if they are not burdened with excess in-process inventories. Certainly Toyota's success with the JIT approach, as evidenced by substantially lower inventory costs per car than are typical for U.S. auto manufacturers, is a testament to the value of JIT.

However, we believe that every new approach must be incorporated carefully into the firm's business and not adopted blindly without evaluating its consequences and appropriateness. JIT, although an important development in material management, is not always the best approach. The principles underlying EOQ (and MRP, discussed in Chapter 8) are sound and should not be ignored. The following example illustrates this point.

Example *4.2*

The Rahway, New Jersey, plant of Metalcase, a manufacturer of office furniture, produces metal desks at a rate of 200 per month. Each desk requires 40 Phillips head metal screws purchased from a supplier in North Carolina. The screws cost 3 cents each. Fixed delivery charges and costs of receiving and storing shipments of the screws amount to about $100 per shipment, independently of the size of the shipment. The firm uses a 25 percent interest rate to determine holding costs. Metalcase would like to establish a standing order with the supplier and is considering several alternatives. What standing order size should they use?

Solution

First we compute the EOQ. The annual demand for screws is

$$(200)(12)(40) = 96{,}000.$$

The annual holding cost per screw is $(0.25)(0.03) = 0.0075$. From the EOQ formula, the optimal lot size is

$$Q^* = \sqrt{\frac{(2)(100)(96{,}000)}{0.0075}} = 50{,}597.$$

Note that the cycle time is $T = Q/\lambda = 50{,}597/96{,}000 = 0.53$ year or about once every six months. Hence the optimal policy calls for replenishment of the screws about twice a year. A JIT approach would be to order the screws as frequently as possible to minimize the inventory held at the plant. Implementing such an approach might suggest a policy of weekly deliveries. Such a policy makes little sense in this context, however. This policy would require 52 deliveries per year, incurring setup costs of $5,200 annually. The EOQ solution gives a total annual cost of both setups and holding of less than $400. For a low-value item such as this with high fixed order costs, small lot sizes in accordance with JIT are inappropriate. The point is that no single approach should be blindly adopted for all situations. The success of a method in one context does not ensure its appropriateness in all other contexts.

Problems for Section 4.5

10. A specialty coffeehouse sells Colombian coffee at a fairly steady rate of 280 pounds annually. The beans are purchased from a local supplier for $2.40 per pound. The coffeehouse estimates that it costs $45 in paperwork and labor to place an order for the coffee, and holding costs are based on a 20 percent annual interest rate.
    a. Determine the optimal order quantity for Colombian coffee.
    b. What is the time between placement of orders?
    c. What is the average annual cost of holding and setup due to this item?
    d. If replenishment lead time is three weeks, determine the reorder level based on the on-hand inventory.
11. For the situation described in Problem 10, draw a graph of the amount of inventory on order. Using your graph, determine the average amount of inventory on order. Also compute the demand during the replenishment lead time. How do these two quantities differ?

12. A large automobile repair shop installs about 1,250 mufflers per year, 18 percent of which are for imported cars. All the imported-car mufflers are purchased from a single local supplier at a cost of $18.50 each. The shop uses a holding cost based on a 25 percent annual interest rate. The setup cost for placing an order is estimated to be $28.
    a. Determine the optimal number of imported-car mufflers the shop should purchase each time an order is placed, and the time between placement of orders.
    b. If the replenishment lead time is six weeks, what is the reorder point based on the level of on-hand inventory?
    c. The current reorder policy is to buy imported-car mufflers only once a year. What are the additional holding and setup costs incurred by this policy?
13. Consider the coffeehouse discussed in Problem 10. Suppose that its setup cost for ordering was really only $15. Determine the error made in calculating the annual cost of holding and setup incurred as a result of its using the wrong value of K. (Note that this implies that its current order policy is suboptimal.)
14. A local machine shop buys hex nuts and molly screws from the same supplier. The hex nuts cost 15 cents each and the molly screws cost 38 cents each. A setup cost of $100 is assumed for all orders. This includes the cost of tracking and receiving the orders. Holding costs are based on a 25 percent annual interest rate. The shop uses an average of 20,000 hex nuts and 14,000 molly screws annually.
    a. Determine the optimal size of the orders of hex nuts and molly screws, and the optimal time between placement of orders of these two items.
    b. If both items are ordered and received simultaneously, the setup cost of $100 applies to the combined order. Compare the average annual cost of holding and setup if these items are ordered separately; if they are both ordered when the hex nuts would normally be ordered; and if they are both ordered when the molly screws would normally be ordered.
15. David's Delicatessen flies in Hebrew National salamis regularly to satisfy a growing demand for the salamis in Silicon Valley. The owner, David Gold, estimates that the demand for the salamis is pretty steady at 175 per month. The salamis cost Gold $1.85 each. The fixed cost of calling his brother in New York and having the salamis flown in is $200. It takes three weeks to receive an order. Gold's accountant, Irving Wu, recommends an annual cost of capital of 22 percent, a cost of shelf space of 3 percent of the value of the item, and a cost of 2 percent of the value for taxes and insurance.
    a. How many salamis should Gold have flown in and how often should he order them?
    b. How many salamis should Gold have on hand when he phones his brother to send another shipment?
    c. Suppose that the salamis sell for $3 each. Are these salamis a profitable item for Gold? If so, what annual profit can he expect to realize from this item? (Assume that he operates the system optimally.)
    d. If the salamis have a shelf life of only 4 weeks, what is the trouble with the policy that you derived in part (*a*)? What policy would Gold have to use in that case? Is the item still profitable?
16. In view of the results derived in the section on sensitivity analysis, discuss the following quotation of an inventory control manager: "If my lot sizes are going to be off the mark, I'd rather miss on the high side than on the low side."

4.6 EXTENSION TO A FINITE PRODUCTION RATE

An implicit assumption of the simple EOQ model is that the items are obtained from an outside supplier. When that is the case, it is reasonable to assume that the entire lot is delivered at the same time. However, if we wish to use the EOQ formula when the units are produced internally, then we are effectively assuming that the production rate is infinite. When the production rate is much larger than the demand rate, this assumption is probably satisfactory as an approximation. However, if the rate of production is comparable to the rate of demand, the simple EOQ formula will lead to incorrect results.

Assume that items are produced at a rate P during a production run. We require that $P > \lambda$ for feasibility. All other assumptions will be identical to those made in the derivation of the simple EOQ. When units are produced internally, the curve describing inventory levels as a function of time is slightly different from the sawtooth pattern of Figure 4–4. The change in the inventory level over time for the finite production rate case is shown in Figure 4–8.

Let Q be the size of each production run. Let T, the cycle length, be the time between successive production startups. Write $T = T_1 + T_2$, where T_1 is uptime (production time) and T_2 is downtime. Note that the maximum level of on-hand inventory during a cycle is *not* Q.

The number of units consumed each cycle is λT, which must be the same as the number of units produced each cycle, which is simply Q. It follows that $Q = \lambda T$, or $T = Q/\lambda$. Define H as the maximum level of on-hand inventory. As items are produced at a rate P for a time T_1, it follows that $Q = PT_1$, or $T_1 = Q/P$. From Figure 4–8 we see that $H/T_1 = P - \lambda$. This follows from the definition of the slope as the rise over the run. Substituting $T_1 = Q/P$ and solving for H gives $H = Q(1 - \lambda/P)$.

We now determine an expression for the average annual cost function. Because the average inventory level is $H/2$, it follows that

$$G(Q) = \frac{K}{T} + \frac{hH}{2} = \frac{K\lambda}{Q} + \frac{hQ}{2}(1 - \lambda/P).$$

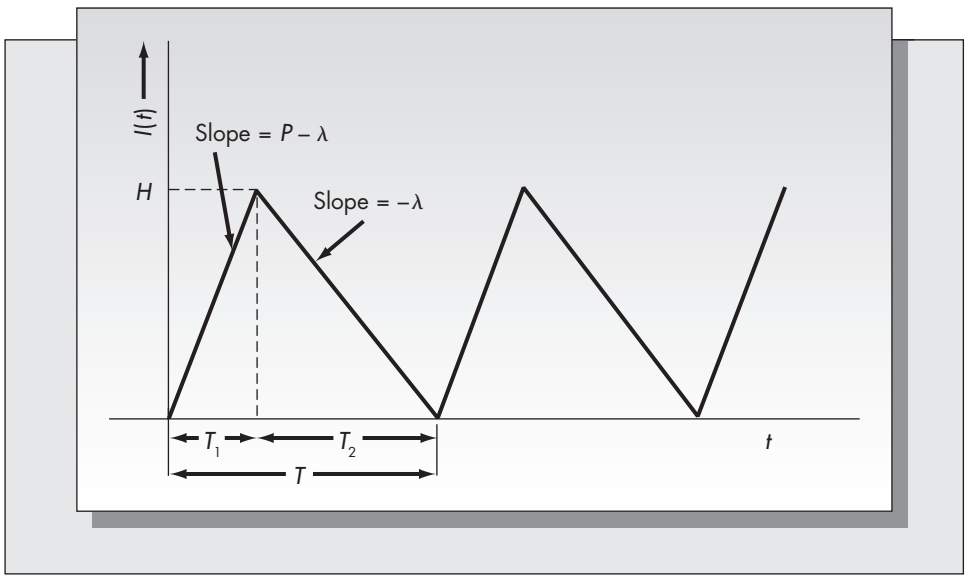


FIGURE 4–8
Inventory levels for finite production rate model

Notice that if we define $h' = h(1 - \lambda/P)$, then this $G(Q)$ is identical to that of the infinite production rate case with h' substituted for h. It follows that

$$Q^* = \sqrt{\frac{2K\lambda}{h'}}.$$

Example *4.3* A local company produces an erasable programmable read-only memory (EPROM) for several industrial clients. It has experienced a relatively flat demand of 2,500 units per year for the product. The EPROM is produced at a rate of 10,000 units per year. The accounting department has estimated that it costs \$50 to initiate a production run, each unit costs the company \$2 to manufacture, and the cost of holding is based on a 30 percent annual interest rate. Determine the optimal size of a production run, the length of each production run, and the average annual cost of holding and setup. What is the maximum level of the on-hand inventory of the EPROMs?

Solution First, we compute $h = (0.3)(2) = 0.6$ per unit per year. The modified holding cost is $h' = h(1 - \lambda/P) = (0.6)(1 - 2{,}500/10{,}000) = 0.45$. Substituting into the EOQ formula and using h' for h, we obtain $Q^* = 745$. Note that the simple EOQ equals 645, all out 14 percent less.

The time between production runs is $T = Q/\lambda = 745/2{,}500 = 0.298$ year. The uptime each cycle is $T_1 = Q/P = 745/10{,}000 = 0.0745$ year, and the downtime each cycle is $T_2 = T - T_1 = 0.2235$ year.

The average annual cost of holding and setup is

$$G(Q^*) = \frac{K\lambda}{Q^*} + \frac{h'Q^*}{2} = \frac{(50)(2{,}500)}{745} + \frac{(0.45)(745)}{2} = 335.41.$$

The maximum level of on-hand inventory is $H = Q^*(1 - \lambda/P) = 559$ units.

Problems for Section 4.6

17. The Wod Chemical Company produces a chemical compound that is used as a lawn fertilizer. The compound can be produced at a rate of 10,000 pounds per day. Annual demand for the compound is 0.6 million pounds per year. The fixed cost of setting up for a production run of the chemical is \$1,500, and the variable cost of production is \$3.50 per pound. The company uses an interest rate of 22 percent to account for the cost of capital, and the costs of storage and handling of the chemical amount to 12 percent of the value. Assume that there are 250 working days in a year.
    a. What is the optimal size of the production run for this particular compound?
    b. What proportion of each production cycle consists of uptime and what proportion consists of downtime?
    c. What is the average annual cost of holding and setup attributed to this item? If the compound sells for \$3.90 per pound, what is the annual profit the company is realizing from this item?
18. Determine the batch size that would result in Problem 17 if you assumed that the production rate was infinite. What is the additional average annual cost that would be incurred using this batch size rather than the one you found in Problem 17?
19. HAL Ltd., discussed in Problem 9, can produce the disk drive housings in the Hamilton, Ontario, plant at a rate of 150 housings per month. The housings cost HAL \$85 each to produce, and the setup cost for beginning a production run is \$700. Assume an annual interest rate of 28 percent for determining the holding cost.
    a. What is the optimal number of housings for HAL to produce in each production run?

b. Find the time between initiation of production runs, the time devoted to production, and the downtime each production cycle.

c. What is the maximum dollar investment in housings that HAL has at any point in time?

20. Filter Systems produces air filters for domestic and foreign cars. One filter, part number JJ39877, is supplied on an exclusive contract basis to Oil Changers at a constant 200 units monthly. Filter Systems can produce this filter at a rate of 50 per hour. Setup time to change the settings on the equipment is 1.5 hours. Worker time (including overhead) is charged at the rate of $55 per hour, and plant idle time during setups is estimated to cost the firm $100 per hour in lost profit.

Filter Systems has established a 22 percent annual interest charge for determining holding cost. Each filter costs the company $2.50 to produce; they are sold for $5.50 each to Oil Changers. Assume 6-hour days, 20 working days per month, and 12 months per year for your calculations.

a. How many JJ39877 filters should Filter Systems produce in each production run of this particular part to minimize annual holding and setup costs?

b. Assuming that it produces the optimal number of filters in each run, what is the maximum level of on-hand inventory of these filters that the firm has at any point in time?

c. What percentage of the working time does the company produce these particular filters, assuming that the policy in part (*a*) is used?

4.7 QUANTITY DISCOUNT MODELS

We have assumed up until this point that the cost c of each unit is independent of the size of the order. Often, however, the supplier is willing to charge less per unit for larger orders. The purpose of the discount is to encourage the customer to buy the product in larger batches. Such quantity discounts are common for many consumer goods.

Although many different types of discount schedules exist, there are two that seem to be the most popular: all-units and incremental. In each case we assume that there are one or more breakpoints defining changes in the unit cost. However, there are two possibilities: either the discount is applied to all the units in an order (all-units), or it is applied only to the additional units beyond the breakpoint (incremental). The all-units case is more common.

Example *4.4* The Weighty Trash Bag Company has the following price schedule for its large trash can liners. For orders of less than 500 bags, the company charges 30 cents per bag; for orders of 500 or more but fewer than 1,000 bags, it charges 29 cents per bag; and for orders of 1,000 or more, it charges 28 cents per bag. In this case the breakpoints occur at 500 and 1,000. The discount schedule is all-units because the discount is applied to all of the units in an order. The order cost function $C(Q)$ is defined as

$$C(Q) = \begin{cases} 0.30Q & \text{for} \quad 0 \le Q < 500, \\ 0.29Q & \text{for} \quad 500 \le Q < 1{,}000, \\ 0.28Q & \text{for} \quad 1{,}000 \le Q \end{cases}$$

The function $C(Q)$ is pictured in Figure 4–9. In Figure 4–10, we consider the same breakpoints, but assume an incremental quantity discount schedule.

Note that the average cost per unit with an all-units schedule will be less than the average cost per unit with the corresponding incremental schedule.

FIGURE 4–9
All-units discount order cost function

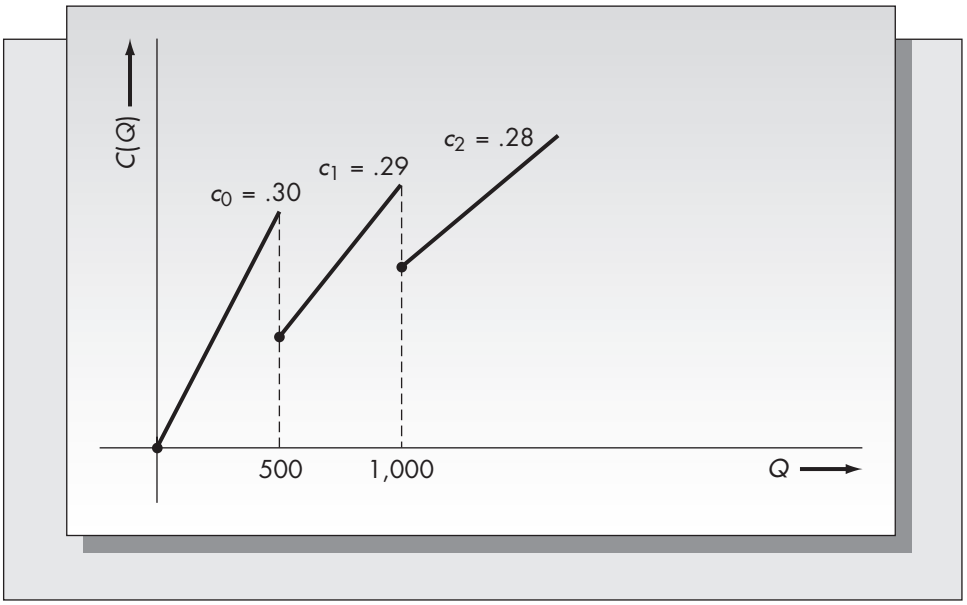
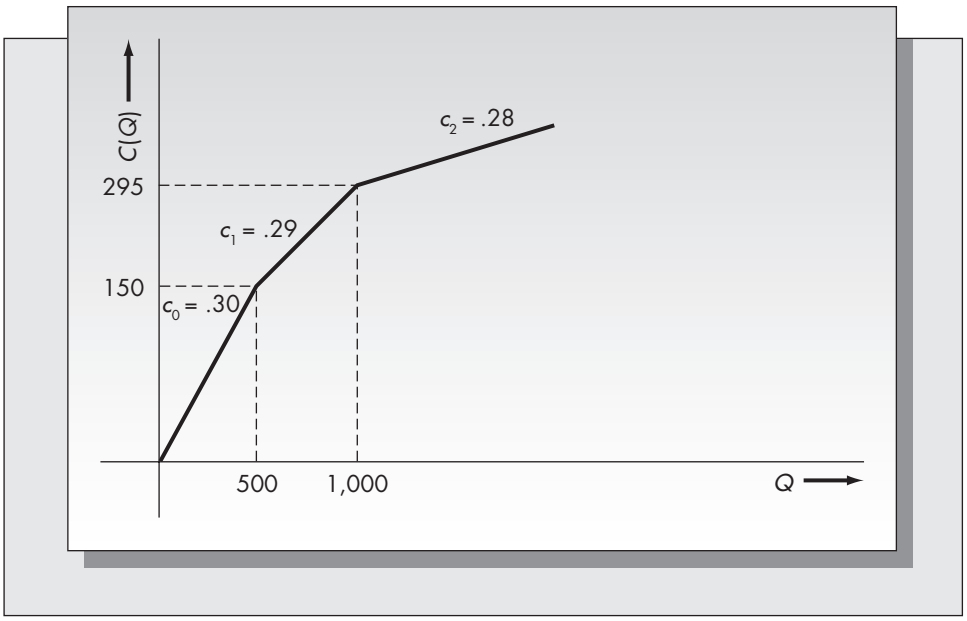


FIGURE 4–10
Incremental discount order cost function



The all-units schedule appears irrational in some respects. In Example 4.4, 499 bags would cost \$149.70, whereas 500 bags would cost only \$145.00. Why would Weighty actually charge less for a larger order? One reason would be to provide an incentive for the purchaser to buy more. If you were considering buying 400 bags, you might choose to move up to the breakpoint to obtain the discount. Furthermore, it is possible that Weighty has stored its bags in lots of 100, so that its savings in handling costs might more than compensate for the lower total cost.

Optimal Policy for All-Units Discount Schedule

We will illustrate the solution technique using Example 4.4. Assume that the company considering what standing order to place with Weighty uses trash bags at a fairly constant

rate of 600 per year. The accounting department estimates that the fixed cost of placing an order is \$8, and holding costs are based on a 20 percent annual interest rate. From Example 4.4, $c_0 = 0.30$, $c_1 = 0.29$, and $c_2 = 0.28$ are the respective unit costs.

The first step toward finding a solution is to compute the EOQ values corresponding to each of the unit costs, which we will label $Q^{(0)}$, $Q^{(1)}$, and $Q^{(2)}$, respectively.

$$Q^{(0)} = \sqrt{\frac{2K\lambda}{Ic_0}} = \sqrt{\frac{(2)(8)(600)}{(0.2)(0.30)}} = 400,$$

$$Q^{(1)} = \sqrt{\frac{2K\lambda}{Ic_1}} = \sqrt{\frac{(2)(8)(600)}{(0.2)(0.29)}} = 406,$$

$$Q^{(2)} = \sqrt{\frac{2K\lambda}{Ic_2}} = \sqrt{\frac{(2)(8)(600)}{(0.2)(0.28)}} = 414.$$

We say that the EOQ value is realizable if it falls within the interval that corresponds to the unit cost used to compute it. Since $0 \leq 400 < 500$, Q^0 is realizable. However, neither $Q^{(1)}$ nor $Q^{(2)}$ is realizable ($Q^{(1)}$ would have to have been between 500 and 1,000, and $Q^{(2)}$ would have to have been 1,000 or more). Each EOQ value corresponds to the minimum of a different annual cost curve. In this example, if $Q^{(2)}$ were realizable, it would necessarily have to have been the optimal solution, as it corresponds to the lowest point on the lowest curve. The three average annual cost curves for this example appear in Figure 4–11. Because each curve is valid only for certain values of Q, the average annual cost function is given by the discontinuous curve shown in heavy shading. The goal of the analysis is to find the minimum of this discontinuous curve.

There are three candidates for the optimal solution: 400, 500, and 1,000. In general, the optimal solution will be either the largest realizable EOQ or one of the breakpoints that exceeds it. The optimal solution is the lot size with the lowest average annual cost.

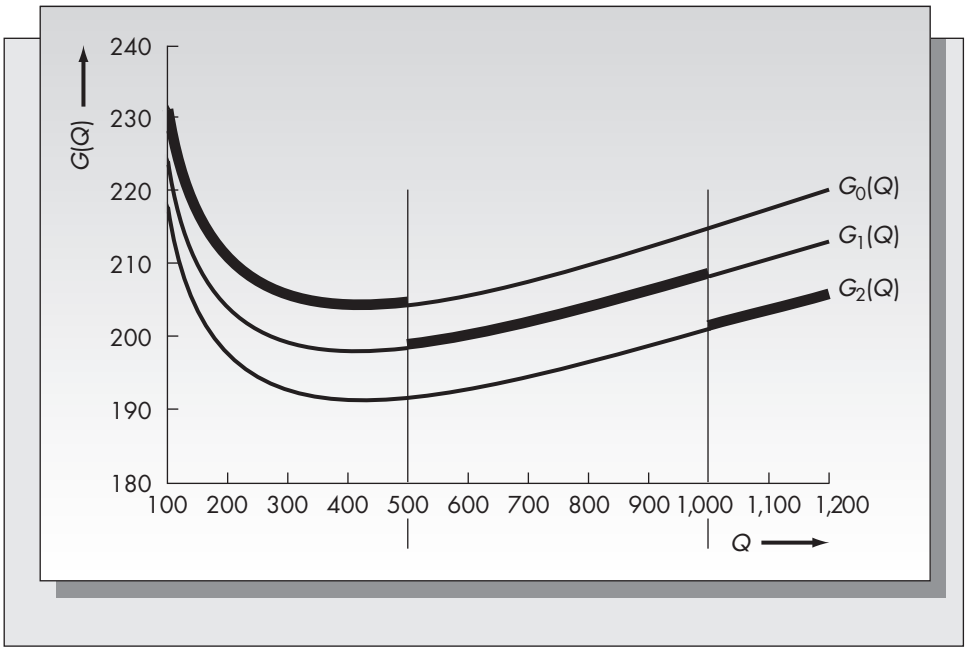


FIGURE 4–11
All-units discount average annual cost function

The average annual cost functions are given by

$$G_j(Q) = \lambda c_j + \lambda K/Q + Ic_j\, Q/2 \qquad \text{for } j = 0, 1, \text{ and } 2.$$

The broken curve pictured in Figure 4–11, $G(Q)$, is defined as

$$G(Q) = \begin{cases} G_0(Q) & \text{for} \quad 0 \le Q < 500, \\ G_1(Q) & \text{for} \quad 500 \le Q < 1{,}000, \\ G_2(Q) & \text{for} \quad 1{,}000 \le Q \end{cases}$$

Substituting Q equals 400, 500, and 1,000, and using the appropriate values of c_j, we obtain

$$\begin{aligned} G(400) &= G_0(400) \\ &= (600)(0.30) + (600)(8)/400 + (0.2)(0.30)(400)/2 = \$204.00 \\ G(500) &= G_1(500) \\ &= (600)(0.29) + (600)(8)/500 + (0.2)(0.29)(500)/2 = \$198.10 \\ G(1{,}000) &= G_2(1{,}000) \\ &= (600)(0.28) + (600)(8)/1{,}000 + (0.2)(0.28)(1{,}000)/2 = \$200.80. \end{aligned}$$

Hence, we conclude that the optimal solution is to place a standing order for 500 units with Weighty at an average annual cost of \$198.10.

Summary of the Solution Technique for All-Units Discounts

1. Determine the largest realizable EOQ value. The most efficient way to do this is to compute the EOQ for the lowest price first, and continue with the next higher price. Stop when the first EOQ value is realizable (that is, within the correct interval).
2. Compare the value of the average annual cost at the largest realizable EOQ and at all the price breakpoints that are greater than the largest realizable EOQ. The optimal Q is the point at which the average annual cost is a minimum.

Incremental Quantity Discounts

Consider Example 4.4, but assume incremental quantity discounts. That is, the trash bags cost 30 cents each for quantities of 500 or fewer; for quantities between 500 and 1,000, the first 500 cost 30 cents each and the remaining amount cost 29 cents each; for quantities of 1,000 and over the first 500 cost 30 cents each, the next 500 cost 29 cents each, and the remaining amount cost 28 cents each. We need to determine a mathematical expression for the function $C(Q)$ pictured in Figure 4–10. From the figure, we see that the first price break corresponds to $C(Q) = (500)(0.30) = \$150$ and the second price break corresponds to $C(Q) = 150 + (0.29)(500) = \$295$. It follows that

$$C(Q) = \begin{cases} 0.30Q & \text{for} \quad 0 \le Q < 500, \\ 150 + 0.29(Q - 500) = 5 + 0.29Q & \text{for} \quad 500 \le Q < 1{,}000, \\ 295 + 0.28(Q - 1{,}000) = 15 + 0.28Q & \text{for} \quad 1{,}000 \le Q \end{cases}$$

so that

$$C(Q)/Q = \begin{cases} 0.30 & \text{for} \quad 0 \le Q < 500, \\ 0.29 + 5/Q & \text{for} \quad 500 \le Q < 1{,}000, \\ 0.28 + 15/Q & \text{for} \quad 1{,}000 \le Q. \end{cases}$$

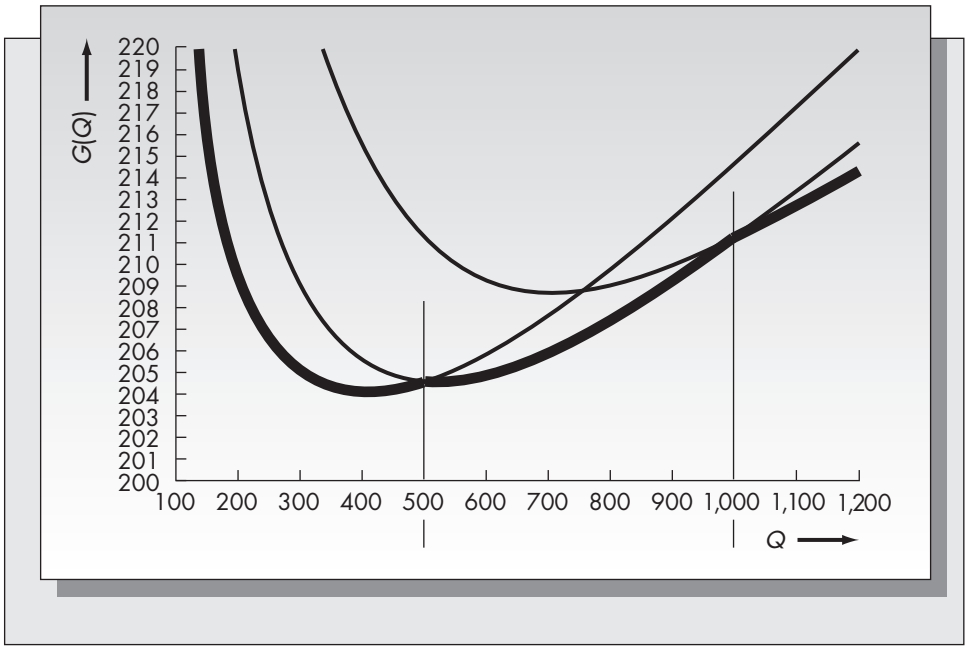


FIGURE 4–12
Average annual cost function for incremental discount schedule

The average annual cost function, $G(Q)$, is

$$G(Q) = \lambda C(Q)/Q + K\lambda/Q + I[C(Q)/Q]Q/2.$$

In this example, $G(Q)$ will have three different algebraic representations [$G_0(Q)$, $G_1(Q)$, and $G_2(Q)$] depending upon into which interval Q falls. Because $C(Q)$ is continuous, $G(Q)$ also will be continuous. The function $G(Q)$ appears in Figure 4–12.

The optimal solution occurs at the minimum of one of the average annual cost curves. The solution is obtained by substituting the three expressions for $C(Q)/Q$ in the defining equation for $G(Q)$, computing the three minima of the curves, determining which of these minima fall into the correct interval, and, finally, comparing the average annual costs at the realizable values. We have that

$$G_0(Q) = (600)(0.30) + (8)(600)/Q + (0.20)(0.30)Q/2$$

which is minimized at

$$Q^{(0)} = \sqrt{\frac{2K\lambda}{Ic_0}} = \sqrt{\frac{(2)(8)(600)}{(0.20)(0.30)}} = 400;$$

$$\begin{aligned} G_1(Q) &= (600)(0.29 + 5/Q) + (8)(600)/Q + (0.20)(0.29 + 5/Q)(Q/2) \\ &= (0.29)(600) + (13)(600)/Q + (0.20)(0.29)Q/2 + (0.20)(5)/2 \end{aligned}$$

which is minimized at

$$Q^{(1)} = \sqrt{\frac{(2)(13)(600)}{(0.20)(0.29)}} = 519;$$

and finally

$$\begin{aligned} G_2(Q) &= (600)(0.28 + 15/Q) + (8)(600)/Q + (0.20)(0.28 + 15/Q)Q/2 \\ &= (0.28)(600) + (23)(600)/Q + (0.20)(0.28)Q/2 + (0.20)(15)/2 \end{aligned}$$

which is minimized at

$$Q^{(2)} = \sqrt{\frac{(2)(23)(600)}{(0.20)(0.28)}} = 702.$$

Both $Q^{(0)}$ and $Q^{(1)}$ are realizable. $Q^{(2)}$ is not realizable because $Q^{(2)} < 1{,}000$. The optimal solution is obtained by comparing $G_0(Q^{(0)})$ and $G_1(Q^{(1)})$. Substituting into the earlier expressions for $G_0(Q)$ and $G_1(Q)$, we obtain

$$G_0(Q^{(0)}) = \$204.00,$$
$$G_1(Q^{(1)}) = \$204.58.$$

Hence, the optimal solution is to place a standing order with the Weighty Trash Bag Company for 400 units at the highest price of 30 cents per unit. The cost of using a standard order of 519 units is only slightly higher. Notice that compared to the all-units case, we obtain a smaller batch size at a higher average annual cost.

Summary of the Solution Technique for Incremental Discounts

1. Determine an algebraic expression for $C(Q)$ corresponding to each price interval. Use that to determine an algebraic expression for $C(Q)/Q$.
2. Substitute the expressions derived for $C(Q)/Q$ into the defining equation for $G(Q)$. Compute the minimum value of Q corresponding to each price interval separately.
3. Determine which minima computed in (2) are realizable (that is, fall into the correct interval). Compare the values of the average annual costs at the realizable EOQ values and pick the lowest.

Other Discount Schedules

Although all-units and incremental discount schedules are the most common, there are a variety of other discount schedules as well. One example is the carload discount schedule pictured in Figure 4–13.

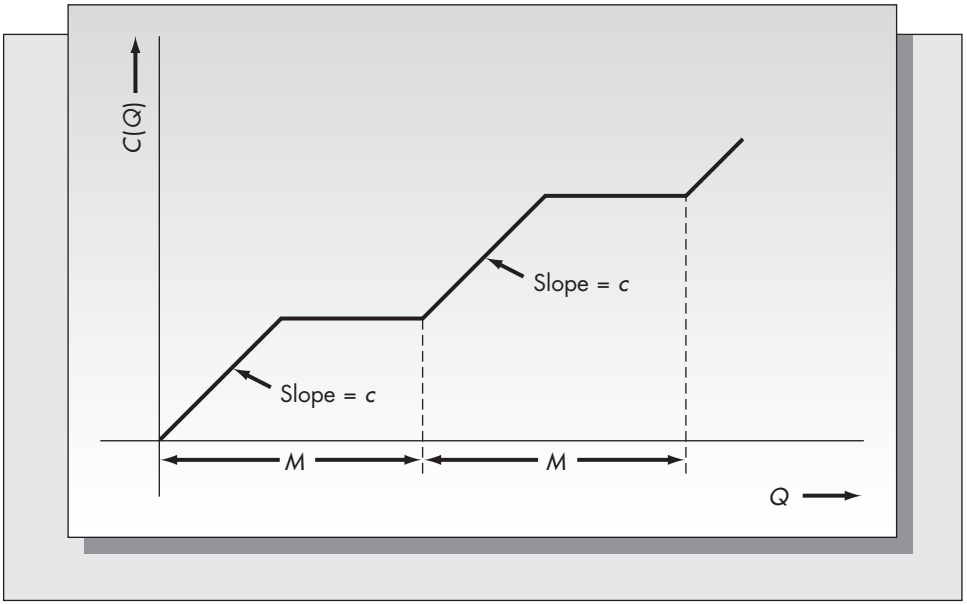


FIGURE 4–13
Order cost function for carload discount schedule

The rationale behind the carload discount schedule is the following. A carload consists of M units. The supplier charges a constant c per unit up until you have paid for the cost of a full carload, at which point there is no charge for the remaining units in that carload. Once the first carload is full, you again pay c per unit until the second carload is full, and so forth.

Determining optimal policies for the carload or other discount schedules could be extremely difficult. Each discount schedule has a unique solution procedure. Some can be extremely complex.

Problems for Section 4.7

21. Your local grocery store stocks rolls of bathroom tissue in single packages and in more economical 12-packs. You are trying to decide which to buy. The single package costs 45 cents and the 12-pack costs $5. You consume bathroom tissue at a fairly steady rate of one roll every three months. Your opportunity cost of money is computed assuming an interest rate of 25 percent and a fixed cost of $1 for the additional time it takes you to buy bathroom tissue when you go shopping. (We are assuming that you shop often enough so that you don't require a special trip when you run out.)
    a. How many single rolls should you be buying in order to minimize the annual holding and setup costs of purchasing bathroom tissue?
    b. Determine if it is more economical to purchase the bathroom tissue in 12-packs.
    c. Are there reasons other than those discussed in the problem that would motivate you not to buy the bathroom tissue in 12-packs?
22. A purchasing agent for a particular type of silicon wafer used in the production of semiconductors must decide among three sources. Source A will sell the silicon wafers for $2.50 per wafer, independently of the number of wafers ordered. Source B will sell the wafers for $2.40 each but will not consider an order for fewer than 3,000 wafers, and Source C will sell the wafers for $2.30 each but will not accept an order for fewer than 4,000 wafers. Assume an order setup cost of $100 and an annual requirement of 20,000 wafers. Assume a 20 percent annual interest rate for holding cost calculations.
    a. Which source should be used, and what is the size of the standing order?
    b. What is the optimal value of the holding and setup costs for wafers when the optimal source is used?
    c. If the replenishment lead time for wafers is three months, determine the reorder point based on the on-hand level of inventory of wafers.
23. Assume that two years have passed, and the purchasing agent mentioned in Problem 22 must recompute the optimal number of wafers to purchase and from which source to purchase them. Source B has decided to accept any size offer, but sells the wafers for $2.55 each for orders of up to 3,000 wafers and $2.25 each for the incremental amount ordered over 3,000 wafers. Source A still has the same price schedule, and Source C went out of business. Now which source should be used?
24. In the calculation of an optimal policy for an all-units discount schedule, you first compute the EOQ values for each of the three order costs, and you obtain: $Q^{(0)} = 800$, $Q^{(1)} = 875$, and $Q^{(2)} = 925$. The all-units discount schedule has breakpoints at 750 and 900. Based on this information only, can you determine what the optimal order quantity is? Explain your answer.

Snapshot Application

SMARTOPS ASSISTS IN DESIGNING CATERPILLAR'S INVENTORY CONTROL SYSTEM

The Caterpillar Corporation is one of the world's leading manufacturers of construction and mining equipment, diesel and natural gas engines, and gas turbines. Caterpillar posted annual sales of nearly $66 billion in 2012, and is a Fortune 100 company. In 1997, management embarked on a project aimed at improving product availability and inventory turns. In 2001, the project team was expanded to include both internal Caterpillar personnel and two analysts from SmartOps, David Alberti and SmartOps founder Sridhar Tayur. The team was charged with examining the following questions: (1) What product availability is possible at what cost and what inventory levels? (2) How much inventory reduction is possible? (3) What mix and deployment of inventory is necessary to achieve the firm's service objectives?

The team focused its attention on backhoe loaders produced by the company's Clayton, North Carolina, facility. A key part of their focus was on lead times. These include both lead times for products being shipped from the plant, and for materials and equipment shipped into the plant from suppliers. While lead times coming out of the plant were relatively short (typically one week), the lead times of products coming from suppliers could be quite long, and also tended to have high variance. The team developed a comprehensive model of the entire inventory supply chain. Inputs to the model included inventory stocking locations, historical forecast errors for different seasons and different products, subassemblies and bill of materials for each product, lead times, and inventory review times. The model developed took into account the multi-echelon (that is, multi-level) nature of the system.

In order to provide an accurate picture of the system, the analysts differentiated three types of orders placed by dealers: (a) sold orders based on firm sales to customers, (b) orders by dealers placed for the purpose of replenishing inventory, and (c) orders placed by dealers to replenish stock of machines rented to customers. Note that replenishment lead times for each of these demand types are different: four to six weeks for sold orders, six weeks for inventory replenishment, and fourteen weeks for replenishment of rental equipment.

The model provided a comprehensive picture of the important service versus inventory trade-offs Caterpillar could expect from the Clayton plant. For example, the team found that the Clayton plant could achieve a consistent 45 and 30 day product availability to dealers with 50 percent less finished goods inventory. However, to achieve a 14 or 10 day availability the finished goods inventory would have to be increased by approximately 90 percent. They also found that current service levels could be maintained with a total supply chain inventory reduction of 30–50 percent by repositioning inventory from finished goods to components.

After careful consideration of the trade-offs involved, management made the following changes to the system: (1) total inventory in the supply chain was reduced by 16 percent, (2) the mean time for orders was reduced by 20 percent and the variance reduced by 50 percent, and (3) order fulfillment was increased 2 percent.

The analysis provided by the team allowed management to better understand that there were no simple answers to their inventory management problems. The Caterpillar supply chain was very complex with interaction among the multiple levels. Seeing what trade-offs were possible allowed management to set priorities which ultimately resulted in lower inventory costs and higher levels of service to the customer.

Source: Keene, et al. "Caterpillar's Building Construction Products Division Improves and Stabilizes Product Availability" *Interfaces* 36 (2006), pp. 283–295.

25. Parasol Systems sells motherboards for personal computers. For quantities up through 25, the firm charges $350 per board; for quantities between 26 and 50, it charges $315 for each board purchased beyond 25; and it charges $285 each for the additional quantities over 50. A large communications firm expects to require these motherboards for the next 10 years at a rate of at least 140 per year. Order setup costs are $30 and holding costs are based on an 18 percent annual interest rate. What should be the size of the standing order?

*4.8 RESOURCE-CONSTRAINED MULTIPLE PRODUCT SYSTEMS

The EOQ model and its extensions apply only to single inventory items. However, these models are often used in companies stocking many different items. Although we could certainly compute optimal order quantities separately for each of the different items,

there could exist constraints that would make the resulting solution infeasible. In Example 4.1, the optimal solution called for purchasing 3,870 pencils every 1.24 years. The bookstore, however, may not have allocated enough space to store that many pencils, nor enough money to purchase that many at one time.

Example *4.5*

Three items are produced in a small fabrication shop. The shop management has established the requirement that the shop never have more than $30,000 invested in the inventory of these items at one time. The management uses a 25 percent annual interest charge to compute the holding cost. The relevant cost and demand parameters are given in the following table. What lot sizes should the shop be producing so that they do not exceed the budget?

	Item		
	1	**2**	**3**
Demand rate λ_j	1,850	1,150	800
Variable cost c_j	50	350	85
Setup cost K_j	100	150	50

Solution

If the budget is not exceeded when using the EOQ values of these three items, then the EOQs are optimal. Hence, the first step is to compute the EOQ values for all items to determine whether or not the constraint is active.

$$\text{EOQ}_1 = \sqrt{\frac{(2)(100)(1{,}850)}{(0.25)(50)}} = 172,$$

$$\text{EOQ}_2 = \sqrt{\frac{(2)(150)(1{,}150)}{(0.25)(350)}} = 63,$$

$$\text{EOQ}_3 = \sqrt{\frac{(2)(50)(800)}{(0.25)(85)}} = 61.$$

If the EOQ value for each item is used, the maximum investment in inventory would be

$$(172)(50) + (63)(350) + (61)(85) = \$35{,}835.^{3}$$

Because the EOQ solution violates the constraint, we need to reduce these lot sizes. But how?

The optimal solution turns out to be very easy to find in this case. We merely multiply each EOQ value by the ratio (30,000)/(35,835) = 0.8372. In order to guarantee that we do not exceed the $30,000 budget, we round each value *down* (adjustments can be made subsequently). Letting Q_1^*, Q_2^*, and Q_3^* be the optimal values, we obtain

$$Q_1^* = (172)(0.8372) \approx 144,$$
$$Q_2^* = (63)(0.8372) \approx 52,$$
$$Q_3^* = (61)(0.8372) \approx 51,$$

where $\approx$ should be interpreted as rounding to the next lower integer in this case.

The total budget required for these lot sizes is $29,735. The remaining $265 can now be used to increase the lot sizes of products 1 and 3 slightly. (For example, $Q_1^* = 147$, $Q_3^* = 52$ results in a budget of $29,970.)

[3] We are assuming for the purposes of this section that the three products are ordered simultaneously. By staggering order cycles, it is possible to meet the constraint with larger lot sizes. We will not consider that case here.

In general, budget- or space-constrained problems are not solved so easily. Suppose that n items have unit costs of $c_1, c_2, \ldots, c_n$, respectively, and the total budget available for them is C. Then the budget constraint can be written

$$c_1Q_1 + c_2Q_2 + \cdots + c_nQ_n \leq C.$$

Let

$$\text{EOQ}_i = \sqrt{\frac{2K_i\lambda_i}{h_i}} \qquad \text{for } i = 1, \ldots, n,$$

where K_i, h_i, and λ_i are the respective cost and demand parameters.

There are two possibilities: either the constraint is active or it is not. If the constraint is not active, then

$$\sum_{i=1}^{n} c_i\text{EOQ}_i \leq C,$$

and the optimal solution is $Q_i = \text{EOQ}_i$. If the constraint is active, then

$$\sum_{i=1}^{n} c_i\text{EOQ}_i > C,$$

and the EOQ solution is no longer feasible. If we include the following assumption, the solution to the active case is relatively easy to find:

$$\textit{Assumption: } c_1/h_1 = c_2/h_2 = \cdots = c_n/h_n.$$

If this assumption holds and the constraint is active, we prove in Appendix 4–A that the optimal solution is

$$Q_i^* = m\text{EOQ}_i,$$

where the multiplier m solves

$$m = C \Big/ \left[\sum_{i=1}^{n} (c_i\text{EOQ}_i)\right].$$

Since $c_i/h_i = c_i/(I_ic_i) = 1/I_i$, the condition that the ratios be equal is equivalent to the requirement that the same interest rate be used to compute the holding cost for each item, which is reasonable in most circumstances.

Suppose that the constraint is on the available space. Let w_i be the space consumed by one unit of product i for $i = 1, 2, \ldots, n$ (this could be floor space measured, say, in square feet, or volume measured in cubic feet), and let W be the total space available. Then the space constraint is of the form

$$w_1Q_1 + w_2Q_2 + \cdots + w_nQ_n \leq W.$$

This is mathematically of the same form as the budget constraint, so the same analysis applies. However, our condition for a simple solution now is that the ratios w_i/h_i are equal. That is, the space consumed by an item should be proportional to its holding cost. When the interest rate is fixed, this is equivalent to the requirement that the space consumed should be proportional to the value of the item. This requirement would probably be too restrictive in most cases. For example, fountain pens take up far less space than legal pads, but are more expensive.

Let us now consider the problem in which the constraint is active, but the proportionality assumption is not met. This problem is far more complex than that just solved. It requires formulating the *Lagrangian* function. The details of the formulation of this problem can be found in Appendix 4–A. As we show, the optimal lot sizes are now of the form

$$Q_i^* = \sqrt{\frac{2K_i\lambda_i}{h_i + 2\theta w_i}},$$

where θ is a constant chosen so that

$$\sum_{i=1}^{n} w_i Q_i^* = W.$$

The constant θ, known as the Lagrange multiplier, reduces the lot sizes by increasing the effective holding cost. The correct value of θ can be found by trial and error or by a search technique such as interval bisection. Note that $\theta > 0$, so that the search can be limited to positive numbers only. The value of θ can be interpreted as the decrease in the average annual cost that would result from adding an additional unit of resource. In this case, it would represent the marginal benefit of an additional square foot of space.

Example *4.6*

Consider the fabrication shop of Example 4.5. In addition to the budget constraint, suppose that there are only 2,000 square feet of floor space available. Assume that the three products consume respectively 9, 12, and 18 square feet per unit.

First, we check to see if the EOQ solution is feasible. Setting $w_1 = 9$, $w_2 = 12$, and $w_3 = 18$, we find that

$$\sum \text{EOQ}_i w_i = (172)(9) + (63)(12) + (61)(18) = 3{,}402,$$

which is obviously infeasible. Next, we check if the budget-constrained solution provides a feasible solution to the space-constrained problem. The budget-constrained solution requires (147)(9) + (52)(12) + (52)(18) = 2,883 square feet of space, which is also infeasible.

The next step is to compute the ratios w_i/h_i for $1 \le i \le 3$. These ratios turn out to be 0.72, 0.14, and 0.85, respectively. Because their values are different, the simple solution obtained by a proportional scaling of the EOQ values will not be optimal. Hence, we must determine the value of the Lagrange multiplier θ.

We can determine upper and lower bounds on the optimal value of θ by assuming equal ratios. If the ratios were equal, the multiplier, m, would be

$$m = W \Big/ \sum(\text{EOQ}_i w_i) = 2{,}000/3{,}402 = 0.5879,$$

which would give the three lot sizes as 101, 37, and 36, respectively. The three values of θ that result in these lot sizes are respectively $\theta = 1.32$, $\theta = 6.86$, and $\theta = 2.33$. (These values were obtained by setting the given expression for Q_1^* equal to the lot sizes 101, 37, and 36, and solving for θ.)

The true value of θ will be between 1.32 and 6.86. If we start with $\theta = 3.5$, we obtain $Q_1^* = 70$, $Q_2^* = 45$, and $Q_3^* = 23$, and $\Sigma w_i Q_i^* = (70)(9) + (45)(12) + (23)(18) = 1{,}584$, which implies $\theta < 3.5$. After considerable experimentation, we finally find that the optimal value of $\theta = 1.75$, and $Q_1^* = 92$, $Q_2^* = 51$, and $Q_3^* = 31$, giving $\Sigma w_i Q_i^* = 1{,}998$. Notice that these lot sizes are in very different proportions from the ones obtained assuming a constant multiplier. Searching for the optimal value of the Lagrange multiplier, although tedious to do by hand, can be accomplished very efficiently using a computer. Spreadsheets are a useful means for effecting such calculations.

We have only touched on some of the complications that could arise when applying EOQ-type models to a realistic system with multiple products. Often, real problems are far too complex to be expressed accurately as a solvable mathematical model. For this reason, simple models such as the ones presented in this and subsequent chapters are used by practitioners. However, any solution recommended by a mathematical model must be considered in the context of the system in which it is to be used.

Problems for Section 4.8

26. A local outdoor vegetable stand has exactly 1,000 square feet of space to display three vegetables: tomatoes, lettuce, and zucchini. The appropriate data for these items are given in the following table.

	Item		
	Tomatoes	**Lettuce**	**Zucchini**
Annual demand (in pounds)	850	1,280	630
Cost per pound	$0.29	$0.45	$0.25

The setup cost for replenishment of the vegetables is $100 in each case, and the space consumed by each vegetable is proportional to its costs, with tomatoes requiring 0.5 square foot per pound. The annual interest rate used for computing holding costs is 25 percent. What are the optimal quantities that should be purchased of these three vegetables?

27. Suppose that the vegetables discussed in Problem 26 are purchased at different times. In what way could that have an effect on the order policy that the stand owner should use?

28. Suppose that in Problem 26 the space consumed by each vegetable is not proportional to its cost. In particular, suppose that one pound of lettuce required 0.4 square foot of space and one pound of zucchini required 1 square foot of space. Determine upper and lower bounds on the optimal values of the order quantities in this case. Test different values of the Lagrange multiplier to find the optimal values of the order quantities. (A spreadsheet is ideally suited for this kind of calculation. If you solve the problem using a spreadsheet, place the Lagrange multiplier in a cell so that its value can be changed easily.)

4.9 EOQ MODELS FOR PRODUCTION PLANNING

Simple lot-sizing models have been successfully applied to a variety of manufacturing problems. In this section we consider an extension of the EOQ model with a finite production rate, discussed in Section 4.6, to the problem of producing n products on a single machine. Following the notation used in this chapter, let

λ_j = Demand rate for product j,

P_j = Production rate for product j,

h_j = Holding cost per unit per unit time for product j,

K_j = Cost of setting up the production facility to produce product j.

The goal is to determine the optimal procedure for producing n products on the machine to minimize the cost of holding and setups, and to guarantee that no stock-outs occur during the production cycle.

We require the assumption that

$$\sum_{j=1}^{n} \lambda_j/P_j \leq 1.$$

This assumption is needed to ensure that the facility has sufficient capacity to satisfy the demand for all products. Notice that this is stronger than the assumption made in Section 4.6 that $\lambda_j < P_j$ for each j. To see why this assumption is necessary, consider the case of two products with identical demand and production rates. In each cycle one would produce product 1 first, then product 2. Clearly $P_1 \geq 2\lambda_1$ and $P_2 \geq 2\lambda_2$, so that enough could be produced to meet the total demand for both products each cycle. This translates to $\lambda_1/P_1 \leq 0.5$ and $\lambda_2/P_2 \leq 0.5$, giving a value of the sum less than or equal to one. Similar reasoning holds for more than two products with nonidentical demand and production rates.

We also will assume that the policy used is a *rotation cycle policy.* That means that in each cycle there is exactly one setup for each product, and products are produced in the same sequence in each production cycle. The importance of this assumption will be discussed further at the end of the section.

At first, one might think that the optimal solution is to sequentially produce lot sizes for each product optimized by treating each product in isolation. From Section 4.6, this would result in a lot size for product j of

$$Q_j = \sqrt{\frac{2K_j\lambda_j}{h_j'}},$$

where $h_j' = h_j(1 - \lambda_j/P_j)$. The problem with this approach is that because we have only a single production facility, it is likely that some of the lot sizes Q_j will not be large enough to meet the demand between production runs for product j, thus resulting in stock-outs.

Let T be the cycle time. During time T we assume that exactly one lot of each product is produced. In order that the lot for product j will be large enough to meet the demand occurring during time T, it follows that the lot size must be

$$Q_j = \lambda_j T.$$

From Section 4.6, the average annual cost associated with product j can be written in the form

$$G(Q_j) = K_j\lambda_j/Q_j + h_j'Q_j/2.$$

The average annual cost for all products is the sum

$$\sum_{j=1}^{n} G(Q_j) = \sum_{j=1}^{n} K_j\lambda_j/Q_j + h_j'Q_j/2.$$

Substituting $T = Q_j/\lambda_j$, we obtain the average annual cost associated with the n products in terms of the cycle time T as

$$G(T) = \sum_{j=1}^{n} [K_j/T + h_j'\lambda_j T/2].$$

The goal is to find T to minimize $G(T)$. The necessary condition for an optimal T is

$$\frac{dG(T)}{dT} = 0.$$

Setting the first derivative with respect to T to zero gives

$$\sum_{j=1}^{n} [-K_j/T^2 + h_j'\lambda_j/2] = 0.$$

Solving for T, we obtain the optimal cycle time T^* as

$$T^* = \sqrt{\frac{2\sum_{j=1}^{n} K_j}{\sum_{j=1}^{n} h_j'\lambda_j}}.$$

If setup times are a factor, we must check that there is enough time each cycle to account for both setup times and production of the n products. Let s_j be the setup time for product j. Ensuring that the total time required for setups and production each cycle does not exceed T leads to the constraint

$$\sum_{j=1}^{n} (s_j + Q_j/P_j) \leq T.$$

Using the fact that $Q_j = \lambda_j T$, this condition translates to

$$\sum_{j=1}^{n} (s_j + \lambda_j T/P_j) \leq T,$$

which gives, after rearranging terms,

$$T \geq \frac{\sum_{j=1}^{n} s_j}{1 - \sum_{j=1}^{n} (\lambda_j/P_j)} = T_{\min}.$$

Because $T_{\min}$ cannot be exceeded without compromising feasibility, the optimal solution is to choose the cycle time T equal to the *larger* of T^* and $T_{\min}$.

Example 4.7

Bali produces several styles of men's and women's shoes at a single facility near Bergamo, Italy. The leather for both the uppers and the soles of the shoes is cut on a single machine. This Bergamo plant is responsible for seven styles and several colors in each style. (The colors are not considered different products for our purposes, because no setup is required when switching colors.) Bali would like to schedule cutting for the shoes using a rotation policy that meets all demand and minimizes setup and holding costs. Setup costs are proportional to setup times.

TABLE 4–1
Relevant Data for Example 4.7

Style	Annual Demand (units/year)	Production Rate (units/year)	Setup Time (hours)	Variable Cost ($/unit)
Women's pump	4,520	35,800	3.2	$40
Women's loafer	6,600	62,600	2.5	26
Women's boot	2,340	41,000	4.4	52
Women's sandal	2,600	71,000	1.8	18
Men's wingtip	8,800	46,800	5.1	38
Men's loafer	6,200	71,200	3.1	28
Men's oxford	5,200	56,000	4.4	31

The firm estimates that setup costs amount to an average of $110 per hour, based on the cost of worker time and the cost of forced machine idle time during setups. Holding costs are based on a 22 percent annual interest charge.

The relevant data for this problem appear in Table 4–1.

Solution

The first step is to verify that the problem is feasible. To do so we compute $\Sigma\, \lambda_j/P_j$. The reader should verify that this sum is equal to 0.69355. Because this is less than one, there will be a feasible solution. Next we compute the value of T^*, but to do so we need to do several intermediate calculations.

First, we compute setup costs. Setup costs are assumed to be $110 times setup times. Second, we compute modified holding costs (h_j'). This is done by multiplying the cost of each product by the annual interest rate (0.22) times the factor $1 - \lambda_j/P_j$. These calculations give

Setup Costs (K_j)	Modified Holding Costs (h_j')
352	7.69
275	5.12
484	10.79
198	3.81
561	6.79
341	5.62
484	6.19

The sum of the setup costs is 2,695, and the sum of the products of the modified holding costs and the annual demands is 230,458.4. Substituting these figures into the formula for T^* gives the optimal cycle time as 0.1529 year. Assuming a 250-day work year, this means that the rotation cycle should repeat roughly every 38 working days. The optimal lot size for each of the shoes is found by multiplying the cycle time by the demand rate for each item. The reader should check that the following lot sizes result:

Style	Optimal Lot Sizes for Each Production Run
Women's pump	691
Women's loafer	1,009
Women's boot	358
Women's sandal	398
Men's wingtip	1,346
Men's loafer	948
Men's oxford	795

The plant would cut the soles and uppers in these lot sizes in sequence (although the sequence does not necessarily have to be this one) and would repeat the rotation cycle roughly

every 38 days (0.1529 year). However, this solution can be implemented only if T^* is at least T_{min}. To determine T_{min} we must express the setup times in years. Assuming 8 working hours per day and 250 working days per year, one would divide the setup times given in hours by 2,000 (250 times 8). The reader should check that the resulting value of T_{min} is 0.04, thus making T^* feasible and, hence, optimal.

The total average annual cost of holding and setups at an optimal policy can be found by computing the value of $G(T)$ when $T = T^*$. It is \$35,244.44. It is interesting to note that if the plant manager chooses to implement this policy, the facility will be idle for a substantial portion of each rotation cycle. The total uptime each rotation cycle is found by dividing the lot sizes by the production rates for each style and summing the results. It turns out to be 0.106 year. Hence, the optimal rotation policy that minimizes total holding and setup costs results in the cutting operation remaining idle about one-third of the time.

The reader should be aware that the relatively simple solution to this problem was the result of two assumptions. One was that the setup costs were not sequence dependent. In Example 4.7, it is possible that the time required to change over from one shoe style to another could depend on the styles. For example, a changeover from a woman's style to a man's style probably takes longer than from one woman's style to another. A second assumption was that the plant used a rotation cycle policy. That is, in each cycle Bali does a single production run of each style. When demand rates and setup costs differ widely, it might be advantageous to do two or more production runs of a product in a cycle. This more general problem has not, to our knowledge, been solved. A discussion of these and related issues can be found in Maxwell (1964). Magee and Boodman (1967) provide some heuristic ways of dealing with the more general problem.

Problems for Section 4.9

29. A metal fabrication shop has a single punch press. There are currently three parts that the shop has agreed to produce that require the press, and it appears that they will be supplying these parts well into the future. You may assume that the press is the critical resource for these parts, so that we need not worry about the interaction of the press with the other machines in the shop. The relevant information here is

Part Number	Annual Contracted Amount (demand)	Setup Cost	Cost (per unit)	Production Rate (per year)
1	2,500	\$80	\$16	45,000
2	5,500	120	18	40,000
3	1,450	60	22	26,000

Holding costs are based on an 18 percent annual interest rate, and the products are to be produced in sequence on a rotation cycle. Setup times can be considered negligible.

a. What is the optimal time between setups for part number 1?

b. What percentage of the time is the punch press idle, assuming an optimal rotation cycle policy?

c. What are the optimal lot sizes of each part put through the press at an optimal solution?

d. What is the total annual cost of holding and setup for these items on the punch press, assuming an optimal rotation cycle?

30. Tomlinson Furniture has a single lathe for turning the wood for various furniture pieces, including bedposts, rounded table legs, and other items. Four forms are turned on the lathe and produced in lots for inventory. To simplify scheduling, one lot of each type will be produced in a cycle, which may include idle time. The four products and the relevant information concerning them appears in the following table.

Piece	Monthly Requirements	Setup Time (hours)	Unit Cost	Production Rate (units/day)
J–55R	125	1.2	$20	140
H–223	140	0.8	35	220
K–18R	45	2.2	12	100
Z–344	240	3.1	45	165

Worker time for setups is valued at $85 per hour, and holding costs are based on a 20 percent annual interest charge. Assume 20 working days per month and 12 months per year for your calculations.

a. Determine the optimal length of the rotation cycle.

b. What are the optimal lot sizes for each product?

c. What are the percentages of uptime and downtime for the lathe, assuming that it is not used for any other purpose?

d. Draw a graph showing the change in the inventory level over a typical cycle for each product.

e. Discuss why the solution you obtained might not be feasible for the firm, or why it might not be desirable even when it is feasible.

4.10 POWER-OF-TWO POLICIES

The inventory models treated in this chapter form the basis of more complex cases. In almost every case treated here, we were able to find relatively straightforward algebraic solutions. However, even when demand is assumed known, there exist several extensions of these basic models whose optimal solutions may be difficult or impossible to find. In those cases, effective approximations are very valuable. Here, we discuss an approach that has proven to be successful in a variety of deterministic environments. The idea is based on choosing the best replenishment interval from a set of possible intervals proportional to powers of two. While the analysis of power-of-two policies in complex environments is beyond the scope of this book, we can illustrate the idea in the context of the basic EOQ model.

From Section 4.5, we know that the order quantity that minimizes average annual holding and setup costs when demand is fixed at λ units per unit time is the EOQ given by

$$Q^* = \sqrt{\frac{2K\lambda}{h}},$$

and the optimal time between placement of orders, say T^*, is given by

$$T^* = Q^*/\lambda = \sqrt{\frac{2K}{\lambda h}}.$$

It is possible, and even likely, that optimal order intervals are inconvenient. For example, the optimal solution might call for ordering every 3.393 weeks. However, one might only want to place orders at the beginning of a day or a week. To account for this, suppose that we impose the constraint that ordering must occur in some multiple of a base time, T_L. To find the optimal solution under the constraint that the order interval must be a multiple of T_L, one would simply compare the costs at the two closest multiples of T_L to T^* and pick the order interval with the lower cost. [That is, find k for which $kT_L \leq T^* \leq (k+1)T_L$ and choose the reorder interval to be either kT_L or $(k+1)T_L$ depending on which results in a lower average annual cost.]

Now suppose we add the additional restriction that the order interval must be a power of two times T_L. That is, the order interval must be of the form 2^kT_L for some integer $k \geq 0$. While it is unlikely one would impose such a restriction in the context of the simple EOQ problem, the ultimate goal is to explore these policies for more complex problems whose optimal solutions are hard to find. The question we wish to address is: Under such a restriction (known as a power-of-two policy), what is the worst cost error we will incur relative to that of the optimal reorder interval T^*? On the surface, it appears that such a restriction would result in serious errors. As k increases, the distance between successive powers of two grows rapidly. For example, if $k = 12$, $2^k = 4{,}096$ and $2^{k+1} = 8{,}192$, a very wide interval. If $T^* = 6{,}000$ and $T_L = 1$, for example, it seems that forcing the order interval to be either 4,096 or 8,192 would result in a large cost error (as the error in T is nearly 30 percent in either direction). However, this turns out not to be the case. In fact, we can prove that the cost error in every case is bounded by slightly more than 6 percent. While the result seems unintuitive at first, recall that the average annual cost function is relatively insensitive to errors in Q, as we saw in Section 4.5. Since Q and T are proportional, a similar cost insensitivity holds with respect to T.

We know from Section 4.5 that the average annual holding and setup cost as a function of the order quantity, Q, is given by

$$G(Q) = \frac{K\lambda}{Q} + \frac{hQ}{2}.$$

Since $Q = \lambda T$, the average annual cost can also be expressed in terms of the reorder interval, T, as

$$G(T) = \frac{K}{T} + \frac{h\lambda T}{2}.$$

In the case of a power-of-two policy, we wish to find the value of k that minimizes $G(2^kT_L)$. Since $G(T)$ is a continuous convex function of T, it follows that $G(2^kT_L)$ is a discrete convex function of k. That means that the optimal value of k, say k^*, satisfies

$$k^* = \min\{k : G(2^{k+1}T_L) \geq G(2^kT_L)\}.$$

Substituting for $G(T)$, the optimality condition becomes

$$\frac{K}{2^{k+1}T_L} - \frac{K}{2^kT_L} \geq \frac{h\lambda 2^kT_L}{2} - \frac{h\lambda 2^{k+1}T_L}{2}$$

which can easily be seen to reduce to

$$\frac{K/2}{2^kT_L} \leq h\lambda 2^{k-1}T_L.$$

Rearranging terms gives

$$2^k \geq \frac{1}{T_L}\sqrt{\frac{K}{h\lambda}},$$

or

$$2^k T_L \geq \frac{1}{\sqrt{2}}T^*.$$

We assume that $T^* \geq T_L$, which means that $\sqrt{2}T^* > T_L$. Hence, to summarize, we seek the smallest value of k that satisfies the simultaneous inequalities

$$\frac{1}{\sqrt{2}}T^* \leq 2^k T_L \leq \sqrt{2}T^*.$$

Note that this implies that as long as $T^* \geq T_L$, the optimal power-of-two solution will always lie between $.707T^*$ and $1.41T^*$. The next question is: What is the worst-case cost error of this policy? Since $Q = \lambda T$, if $T = T^*/\sqrt{2}$, then $Q = Q^*/\sqrt{2}$, and similarly if $T = \sqrt{2}T^*$, then $Q = \sqrt{2}Q^*$. It follows that the worst-case cost error of the power-of-two policy is given by

$$\frac{G(Q)}{G(Q^*)} = \frac{1}{2}\left(\frac{Q^*}{Q} + \frac{Q}{Q^*}\right) = \frac{1}{2}\left(\frac{1}{\sqrt{2}} + \sqrt{2}\right) = 1.0607,$$

or slightly more than 6 percent. (Because of symmetry, we obtain the same result when $Q = Q^*/\sqrt{2}$ or when $Q = \sqrt{2}Q^*$.)

The real "power" of power-of-two policies occurs when trying to solve more complex problems whose optimal policies are difficult to find. Consider the following scenario. A single warehouse is the sole supplier of N retailers. The demand rate experienced by each retailer is known and constant. As with the simple EOQ problem, assume that shortages at both the warehouse and the retailers are not permitted. There are fixed costs for ordering at both the warehouse and the retailers, and holding costs at these locations. These costs do not necessarily need to be the same. It is assumed that there is no lead time for placement or arrival of orders.

Unlike the simple EOQ problem, determining an optimal policy (that is, one that minimizes long-run average costs) for this problem could be extremely difficult, or even close to impossible. In many cases, even the form of an optimal policy may not be known. Clearly, effective approximations are very important.

One approximation for this problem is a so-called nested policy. In this case, a retailer would automatically order whenever the warehouse does, and possibly at other times as well. Although nested policies can have arbitrarily large cost errors, various adaptations of power-of-two policies can be shown to have 94 percent or even 98 percent guaranteed effectiveness. Power-of-two approximations have also been shown to be equally effective in serial production systems, and more complex arborescent assembly systems. We refer the interested reader to Roundy (1985) and Muckstadt and Roundy (1993).

Case Study. Betty Buys a Business

Betty Robinson decided it was time for a change. She had been a social worker for 10 years, and although she found the work rewarding, it was time for something else. So, when Herbie's Hut came up for sale, she made some inquiries.

Herbie's Hut was a small gift shop and toy store in her old neighborhood of Skokie, a Chicago suburb. Herb Gold had been running the business for 40 years, but it was a demanding job and he decided it was time to hang it up. The business was pretty steady, so Herb figured that there would buyers. He was delighted to hear from Betty,

whom he had known as a child growing up in the neighborhood. The money was less important to Herb than making sure the store was in good hands. He had a good feeling about Betty, so he agreed to let the business go for a modest price.

After Betty purchased the business, she decided to remodel the store. Herbie's Hut would reopen as Betty's Best. Skylights, new wallpaper, and a fresh layout gave the store a sprightlier look that reflected Betty's personality. During this time, Betty took a careful look at the stock that came with the purchase. She noticed that several items were severely understocked. Two of those were ID bracelets and Lego sets. She spoke with Herb about it and he apologized. These were popular items and he hadn't had a chance to reorder them.

Betty was concerned that she might run out of these items quickly and that it would hurt her reputation. When she called the suppliers she realized that she had no idea how much to order. She asked Herb and he said that his rule of thumb was to order about one month's supply. This sounded reasonable to Betty; she could figure out what a month's supply was from the store's sales records, which Herb graciously turned over to her.

The store stocked hundreds of different items. It occurred to Betty that if she reordered every item monthly, she would be spending a lot of time processing orders. Betty thought that perhaps there was a better way. She consulted her boyfriend Bob who had an MBA degree. Bob fished up his old class notes and showed Betty the EOQ formula, $\sqrt{\frac{2K\lambda}{h}}$, which was supposed to be used for determining order sizes. "What are these symbols supposed be?" asked Betty. Bob had reviewed his notes and explained that K is the fixed cost of each order, a Greek letter (λ) is the sales rate, and h is the cost of holding. Betty still wasn't sure what it all meant.

"Let's take an item you need to order now and see how this works out," suggested Bob. So they considered the bracelets.

"First, what's the process you would go through to place a new order?" asked Bob. "Well," replied Betty, "I would call the supplier to place the order, and when it arrived I would unpack the shipment, possibly put some of the items on display, and store the rest." "How long does that take?" inquired Bob. "I don't know," she responded, "maybe a couple of hours. Also, that supplier charges a fixed cost of $50 per order in addition to the cost of each bracelet." Figuring Betty's time at $50 per hour Bob computed a total fixed cost of (2)(50) + 50 = $150.

"Ok. Let's look at the sales rate," said Bob. Based on Herb's records, she estimated that he had sold an average of about 75 bracelets a month. That would give the value of λ. Finally, they needed to estimate the holding cost. Bob had spent some time thinking about this. "Holding cost can be thought of in two ways. The symbol h refers to the cost of holding a single unit of inventory for some fixed time, typically one year. It can also be thought of as an interest rate, which is then multiplied by the unit cost of the item. I think the most appropriate value of the interest rate is the rate of return you expect to earn in this business. I did a little research on the web and it seems that for businesses of this type, a 15 percent annual return on investment seems to be about right. We'll use that to figure the holding cost." Since Betty's cost of each bracelet was $30, this resulted in an annual holding cost of h = (.15)(30) = $4.50 annually.

Since the holding cost was based an annual interest rate, the sales rate had to be yearly as well. This translated to an annual sales rate of (12)(75) = 900. Plugging these numbers into the formula gave a lot size of $\sqrt{\frac{(2)(150)(900)}{4.5}} = 245$. This made more sense to Betty than ordering 75 bracelets every month. This translates to a little more than a three month supply.

Once Betty got the idea, it didn't take her long to set up a spreadsheet for all the items in the store. For example, a similar analysis of the Lego sets resulted in an EOQ value of 80 Lego sets. That seemed all well and good until her first order arrived.

Uh oh, she thought, these things are pretty bulky. They filled up almost half of her backroom storage area. She realized that there was a little more to figuring out the right lot sizes than just applying the EOQ formula.

Betty asked Bob what she should do about the fact that the Lego order took up so much space. "Well, I'm not sure," answered Bob. "I guess you shouldn't order so many Lego sets." "Well thanks for nothing, genius." Betty responded. "I know that now, but just how many should I order?"

So Bob did his research, and found out this was a tougher problem than figuring the EOQ. As a first cut, Bob read that if the value of each item were proportional to the space it consumed, the solution would be pretty simple. Realizing that this was only an approximation, he figured it would be better than nothing, and would hopefully satisfy Betty.

"Ok," Bob said, "let's consider your spreadsheet that computes the EOQ values for all of the items in the store. Now let's add a column to the spreadsheet that indicates the cubic feet of space consumed by each item. Multiply the two for each item and add up the total."

This was an easy calculation once Betty was able to approximate the space requirements for each of the items. Multiplying the two columns and adding resulted in a total space requirement of 120,000 cubic feet. Betty had only 50,000 cubic feet of storage space, so Bob suggested that each of the order quantities be reduced by multiplying by the constant 50,000/120,000 = .4167. "If you reduce all of your order quantities by around 60 percent, you shouldn't run of the space," Bob recommended.

"Bob, that's great! That will be a big help," said Betty. Betty added another column to her spreadsheet that reduced all of the lot sizes by 60 percent. But as she started looking at the numbers, something bothered her.

"You know, Bob, this will solve my space problem, but there's something about it that doesn't make sense to me. Reducing the lot size of the Lego sets from 80 to 32 sounds about right, but why should I reduce the lot size for the bracelets? They hardly take up any room at all."

Jeez, she's absolutely right on the money about that, thought Bob. You really wanted to reduce the lot sizes of the bulky items a lot more than the small items. But how do you do that? Bob consulted his friend Phil, who has a PhD in operations research. Phil explained that this was a nontrivial optimization problem that involved finding the value of something called the Lagrange multiplier. Given Betty's spreadsheet, this was a piece of cake for Phil. Once the calculation was completed, the lot sizes for the bulky items were indeed reduced much more than those of the smaller items.

By applying a few basic concepts from inventory control, Betty was able to get a handle on the problem of managing her stock. Her willingness to apply scientific principles meant that Betty was well on her way to creating a successful business.

4.11 HISTORICAL NOTES AND ADDITIONAL TOPICS

The interest in using mathematical models to control the replenishment of inventories dates back to the early part of the 20th century. Ford Harris (1915) is generally credited with the development of the simple EOQ model. R. H. Wilson (1934) is also recognized for his analysis of the model. The procedure we suggest for the all-units quantity discount model appears to be due to Churchman, Ackoff, and Arnoff (1957), and the incremental discount model appears to be due to Hadley and Whitin (1963). Kenneth Arrow provides an excellent discussion of the economic motives for holding inventories in Chapter 1 of Arrow, Karlin, and Scarf (1958).

Hadley and Whitin (1963) also seem to have been the first to consider budget- and space-constrained problems, although several researchers have studied the problem

subsequently. Rosenblatt (1981) derives results similar to those of Section 4.8 and considers several extensions not treated in this section.

Academic interest in inventory management problems took a sudden upturn in the late 1950s and early 1960s with the publication of a number of texts in addition to those just mentioned, including Whitin (1957); Magee and Boodman (1967); Bowman and Fetter (1961); Fetter and Dalleck (1961); Hanssmann (1961); Star and Miller (1962); Wagner (1962); and Scarf, Gilford, and Shelly (1963).

By and large, the vast majority of the published research on inventory systems since the 1960s has been on stochastic models, which will be the subject of Chapter 5. Extensions of the EOQ model have been considered more recently by Barbosa and Friedman (1978) and Schwarz and Schrage (1971), for example.

4.12 Summary

This chapter presented several popular models used to control inventories when the demand is known. We discussed the various *economic motives for holding inventories*, which include economies of scale, uncertainties, speculation, and logistics. In addition, we mentioned some of the *physical characteristics of inventory systems* that are important in determining the complexity and applicability of the models to real problems. These include demand, lead time, review time, back-order/lost-sales assumptions, and the changes that take place in the inventory over time.

There are three significant classes of costs in inventory management. These are *holding or carrying costs*, *order or setup costs*, and *penalty cost* for not meeting demand. The holding cost is usually expressed as the product of an interest rate and the cost of the item.

The grandfather of all inventory control models is the simple *EOQ model*, in which demand is assumed to be constant, no stock-outs are permitted, and only holding and order costs are present. The optimal batch size is given by the classic square root formula. The first extension of the EOQ model we considered was to the case in which items are produced internally at a finite production rate. We showed that the optimal batch size in this case could be obtained from the EOQ formula with a modified holding cost.

Two types of *quantity discounts* were considered: *all-units*, in which the discounted price is valid on all the units in an order, and *incremental*, in which the discount is applied only to the additional units beyond the breakpoint. The optimal solution to the all-units case will often occur at a breakpoint, whereas the optimal solution to the incremental case will almost never occur at a breakpoint.

In most real systems, the inventory manager cannot ignore the interactions that exist between products. These interactions impose *constraints* on the system; these might arise because of limitations in the space available to store inventory or in the budget available to purchase items. We considered both cases, and showed that when the ratio of the item value or space consumed by the item over the holding cost is the same for all items, a solution to the constrained problem can be obtained easily. When this condition is not met, the formulation requires introducing a *Lagrange multiplier*. The correct value of the Lagrangian multiplier can be found by trial and error or by some type of search. The presence of the multiplier reduces lot sizes by effectively increasing holding costs.

The finite production rate model of Section 4.6 was extended to multiple products under the assumption that all the products are produced on a single machine. Assuming a *rotation cycle policy* in which one lot of each product is produced each cycle, we showed how to determine the cycle time minimizing the sum of annual setup and holding costs for all products. Rotation cycles provide a straightforward way to schedule production on one machine, but may be suboptimal when demand or production rates differ widely or setup costs are sequence dependent.

The chapter concluded with a brief overview of several commercial inventory control systems. There are dozens of products available ranging in price from under $100 to tens of thousands of dollars. Many of the products designed to run on personal computers are parts of integrated accounting systems. Most of these products do not include item forecasting, lot sizing, and reorder point calculations.

Additional Problems on Deterministic Inventory Models

31. Peet's Coffees in Menlo Park, California, sells Melitta Number 101 coffee filters at a fairly steady rate of about 60 boxes of filters monthly. The filters are ordered from a supplier in Trenton, New Jersey. Peet's manager is interested in applying some inventory theory to determine the best replenishment strategy for the filters.

    Peet's pays $2.80 per box of filters and estimates that fixed costs of employee time for placing and receiving orders amount to about $20. Peet's uses a 22 percent annual interest rate to compute holding costs.

    a. How large a standing order should Peet's have with its supplier in Trenton, and how often should these orders be placed?

    b. Suppose that it takes three weeks to receive a shipment. What inventory of filters should be on hand when an order is placed?

    c. What are the average annual holding and fixed costs associated with these filters, assuming they adopt an optimal policy?

    d. The Peet's store in Menlo Park is rather small. In what way might this affect the solution you recommended in part (*a*)?

32. A local supermarket sells a popular brand of shampoo at a fairly steady rate of 380 bottles per month. The cost of each bottle to the supermarket is 45 cents, and the cost of placing an order has been estimated at $8.50. Assume that holding costs are based on a 25 percent annual interest rate. Stock-outs of the shampoo are not allowed.

    a. Determine the optimal lot size the supermarket should order and the time between placements of orders for this product.

    b. If the procurement lead time is two months, find the reorder point based on the on-hand inventory.

    c. If the item sells for 99 cents, what is the annual profit (exclusive of overhead and labor costs) from this item?

33. Diskup produces a variety of personal computer products. High-density 3.5-inch disks are produced at a rate of 1,800 per day and are shipped out at a rate of 800 per day. The disks are produced in batches. Each disk costs the company 20 cents, and the holding costs are based on an 18 percent annual interest rate. Shortages are not permitted. Each production run of a disk type requires recalibration of the equipment. The company estimates that this step costs $180.

    a. Find the optimal size of each production run and the time between runs.

    b. What fraction of the time is the company producing high-density 3.5-inch disks?

    c. What is the maximum dollar investment in inventory that the company has in these disks?

34. Berry Computer is considering moving some of its operations overseas in order to reduce labor costs. In the United States, its main circuit board costs Berry $75 per unit to produce, while overseas it costs only $65 to produce. Holding costs are based on a 20 percent annual interest rate, and the demand has been a fairly steady 200 units per week. Assume that setup costs are $200 both locally and overseas. Production lead times are one month locally and six months overseas.

    a. Determine the average annual costs of production, holding, and setup at each location, assuming that an optimal solution is employed in each case. Based on these results only, which location is preferable?

    b. Determine the value of the pipeline inventory in each case. (The pipeline inventory is the inventory on order.) Does comparison of the pipeline inventories alter the conclusion reached in part (*a*)?

c. Might considerations other than cost favor local over overseas production?

35. A large producer of household products purchases a glyceride used in one of its deodorant soaps from outside of the company. It uses the glyceride at a fairly steady rate of 40 pounds per month, and the company uses a 23 percent annual interest rate to compute holding costs. The chemical can be purchased from two suppliers, A and B. A offers the following all-units discount schedule:

Order Size	Price per Pound
$0 \leq Q < 500$	\$1.30
$500 \leq Q < 1{,}000$	1.20
$1{,}000 \leq Q$	1.10

whereas B offers the following incremental discount schedule: \$1.25 per pound for all orders less than or equal to 700 pounds, and \$1.05 per pound for all incremental amounts over 700 pounds. Assume that the cost of order processing for each case is \$150. Which supplier should be used?

36. The president of Value Filters became very enthusiastic about using EOQs to plan the sizes of her production runs, and instituted lot sizing based on EOQ values before she could properly estimate costs. For one particular filter line, which had an annual demand of 1,800 units per year and which was valued at \$2.40 per unit, she assumed a holding cost based on a 30 percent annual interest rate and a setup cost of \$100. Some time later, after the cost accounting department had time to perform an analysis, it found that the appropriate value of the interest rate was closer to 20 percent and the setup cost was about \$40. What was the additional average annual cost of holding and setup incurred from the use of the wrong costs?

37. Consider the carload discount schedule pictured in Figure 4–13 (page 225). Suppose $M = 500$ units, $C = \$10$ per unit, and a full carload of 500 units costs \$3,000.

a. Develop a graph of the average cost per unit, $C(Q)/Q$, assuming this schedule.

b. Suppose that the units are consumed at a rate of 800 per week, order setup cost is \$2,500, and holding costs are based on an annual interest charge of 22 percent. Graph the function $G(Q) = \lambda C(Q)/Q + K\lambda/Q + I(C(Q)/Q)Q/2$ and find the optimal value of Q. (Assume that 1 year = 50 weeks.)

c. Repeat part (*b*) for $\lambda = 1{,}000$ per week and $K = \$1{,}500$.

38. Harold Gwynne is considering starting a sandwich-making business from his dormitory room in order to earn some extra income. However, he has only a limited budget of \$100 to make his initial purchases. Harold divides his needs into three areas: breads, meats and cheeses, and condiments. He estimates that he will be able to use all the products he purchases before they spoil, so perishability is not a relevant issue. The demand and cost parameters are as follows.

	Breads	Meats and Cheeses	Condiments
Weekly demand	6 packages	12 packages	2 pounds
Cost per unit	\$0.85	\$3.50	\$1.25
Fixed order cost	\$12	\$8	\$10

The choice of these fixed costs is based on the fact that these items are purchased at different locations in town. They include the cost of Harold's time in making the purchase. Assume that holding costs are based on an annual interest rate of 25 percent.

a. Find the optimal quantities that Harold should purchase of each type of product so that he does not exceed his budget.

b. If Harold could purchase all the items at the same location, would that alter your solution? Why?

39. Mike's Garage, a local automotive service and repair shop, uses oil filters at a fairly steady rate of 2,400 per year. Mike estimates that the cost of his time to make and process an order is about $50. It takes one month for the supplier to deliver the oil filters to the garage, and each one costs Mike $5. Mike uses an annual interest rate of 25 percent to compute his holding cost.

a. Determine the optimal number of oil filters that Mike should purchase, and the optimal time between placement of orders.

b. Determine the level of on-hand inventory at the time a reorder should be placed.

c. Assuming that Mike uses an optimal inventory control policy for oil filters, what is the annual cost of holding and order setup for this item?

40. An import/export firm has leased 20,000 cubic feet of storage space to store six items that it imports from the Far East. The relevant data for the six items it plans to store in the warehouse are

Item	Annual Demand (units)	Cost per Unit	Space per Unit ($feet^3$)
DVD	800	$200.00	12
32-inch flat screen TV	1,600	150.00	18
Blank DVD's (box of 10)	8,000	30.00	3
Blank CD's (box of 50)	12,000	18.00	2
Compact stereo	400	250.00	24
Telephone	1,200	12.50	3

Setup costs for each product amount to $2,000 per order, and holding costs are based on a 25 percent annual interest rate. Find the optimal order quantities for these six items so that the storage space is never exceeded. (Hint: Use a cell location for the Lagrange multiplier and experiment with different values until the storage constraint is satisfied as closely as possible.)

41. A manufacturer of greeting cards must determine the size of production runs for a certain popular line of cards. The demand for these cards has been a fairly steady 2 million per year, and the manufacturer is currently producing the cards in batch sizes of 50,000. The cost of setting up for each production run is $400.

Assume that for each card the material cost is 35 cents, the labor cost is 15 cents, and the distribution cost is 5 cents. The accounting department of the firm has established an interest rate to represent the opportunity cost of alternative investment and storage costs at 20 percent of the value of each card.

a. What is the optimal value of the EOQ for this line of greeting cards?

b. Determine the additional annual cost resulting from using the wrong production lot size.

42. Suppose that in Problem 41 the firm decides to account for the fact that the production rate of the cards is not infinite. Determine the optimal size of each production run assuming that cards are produced at the rate of 75,000 per week.

43. Pies 'R' Us bakes its own pies on the premises in a large oven that holds 100 pies. They sell the pies at a fairly steady rate of 86 per month. The pies cost $2 each to make. Prior to each baking, the oven must be cleaned out, which requires one hour's time for four workers, each of whom is paid $8 per hour. Inventory costs are based on an 18 percent annual interest rate. The pies have a shelf life of three months.

    a. How many pies should be baked for each production run? What is the annual cost of setup and holding for the pies?

    b. The owner of Pies 'R' Us is thinking about buying a new oven that requires one-half the cleaning time of the old oven and has a capacity twice as large as the old one. What is the optimal number of pies to be baked each time in the new oven?

    c. The net cost of the new oven (after trading in the old oven) is $350. How many years would it take for the new oven to pay for itself?

44. The Kirei-Hana Japanese Steak House in San Francisco consumes 3,000 pounds of sirloin steak each month. Yama Hirai, the new restaurant manager, recently completed an MBA degree. He learned that the steak was replenished using an EOQ value of 2,000 pounds. The EOQ value was computed assuming an interest rate of 36 percent per year. Assume that the current cost of the sirloin steak to the steak house is $4 per pound.

    a. What is the setup cost used in determining the EOQ value?

    Mr. Hirai received an offer from a meat wholesaler in which a discount of 5 percent would be given if the steak house purchased the steak in quantities of 3,000 pounds or more.

    b. Should Mr. Hirai accept the offer from the wholesaler? If so, how much can be saved?

    c. Because of Mr. Hirai's language problems, he apparently misunderstood the offer. In fact, the 5 percent discount is applied only to the amounts ordered over 3,000 pounds. Should this offer be accepted, and if so, how much is now saved?

45. Green's Buttons of Rolla, Missouri, supplies all the New Jersey Fabrics stores with eight different styles of buttons for men's dress shirts. The plastic injection molding machine can produce only one button style at a time and substantial time is required to reconfigure the machine for different button styles. As Green's has contracted to supply fixed quantities of buttons for the next three years, its demand can be treated as fixed and known. The relevant data for this problem appear in the following table.

Button Type	Annual Sales	Production Rate (units/day)	Setup Time (hours)	Variable Cost
A	25,900	4,500	6	$0.003
B	42,000	5,500	4	0.002
C	14,400	3,300	8	0.008
D	46,000	3,200	4	0.002
E	12,500	1,800	3	0.010
F	75,000	3,900	6	0.005
G	30,000	2,900	1	0.004
H	18,900	1,200	3	0.007

Assume 250 working days per year. Green's accounting department has established an 18 percent annual interest rate for the cost of capital and a 3 percent annual interest rate to account for storage space. Setup costs are $20 per hour required to reconfigure the equipment for a new style. Suppose that the firm decides to use a rotation cycle policy for production of the buttons.

a. What is the optimal rotation cycle time?

b. How large should the lots be?

c. What is the average annual cost of holding and setups at the optimal solution?

d. What contractual obligations might Green's have with New Jersey Fabrics that would prevent them from implementing the policy you determined in parts (*a*) and (*b*)? More specifically, if Green's agreed to make three shipments per year for each button style, what production policy would you recommend?

Appendix 4–A

Mathematical Derivations for Multiproduct Constrained EOQ Systems

Consider the standard multiproduct EOQ system with a budget constraint, which was discussed in Section 4.8. Mathematically, the problem is to find values of the variables $Q_1, Q_2, \ldots, Q_n$ to

$$\text{Minimize} \sum_{i=1}^{n} \left[\frac{h_i Q_i}{2} + \frac{K_i \lambda_i}{Q_i} \right]$$

subject to

$$\sum_{i=1}^{n} c_i Q_i \leq C.$$

Let EOQ_i be the respective unconstrained EOQ values. Then there are two possibilities:

$$\sum_{i=1}^{n} c_i \text{EOQ}_i \leq C, \tag{1}$$

$$\sum_{i=1}^{n} c_i \text{EOQ}_i > C. \tag{2}$$

If Equation (1) holds, the optimal solution is the trivial solution; namely, set $Q_i = \text{EOQ}_i$. If Equation (2) holds, then we are guaranteed that the constraint is binding at the optimal solution. That means that the constraint may be written

$$\sum_{i=1}^{n} c_i Q_i = C.$$

In this case, we introduce the Lagrange multiplier θ, and the problem is now to find $Q_1, Q_2, \ldots, Q_n$, and θ to solve the unconstrained problem:

$$\text{Minimize } G(Q_1, Q_2, \ldots, Q_n, \theta) = \sum_{i=1}^{n} \left(\frac{h_i Q_i}{2} + \frac{K_i \lambda_i}{Q_i} \right) + \theta \sum_{i=1}^{n} (c_i Q_i - C).$$

Necessary conditions for optimality are that

$$\frac{\partial G}{\partial Q_i} = 0 \qquad \text{for } i = 1, \ldots, n$$

and

$$\frac{\partial G}{\partial \theta} = 0.$$

The first n conditions give

$$\frac{h_i}{2} - \frac{K_i \lambda_i}{Q_i^2} + \theta c_i = 0 \quad \text{for } i = 1, \ldots, n.$$

Rearranging terms, we get

$$Q_i = \sqrt{\frac{2K_i \lambda_i}{h_i + 2\theta c_i}} \qquad \text{for } i = 1, \ldots, n,$$

and we also have the final condition

$$\sum_{i=1}^{n} c_i Q_i = C.$$

Now consider the case where $c_i/h_i = c/h$ independent of i. By dividing the numerator and the denominator by h_i, we may write

$$\begin{aligned} Q_i &= \sqrt{\frac{2K_i \lambda_i}{h_i}} \sqrt{\frac{1}{1 + 2\theta c/h}} \\ &= \text{EOQ}_i \sqrt{\frac{1}{1 + 2\theta c/h}} \\ &= \text{EOQ}_i m, \end{aligned}$$

where

$$m = \sqrt{\frac{1}{1 + 2\theta c/h}}.$$

Substituting this expression for Q_i into the constraint gives

$$\sum_{i=1}^{n} c_i \text{EOQ}_i m = C$$

or

$$m = \frac{C}{\sum_{i=1}^{n} c_i \text{EOQ}_i}.$$

Appendix 4–B

Glossary of Notation for Chapter 4

c = Proportional order cost.
EOQ = Economic order quantity (optimal lot size).
$G(Q)$ = Average annual cost associated with lot size Q.
h = Holding cost per unit time.
h' = Modified holding cost for finite production rate model.
I = Annual interest rate used to compute holding cost.
K = Setup cost or fixed order cost.
λ = Demand rate (units per unit time).
P = Production rate for finite production rate model.
Q = Lot size or size of the order.
s_i = Setup time for product i (refer to Section 4.9).
T = Cycle time; time between placement of successive orders.
τ = Order lead time.
θ = Lagrange multiplier for space-constrained model (refer to Section 4.8).
w_i = Space consumed by one unit of product i (refer to Section 4.8).
W = Total space available (refer to Section 4.8).

Bibliography

Arrow, K. A.; S. Karlin; and H. Scarf, eds. *Studies in the Mathematical Theory of Inventory Production*. Stanford, CA: Stanford University Press, 1958.

Barbosa, L. C., and M. Friedman. "Deterministic Inventory Lot Size Models—A General Root Law." *Management Science* 23 (1978), pp. 820–29.

Bowman, E. H., and R. B. Fetter. *Analysis for Production Management*. New York: McGraw-Hill/Irwin, 1961.

Churchman, C. W.; R. L. Ackoff; and E. L. Arnoff. *Introduction to Operations Research*. New York: John Wiley & Sons, 1957.

Donnelly, H. "Technology Awards: Recognizing Retailing's Best." *Stores* 76, no. 10 (1994), pp. 52–56.

Fetter, R. B., and W. C. Dalleck. *Decision Models for Inventory Management*. New York: McGraw-Hill/Irwin, 1961.

Hadley, G. J., and T. M. Whitin. *Analysis of Inventory Systems*. Englewood Cliffs, NJ: Prentice Hall, 1963.

Hanssmann, F. "A Survey of Inventory Theory from the Operations Research Viewpoint." In *Progress in Operations Research* 1, ed. R. L. Ackoff. New York: John Wiley & Sons, 1961.

Harris, F. W. *Operations and Cost* (Factory Management Series). Chicago: Shaw, 1915.

Magee, J. F., and D. M. Boodman. *Production Planning and Inventory Control*. 2nd ed. New York: McGraw-Hill, 1967.

Maxwell, W. L. "The Scheduling of Economic Lot Sizes." *Naval Research Logistics Quarterly* 11 (1964), pp. 89–124.

Muckstadt, J. A., and R. O. Roundy. "Analysis of Multi-Stage Production Systems." Chapter 2 in *Logistics of Production and Inventory*, eds. S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin. Volume 4 of *Handbooks in Operations Research and Management Science*, Amsterdam: North-Holland, 1993.

Rosenblatt, M. J. "Multi-Item Inventory System with Budgetary Constraint: A Comparison between the Lagrangian and the Fixed Cycle Approach." *International Journal of Production Research* 19 (1981), pp. 331–39.

Roundy, R. O. "98%-Effective Integer-Ratio Lot-Sizing for One Warehouse Multi-Retailer Systems." *Management Science* 31 (1985), pp. 1416–30.

Scarf, H. E.; D. M. Gilford; and M. W. Shelly. *Multistage Inventory Models and Techniques*. Stanford, CA: Stanford University Press, 1963.

Schwarz, L. B., and L. Schrage. "Optimal and Systems Myopic Policies for Multiechelon Production/Inventory Assembly Systems." *Management Science* 21 (1971), pp. 1285–94.

Starr, M. K., and D. W. Miller. *Inventory Control: Theory and Practice*. Englewood Cliffs, NJ: Prentice Hall, 1962.

Wagner, H. M. *Statistical Management of Inventory Systems*. New York: John Wiley & Sons, 1962.

Whitin, T. M. *The Theory of Inventory Management*. Rev. ed. Princeton, NJ: Princeton University Press, 1957.

Wilson, R. H. "A Scientific Routine for Stock Control." *Harvard Business Review* 13 (1934), pp. 116–28.

Chapter **Five**

Inventory Control Subject to Uncertain Demand

> "Knowing what you've got, knowing what you need, knowing what you don't—that's inventory control."
>
> —Frank Wheeler in *Revolutionary Road*

Chapter Overview

Purpose

To understand how one deals with uncertainty (randomness) in the demand when computing replenishment policies for a single inventory item.

Key Points

1. *What is uncertainty and when should it be assumed?* Uncertainty means that demand is a random variable. A random variable is defined by its probability distribution, which is generally estimated from a past history of demands. In practice, it is common to assume that demand follows a normal distribution. When demand is assumed normal, one only needs to estimate the mean, μ, and variance, σ^2. Clearly, demand is uncertain to a greater or lesser extent in all real-world applications. What value, then, does the analysis of Chapters 3 and 4 have, where demand was assumed known? Chapter 3 focused on *systematic* or predictable changes in the demand pattern, such as peaks and valleys. Chapter 4 results for single items are useful if the variance of demand is low relative to the mean. In this chapter we consider items whose primary variation is due to uncertainty rather than predictable causes.

   If demand is described by a random variable, it is unclear what the optimization criterion should be, since the cost function is a random variable as well. To handle this, we assume that the objective is to minimize *expected* costs. The use of the expectation operator is justified by the law of large numbers from probability, since an inventory control problem invariably spans many planning periods. The law of large numbers guarantees that the arithmetic average of the incurred costs and the expected costs grow close as the number of planning periods gets large.
2. *The newsvendor model.* Consider a news vendor that decides each morning how many papers to buy to sell during the day. Since daily demand is highly variable, it is modeled with a random variable, *D.* Suppose that *Q* is the number of papers he purchases. If *Q* is too large, he is left with unsold papers, and if *Q* is too small, some demands go unfilled. If we let c_o be the unit overage cost, and c_u be the

unit underage cost, then we show that the optimal number of papers he should purchase at the start of a day, say Q^*, satisfies:

$$F(Q^*) = c_u/(c_u + c_o)$$

where $F(Q^*)$ is the cumulative distribution function of D evaluated at Q^* (which is the same as the probability that demand is less than or equal to Q^*).

3. *Lot size–reorder point systems.* The newsvendor model is appropriate for a problem that essentially restarts from scratch every period. Yesterday's newspaper has no value in the market, save for the possible scrap value of the paper itself. However, most inventory control situations that one encounters in the real world are not like this. Unsold items continue to have value in the marketplace for many periods. For these cases we use an approach that is essentially an extension of the EOQ model of Chapter 4.

   The lot size–reorder point system relies on the assumption that inventories are reviewed continuously rather than periodically. That is, the state of the system is known at all times. The system consists of two decision variables: Q and R. Q is the order size and R is the reorder point. That is, when the inventory of stock on hand reaches R, an order for Q units is placed. The model also allows for a positive order lead time, τ. It is the demand over the lead time that is the key uncertainty in the problem, since the lead time is the response time of the system. Let D represent the demand over the lead time, and let $F(t)$ be the cumulative distribution function of D. Cost parameters include a fixed order cost K, a unit penalty cost for unsatisfied demand p, and a per unit per unit time holding cost h. Interpret λ as the average annual demand rate (that is, the expected demand over a year). Then we show in this section that the optimal values of Q and R satisfy the following two simultaneous nonlinear equations:

$$Q = \sqrt{\frac{2\lambda[K + pn(R)]}{h}}$$

$$1 - F(R) = Qh/p\lambda.$$

   The solution to these equations requires a back-and-forth iterative solution method. We provide details of the method only when the lead time demand distribution is normal. Convergence generally occurs quickly. A quick and dirty approximation is to set Q = EOQ and solve for R in the second equation. This will give good results in most cases.

4. *Service levels in (Q, R) systems.* We assume two types of service: Type 1 service is the probability of not stocking out in the lead time and is represented by the symbol α. Type 2 service is the proportion of demands that are filled from stock (also known as the fill rate) and is represented by the symbol β. Finding the optimal (Q, R) subject to a Type 1 service objective is very easy. One merely finds R from $F(R) = \alpha$ and sets Q = EOQ. Unfortunately, what one generally means by service is the Type 2 criterion, and finding (Q, R) in that case is more difficult. For Type 2 service, we only consider the normal distribution. The solution requires using standardized loss tables, $L(z)$, which are supplied in the back of the book. As with the cost model, setting Q = EOQ and solving for R will usually give good results if one does not want to bother with an iterative procedure.

In this chapter, we consider the link between inventory control and forecasting, and how one typically updates estimates of the mean and standard deviation of demand using exponential smoothing. The section concludes with a discussion of lead time variability, and how that additional uncertainty is taken into account.

5. *Periodic review systems under uncertainty*. The newsvendor model treats a product that perishes quickly (after one period). However, periodic review models also make sense when unsold product can be used in future periods. In this case the form of the optimal policy is known as an (s, S) policy. Let u be the starting inventory in any period. Then the (s, S) policy is

   If $u \leq s$, order to S (that is, order $S - u$).
   If $u > s$, don't order.

   Unfortunately, finding the optimal values of (s, S) each period is much more difficult than finding the optimal (Q, R) policy, and is beyond the scope of this book. We also briefly discuss service levels in periodic review systems.
6. *Multiproduct systems*. Virtually all inventory control problems occurring in the operations planning context involve multiple products. One issue that arises in multiproduct systems is determining the amount of effort one should expend managing each item. Clearly, some items are more valuable to the business than others. The ABC classification system is one means of ranking items. Items are sequenced in decreasing order of annual dollar volume of sales or usage. Ordering the items in this way, and graphing the cumulative dollar volume gives an exponentially increasing curve known as a Pareto curve. Typically, 20 percent of the items account for 80 percent of the annual dollar volume (A items), the next 30 percent of the items typically account for the next 15 percent of the dollar volume (B items), and the final 50 percent of the items account for the final 5 percent of the dollar volume (C items). A items should receive the most attention. Their inventory levels should be reviewed often, and they should carry a high service level. B items do not need such close scrutiny, and C items are typically ordered infrequently in large quantities.
7. *Other issues.* The discussion of stochastic inventory models in this chapter barely reveals the tip of the iceberg in terms of the vast quantity of research done on this topic. Two important areas of research are multi-echelon inventory systems, and perishable inventory systems. A multi-echelon inventory system is one in which items are stored at multiple locations linked by a network. Supply chains, discussed in detail in Chapter 6, are such a system. Another important area of research are items that change during storage, thus affecting their useful lifetime. One class of such items are perishable items. Perishable items have a fixed lifetime known in advance, and include food, pharmaceuticals and photographic film. A related problem is managing items subject to obsolescence. Obsolescence differs from perishability in that the useful lifetime of an item subject to obsolescence cannot be predicted in advance. Mathematical models for analyzing such problems are quite complex and well beyond the scope of this book.

The management of uncertainty plays an important role in the success of any firm. What are the sources of uncertainty that affect a firm? A partial list includes uncertainty in consumer preferences and trends in the market, uncertainty in the availability

and cost of labor and resources, uncertainty in vendor resupply times, uncertainty in weather and its ramifications on operations logistics, uncertainty of financial variables such as stock prices and interest rates, and uncertainty of demand for products and services.

Before the terrible tragedy of September 11, 2001, many of us would arrive at the airport no more than 30 or 40 minutes before the scheduled departure time of our flight. Now we might arrive two hours before the flight. Increased airport security has not only increased the average time required, it has increased the uncertainty of this time. To compensate for this increased uncertainty, we arrive far in advance of our scheduled departure time to provide a larger buffer time. This same principle will apply when managing inventories.

The uncertainty of demand and its effect on inventory management strategies are the subjects of this chapter. When a quantity is uncertain, it means that we cannot predict its value exactly in advance. For example, a department store cannot exactly predict the sales of a particular item on any given day. An airline cannot exactly predict the number of people that will choose to fly on any given flight. How, then, can these firms choose the number of items to keep in inventory or the number of flights to schedule on any given route?

Although exact sales of an item or numbers of seats booked on a plane cannot be predicted in advance, one's past experience can provide useful information for planning. As shown in Section 5.1, previous observations of any random phenomenon can be used to estimate its *probability distribution*. By properly quantifying the consequences of incorrect decisions, a well-thought-out mathematical model of the system being studied will result in intelligent strategies. When uncertainty is present, the objective is almost always to minimize expected cost or maximize expected profits.

Demand uncertainty plays a key role in many industries, but some are more susceptible to business cycles than others. The world economy saw one of the worst recessions in recent times in 2008. The stock market eventually dropped to about half of its high in 2007 and unemployment levels soared. The retailing industry in particular is very sensitive to the vicissitudes of consumer demand. Low cost providers such as Costco and Walmart fared well, but high-end retailers such as Nordstrom and Bloomingdales suffered. Matching supply and demand becomes even more critical in times of recession.

As some level of demand uncertainty seems to characterize almost all inventory management problems in practice, one might question the value of the deterministic inventory control models discussed in Chapter 4. There are two reasons for studying deterministic models. One is that they provide a basis for understanding the fundamental trade-offs encountered in inventory management. Another is that they may be good approximations depending on the degree of uncertainty in the demand.

To better understand the second point, let D be the demand for an item over a given period of time. We express D as the sum of two parts, D_{Det} and D_{Ran}. That is,

$$D = D_{\text{Det}} + D_{\text{Ran}},$$

where

$$D_{\text{Det}} = \text{Deterministic component of demand,}$$
$$D_{\text{Ran}} = \text{Random component of demand.}$$

There are a number of circumstances under which it would be appropriate to treat D as being deterministic even though D_{Ran} is not zero. Some of these are

1. When the variance of the random component, D_{Ran}, is small relative to the magnitude of D.
2. When the predictable variation is more important than the random variation.
3. When the problem structure is too complex to include an explicit representation of randomness in the model.

An example of circumstance 2 occurs in the aggregate planning problem. Although the forecast error of the aggregate demands over the planning horizon may not be zero, we are more concerned with planning for the anticipated changes in the demand than for the unanticipated changes. An example of circumstance 3 occurs in material requirements planning (treated in detail in Chapter 7). The intricacies of the relationships among various component levels and end items make it difficult to incorporate demand uncertainty into the analysis.

However, for many items, the random component of the demand is too significant to ignore. As long as the expected demand per unit time is relatively constant and the problem structure not too complex, explicit treatment of demand uncertainty is desirable. In this chapter we will examine several of the most important stochastic inventory models and the key issues surrounding uncertainty.[1]

Overview of Models Treated in This Chapter

Inventory control models subject to uncertainty are basically of two types: (1) **periodic review** and (2) **continuous review.** (Recall the discussion at the start of Chapter 4. Periodic review means that the inventory level is known at discrete points in time only, and continuous review means that the inventory level is known at all times.) Periodic review models may be for one planning period or for multiple planning periods. For one-period models, the objective is to properly balance the costs of overage (ordering too much) and underage (ordering too little). Single-period models are useful in several contexts: planning for initial shipment sizes for high-fashion items, ordering policies for food products that perish quickly, or determining run sizes for items with short useful lifetimes, such as newspapers. Because of this last application, the single-period stochastic inventory model has come to be known as the **newsvendor model.** The newsvendor model will be the first one considered in this chapter.

From this writer's experience, the vast majority of computer-based inventory control systems on the market use some variant of the continuous review models treated in the remainder of this chapter. They are, in a sense, extensions of the EOQ model to incorporate uncertainty. Their popularity in practice is attributable to several factors. First, the policies are easy to compute and easy to implement. Second, the models accurately describe most systems in which there is ongoing replenishment of inventory items under uncertainty. A detailed discussion of service levels is included as well. Because estimating penalty costs is difficult in practice, service level approaches are more frequently implemented than penalty cost approaches.

Multiperiod stochastic inventory models dominate the professional literature on inventory theory. There are enough results in this fascinating research area to compose a

[1] For those unfamiliar with it, the word *stochastic* is merely a synonym for *random.*

volume in its own right. However, our goal in this book is to concentrate on methodology that has been applied in the real world. Although they provide insight, multiperiod stochastic models are rarely implemented. In addition, the level of mathematical sophistication they require is beyond that of this book. For these two reasons, multiperiod inventory models subject to uncertainty are not considered here.

5.1 THE NATURE OF RANDOMNESS

In order to clarify what the terms *randomness* and *uncertainty* mean in the context of inventory control, we begin with an example.

Example *5.1*

On consecutive Sundays, Mac, the owner of a local newsstand, purchases a number of copies of *The Computer Journal,* a popular weekly magazine. He pays 25 cents for each copy and sells each for 75 cents. Copies he has not sold during the week can be returned to his supplier for 10 cents each. The supplier is able to salvage the paper for printing future issues. Mac has kept careful records of the demand each week for the *Journal.* (This includes the number of copies actually sold plus the number of customer requests that could not be satisfied.) The observed demands during each of the last 52 weeks were

15	19	9	12	9	22	4	7	8	11
14	11	6	11	9	18	10	0	14	12
8	9	5	4	4	17	18	14	15	8
6	7	12	15	15	19	9	10	9	16
8	11	11	18	15	17	19	14	14	17
13	12								

There is no discernible pattern to these data, so it is difficult to predict the demand for the *Journal* in any given week. However, we can represent the demand experience of this item as a frequency histogram, which gives the number of times each weekly demand occurrence was observed during the year. The histogram for this demand pattern appears in Figure 5–1.

One uses the frequency histogram to estimate the probability that the number of copies of the *Journal* sold in any week is a specific value. These probability estimates are obtained by dividing the number of times that each demand occurrence was observed during the year by 52. For example, the probability that demand is 10 is estimated to be $2/52 = .0385$, and the

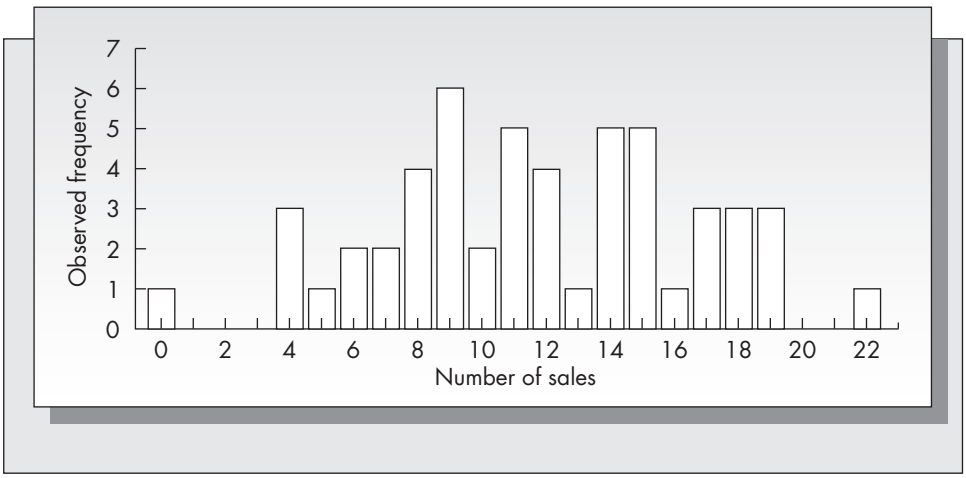


FIGURE 5–1
Frequency histogram for a 52-week history of sales of *The Computer Journal* at Mac's

probability that the demand is 15 is 5/52 = .0962. The collection of all the probabilities is known as the *empirical probability distribution*. Cumulative probabilities also can be estimated in a similar way. For example, the probability that there are nine or fewer copies of the *Journal* sold in any week is (1 + 0 + 0 + 0 + 3 + 1 + 2 + 2 + 4 + 6) = 19/52 = .3654.

Although empirical probabilities can be used in subsequent analysis, they are inconvenient for a number of reasons. First, they require maintaining a record of the demand history for every item. This can be costly and cumbersome. Second, the distribution must be expressed (in this case) as 23 different probabilities. Other items may have an even wider range of past values. Finally, it is more difficult to compute optimal inventory policies with empirical distributions.

For these reasons, we generally approximate the demand history using a continuous distribution. The form of the distribution chosen depends upon the history of past demand and its ease of use. By far the most popular distribution for inventory applications is the normal. One reason is the frequency with which it seems to accurately model demand fluctuations. Another is its convenience. The normal model of demand must be used with care, however, as it admits the possibility of negative values. When using the normal distribution to describe a nonnegative phenomenon such as demand, the likelihood of a negative observation should be sufficiently small (less than .01 should suffice for most applications) so as not to be a factor.

A normal distribution is determined by two parameters: the mean μ and the variance σ^2. These can be estimated from a history of demand by the sample mean $\overline{D}$ and the sample variance s^2. Let $D_1, D_2, \ldots, D_n$ be n past observations of demand. Then

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i,$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (D_i - \overline{D})^2.$$

For the data pictured in Figure 5–1 we obtain

$$\overline{D} = 11.73,$$

$$s = 4.74.$$

The normal density function, $f(x)$, is given by the formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \text{for } -\infty < x < +\infty.$$

We substitute $\overline{D}$ as the estimator for μ and s as the estimator for σ.

The *relative frequency histogram* is the same as the frequency histogram pictured in Figure 5–1, except that the y-axis entries are divided by 52. In Figure 5–2 we show the normal density function that results from the substitution we made, superimposed on the relative frequency histogram.

In practice, exponential smoothing is used to recursively update the estimates of the mean and the standard deviation of demand. The standard deviation is estimated using the mean absolute deviation (MAD). Both exponential smoothing and MAD are discussed in detail in Chapter 2. Let $\overline{D}_t$ be the estimate of the mean after observing demand D_t, and let MAD_t be the estimate of the MAD. Then

$$\overline{D}_t = \alpha D_t + (1-\alpha)\overline{D}_{t-1},$$

$$\text{MAD}_t = \alpha|D_t - \overline{D}_{t-1}| + (1-\alpha)\text{MAD}_{t-1},$$

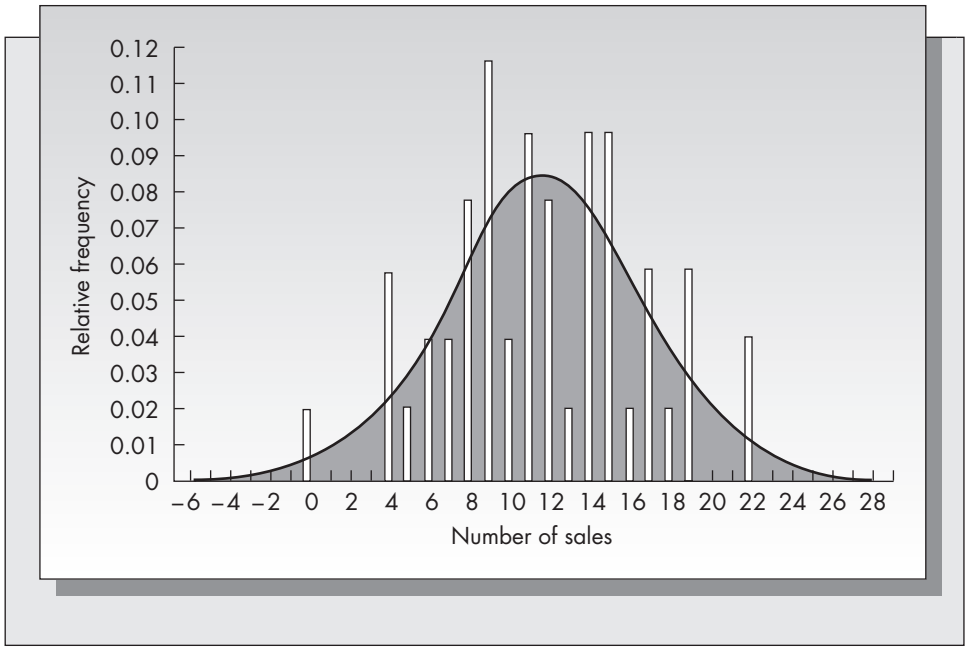


FIGURE 5–2
Frequency histogram and normal approximation

where $0 < \alpha < 1$ is the smoothing constant. For normally distributed demand

$$\sigma \approx 1.25 * \text{MAD}.$$

A smoothing constant of $\alpha \approx .1$ is generally used to ensure stability in the estimates (see Chapter 2 for a more complete discussion of the issues surrounding the choice of the smoothing constant).

5.2 OPTIMIZATION CRITERION

In general, optimization in production problems means finding a control rule that achieves minimum cost. However, when demand is random, the cost incurred is itself random, and it is no longer obvious what the optimization criterion should be. Virtually all stochastic optimization techniques applied to inventory control assume that the goal is to minimize *expected* costs.

The motivation for using the expected value criterion is that inventory control problems are generally ongoing problems. Decisions are made repetitively. The law of large numbers from probability theory says that the arithmetic average of many observations of a random variable will converge to the expected value of that random variable. In the context of the inventory problem, if we follow a control rule that minimizes expected costs, then the arithmetic average of the actual costs incurred over many periods will also be a minimum.

In certain circumstances, the expected value may not be the best optimization criterion. When a product is acquired only once and not on an ongoing basis, it is not clear that minimizing expected costs is appropriate. In such a case, maximizing the

probability of some event (such as satisfying a proportion of the demand) is generally more suitable. However, because of the ongoing nature of most production problems, the expected value criterion is used in virtually all stochastic inventory control applications.

Problems for Sections 5.1 and 5.2

1. Suppose that Mac only kept track of the number of magazines sold. Would this give an accurate representation of the demand for the *Journal*? Under what circumstances would the actual demand and the number sold be close, and under what circumstances would they differ by a substantial amount?
2. What is the difference between deterministic and random variations in the pattern of demands? Provide an example of a real problem in which predictable variation would be important and an example in which random variation would be important.
3. Oakdale Furniture Company uses a special type of woodworking glue in the assembly of its furniture. During the past 36 weeks, the following amounts of glue (in gallons) were used by the company:

25	38	26	31	21	46	29	19	35	39	24	21
17	42	46	19	50	40	43	34	31	51	36	32
18	29	22	21	24	39	46	31	33	34	30	30

   a. Compute the mean and the standard deviation of this sample.

   b. Consider the following class intervals for the number of gallons used each week:

   Less than 20.
   20–27.
   28–33.
   34–37.
   38–43.
   More than 43.

   Determine the proportion of data points that fall into each of these intervals. Compare these proportions to the probabilities that a normal variate with the mean and the standard deviation you computed in part (*a*) falls into each of these intervals. Based on the comparison of the observed proportions and those obtained from assuming a normal distribution, would you conclude that the normal distribution provides an adequate fit of these data? (This procedure is essentially the same as a chi-square goodness-of-fit test.)

   c. Assume that the numbers of gallons of glue used each week are independent random variables, having the normal distribution with mean and standard deviation computed in part (*a*). What is the probability that the total number of gallons used in six weeks does not exceed 200 gallons? (Hint: The mean of a sum of random variables is the sum of the means and the *variance* of a sum of *independent* random variables is the sum of the variances.)

4. In Problem 3, what other probability distributions might accurately describe Oakdale's weekly usage of glue?
5. Rather than keeping track of each demand observation, Betty Sucasas, a member of the marketing staff with a large company that produces a line of switches, has kept only grouped data. For switch C9660Q, used in small power supplies, she has observed the following numbers of units of the switch shipped over the last year.

Units Shipped	Number of Weeks
0–2,000	3
2,001–5,000	6
5,001–9,000	12
9,001–12,000	17
12,001–18,000	10
18,001–20,000	4

Based on these observations, estimate the mean and the standard deviation of the weekly shipments. (Hint: This is known as grouped data. For the purposes of your calculation, assume that all observations occur at the midpoint of each interval.)

6. *a.* Consider the Oakdale Furniture Company described in Problem 3. Under what circumstances might the major portion of the usage of the glue be predictable?
   b. If the demand were predictable, would you want to use a probability law to describe it? Under what circumstances might the use of a probability model of demand be justified even if the demand could be predicted exactly?

5.3 THE NEWSVENDOR MODEL

Let us return to Example 5.1, in which Mac wishes to determine the number of copies of *The Computer Journal* he should purchase each Sunday. A study of the historical data showed that the demand during any week is a random variable that is approximately normally distributed, with mean 11.73 and standard deviation 4.74. Each copy is purchased for 25 cents and sold for 75 cents, and he is paid 10 cents for each unsold copy by his supplier. One obvious solution is that he should buy enough to meet the mean demand, which is approximately 12 copies. There is something wrong with this solution. Suppose Mac purchases a copy that he does not sell. His out-of-pocket expense is 25 cents − 10 cents = 15 cents. Suppose, on the other hand, that he is unable to meet the demand of a customer. In that case, he loses 75 cents − 25 cents = 50 cents profit. Hence, there is a significantly greater penalty for not having enough than there is for having too much. If he only buys enough to satisfy mean demand, he will stock-out with the same frequency that he has an oversupply. Our intuition tells us that he should buy more than the mean, but how much more? This question is answered in this section.

Notation

This problem is an example of the newsvendor model, in which a single product is to be ordered at the beginning of a period and can be used only to satisfy demand during that

period. Assume that all relevant costs can be determined on the basis of ending inventory. Define

c_o = Cost per unit of positive inventory remaining at the end of the period (known as the *overage cost*).

c_u = Cost per unit of unsatisfied demand. This can be thought of as a cost per unit of negative ending inventory (known as the *underage cost*).

In the development of the model, we will assume that the demand D is a continuous nonnegative random variable with density function $f(x)$ and cumulative distribution function $F(x)$. [A brief review of probability theory is given in Appendix 5–A. In particular, both $F(x)$ and $f(x)$ are defined there.]

The *decision variable* Q is the number of units to be purchased at the beginning of the period. The goal of the analysis is to determine Q to minimize the expected costs incurred at the end of the period.

Development of the Cost Function

A general outline for analyzing most stochastic inventory problems is the following:

1. Develop an expression for the cost incurred as a function of both the random variable D and the decision variable Q.
2. Determine the expected value of this expression with respect to the density function or probability function of demand.
3. Determine the value of Q that minimizes the expected cost function.

Define $G(Q, D)$ as the total overage and underage cost incurred at the end of the period when Q units are ordered at the start of the period and D is the demand. If Q units are purchased and D is the demand, $Q - D$ units are left at the end of the period as long as $Q \geq D$. If $Q < D$, then $Q - D$ is negative and the number of units remaining on hand at the end of the period is 0. Notice that

$$\max\{Q - D, 0\} = \begin{cases} Q - D & \text{if } Q \geq D, \\ 0 & \text{if } Q \leq D. \end{cases}$$

In the same way, $\max\{D - Q, 0\}$ represents the excess demand over the supply, or the unsatisfied demand remaining at the end of the period. For any realization of the random variable D, either one or the other of these terms will be zero.

Hence, it now follows that

$$G(Q, D) = c_o \max(0, Q - D) + c_u \max(0, D - Q).$$

Next, we derive the expected cost function. Define

$$G(Q) = E(G(Q, D)).$$

Using the rules outlined in Appendix 5–A for taking the expected value of a function of a random variable, we obtain

$$\begin{aligned} G(Q) &= c_o \int_0^\infty \max(0, Q - x) f(x)\, dx + c_u \int_0^\infty \max(0, x - Q) f(x)\, dx \\ &= c_o \int_0^Q (Q - x) f(x)\, dx + c_u \int_Q^\infty (x - Q) f(x)\, dx. \end{aligned}$$

Determining the Optimal Policy

We would like to determine the value of Q that minimizes the expected cost $G(Q)$. In order to do so, it is necessary to obtain an accurate description of the function $G(Q)$. We have that

$$\frac{dG(Q)}{dQ} = c_o \int_0^Q 1 f(x)\, dx + c_u \int_Q^\infty (-1) f(x)\, dx$$

$$= c_o F(Q) - c_u(1 - F(Q)).$$

(This is a result of Leibniz's rule, which indicates how one differentiates integrals. Leibniz's rule is stated in Appendix 5–A.)

It follows that

$$\frac{d^2G(Q)}{dQ^2} = (c_o + c_u) f(Q) \geq 0 \qquad \text{for all } Q \geq 0.$$

Because the second derivative is nonnegative, the function $G(Q)$ is said to be *convex* (bowl shaped). We can obtain additional insight into the shape of $G(Q)$ by further analysis. Note that

$$\left.\frac{dG(Q)}{dQ}\right|_{Q=0} = c_o F(0) - c_u(1 - F(0))$$

$$= -c_u < 0 \qquad \text{since } F(0) = 0.$$

Since the slope is negative at $Q = 0$, $G(Q)$ is decreasing at $Q = 0$. The function $G(Q)$ is pictured in Figure 5–3.

It follows that the optimal solution, say Q^*, occurs where the first derivative of $G(Q)$ equals zero. That is,

$$G'(Q^*) = (c_o + c_u) F(Q^*) - c_u = 0.$$

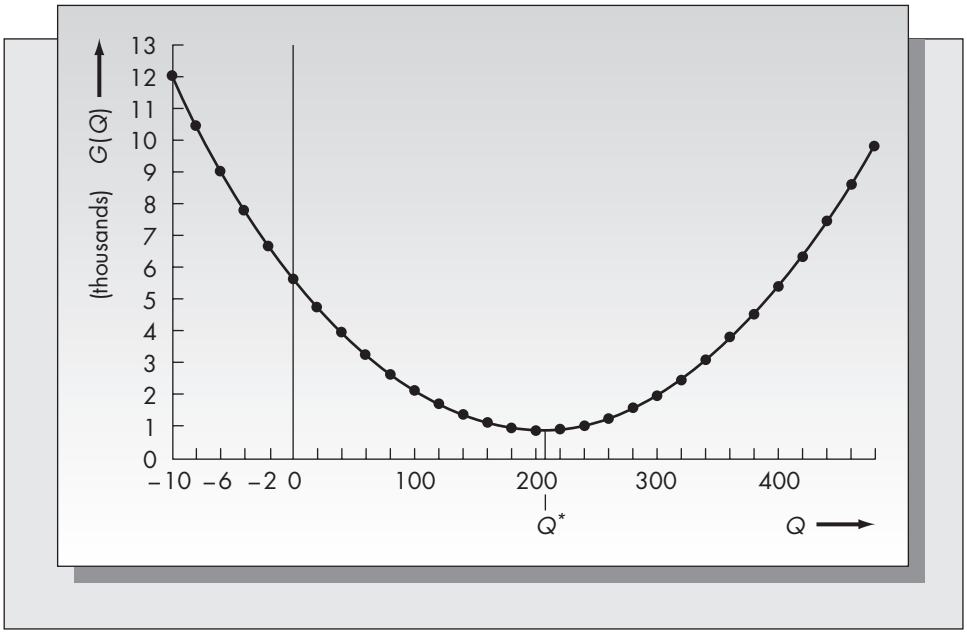


FIGURE 5–3
Expected cost function for newsvendor model

Rearranging terms gives

$$F(Q^*) = c_u/(c_o + c_u).$$

We refer to the right-hand side of the last equation as the *critical ratio.* Because c_u and c_o are positive numbers, the critical ratio is strictly between zero and one. This implies that for a continuous demand distribution this equation is always solvable.

As $F(Q^*)$ is defined as the probability that the demand does not exceed Q^*, the critical ratio is the probability of satisfying all the demand during the period if Q^* units are purchased at the start of the period. It is important to understand that this is *not* the same as the proportion of satisfied demands. When underage and overage costs are equal, the critical ratio is exactly one-half. In that case Q^* corresponds to the *median* of the demand distribution. When the demand density is symmetric (such as the normal density), the mean and the median are the same.

Example *5.1* *(continued)*

Consider the example of Mac's newsstand. From past experience, we saw that the weekly demand for the *Journal* is approximately normally distributed with mean $\mu = 11.73$ and standard deviation $\sigma = 4.74$. Because Mac purchases the magazines for 25 cents and can salvage unsold copies for 10 cents, his overage cost is $c_o = 25 - 10 = 15$ cents. His underage cost is the profit on each sale, so that $c_u = 75 - 25 = 50$ cents. The critical ratio is $c_u/(c_o + c_u) = 0.50/0.65 = .77$. Hence, he should purchase enough copies to satisfy all the weekly demand with probability .77. The optimal Q^* is the 77th percentile of the demand distribution (see Figure 5–4).

Using either Table A–1 or Table A–4 at the back of the book, we obtain a standardized value of $z = 0.74$. The optimal Q is

$$Q^* = \sigma z + \mu = (4.74)(0.74) + 11.73$$
$$= 15.24 \approx 15.$$

Hence, he should purchase 15 copies every week.

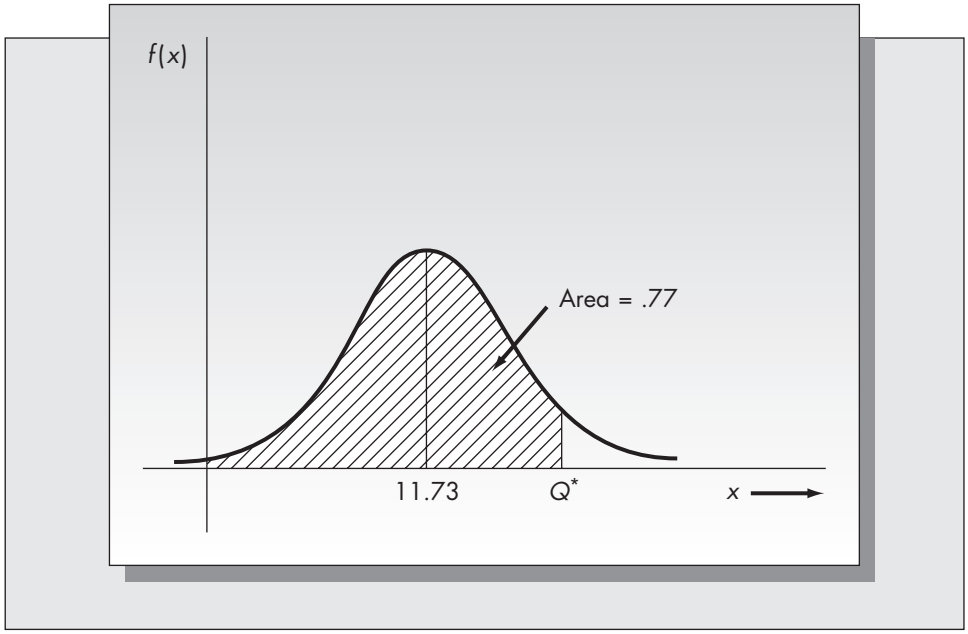


FIGURE 5–4
Determination of the optimal order quantity for the newsvendor example

Optimal Policy for Discrete Demand

Our derivation of the newsvendor formula was based on the assumption that the demand in the period was described by a continuous probability distribution. We noted a number of reasons for the desirability of working with continuous distributions. However, in some cases, and particularly when the mean demand is small, it may not be possible to obtain an accurate representation of the observed pattern of demand using a continuous distribution. For example, the normal approximation of the 52-week history of the demand for the *Journal* pictured in Figure 5–2 may not be considered sufficiently accurate for our purposes.

The procedure for finding the optimal solution to the newsvendor problem when the demand is assumed to be discrete is a natural generalization of the continuous case. In the continuous case, the optimal solution is the value of Q that makes the distribution function equal to the critical ratio $c_u/(c_u + c_o)$. In the discrete case, the distribution function increases by jumps; it is unlikely that any of its values exactly equal the critical ratio. The critical ratio will generally fall between two values of $F(Q)$. The optimal solution procedure is to locate the critical ratio between two values of $F(Q)$ and choose the Q corresponding to the *higher* value. (The fact that you always round Q up rather than simply round it off can be easily proven mathematically. This is different from assuming that units are ordered in discrete amounts but demand is continuous, in which case Q is rounded to the closest integer, as we did earlier.)

Example *5.2*

We will solve the problem faced by Mac's newsstand using the empirical distribution derived from a one-year history of the demand, rather than the normal approximation of that demand history. The empirical probabilities are obtained from Figure 5–1 by dividing each of the heights by 52. We obtain

Q	*f(Q)*	*F(Q)*	*Q*	*f(Q)*	*F(Q)*
0	1/52	1/52 (.0192)	12	4/52	30/52 (.5769)
1	0	1/52 (.0192)	13	1/52	31/52 (.5962)
2	0	1/52 (.0192)	14	5/52	36/52 (.6923)
3	0	1/52 (.0192)	15	5/52	41/52 (.7885)
4	3/52	4/52 (.0769)	16	1/52	42/52 (.8077)
5	1/52	5/52 (.0962)	17	3/52	45/52 (.8654)
6	2/52	7/52 (.1346)	18	3/52	48/52 (.9231)
7	2/52	9/52 (.1731)	19	3/52	51/52 (.9808)
8	4/52	13/52 (.2500)	20	0	51/52 (.9808)
9	6/52	19/52 (.3654)	21	0	51/52 (.9808)
10	2/52	21/52 (.4038)	22	1/52	52/52 (1.0000)
11	5/52	26/52 (.5000)			

The critical ratio for this problem was .77, which corresponds to a value of $F(Q)$ between $Q = 14$ and $Q = 15$. Because we round up, the optimal solution is $Q^* = 15$. Notice that this is exactly the same order quantity obtained using the normal approximation.

Extension to Include Starting Inventory

In the derivation of the newsvendor model, we assumed that the starting inventory at the beginning of the period was zero. Suppose now that the starting inventory is some value u and $u > 0$. The optimal policy for this case is a simple modification of the case $u = 0$. Note that this extension would not apply to newspapers, but would be appropriate for a product with a shelf life that exceeds one period.

Snapshot Application

USING INVENTORY MODELS TO MANAGE THE SEED-CORN SUPPLY CHAIN AT SYNGENTA

Each year farmers plant tens of thousands of acres of corn worldwide to meet the demands of a growing population. Companies such as Minnesota-based Syngenta Seeds supply seed to these farmers. In the United States alone, the market for corn seed is approximately $2.3 billion annually. Syngenta is one of eight firms that accounts for 73 percent of this market. Where do Syngenta and its competitors obtain this seed? The answer is that it is produced by growing corn and harvesting the seeds. Corn grown for the purpose of harvesting seed is known as seed-corn.

The problem of determining how much seed-corn to plant is complicated by several factors. One is that there are hundreds of different seed hybrids. Some hybrids do better in warmer, more humid climates, while others do better in cooler, dryer climates. The color, texture, sugar content, and so forth, of the corn produced by different hybrids varies as well. Farmers will not reuse a hybrid that yielded disappointing results. Hence, annual demand is hard to predict. In addition to facing uncertain demand, seed-corn producers also face uncertain yields. Their seed-corn plantings are subject to the same set of risks faced by all farmers: frost, draught, and heat spells.

Syngenta must decide each season how much seed-corn to plant. Since demand for the seeds is uncertain, the problem sounds like a straightforward application of the newsvendor model with uncertain demand and uncertain supply. However, Syngenta's decision problem has an additional feature that makes it more complicated than an ordinary newsvendor problem. Syngenta, along with many of its competitors, plants seed-corn in both northern and southern hemispheres. Since the hemispheric seasons run counter to each other, the plantings are done at different times of the year. In particular, the seed-corn is planted in the spring season in each hemisphere, so that the South American planting occurs about six months after the North American planting. This gives the company a second chance to increase production levels in South America to make up for shortfalls in North America, or decrease production levels in South America when there are surpluses in North America.

The problem of planning the size of the seed-corn planting was tackled by a team of researchers from the University of Iowa in collaboration with a vice president in charge of supply at Syngenta.[1] Using discrete approximations of the demand and yield distributions, they were able to formulate the planning problem as a linear program, so that it could be solved on a firmwide scale. A retrospective analysis showed that the company could have saved upwards of $5 million using the model. Also, the analysts were able to identify a systematic bias in the forecasts for seed generated by the firm that resulted in consistent overproduction. The mathematical model is now used to help guide the firm on its planting decisions each year.

[1]Jones, P. C.; Kegler, G.; Lowe, T. J.; and Traub, R. D. "Managing the Seed-Corn Supply Chain at Syngenta," *Interfaces* 33 (1), January-February 2003, pp. 80–90.

Consider the expected cost function $G(Q)$, pictured in Figure 5–3. If $u > 0$, it just means that we are starting at some point other than 0. We still want to be at Q^* after ordering, as that is still the lowest point on the cost curve. If $u < Q^*$, this is accomplished by ordering $Q^* - u$. If $u > Q^*$, we are past where we want to be on the curve. Ordering any additional inventory merely moves us up the cost curve, which results in higher costs. In this case it is optimal to simply not order.

Hence the optimal policy when there is a starting inventory of $u > 0$ is

Order $Q^* - u$ if $u < Q^*$.
Do not order if $u \geq Q^*$.

Note that Q^* should be interpreted as the order-up-to point rather than the order quantity when $u > 0$. It is also known as a target or base stock level.

Example *5.2 (continued)*

Let us suppose that in Example 5.2, Mac has received 6 copies of the *Journal* at the beginning of the week from another supplier. The optimal policy still calls for having 15 copies on hand after ordering, so now he would order the difference $15 - 6 = 9$ copies. (Set $Q^* = 15$ and $u = 6$ to get the order quantity of $Q^* - u = 9$.)

Extension to Multiple Planning Periods

The underlying assumption made in the derivation of the newsvendor model was that the item "perished" quickly and could not be used to satisfy demand in subsequent periods. In most industrial and retail environments, however, products are durable and inventory left at the end of a period can be stored and used to satisfy future demand.

This means that the ending inventory in any period becomes the starting inventory in the next period. Previously, we indicated how the optimal policy is modified when starting inventory is present. However, when the number of periods remaining exceeds one, the value of Q^* also must be modified. In particular, the interpretation of both c_o and c_u will be different. We consider only the case in which there are infinitely many periods remaining. The optimal value of the order-up-to point when a finite number of periods remain will fall between the one-period and the infinite-period solutions.

In our derivation and subsequent analysis of the EOQ formula in Chapter 4, we saw that the variable order cost c only entered into the optimization to determine the holding cost ($h = Ic$). In addition, we saw that all feasible operating policies incurred the same average annual cost of replenishment, λc. It turns out that essentially the same thing applies in the infinite horizon newsvendor problem. As long as excess demand is back-ordered, all feasible policies will just order the demand over any long period of time. Similarly, as long as excess demand is back-ordered, the number of units sold will just be equal to the demand over any long period of time. Hence, both c_u and c_o will be independent of both the proportional order cost c and the selling price of the item. Interpret c_u as the loss-of-goodwill cost and c_o as the holding cost in this case. That this is the correct interpretation of the underage and overage costs is established rigorously in Appendix 5–B.

Example *5.3*

Let us return to Mac's newsstand, described in Examples 5.1 and 5.2. Suppose that Mac is considering how to replenish the inventory of a very popular paperback thesaurus that is ordered monthly. Copies of the thesaurus unsold at the end of a month are still kept on the shelves for future sales. Assume that customers who request copies of the thesaurus when they are out of stock will wait until the following month. Mac buys the thesaurus for \$1.15 and sells it for \$2.75. Mac estimates a loss-of-goodwill cost of 50 cents each time a demand for a thesaurus must be back-ordered. Monthly demand for the book is fairly closely approximated by a normal distribution with mean 18 and standard deviation 6. Mac uses a 20 percent annual interest rate to determine his holding cost. How many copies of the thesaurus should he purchase at the beginning of each month?

Solution

The overage cost in this case is just the cost of holding, which is $(1.15)(0.20)/12 = 0.0192$. The underage cost is just the loss-of-goodwill cost, which is assumed to be 50 cents. Hence, the critical ratio is $0.5/(0.5 + 0.0192) = .9630$. From Table A–1 at the back of this book, this corresponds to a z value of 1.79. The optimal value of the order-up-to point $Q^* = \sigma z + \mu = (6)(1.79) + 18 = 28.74 \approx 29$.

Example *5.3 (continued)*

Assume that a local bookstore also stocks the thesaurus and that customers will purchase the thesaurus there if Mac is out of stock. In this case excess demands are lost rather than back-ordered. The order-up-to point will be different from that obtained assuming full back-ordering of demand. In Appendix 5–B we show that in the lost sales case the underage cost should be interpreted as the loss-of-goodwill cost plus the lost profit. The overage cost should still be interpreted as the holding cost only. Hence, the lost sales solution for this example gives $c_u = 0.5 + 1.6 = 2.1$. The critical ratio is $2.1/(2.1 + 0.0192) = .9909$, giving a z value of 2.36. The optimal value of Q in the lost sales case is $Q^* = \sigma z + \mu = (6)(2.36) + 18 = 32.16 \approx 32$.

Although the multiperiod solution appears to be sufficiently general to cover many types of real problems, it suffers from one serious limitation: there is no fixed cost of ordering. This means that the optimal policy, which is to order up to Q^*, requires that ordering take place in every period. In most real systems, however, there are fixed costs associated with ordering, and it is not optimal to place orders each period. Unfortunately, if we include a fixed charge for placing an order, it becomes extremely difficult to determine optimal operating policies. For this reason, we approach the problem of random demand when a fixed charge for ordering is present in a different way. We will assume that inventory levels are reviewed continuously and develop a generalization of the EOQ analysis presented in Chapter 4. This analysis is presented in Section 5.4.

Problems for Section 5.3

7. A newsvendor keeps careful records of the number of papers he sells each day and the various costs that are relevant to his decision regarding the optimal number of newspapers to purchase. For what reason might his results be inaccurate? What would he need to do in order to accurately measure the daily demand for newspapers?
8. Billy's Bakery bakes fresh bagels each morning. The daily demand for bagels is a random variable with a distribution estimated from prior experience given by

Number of Bagels Sold in One Day	Probability
0	.05
5	.10
10	.10
15	.20
20	.25
25	.15
30	.10
35	.05

The bagels cost Billy's 8 cents to make, and they are sold for 35 cents each. Bagels unsold at the end of the day are purchased by a nearby charity soup kitchen for 3 cents each.

a. Based on the given discrete distribution, how many bagels should Billy's bake at the start of each day? (Your answer should be a multiple of 5.)

b. If you were to approximate the discrete distribution with a normal distribution, would you expect the resulting solution to be close to the answer that you obtained in part (*a*)? Why or why not?

c. Determine the optimal number of bagels to bake each day using a normal approximation. (Hint: You must compute the mean μ and the variance σ^2 of the demand from the given discrete distribution.)

9. The Crestview Printing Company prints a particularly popular Christmas card once a year and distributes the cards to stationery and gift shops throughout the United States. It costs Crestview 50 cents to print each card, and the company receives 65 cents for each card sold.

Because the cards have the current year printed on them, those cards that are not sold are generally discarded. Based on past experience and forecasts of current

buying patterns, the probability distribution of the number of cards to be sold nationwide for the next Christmas season is estimated to be

Quantity Sold	Probability
100,000–150,000	.10
150,001–200,000	.15
200,001–250,000	.25
250,001–300,000	.20
300,001–350,000	.15
350,001–400,000	.10
400,001–450,000	.05

Determine the number of cards that Crestview should print this year.

10. Happy Henry's car dealer sells an imported car called the EX123. Once every three months, a shipment of the cars is made to Happy Henry's. Emergency shipments can be made between these three-month intervals to resupply the cars when inventory falls short of demand. The emergency shipments require two weeks, and buyers are willing to wait this long for the cars, but will generally go elsewhere before the next three-month shipment is due.

    From experience, it appears that the demand for the EX123 over a three-month interval is normally distributed with a mean of 60 and a variance of 36. The cost of holding an EX123 for one year is $500. Emergency shipments cost $250 per car over and above normal shipping costs.

    a. How many cars should Happy Henry's be purchasing every three months?

    b. Repeat the calculations, assuming that excess demands are back-ordered from one three-month period to the next. Assume a loss-of-goodwill cost of $100 for customers having to wait until the next three-month period and a cost of $50 per customer for bookkeeping expenses.

    c. Repeat the calculations, assuming that when Happy Henry's is out of stock of EX123s, the customer will purchase the car elsewhere. In this case, assume that the cars cost Henry an average of $10,000 and sell for an average of $13,500. Ignore loss-of-goodwill costs for this calculation.

11. Irwin's sells a particular model of fan, with most of the sales being made in the summer months. Irwin's makes a one-time purchase of the fans prior to each summer season at a cost of $40 each and sells each fan for $60. Any fans unsold at the end of the summer season are marked down to $29 and sold in a special fall sale. Virtually all marked-down fans are sold. The following is the number of sales of fans during the past 10 summers: 30, 50, 30, 60, 10, 40, 30, 30, 20, 40.

    a. Estimate the mean and the variance of the demand for fans each summer.

    b. Assume that the demand for fans each summer follows a normal distribution, with mean and variance given by what you obtained in part (*a*). Determine the optimal number of fans for Irwin's to buy prior to each summer season.

    c. Based on the observed 10 values of the prior demand, construct an empirical probability distribution of summer demand and determine the optimal number of fans for Irwin's to buy based on the empirical distribution.

    d. Based on your results for parts (*b*) and (*c*), would you say that the normal distribution provides an adequate approximation?

12. The buyer for Needless Markup, a famous "high end" department store, must decide on the quantity of a high-priced woman's handbag to procure in Italy for the following Christmas season. The unit cost of the handbag to the store is $28.50 and the handbag will sell for $150.00. Any handbags not sold by the end of the season are purchased by a discount firm for $20.00. In addition, the store accountants estimate that there is a cost of $0.40 for each dollar tied up in inventory, as this dollar invested elsewhere could have yielded a gross profit. Assume that this cost is attached to unsold bags only.
    a. Suppose that the sales of the bags are equally likely to be anywhere from 50 to 250 handbags during the season. Based on this, how many bags should the buyer purchase? (Hint: This means that the correct distribution of demand is uniform. You may solve this problem assuming either a discrete or a continuous uniform distribution.)
    b. A detailed analysis of past data shows that the number of bags sold is better described by a normal distribution, with mean 150 and standard deviation 20. Now what is the optimal number of bags to be purchased?
    c. The expected demand was the same in parts (*a*) and (*b*), but the optimal order quantities should have been different. What accounted for this difference?

5.4 LOT SIZE–REORDER POINT SYSTEMS

The form of the optimal solution for the simple EOQ model with a positive lead time analyzed in Chapter 4 is: When the level of on-hand inventory hits R, place an order for Q units. In that model the only independent decision variable was Q, the order quantity. The value of R was determined from Q, λ, and τ. In what follows, we also assume that the operating policy is of the (Q, R) form. However, when generalizing the EOQ analysis to allow for random demand, we treat Q and R as independent decision variables.

The multiperiod newsvendor model was unrealistic for two reasons: it did not include a setup cost for placing an order and it did not allow for a positive lead time. In most real systems, however, both a setup cost and a lead time are present. For these reasons, the kinds of models discussed in this section are used much more often in practice and, in fact, form the basis for the policies used in many commercial inventory systems.

Note that Q in this section is the amount ordered, whereas Q in Section 5.3 was the order-up-to point.

We make the following assumptions:

1. The system is continuous review. That is, demands are recorded as they occur, and the level of on-hand inventory is known at all times.
2. Demand is random and stationary. That means that although we cannot predict the value of demand, the *expected* value of demand over any time interval of fixed length is constant. Assume that the expected demand rate is λ units per year.
3. There is a fixed positive lead time τ for placing an order.
4. The following costs are assumed:

    Setup cost at $\$K$ per order.

    Holding cost at $\$h$ per unit held per year.

    Proportional order cost of $\$c$ per item.

    Stock-out cost of $\$p$ per unit of unsatisfied demand. This is also called the shortage cost or the penalty cost.

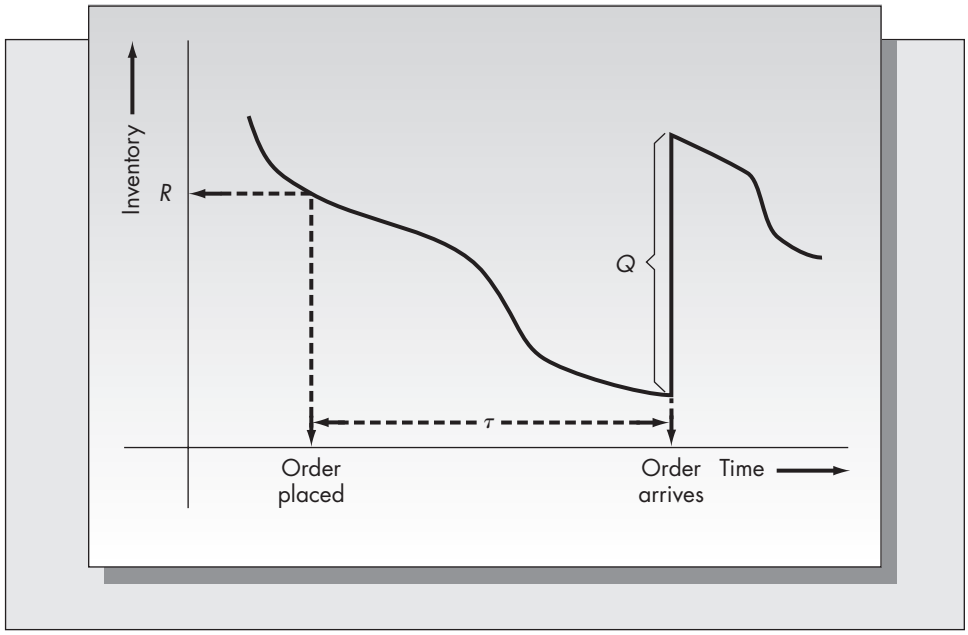


FIGURE 5–5
Changes in inventory over time for continuous-review (Q, R) system

Describing Demand

In the newsvendor problem, the appropriate random variable is the demand during the period. One period is the amount of time required to effect a change in the on-hand inventory level. This is known as the response time of the system. In the context of our current problem, the response time is the reorder lead time τ. Hence, the random variable of interest is the demand during the lead time. We will assume that the demand during the lead time is a continuous random variable D with probability density function (or pdf) $f(x)$ and cumulative distribution function (or cdf) $F(x)$. Let $\mu = E(D)$ and $\sigma = \sqrt{\text{var}(D)}$ be the mean and the standard deviation of demand during the lead time.

Decision Variables

There are two decision variables for this problem, Q and R, where Q is the lot size or order quantity and R is the reorder level in units of inventory.[2] Unlike the EOQ model, this problem treats Q and R as independent decision variables. The policy is implemented as follows: when the level of on-hand inventory reaches R, an order for Q units is placed that will arrive in τ units of time. The operation of this system is pictured in Figure 5–5.

Derivation of the Expected Cost Function

The analytical approach we will use to solve this problem is, in principle, the same as that used in the derivation of the newsvendor model. Namely, we will derive an expression for the expected average annual cost in terms of the decision variables (Q, R) and search for the optimal values of (Q, R) to minimize this cost.

[2] When lead times are very long, it may happen that an order should be placed again before a prior order arrives. In that case, the reorder decision variable R should be interpreted as the inventory position (on-hand plus on-order) when a reorder is placed, rather than the inventory level.

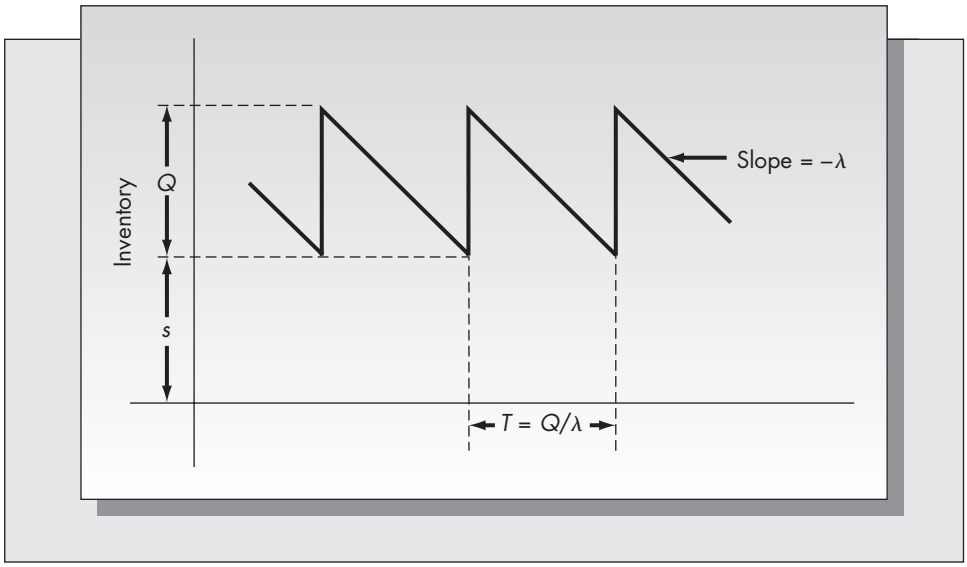


FIGURE 5–6
Expected inventory level for (Q, R) inventory model

The Holding Cost

We assume that the mean rate of demand is λ units per year. The expected inventory level varies linearly between s and $Q + s$. We call s the safety stock; it is defined as the expected level of on-hand inventory just before an order arrives, and is given by the formula $s = R - \lambda\tau$. The expected inventory level curve appears in Figure 5–6.

We estimate the holding cost from the average of the expected inventory curve. The average of the function pictured in Figure 5–6 is $s + Q/2 = R - \lambda\tau + Q/2$. An important point to note here is that this computation is only an approximation. When computing the average inventory level, we include both the cases when inventory is positive and the cases when it is negative. However, the holding cost should *not* be charged against the inventory level when it is negative, so that we are underestimating the true value of expected holding cost. An exact expression for the true average inventory is quite complex and has been derived only for certain specific demand distributions. In most real systems, however, the proportion of time spent out of stock is generally small, so this approximation should be reasonably accurate.

Setup Cost

A cycle is defined as the time between the arrival of successive orders of size Q. Consistent with the notation used in Chapter 4, let T represent the expected cycle length. Because the setup cost is incurred exactly once each cycle, we need to obtain an expression for the average length of a cycle in order to accurately estimate the setup cost per unit of time.

There are a number of ways to derive an expression for the expected cycle length. From Figure 5–6, we see that the distance between successive arrivals of orders is Q/λ. Another argument is the following. The expected demand during T is clearly λT. However, because the number of units that are entering inventory each cycle is Q and there is conservation of units, the number of units demanded each cycle on average also must be Q. Setting $Q = \lambda T$ and solving for T gives the same result.

It follows, therefore, that the average setup cost incurred per unit time is $K/T = K\lambda/Q$.

Penalty Cost

From Figure 5–5 we see that the only portion of the cycle during which the system is exposed to shortages is between the time that an order is placed and the time that it arrives (the lead time). The number of units of excess demand is simply the amount by which the demand over the lead time, D, exceeds the reorder level, R. It follows that the expected number of shortages that occur in one cycle is given by the expression

$$E(\max(D - R, 0)) = \int_R^\infty (x - R)f(x)\,dx,$$

which is defined as $n(R)$.

Note that this is essentially the same expression that we derived for the expected number of stock-outs in the newsvendor model. As $n(R)$ represents the expected number of stock-outs incurred in a cycle, it follows that the expected number of stock-outs incurred per unit of time is $n(R)/T = \lambda n(R)/Q$.

Proportional Ordering Cost Component

Over a long period of time, the number of units that enter inventory and the number that leave inventory must be the same. This means that every feasible policy will necessarily order a number of units equal to the demand over any long interval of time. That is, every feasible policy, on average, will replenish inventory at the rate of demand. It follows that the expected proportional order cost per unit of time is λc. Because this term is independent of the decision variables Q and R, it does not affect the optimization. We will henceforth ignore it.

It should be pointed out, however, that the proportional order cost will generally be part of the optimization in an indirect way. The holding cost h is usually computed by multiplying an appropriate value of the annual interest rate I by the value of the item c. For convenience we use the symbol h to represent the holding cost, but keep in mind that it could also be written in the form Ic.

The Cost Function

Define $G(Q, R)$ as the expected average annual cost of holding, setup, and shortages. Combining the expressions derived for each of these terms gives

$$G(Q, R) = h(Q/2 + R - \lambda\tau) + K\lambda/Q + p\lambda n(R)/Q.$$

The objective is to choose Q and R to minimize $G(Q, R)$. We present the details of the optimization in Appendix 5–C. As shown, the optimal solution is to iteratively solve the two equations

$$Q = \sqrt{\frac{2\lambda[K + pn(R)]}{h}} \tag{1}$$

$$1 - F(R) = Qh/p\lambda. \tag{2}$$

The solution procedure requires iterating between Equations (1) and (2) until two successive values of Q and R are (essentially) the same. The procedure is started by using $Q_0 = \text{EOQ}$ (as defined in Chapter 4). One then finds R_0 from Equation (2). That value of R is used to compute $n(R)$, which is substituted into Equation (1) to find Q_1,

which is then substituted into Equation (2) to find R_1, and so on. Convergence generally occurs within two or three iterations. When units are integral, the computations should be continued until successive values of both Q and R are within a single unit of their previous values. When units are continuous, a convergence requirement of less than one unit may be required depending upon the level of accuracy desired.

When the demand is normally distributed, $n(R)$ is computed by using the standardized loss function. The standardized loss function $L(z)$ is defined as

$$L(z) = \int_z^\infty (t - z)\phi(t)\, dt$$

where $\phi(t)$ is the standard normal density. If lead time demand is normal with mean μ and standard deviation σ, then it can be shown that

$$n(R) = \sigma L\left(\frac{R - \mu}{\sigma}\right) = \sigma L(z).$$

The standardized variate z is equal to $(R - \mu)/\sigma$. Calculations of the optimal policy are carried out using Table A–4 at the back of this book.[3]

Example *5.4*

Harvey's Specialty Shop is a popular spot that specializes in international gourmet foods. One of the items that Harvey sells is a popular mustard that he purchases from an English company. The mustard costs Harvey $10 a jar and requires a six-month lead time for replenishment of stock. Harvey uses a 20 percent annual interest rate to compute holding costs and estimates that if a customer requests the mustard when he is out of stock, the loss-of-goodwill cost is $25 a jar. Bookkeeping expenses for placing an order amount to about $50. During the six-month replenishment lead time, Harvey estimates that he sells an average of 100 jars, but there is substantial variation from one six-month period to the next. He estimates that the standard deviation of demand during each six-month period is 25. Assume that demand is described by a normal distribution. How should Harvey control the replenishment of the mustard?

Solution

We wish to find the optimal values of the reorder point R and the lot size Q. In order to get the calculation started we need to find the EOQ. However, this requires knowledge of the annual rate of demand, which does not seem to be specified. But notice that if the order lead time is six months and the mean lead time demand is 100, that implies that the mean yearly demand is 200, giving a value of $\lambda = 200$. It follows that the EOQ $= \sqrt{2K\lambda/h} = \sqrt{(2)(50)(200)/(0.2)(10)} = 100$.

The next step is to find R_0 from Equation (2). Substituting $Q = 100$, we obtain

$$1 - F(R_0) = Q_0 h/p\lambda = (100)(2)/(25)(200) = .04.$$

From Table A–4 we find that the z value corresponding to a right tail of .04 is $z = 1.75$. Solving $R = \sigma z + \mu$ gives $R = (25)(1.75) + 100 = 144$. Furthermore, $z = 1.75$ results in $L(z) = 0.0162$. Hence, $n(R) = \sigma L(z) = (25)(0.0162) = 0.405$.

We can now find Q_1 from Equation (1):

$$Q_1 = \sqrt{\frac{(2)(200)}{2}\,[50 + (25)(0.405)]} = 110.$$

This value of Q is compared with the previous one, which is 100. They are not close enough to stop. Substituting $Q = 110$ into Equation (2) results in $1 - F(R_1) = (110)(2)/(25)(200) = .044$. Table A–4 now gives $z = 1.70$ and $L(z) = 0.0183$. Furthermore, $R_1 = (25)(1.70) + 100 = 143$. We

[3] Note that in rare cases Equations (1) and (2) may not be solvable. This can occur when the penalty cost p is small compared to the holding cost h. The result is either diverging values of Q and R or the right side of Equation (2) having a value exceeding 1 at some point in the calculation. In this case, the recommended solution is to set $Q =$ EOQ and $R = R_0$ as long as the right side of Equation (2) is less than 1. This often leads to negative safety stock, discussed later in this chapter.

now obtain $n(R_1) = (25)(0.0183) = 0.4575$, and $Q_2 = \sqrt{(200)[50 + (25)(0.4575)]} = 110.85 \approx 111$. Substituting $Q_2 = 111$ into Equation (2) gives $1 - F(R_2) = .0444$, $z = 1.70$, and $R_2 = R_1 = 143$. Because both Q_2 and R_2 are within one unit of Q_1 and R_1, we may terminate computations.

We conclude that the optimal values of Q and R are $(Q, R) = (111, 143)$. Hence, each time that Harvey's inventory of this type of mustard hits 143 jars, he should place an order for 111 jars.

Example *5.4 (continued)*

For the same example, determine the following:

1. Safety stock.
2. The average annual holding, setup, and penalty costs associated with the inventory control of the mustard.
3. The average time between placement of orders.
4. The proportion of order cycles in which no stock-outs occur.
5. The proportion of demands that are not met.

Solution

1. The safety stock is $s = R - \mu = 143 - 100 = 43$ jars.
2. We will compute average annual holding, setup, and penalty costs separately.

   The holding cost is $h[Q/2 + R - \mu] = 2[111/2 + 143 - 100] = \$197$ per year.
   The setup cost is $K\lambda/Q = (50)(200)/111 = \$90.09$ per year.
   The stock-out cost is $p\lambda n(R)/Q = (25)(200)(0.4575)/111 = \$20.61$ per year.

   Hence, the total average annual cost associated with the inventory control of the mustard, assuming an optimal control policy, is \$307.70 per year.
3. $T = Q/\lambda = 111/200 = 0.556$ year $= 6.7$ months.
4. Here we need to compute the probability that no stock-out occurs in the lead time. This is the same as the probability that the lead time demand does not exceed the reorder point. We have $P\{D \leq R\} = F(R) = 1 - .044 = .956$. We conclude that there will be no stock-outs in 95.6 percent of the order cycles.
5. The expected demand per cycle must be Q (see the argument in the derivation of the expected setup cost). The expected number of stock-outs per cycle is $n(R)$. Hence, the proportion of demands that stock out is $n(R)/Q = 0.4575/111 = .004$. Another way of stating this result is that on average 99.6 percent of the demands are satisfied as they occur.

It should be noted in ending this section that Equations (1) and (2) are derived under the assumption that all excess demand is back-ordered. That is, when an item is demanded that is not immediately available, the demand is filled at a later time. However, in many competitive situations, such as retailing, a more accurate assumption is that excess demand is lost. This case is known as *lost sales*. As long as the likelihood of being out of stock is relatively small, Equations (1) and (2) will give adequate solutions to both lost sales and back-order situations. If this is not the case, a slight modification to Equation (2) is required. The lost-sales version of Equation (2) is

$$1 - F(R) = Qh/(Qh + p\lambda). \qquad (2')$$

The effect of solving Equations (1) and (2′) simultaneously rather than Equations (1) and (2) will be to increase the value of R slightly and decrease the value of Q slightly.

Inventory Level versus Inventory Position

An implicit assumption required in the analysis of (Q, R) policies was that each time an order of Q units arrives, it increases the inventory level to a value greater than the reorder point R. If this were not the case, one would never reach R, and one would never order again. This will happen if the demand during the lead time exceeds Q, which is certainly

Snapshot Application

INVENTORY MANAGEMENT SOFTWARE FOR THE SMALL BUSINESS

The software for PC-based inventory management continues to grow at an increasing rate. As personal computers have become more powerful, the inventory management needs of larger businesses can be handled by software designed to run on PCs. One example is an inventory system called inFlow, which can handle inventories stored at multiple locations. It is relatively expensive at $299. A less expensive alternative is Inventoria, which costs $79.99, which might be more appropriate for single location businesses, such as one-off retail stores. Another choice for small businesses is Small Business Inventory Control, which has the advantage of allowing the user to enter items into the system via barcode scanning. At $99, this package is also relatively inexpensive. Some other choices include iMagic Inventory, Inventory Power, Inventory Tracker Plus, Inventory Executive System, and many others. Each of the packages have certain advantages, including compatibility with other software systems, additional features such as financial functions, single versus multiple locations, industry specific applications, etc. Of course, inventory management modules are also available in large ERP systems, such as those sold by United States-based Oracle Corporation and German-based SAS.

An application that makes use of the most modern hardware and software is ShopKeep POS. ShopKeep is designed to run on the Apple iPad, and all of the records are stored in the cloud. This means that inventory data can be accessed from any iPad in the system at any time. One customer, The Bean, a small chain of coffee shops based in New York City, ported their inventory control function from a Windows-based system to ShopKeep, running on iPads and iPhones. The system gives them the opportunity to run reports from any iPad that's connected to the system. According to owner Ike Estavo, one of the most useful functions of ShopKeep is identifying the most profitable items per square foot of display space.

The personal computer has become ubiquitous since the first edition of this book was published in 1989. The huge growth in personal computers has been accompanied by an explosion in software choices for almost every conceivable application. Inventory control is no exception. A challenge for operations managers is to choose wisely among the multitude of applications, only a few of which are mentioned here. One must be able to pick the level of generality and complexity that's appropriate to the particular application. But no matter what software package is chosen, it is vitally important to understand the basic principles of inventory management discussed here. Hoping that the software will solve one's problem without really understanding the key elements of that problem is a big mistake.

Source: The information discussed here was obtained from the websites of the vendors mentioned.

possible. On the surface, this seems like a serious flaw in the approach, but it isn't. To avoid this problem, one bases the reorder decision on the inventory position rather than on the inventory level. The inventory position is defined as the total stock on hand plus on order. The inventory position varies from R to $R + Q$ as shown in Figure 5–7.

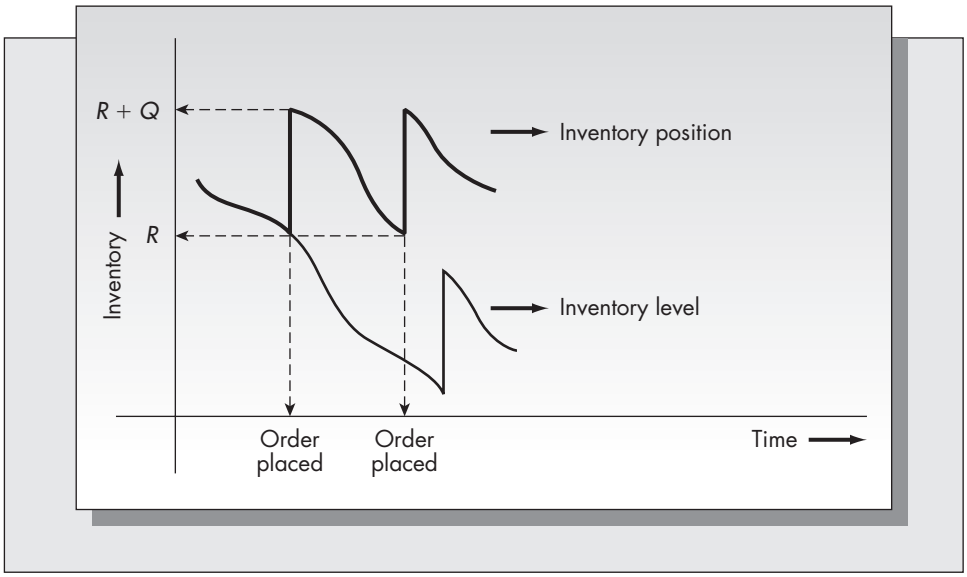


FIGURE 5–7
Inventory level versus inventory position

5.5 SERVICE LEVELS IN (*Q*, *R*) SYSTEMS

Although the inventory model described in Section 5.4 is quite realistic for describing many real systems, managers often have a difficult time determining an exact value for the stock-out cost p. In many cases, the stock-out cost includes intangible components such as loss of goodwill and potential delays to other parts of the system. A common substitute for a stock-out cost is a service level. Although there are a number of different definitions of service, it generally refers to the probability that a demand or a collection of demands is met. Service levels can be applied in both periodic review and (Q, R) systems. The application of service levels to periodic-review systems will be discussed in Section 5.6. Service levels for continuous-review systems are considered here.

Two types of service are considered, labeled Type 1 and Type 2, respectively.

Type 1 Service

In this case we specify the probability of not stocking out in the lead time. We will use the symbol α to represent this probability. As specification of α completely determines the value of R, the computation of R and Q can be decoupled. The computation of the optimal (Q, R) values subject to a Type 1 service constraint is very straightforward.

a. Determine R to satisfy the equation $F(R) = \alpha$.

b. Set $Q = \text{EOQ}$.

Interpret α as the proportion of cycles in which no stock-out occurs. A Type 1 service objective is appropriate when a shortage occurrence has the same consequence independent of its time or amount. One example would be where a production line is stopped whether 1 unit or 100 units are short. However, Type 1 service is not how service is interpreted in most applications. Usually when we say we would like to provide 95 percent service, we mean that we would like to be able to fill 95 percent of the demands when they occur, not fill all the demands in 95 percent of the order cycles. Also, as different items have different cycle lengths, this measure will not be consistent among different products, making the proper choice of α difficult.

Type 2 Service

Type 2 service measures the proportion of demands that are met from stock. We will use the symbol β to represent this proportion. As we saw in part 5 of Example 5.4, $n(R)/Q$ is the average fraction of demands that stock out each cycle. Hence, specification of β results in the constraint $n(R)/Q = 1 - \beta$.

This constraint is more complex than the one arising from Type 1 service, as it involves both Q and R. It turns out that although the EOQ is not optimal in this case, it usually gives pretty good results. If we use the EOQ to estimate the lot size, then we would find R to solve $n(R) = \text{EOQ}(1 - \beta)$.

Example *5.5* Consider again Harvey's Specialty Shop, described in Example 5.4. Harvey feels uncomfortable with the assumption that the stock-out cost is $25 and decides to use a service level criterion instead. Suppose that he chooses to use a 98 percent service objective.

1. *Type 1 service.* If we assume an α of .98, then we find R to solve $F(R) = 0.98$. From Table A–1 or A–4, we obtain $z = 2.05$. Setting $R = \sigma z + \mu$ gives $R = 151$.

2. *Type 2 service.* Here $\beta = 0.98$. We are required to solve the equation

$$n(R) = \text{EOQ}(1 - \beta),$$

which is equivalent to

$$L(z) = \text{EOQ}(1 - \beta)/\sigma.$$

Substituting EOQ = 100 and $\beta = .98$, we obtain

$$L(z) = (100)(0.02)/25 = 0.08.$$

From Table A–4 of the unit normal partial expectations, we obtain $z = 1.02$. Setting $R = \sigma z + \mu$ gives $R = 126$. Notice that the same values of α and β give *considerably* different values of R.

In order to understand more clearly the difference between these two measures of service, consider the following example. Suppose that we have tracked the demands and stock-outs over 10 consecutive order cycles with the following results:

Order Cycle	Demand	Stock-Outs
1	180	0
2	75	0
3	235	45
4	140	0
5	180	0
6	200	10
7	150	0
8	90	0
9	160	0
10	40	0

Based on a Type 1 measure of service, we find that the fraction of periods in which there is no stock-out is $8/10 = 80$ percent. That is, the probability that all the demands are met in a single order cycle is 0.8, based on these observations. However, the Type 2 service provided here is considerably better. In this example, the total number of demands over the 10 periods is 1,450 (the sum of the numbers in the second column), and the total number of demands that result in a stock-out is 55. Hence, the number of satisfied demands is $1{,}450 - 55 = 1{,}395$. The proportion of satisfied demands is $1{,}395/1{,}450 = .9621$, or roughly 96 percent.

The term *fill rate* is often used to describe Type 2 service, and is generally what most managers mean by service. (The fill rate in this example is 96 percent.) We saw in the example that there is a significant difference between the proportion of cycles in which all demands are satisfied (Type 1 service) and the fill rate (Type 2 service). *Even though it is easier to determine the best operating policy satisfying a Type 1 service objective, this policy will not accurately approximate a Type 2 service objective and should not be used in place of it.*

Optimal (*Q*, *R*) Policies Subject to Type 2 Constraint

Using the EOQ value to estimate the lot size gives reasonably accurate results when using a fill rate constraint, but the EOQ value is only an approximation of the optimal lot size. A more accurate value of Q can be obtained as follows. Consider the pair of equations (1) and (2) we solved for the optimal values of Q and R when a stock-out cost was present. Solving for p in Equation (2) gives

$$p = Qh/[(1 - F(R))\lambda],$$

which now can be substituted for p in Equation (1), resulting in

$$Q = \sqrt{\frac{2\lambda\{K + Qhn(R)/[(1 - F(R))\lambda]\}}{h}},$$

which is a quadratic equation in Q. It can be shown that the positive root of this equation is

$$Q = \frac{n(R)}{1 - F(R)} + \sqrt{\frac{2K\lambda}{h} + \left(\frac{n(R)}{1 - F(R)}\right)^2} \tag{3}$$

Equation (3) will be called the SOQ formula (for service level order quantity).[4] This equation is solved simultaneously with

$$n(R) = (1 - \beta)Q \tag{4}$$

to obtain optimal values of (Q, R) satisfying a Type 2 service constraint.

The reader should note that the version of Equation (4) used in the calculations is in terms of the standardized variate z and is given by

$$L(z) = (1 - \beta)Q/\sigma.$$

The solution procedure is essentially the same as that required to solve Equations (1) and (2) simultaneously. Start with $Q_0 =$ EOQ, find R_0 from (4), use R_0 in (3) to find Q_1, and so on, and stop when two successive values of Q and R are sufficiently close (within one unit is sufficient for most problems).

Example *5.5 (continued)*

Returning to Example 5.5, $Q_0 = 100$ and $R_0 = 126$. Furthermore, $n(R_0) = (0.02)(100) = 2$. Using $z = 1.02$ gives $1 - F(R_0) = 0.154$. Continuing with the calculations,

$$Q_1 = \frac{2}{0.154} + \sqrt{(100)^2 + \left(\frac{2}{0.154}\right)^2}$$
$$= 114.$$

Solving Equation (4) gives $n(R_1) = (114)(0.02)$, which is equivalent to

$$L(z) = (114)(0.02)/25 = 0.0912.$$

From Table A–4, $z = 0.95$, so that

$$1 - F(R_1) = 0.171$$

and

$$R_1 = \sigma z + \mu = 124.$$

Carrying the computation one more step gives $Q_2 = 114$ and $R_2 = 124$. As both Q and R are within one unit of their previous values, we terminate computations. Hence, we conclude that the optimal values of Q and R satisfying a 98 percent fill rate constraint are $(Q, R) = (114, 124)$.

Consider the cost error resulting from the EOQ substituted for the SOQ. In order to compare these policies, we compute the average annual holding and setup costs (notice that there is no stock-out cost) for the policies $(Q, R) = (100, 126)$ and $(Q, R) = (114, 124)$.

[4] The SOQ formula also could have been derived by more conventional Lagrange multiplier techniques. We include this derivation to demonstrate the relationship between the fill rate objective and the stock-out cost model.

Recall the formulas for average annual holding and setup costs.

$$\text{Holding cost} = h(Q/2 + R - \mu).$$
$$\text{Setup cost} = K\lambda/Q.$$

For (100, 126):

$$\left.\begin{aligned}\text{Holding cost} &= 2(100/2 + 126 - 100) = \$152\\ \text{Setup cost} &= (50)(200)/100 = \$100\end{aligned}\right\} \quad \text{Total} = \$252$$

For (114, 124):

$$\left.\begin{aligned}\text{Holding cost} &= 2(114/2 + 124 - 100) = \$162\\ \text{Setup cost} &= (50)(200)/114 = \$88\end{aligned}\right\} \quad \text{Total} = \$250$$

We see that the EOQ approximation gives costs close to the optimal in this case.

Imputed Shortage Cost

Consider the solutions that we obtained for (Q, R) in Example 5.5 when we used a service level criterion rather than a shortage cost. For a Type 2 service of $\beta = 0.98$ we obtained the solution (114, 124). Although no shortage cost was specified, this solution clearly corresponds to *some* value of p. That is, there is some value of p such that the policy (114, 124) satisfies Equations (1) and (2). This particular value of p is known as the imputed shortage cost.

The imputed shortage cost is easy to find. One solves for p in Equation (2) to obtain $p = Qh/[(1 - F(R))\lambda]$. The imputed shortage cost is a useful way to determine whether the value chosen for the service level is appropriate.

Example *5.5 (continued)*

Consider again Harvey's Specialty Shop. Using a value of $\alpha = .98$ (Type 1 service), we obtained the policy (100, 151). The imputed shortage cost is $p = (100)(2)/[(0.02)(200)] = \$50$.

Using a value of $\beta = 0.98$ (Type 2 service) we obtained the policy (114, 124). In this case the imputed cost of shortage is $p = (114)(2)/[(0.171)(200)] = \$6.67$.

Scaling of Lead Time Demand

In all previous examples the demand during the lead time was given. However, in most applications demand would be forecast on a periodic basis, such as monthly. In such cases one would need to convert the demand distribution to correspond to the lead time.

Assume that demands follow a normal distribution. Because sums of independent normal random variables are also normally distributed, the form of the distribution of lead time demand is normal. Hence, all that remains is to determine the mean and the standard deviation. Let the periodic demand have mean λ and standard deviation ν, and let τ be the lead time in periods. As both the means and the *variances* (not standard deviations) are additive, the mean demand during lead time is $\mu = \lambda\tau$ and the variance of demand during lead time is $\nu^2\tau$. Hence, the standard deviation of demand during lead time is $\sigma = \nu\sqrt{\tau}$ (although the square root may not always be appropriate).[5]

[5] Often in practice it turns out that there is more variation in the demand process than is described by a pure normal distribution. For that reason the standard deviation of demand is generally expressed in the form $\nu\tau^q$ where the correct value of q, generally between 0.5 and 1, must be determined for each item or group of items by an analysis of historical data.

Example *5.6* Weekly demand for a certain type of automotive spark plug in a local repair shop is normally distributed with mean 34 and standard deviation 12. Procurement lead time is six weeks. Determine the lead time demand distribution.

Solution The demand over the lead time is also normally distributed with mean (34)(6) = 204 and standard deviation $(12)\sqrt{6} = 29.39$. These would be the values of μ and σ that one would use for all remaining calculations.

Estimating Sigma When Inventory Control and Forecasting Are Linked

Thus far in this chapter we have assumed that the distribution of demand over the lead time is known. In practice, one assumes a *form* for the distribution, but its parameters must be estimated from real data. Assuming a normal distribution for lead time demand (which is the most common assumption), one needs to estimate the mean and the standard deviation. When a complete history of past data is available, the standard statistical estimates for the mean and the standard deviation (i.e., those suggested in Section 5.1) are fine. However, most forecasting schemes do *not* use all past data. Moving averages use only the past N data values and exponential smoothing places declining weights on past data.

In these cases, it is unclear exactly what are the right estimators for the mean and the standard deviation of demand. This issue was discussed in Section 2.13. The best estimate of the mean is simply the forecast of demand for the next period. For the variance, one should use the estimator for the variance of forecast error. The rationale for this is rarely understood. The variance of forecast error and the variance of demand are *not* the same thing. This was established rigorously in Appendix 2–A.

Why is it appropriate to use the standard deviation of forecast error to estimate σ? The reason is that it is the forecast that we are using to estimate demand. Safety stock is held to protect against errors in forecasting demand. In general, the variance of forecast error will be higher than the variance of demand. This is the result of the additional sampling error introduced by a forecasting scheme that uses only a portion of past data.[6]

R. G. Brown (1959) was apparently the first to recommend using the standard deviation of forecast error in safety stock calculations. The method he recommended, which is still in widespread use today, is to track the MAD (mean absolute deviation) of forecast error using the formula

$$\text{MAD}_t = \alpha\,\text{MAD}_{t-1} + (1-\alpha)|F_t - D_t|$$

where F_t is the forecast of demand at time t and D_t is the actual observed demand at time t. The estimator for the standard deviation of forecast error at time t is $1.25\,\text{MAD}_t$. While this method is very popular in commercial inventory control systems, apparently few realize that by using this approach they are estimating not demand variance but forecast error variance, and that these are not the same quantities.

[6] In cases where the underlying demand process is nonstationary, that is, where there is trend or seasonality, the variance of demand due to *systematic* changes could be higher than the variance of forecast error. All of our analysis assumes stationary (constant mean and variance) demand patterns, however.

*Lead Time Variability

Thus far, we have assumed that the lead time τ is a known constant. However, lead time uncertainty is common in practice. For example, the time required to transport commodities, such as oil, that are shipped by sea depends upon weather conditions. In general, it is very difficult to incorporate the variability of lead time into the calculation of optimal inventory policies. The problem is that if we assume that successive lead times are independent random variables, then it is possible for lead times to cross; that is, two successive orders would not necessarily be received in the same sequence that they were placed.

Order crossing is unlikely when a single supplier is used. If we are willing to make the simultaneous assumptions that orders do not cross and that successive lead times are independent, the variability of lead time can be easily incorporated into the analysis. Suppose that the lead time τ is a random variable with mean μ_τ and variance σ_τ^2. Furthermore, suppose that demand in any time t has mean λt and variance $\nu^2 t$. Then it can be shown that the demand during lead time has mean and variance[7]

$$\mu = \lambda \mu_\tau,$$
$$\sigma^2 = \mu_\tau \nu^2 + \lambda^2 \sigma_\tau^2.$$

Example *5.7* Harvey Gold, the owner of Harvey's Specialty Shop, orders an unusual olive from the island of Santorini, off the Greek coast. Over the years, Harvey has noticed considerable variability in the time it takes to receive orders of these olives. On average, the order lead time is four months and the standard deviation is six weeks (1.5 months). Monthly demand for the olives is normally distributed with mean 15 (jars) and standard deviation 6.

Setting $\mu_\tau = 4$, $\sigma_\tau = 1.5$, $\lambda = 15$, and $\nu = 6$, we obtain

$$\mu = \mu_\tau \lambda = (4)(15) = 60,$$
$$\sigma^2 = \mu_\tau \nu^2 + \lambda^2 \sigma_\tau^2 = (4)(36) + (225)(2.25) = 650.25.$$

One would proceed with the calculations of optimal inventory policies using $\mu = 60$ and $\sigma^2 = 650.25$ as the mean and the variance of lead time demand.

Calculations in Excel

The standardized loss function, $L(z)$, can be computed in Excel. To do so, write $L(z)$ in the following way:

$$L(z) = \int_z^\infty (t - z)\phi(t)\, dt = \int_z^\infty t\phi(t)\, dt - z(1 - \Phi(z)) = \phi(z) - z(1 - \Phi(z)).$$

The first equality results from the definition of the cumulative distribution function, and the second equality is a consequence of a well-known property of the standard normal distribution (see, for example, Hadley and Whitin, 1963, p. 444). To program this formula into Excel, use the definition of the standard normal density function, which is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-0.5z^2),$$

and the built-in Excel function *normsdist( )*, which returns the value of $\Phi(\ )$. In this way one could compute (Q, R) policies within the spreadsheet. Alternatively, one could build a table of $L(z)$, which can then be embedded into a search routine.

[7] Hadley and Whitin (1963), p. 153.

Negative Safety Stock

When the shortage cost or service level is relatively low, it is possible for a negative safety stock situation to arise. Recall that safety stock is the expected inventory on hand at the arrival of an order. If the safety stock is negative, there would be a back-order situation in more than 50 percent of the order cycles (that is, the Type 1 service would be under 50 percent). When this occurs, the optimal z value is negative, and the optimal $L(z)$ value would exceed L(0) = .3989. We illustrate with an example.

Example *5.8*

Consider once again Harvey's Specialty Shop from Examples 5.4 to 5.7. Harvey sells a high-end espresso machine, an expensive and bulky item. For this reason, Harvey attaches a very high holding cost of $50 per year to each machine. Harvey sells about 20 of these yearly, and estimates the variance of yearly demand to be 50. Annual demand follows the normal distribution. Since customers are willing to wait for the machine when he is out of stock, the shortage penalty is low. He estimates it to be $25 per unit. Order lead time is six months. The fixed cost of ordering is $80. Assume that the lot size used for reordering is the EOQ value. Determine the following:

1. Optimal reorder level
2. Safety stock
3. Resulting Type 1 service level
4. Resulting Type 2 service level

Solution

1. We have that $K = \$80$, $p = \$25$, $h = \$50$, and $\lambda = 20$. Scaling lead time demand to 0.5 year, we obtain $\mu = 10$ and $\sigma\sqrt{(50)(0.5)} = 5$. The reader can check that the EOQ rounded to the nearest integer is $Q = 8$. The equation for determining R_0 is

$$1 - F(R_0) = \frac{Q_0 h}{p\lambda} = \frac{(8)(50)}{(25)(20)} = 0.8.$$

   Since $1 - F(R_0)$ exceeds 0.5, the resulting z value is negative. In this case $z = -0.84$, $L(z) = 0.9520$, and $R_0 = \sigma z + \mu = (5)(-.84) + 10 = 5.8$, which we round to 6. Hence, the optimal solution based on the EOQ is to order eight units when the on-hand inventory falls to six units.
2. The safety stock is $S = R - \mu = 6 - 10 = -4$.
3. The Type 1 service level is $\alpha = F(R) = .20$ (or 20 percent).
4. The Type 2 service level is $\beta = 1 - n(R)/Q = 1 - \sigma L(z)/Q = 1 - (5)(.9520)/8 = 0.405$ (or 41 percent).

Note: Had we tried to solve this problem for the optimal (Q, R) iteratively using Equations (1) and (2), the equations would have diverged and not yielded a solution. This is a result of having a very low shortage cost compared to the holding cost. Furthermore, if Q were larger, the problem could be unsolvable. For example, if one assumes the setup cost $K = 150$, then EOQ = 11 and $1 - F(R_0) = Q_0 h/p\lambda = (11)(50)/(25)(20) = 1.1$, which is obviously not solvable. Such circumstances are very rare in practice, but, as we see from this example, it is possible for the model to fail. This is a consequence of the fact that the model is not exact. Exact (Q, R) models are beyond the scope of this book, and are known only for certain demand distributions.

Problems for Sections 5.4 and 5.5

13. An automotive warehouse stocks a variety of parts that are sold at neighborhood stores. One particular part, a popular brand of oil filter, is purchased by the warehouse for $1.50 each. It is estimated that the cost of order processing and

receipt is $100 per order. The company uses an inventory carrying charge based on a 28 percent annual interest rate.

The monthly demand for the filter follows a normal distribution with mean 280 and standard deviation 77. Order lead time is assumed to be five months.

Assume that if a filter is demanded when the warehouse is out of stock, then the demand is back-ordered, and the cost assessed for each back-ordered demand is $12.80. Determine the following quantities:

a. The optimal values of the order quantity and the reorder level.

b. The average annual cost of holding, setup, and stock-out associated with this item assuming that an optimal policy is used.

c. Evaluate the cost of uncertainty for this process. That is, compare the average annual cost you obtained in part (*b*) with the average annual cost that would be incurred if the lead time demand had zero variance.

14. Weiss's paint store uses a (Q, R) inventory system to control its stock levels. For a particularly popular white latex paint, historical data show that the distribution of monthly demand is approximately normal, with mean 28 and standard deviation 8. Replenishment lead time for this paint is about 14 weeks. Each can of paint costs the store $6. Although excess demands are back-ordered, the store owner estimates that unfilled demands cost about $10 each in bookkeeping and loss-of-goodwill costs. Fixed costs of replenishment are $15 per order, and holding costs are based on a 30 percent annual rate of interest.

    a. What are the optimal lot sizes and reorder points for this brand of paint?

    b. What is the optimal safety stock for this paint?

15. After taking a production seminar, Al Weiss, the owner of Weiss's paint store mentioned in Problem 14, decides that his stock-out cost of $10 may not be very accurate and switches to a service level model. He decides to set his lot size by the EOQ formula and determines his reorder point so that there is *no stock-out* in 90 percent of the order cycles.

    a. Find the resulting (Q, R) values.

    b. Suppose that, unfortunately, he really wanted to satisfy 90 percent of his demands (that is, achieve a 90 percent fill rate). What fill rate did he actually achieve from the policy determined in part (*a*)?

16. Suppose that in Problem 13 the stock-out cost is replaced with a Type 1 service objective of 95 percent. Find the optimal values of (Q, R) in this case.

17. Suppose that in Problem 13 a Type 2 service objective of 95 percent is substituted for the stock-out cost of $12.80. Find the resulting values of Q and R. Also, what is the imputed cost of shortage for this case?

18. Suppose that the warehouse mistakenly used a Type 1 service objective when it really meant to use a Type 2 service objective (see Problems 16 and 17). What is the additional holding cost being incurred each year for this item because of this mistake?

19. Disk Drives Limited (DDL) produces a line of internal Winchester disks for microcomputers. The drives use a 3.5-inch platter that DDL purchases from an outside supplier. Demand data and sales forecasts indicate that the weekly demand for the platters seems to be closely approximated by a normal distribution with mean 38 and variance 130. The platters require a three-week lead time for receipt. DDL has been using a 40 percent annual interest charge to compute holding costs.

The platters cost $18.80 each, order cost is $75.00 per order, and the company is currently using a stock-out cost of $400.00 per platter. (Because the industry is so competitive, stock-outs are very costly.)

a. Because of a prior contractual agreement with the supplier, DDL must purchase the platters in lots of 500. What is the reorder point that it should be using in this case?

b. When DDL renegotiates its contract with the supplier, what lot size should it write into the agreement?

c. How much of a penalty in terms of setup, holding, and stock-out cost is DDL paying for contracting to buy too large a lot?

d. DDL's president is uncomfortable with the $400 stock-out cost and decides to substitute a 99 percent fill rate criterion. If DDL used a lot size equal to the EOQ, what would its reorder point be in this case? Also, find the imputed cost of shortage.

20. Bobbi's Restaurant in Boise, Idaho, is a popular place for weekend brunch. The restaurant serves real maple syrup with french toast and pancakes. Bobbi buys the maple syrup from a company in Maine that requires three weeks for delivery. The syrup costs Bobbi $4 a bottle and may be purchased in any quantity. Fixed costs of ordering amount to about $75 for bookkeeping expenses, and holding costs are based on a 20 percent annual rate. Bobbi estimates that the loss of customer goodwill for not being able to serve the syrup when requested amounts to $25. Based on past experience, the weekly demand for the syrup is normal with mean 12 and variance 16 (in bottles). For the purposes of your calculations, you may assume that there are 52 weeks in a year and that all excess demand is back-ordered.

a. How large an order should Bobbi be placing with her supplier for the maple syrup, and when should those orders be placed?

b. What level of Type 1 service is being provided by the policy you found in part (*a*)?

c. What level of Type 2 service is being provided by the policy you found in part (*a*)?

d. What policy should Bobbi use if the stock-out cost is replaced with a Type 1 service objective of 95 percent?

e. What policy should Bobbi use if the stock-out cost is replaced with a Type 2 service objective of 95 percent? (You may assume an EOQ lot size.)

f. Suppose that Bobbi's supplier requires a minimum order size of 500 bottles. Find the reorder level that Bobbi should use if she wishes to satisfy 99 percent of her customer demands for the syrup.

5.6 ADDITIONAL DISCUSSION OF PERIODIC-REVIEW SYSTEMS

(*s*, *S*) Policies

In our analysis of the newsvendor problem, we noted that a severe limitation of the model from a practical viewpoint is that there is no setup cost included in the formulation. The (*Q*, *R*) model treated in the preceding sections included an order setup cost, but assumed that the inventory levels were reviewed continuously; that is, known at all times. How should the system be managed when there is a setup cost for ordering, but inventory levels are known only at discrete points in time?

The difficulty that arises from trying to implement a continuous-review solution in a periodic-review environment is that the inventory level is likely to overshoot the reorder point *R* during a period, making it impossible to place an order the instant

the inventory reaches R. To overcome this problem, the operating policy is modified slightly. Define two numbers, s and S, to be used as follows: When the level of on-hand inventory is *less than or equal to s,* an order for the difference between the inventory and S is placed. If u is the starting inventory in any period, then the (s, S) policy is

If $u \leq s$, order $S - u$.

If $u > s$, do not order.

Determining optimal values of (s, S) is extremely difficult, and for that reason few real operating systems use optimal (s, S) values. Several approximations have been suggested. One such approximation is to compute a (Q, R) policy using the methods described earlier, and set $s = R$ and $S = R + Q$. This approximation will give reasonable results in many cases, and is probably the most commonly used. The reader interested in a comprehensive comparison of several approximate (s, S) policies should refer to Porteus (1985).

*Service Levels in Periodic-Review Systems

Service levels also may be used when inventory levels are reviewed periodically. Consider first a Type 1 service objective. That is, we wish to find the order-up-to point Q so that all the demand is satisfied in a given percentage of the periods. Suppose that the value of Type 1 service is α. Then Q should solve the equation

$$F(Q) = \alpha.$$

This follows because $F(Q)$ is the probability that the demand during the period does not exceed Q. Notice that one simply substitutes α for the critical ratio in the newsvendor model. To find Q to satisfy a Type 2 service objective of β, it is necessary to obtain an expression for the fraction of demands that stock out each period. Using essentially the same notation as that used for (Q, R) systems, define

$$n(Q) = \int_Q^\infty (x - Q) f(x)\, dx.$$

Note that $n(Q)$, which represents the expected number of demands that stock out at the end of the period, is the same as the term multiplying c_u in the expression for the expected cost function for the newsvendor model discussed in Section 5.3. As the demand per period is μ, it follows that the proportion of demands that stock out each period is $n(Q)/\mu$. Hence, the value of Q that meets a fill rate objective of β solves

$$n(Q) = (1 - \beta)\mu.$$

The specification of either a Type 1 or Type 2 service objective completely determines the order quantity, independent of the cost parameters.

Example *5.9*

Mac, the owner of the newsstand described in Example 5.1, wishes to use a Type 1 service level of 90 percent to control his replenishment of *The Computer Journal.* The z value corresponding to the 90th percentile of the unit normal is $z = 1.28$. Hence,

$$Q^* = \sigma z + \mu = (4.74)(1.28) + 11.73 = 17.8 \approx 18.$$

Using a Type 2 service of 90 percent, we obtain

$$n(Q) = (1 - \beta)\mu = (0.1)(11.73) = 1.173.$$

It follows that $L(z) = n(Q)/\sigma = 1.173/4.74 = 0.2475$. From Table A–4 at the back of this book, we find

$$z \approx 0.35;$$

then

$$Q^* = \sigma z + \mu = (4.74)(0.35) + 11.73 = 13.4 \approx 13.$$

As with (Q, R) models, notice the striking difference between the resulting values of Q^* for the same levels of Type 1 and Type 2 service.

Fixed Order Size Model

If a positive order lead time is included, only a slight modification of these equations is required. In particular, the response time of the system is now the order lead time plus one period. Hence, we would now use the distribution of demand over $\tau + T$, where T is the time between inventory reviews.

This periodic review service level model is very useful in retail settings, in particular. It is common in retailing to place orders at fixed points in time to take advantage of bundling multiple orders together.

Example *5.10* Stroheim's is a dry goods store located in downtown Milwaukee, Wisconsin. Stroheim's places orders weekly with their suppliers for all of their reorders. The lead time for men's briefs is 4 days. Stroheim's uses a 95 percent service level. Assuming a Type 1 service, what is the order up to level for the briefs? Assume demands for briefs are uncertain with daily mean demand of 20 and daily standard deviation of 12.

Solution The total response time (review time plus lead time) is $7 + 4 = 11$ days. No matter what the form of the daily demand distribution, the Central Limit Theorem indicates that the demand over 11 days should be close to normal. The parameters are $\mu = (11)(20) = 220$ and $\sigma = 12\sqrt{11} = 39.80$. Hence it follows that for a type 1 service objective of 95 percent, the order-up-to-point should be $Q = \sigma z + \mu = (39.80)(1.645) + 220 = 286$.

If Stroheim's was interested in a Type 2 service objective, the z value would be lower, and the corresponding order up to point would be lower as well. In particular, $n(Q) = (1 - \beta)\,\mu = (.05)(220) = 11$, and $L(z) = n(Q)/\sigma = 11/39.80 = .2764$, which gives a z value of 0.27 and a corresponding order up to level of 231.

Problems for Section 5.6

21. Consider the Crestview Printing Company mentioned in Problem 9. Suppose that Crestview wishes to produce enough cards to satisfy all Christmas demand with probability 90 percent. How many cards should they print? Suppose the probability is 97 percent. What would you recommend in this case? (Your answer will depend upon the assumption you make concerning the shape of the cumulative distribution function.)
22. Consider Happy Henry's car dealer described in Problem 10.
    a. How many EX123s should Happy Henry's purchase to satisfy all the demand over a three-month interval with probability .95?
    b. How many cars should be purchased if the goal is to satisfy 95 percent of the demands?

Snapshot Application

TROPICANA USES SOPHISTICATED MODELING FOR INVENTORY MANAGEMENT

Tropicana, based in Bradenton, Florida, is one of the world's largest suppliers of citrus-based juice products. The company was founded by an Italian immigrant, Anthony Rossi, in 1947, and was acquired by PepsiCo in 1998. From its production facilities in Florida, Tropicana makes daily rail shipments to its regional distribution centers (DCs).

The focus of this application is the largest DC located in Jersey City, New Jersey, that services the northeast United States and Canada. Shipments from Bradenton to Jersey City require four days: one day for loading, two days in transit, and one day for unloading. This is the order lead time as seen from the DC. Lead time variability is not considered to be a significant issue.

Based on a statistical analysis of past data, Tropicana planners have determined that daily demands on the DC are closely approximated by a normal distribution for each of their products. Product classes are assumed to be independent.

Trains are sent from Florida five times per week; arrivals coincide with each business day at the Jersey City DC. Demand data and inventory levels are reviewed very frequently, a reorder point R triggers replenishment, and lot size Q is defined by the user. Hence, they operate a standard (Q, R) continuous-review system with a positive lead time. The state variable is the inventory position defined as the total amount of stock on hand at, and in transit to, the DC.

Let μ_D and σ_D be the mean and standard deviation of daily demand for a particular product line. According to the theory outlined in this chapter, the mean and standard deviation of demand over the four-day lead time should be

$$\mu = \mu_D\tau = 4\mu_D,$$
$$\sigma = \sigma_D\sqrt{\tau} = 2\sigma_D.$$

However, analysis of Tropicana data showed that the standard deviation of lead time demand is closer to $\sigma_D\tau^{0.7} = 2.64\sigma_D$.

The firm's objective is to maintain a 99.5 percent Type 2 service level at the DC. Values of (Q, R) are computed using the methodology of this chapter. Planners check the inventory on hand and in transit, and place an order equal to the EOQ, when this value falls below the reorder level, R. This analysis is carried out for a wide range of the company's products and individually determined for each regional DC.

Source: Based on joint work between the author and Tim Rowell of Tropicana.

23. For the problem of controlling the inventory of white latex paint at Weiss's paint store, described in Problem 14, suppose that the paint is reordered on a monthly basis rather than on a continuous basis.
    a. Using the (Q, R) solution you obtained in part (a) of Problem 14, determine appropriate values of (s, S).
    b. Suppose that the demands during the months of January to June were

Month	Demand	Month	Demand
January	37	April	31
February	33	May	14
March	26	June	40

If the starting inventory in January was 26 cans of paint, determine the number of units of paint ordered in each of the months January to June following the (s, S) policy you found in part (a).

5.7 MULTIPRODUCT SYSTEMS

ABC Analysis

One issue that we have not discussed is the cost of implementing an inventory control system and the trade-offs between the cost of controlling the system and the potential benefits that accrue from that control. In multiproduct inventory systems, not all products are equally profitable. Control costs may be justified in some cases and not in others. For example, spending $200 annually to monitor an item that contributes only $100 a year to profits is clearly not economical.

For this reason, it is important to differentiate profitable from unprofitable items. To do so, we borrow a concept from economics. The economist Vilfredo Pareto, who studied the distribution of wealth in the 19th century, noted that a large portion of the wealth was owned by a small segment of the population. This *Pareto effect* also applies to inventory systems: a large portion of the total dollar volume of sales is often accounted for by a small number of inventory items. Assume that items are ranked in decreasing order of the dollar value of annual sales. The *cumulative* value of sales generally results in a curve much like the one pictured in Figure 5–8.

Typically, the top 20 percent of the items account for about 80 percent of the annual dollar volume of sales, the next 30 percent of the items for the next 15 percent of sales, and the remaining 50 percent for the last 5 percent of dollar volume. These figures are

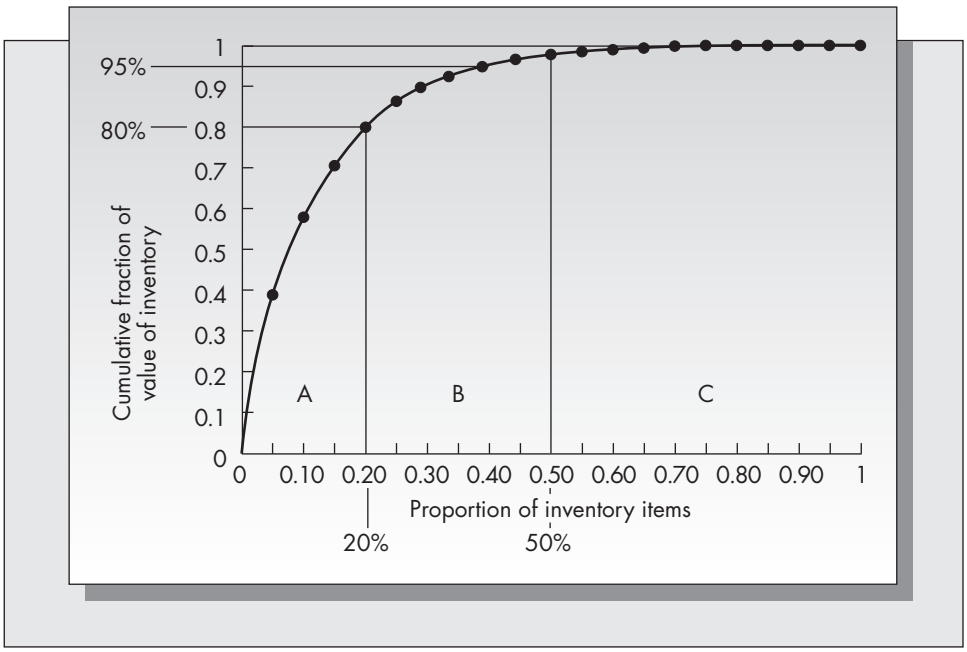


FIGURE 5–8
Pareto curve: Distribution of inventory by value

only approximate and will vary slightly from one system to another. The three item groups are labeled A, B, and C, respectively. When a finer distinction is needed, four or five categories could be used. Even when using only three categories, the percentages used in defining A, B, and C items could be different from the 80 percent, 15 percent, and 5 percent recommended.

Because A items account for the lion's share of the yearly revenue, these items should be watched most closely. Inventory levels for A items should be monitored continuously. More sophisticated forecasting procedures might be used and more care would be taken in the estimation of the various cost parameters required in calculating operating policies. For B items inventories could be reviewed periodically, items could be ordered in groups rather than individually, and somewhat less sophisticated forecasting methods could be used. The minimum degree of control would be applied to C items. For very inexpensive C items with moderate levels of demand, large lot sizes are recommended to minimize the frequency that these items are ordered. For expensive C items with very low demand, the best policy is generally not to hold any inventory. One would simply order these items as they are demanded.

Example *5.10 (continued)*

A sample of 20 different stock items from Harvey's Specialty Shop is selected at random. These items vary in price from \$0.25 to \$24.99 and in average yearly demand from 12 to 786. The results of the sampling are presented in Table 5–1. In Table 5–2 the items are ranked in decreasing order of the annual dollar volume of sales. Notice that only 4 of the 20 stock items account for over 80 percent of the annual dollar volume generated by the entire group. Also notice that there are high-priced items in both categories A and C.

This report was very illuminating to Harvey, who had assumed that R077, a packaged goat cheese from the south of France, was a profitable item because of its cost, and had been virtually ignoring TTR77, a domestic chocolate bar.

TABLE 5–1
Performance of 20 Stock Items Selected at Random

Part Number	Price	Yearly Demand	Dollar Volume
5497J	\$2.25	260	\$ 585.00
3K62	2.85	43	122.55
88450	1.50	21	31.50
P001	0.77	388	298.76
2M993	4.45	612	2,723.40
4040	6.10	220	1,342.00
W76	3.10	110	341.00
JJ335	1.32	786	1,037.52
R077	12.80	14	179.20
70779	24.99	334	8,346.66
4J65E	7.75	24	186.00
334Y	0.68	77	52.36
8ST4	0.25	56	14.00
16113	3.89	89	346.21
45000	7.70	675	5,197.50
7878	6.22	66	410.52
6193L	0.85	148	125.80
TTR77	0.77	690	531.30
39SS5	1.23	52	63.96
93939	4.05	12	48.60

TABLE 5–2
Twenty Stock Items Ranked in Decreasing Order of Annual Dollar Volume

Part Number	Price	Yearly Demand	Dollar Volume	Cumulative Dollar Volume	
70779	$24.99	334	$8,346.66	$ 8,346.66	A items: 20% of items account for 80.1% of total value.
45000	7.70	675	5,197.50	13,544.16	
2M993	4.45	612	2,723.40	16,267.56	
4040	6.10	220	1,342.00	17,609.56	
JJ335	1.32	786	1,037.52	18,647.08	
5497J	2.25	260	585.00	19,232.08	B items: 30% of items account for 14.8% of total value.
TTR77	0.77	690	531.30	19,763.38	
7878	6.22	66	410.52	20,173.90	
16113	3.89	89	346.21	20,520.11	
W76	3.10	110	341.00	20,861.11	
P001	0.77	388	298.76	21,159.87	
4J65E	7.75	24	186.00	21,345.87	
R077	12.80	14	179.20	21,525.07	C items: 50% of items account for 5.1% of total value.
6193L	0.85	148	125.80	21,650.87	
3K62	2.85	43	122.55	21,773.42	
39SS5	1.23	52	63.96	21,837.38	
334Y	0.68	77	52.36	21,889.74	
93939	4.05	12	48.60	21,938.34	
88450	1.50	21	31.50	21,969.84	
8ST4	0.25	56	14.00	21,983.84	

Exchange Curves

Much of our analysis assumes a single item in isolation and that the relevant cost parameters K, h, and p (or just K and h in the case of service levels) are constants with "correct" values that can be determined. However, it may be more appropriate to think of one or all of the cost parameters as policy variables. The correct values are those that result in a control system with characteristics that meet the needs of the firm and the goals of management. In a typical multiproduct system, the same values of setup cost K and interest rate I are used for all items. We can treat the ratio K/I as a policy variable; if this ratio is large, lot sizes will be larger and the average investment in inventory will be greater. If this ratio is small, the number of annual replenishments will increase.

To see exactly how a typical exchange curve is derived, consider a deterministic system consisting of n products with varying demand rates $\lambda_1, \ldots, \lambda_n$ and item values $c_1, \ldots, c_n$. If EOQ values are used to replenish stock for each item, then

$$Q_i = \sqrt{\frac{2K\lambda_i}{Ic_i}} \qquad \text{for } 1 \leq i \leq n.$$

For item i, the cycle time is Q_i/λ_i, so that λ_i/Q_i is the number of replenishments in one year. The total number of replenishments for the entire system is $\Sigma\lambda_i/Q_i$. The average on-hand inventory of item i is $Q_i/2$, and the value of this inventory in dollars is $c_iQ_i/2$. Hence, the total value of the inventory is $\Sigma c_iQ_i/2$.

Each choice of the ratio K/I will result in a different value of the number of replenishments per year and the dollar value of inventory. As K/I is varied, one traces out a curve such as the one pictured in Figure 5–9. An exchange curve such as this one allows management to easily see the trade-off between the dollar investment in inventory and the frequency of stock replenishment.

Exchange curves also can be used to compare various safety stock and service level strategies. As an example, consider a system in which a fill rate constraint is used (i.e., Type 2 service) for all items. Furthermore, suppose that the lead time demand distribution for all items is normal, and each item gets equal service. The dollar value of the safety stock is $\Sigma c_i(R_i - \mu_i)$, and the annual value of back-ordered demand is $\Sigma c_i \lambda_i n(R_i)/Q_i$. A fixed value of the fill rate β will result in a set of values of the control variables $(Q_1, R_1), \ldots, (Q_n, R_n)$, which can be computed by the methods discussed

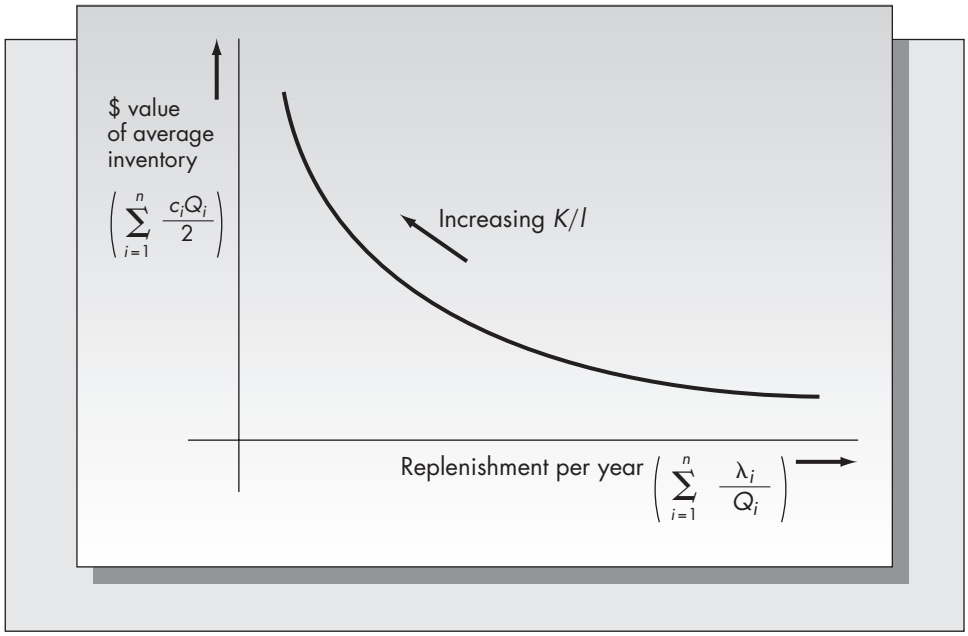


FIGURE 5–9
Exchange curve of replenishment frequency and inventory value

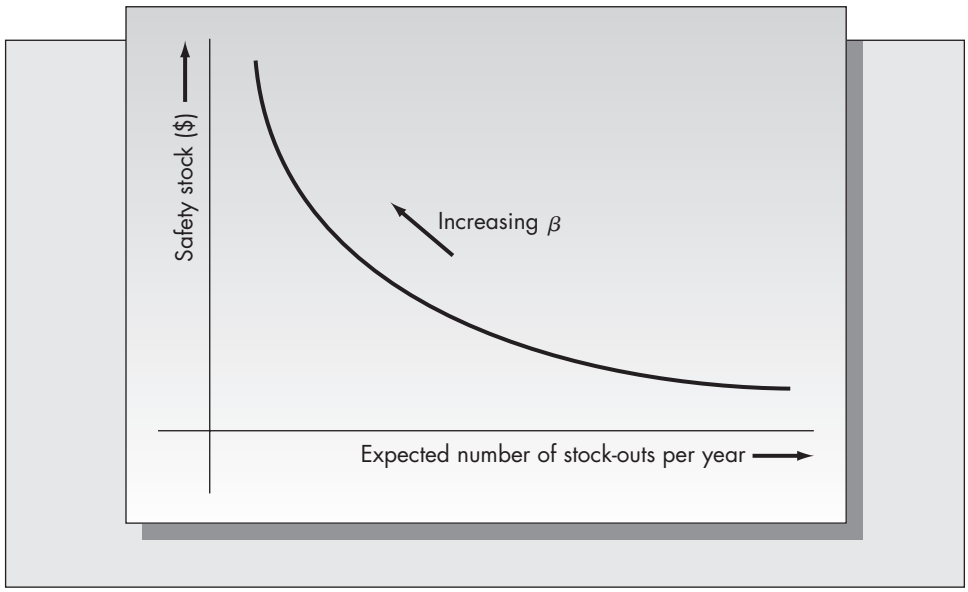


FIGURE 5–10
Exchange curve of the investment in safety stock and β

earlier in the chapter. Each set of (Q, R) values yields a pair of values for the safety stock and the back-ordered demand. As the fill rate is increased, the investment in safety stock increases and the value of back-ordered demand decreases. The exchange curve one would obtain is pictured in Figure 5–10. Such an exchange curve is a useful way for management to assess the dollar impact of various service levels.

Example *5.11* Consider the 20 stock items listed in Tables 5–1 and 5–2. Suppose that Harvey, the owner of Harvey's Specialty Shop, is reconsidering his choices of the setup cost of \$50 and interest charge of 20 percent. Harvey uses the EOQ formula to compute lot sizes for the 20 items for a range of values of K/I from 50 to 500. The resulting exchange curve appears in Figure 5–11.

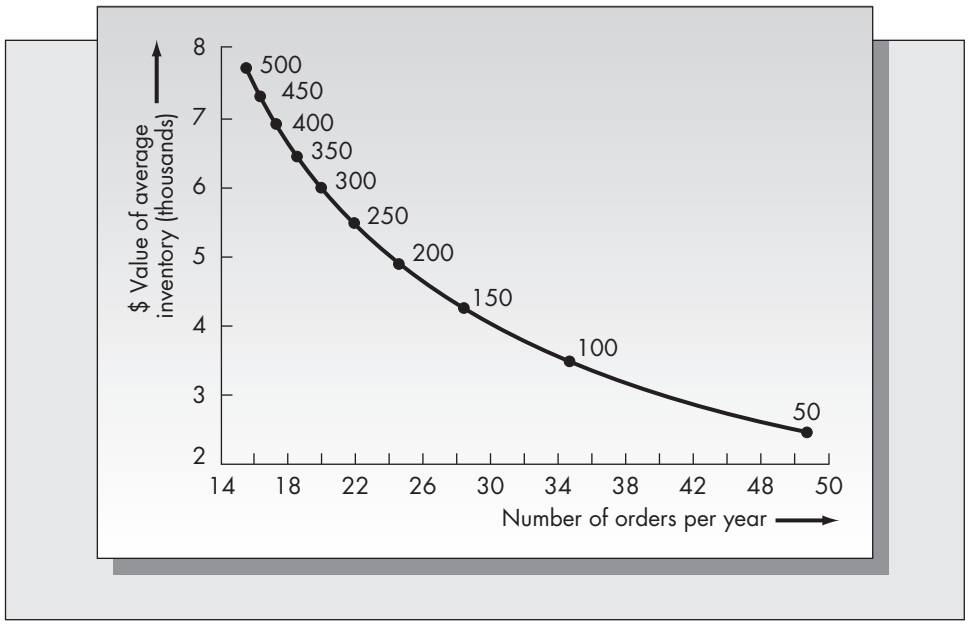


FIGURE 5–11
Exchange curve: Harvey's Specialty Shop

Harvey is currently operating at $K/I = 50/0.2 = 250$, which results in approximately 22 orders per year and an average inventory cost of \$5,447 annually. By reducing K/I to 100, the inventory cost for these 20 items is reduced to \$3,445 and the order frequency is increased to 34 orders a year. After some thought, Harvey decides that the additional time and bookkeeping expenses required to track an additional 12 orders annually is definitely worth the \$2,000 savings in inventory cost. (He is fairly comfortable with the 20 percent interest rate, which means that the true value of his setup cost for ordering is closer to \$20 than \$50. In this way the exchange curve can assist in determining the correct value of cost parameters that may otherwise be difficult to estimate.) He also considers moving to $K/I = 50$, but decides that the additional savings of about \$1,000 are not worth having to process almost 50 orders a year.

Problems for Section 5.7

24. Describe the ABC classification system. What is the purpose of classifying items in this fashion? What would be the primary value of ABC analysis to a retailer? To a manufacturer?

25. Consider the following list of retail items sold in a small neighborhood gift shop.

Item	Annual Volume	Average Profit per Item
Greeting cards	3,870	$ 0.40
T-shirts	1,550	1.25
Men's jewelry	875	4.50
Novelty gifts	2,050	12.25
Children's clothes	575	6.85
Chocolate cookies	7,000	0.10
Earrings	1,285	3.50
Other costume jewelry	1,900	15.00

   a. Rank the item categories in decreasing order of the annual profit. Classify each in one of the categories as A, B, or C.

   b. For what reason might the store proprietor choose to sell the chocolate cookies even though they might be her least profitable item?

26. From management's point of view, what is the primary value of an exchange curve? Discuss both the exchange curve for replenishment frequency and inventory value and the exchange curve for expected number of stock-outs per year and the investment in safety stock.

27. Consider the eight stock items listed in Problem 25. Suppose that the average costs of these item categories are

Item	Cost
Greeting cards	$ 0.50
T-shirts	3.00
Men's jewelry	8.00
Novelty gifts	12.50
Children's clothes	8.80
Chocolate cookies	0.40
Earrings	4.80
Other costume jewelry	12.00

Compare the total number of replenishments per year and the dollar value of the inventory of these items for the following values of the ratio of K/I: 100, 200, 500, 1,000. From the four points obtained, estimate the exchange curve of replenishment frequency and inventory value.

*5.8 OVERVIEW OF ADVANCED TOPICS

This chapter has treated only a small portion of inventory models available. Considerable research has been devoted to analyzing far more complex stochastic inventory control problems, but most of this research is beyond the scope of our coverage. This section presents a brief overview of two areas not discussed in this chapter that account for a large portion of the recent research on inventory management.

Multi-echelon Systems

Firms involved in the manufacture and distribution of consumer products must take into account the interactions of the various levels in the distribution chain. Typically,

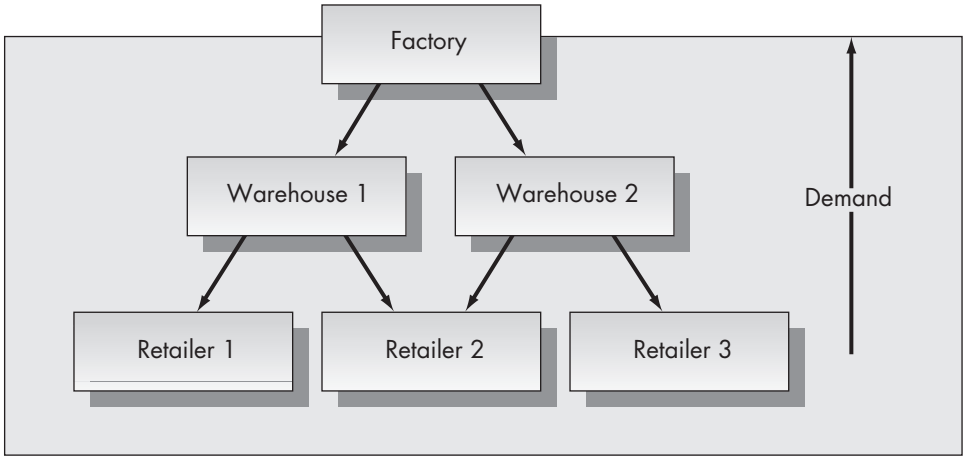


FIGURE 5–12
Typical three-level distribution system

items are produced at one or more factories, shipped to warehouses for intermediate storage, and subsequently shipped to retail outlets to be sold to the consumer. We refer to the levels of the system as *echelons.* For example, the factory–warehouse–retailer three-echelon system is pictured in Figure 5–12.

A system of the type pictured in Figure 5–12 is generally referred to as a *distribution* system. In such a system, the demand arises at the lowest echelon (the retailer in this case) and is transmitted up to the higher echelons. Production plans at the factory must be coordinated with orders placed at the warehouse by the retailers. Another type of multi-echelon system arising in manufacturing contexts is known as an *assembly* system. In an assembly system, components are combined to form subassemblies, which are eventually combined to form end items (that is, final products). A typical assembly system is pictured in Figure 5–13.

In the assembly system, external demand originates at the end-item level only. Demand for the components arises only when "orders" are placed at higher levels of the system, which are the consequence of end-item production schedules. Materials requirements planning methodology is designed to model systems of this type, but does not consider the effect of uncertainty on the final demand. Although there has been some research on this problem, precisely how the uncertainty in the final demand

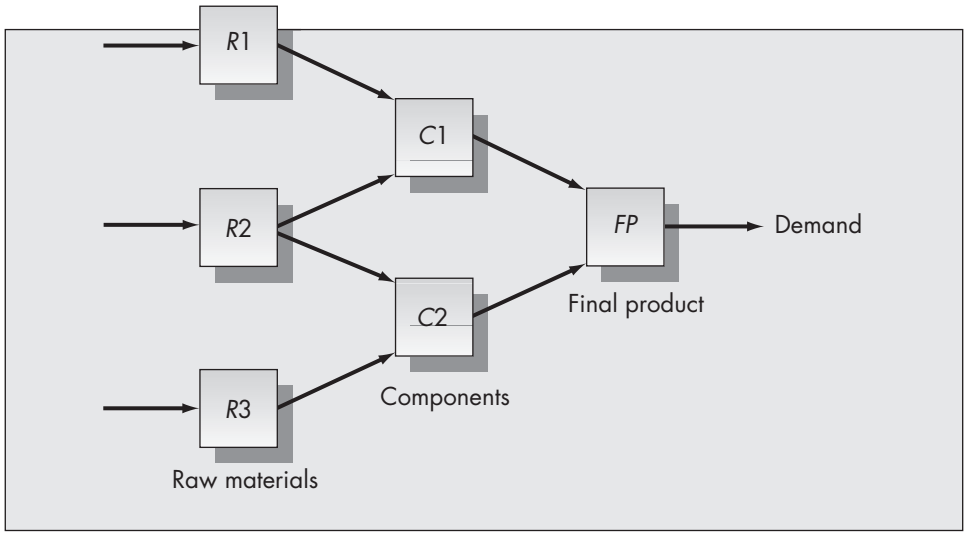


FIGURE 5–13
Typical three-level assembly system

affects the optimal production and replenishment policies of components remains an open question.

One multi-echelon system that has received a great deal of attention in the literature (Sherbrooke, 1968) was designed for an application in military logistics. Consider the replenishment problem that arises in maintaining a supply of spare parts to support jet aircraft engine repair. When an engine (or engine component) fails, it is sent to a central depot for repair. The depot maintains its own inventory of engines, so that it can ship a replacement to the base before the repair is completed. The problem is to determine the correct levels of spares at the base and at the depot given the conflicting goals of budget limitations and the desirability of a high system fill rate.

The U.S. military's investment in spare parts is enormous. As far back as 1968 it was estimated that the U.S. Air Force alone has $10 billion invested in repairable item inventories. Considering the effects of inflation and the sizes of the other armed services, the current investment in repairable items in the U.S. military probably exceeds $100 billion. Mathematical inventory models have made a significant impact on the management of repairable item inventories, both in military and nonmilitary environments.

In the private sector, retailing is an area in which multi-echelon distribution systems are common. Most major retail chains utilize a distribution center (DC) as an intermediate storage point between the manufacturer and the retail store. Determining how to allocate inventory between the DC and the stores is an important strategic issue. DC inventory allows "risk pooling" among the stores and facilitates redistribution of store inventories that might grow out of balance. Several commercial computer-based inventory packages are available for retail inventory management, perhaps the best known being Inforem, designed and marketed by IBM. Readers interested in a comprehensive review of the research on the retailer inventory management problem should refer to the recent book by Agrawal and Smith (2008) and Nahmias and Smith (1992).

Perishable Inventory Problems

Although certain types of decay can be accounted for by adjustments to the holding cost, the mathematical models that describe perishability, decay, or obsolescence are generally quite complex.

Typical examples of fixed-life perishable inventories include food, blood, drugs, and photographic film. In order to keep track of the age of units in stock, it is necessary to know the amount of each age level of inventory on hand. This results in a complicated multidimensional system state space. Perishability adds an additional penalty for holding inventory, thus reducing the size of orders placed. The most notable application of the theory of perishable inventory models has been to the management of blood banks.

A somewhat simpler but related problem is that of exponential decay, in which it is assumed that a fixed fraction of the inventory in stock is lost each period. Exponential decay is an accurate description of the changes that take place in volatile liquids such as alcohol and gasoline, and is an exact model of decline in radioactivity of radioactive pharmaceuticals and nuclear fuels.

Obsolescence has been modeled by assuming that the length of the planning horizon is a random variable with a known probability distribution. In practice, such models

Snapshot Application

INTEL USES MULTIECHELON INVENTORY MODELLING TO MANAGE THE SUPPLY CHAIN FOR BOXED CPUS

Intel Corporation, headquartered in Santa Clara, California, is the largest producer of integrated circuits in the world. Intel microprocessors have been the "guts" of most personal computers ever since IBM first adopted the Intel 8088 as the CPU in the original IBM PC in 1981. In 2005, a group at Intel and one outside consultant was charged with the problem of managing the supply chain of its branded boxed CPUs. The division typically sells 100–150 configurations of these products.

The group chose to model the problem as a three echelon inventory system: (1) global supply, (2) boxing sites, and (3) boxed CPU warehouses. The group quickly realized that there were many factors affecting the supply chain, but the primary driver of system effectiveness was demand uncertainty. Predicting demand uncertainty is difficult in this environment since product life cycles are often relatively short, resulting in very little historical data for some products.

Another complicating factor is that the three echelons are far apart geographically, resulting in a very long cumulative lead time for the system. Hence, a make-to-order business model is simply not feasible. For that reason, proper positioning of inventory along the entire supply chain is critical. The mathematical model developed for this application was a nonlinear mathematical programming model that could be solved by standard dynamic programming techniques.

A key issue is metrics. That is, how does one evaluate the quality of the model's results? One metric is the total inventory in the system. This translates to the total dollar investment required by Intel. A second metric is the quality of the service provided to the customer. Both Type 1 and Type 2 service level metrics (as discussed in detail in this chapter) were considered. Intel management chose a modified version of Type 2 service (that is, the fill rate). Their metric was the percentage of an order filled on the date the customer asks for delivery of the product.

The team decided that it was not worth trying to apply their optimization tool to all 150 products, so they embarked on an ABC analysis, and ultimately chose to apply the model only to products with more than 3,000 sales per month and 12,000 sales per quarter. Also, it became clear that in order to be responsive to customer demand, the vast majority of the inventory would be held at the third echelon, namely the boxed CPU warehouses. This inventory has the most value added, so is the most expensive. That expense is offset by the fact that keeping the inventory closer to the customer significantly improves service levels. The system was implemented at Intel in late 2005. Comparisons with pre and post implementation metrics showed that the model resulted in a modest decline in system-wide inventory levels, but significantly improved level of customer service.

An important aspect of this study was the priorities set by the project team. These priorities provide a good blueprint on how to go about a project of this type to maximize the likelihood of implementation. The four goals they set were: (1) keep models and processes simple, (2) make things better now, (3) implement changes in a phased manner, and (4) be clear about what success is.

Source: Wieland, B. et al. "Optimizing Inventory Levels within Intel's Channel Supply Demand Operations." *Interfaces* 42 (2012), pp. 517–527.

are valuable only if the distribution of the useful lifetime of the item can be accurately estimated in advance.

A related problem is the management of style goods, such as fashion items in the garment industry. The style goods problem differs from the inventory models considered in this chapter in that not only is the demand itself uncertain, but the distribution of demand is uncertain as well. This is typical of problems for which there is no prior demand history. The procedures suggested to deal with the style goods problem generally involve some type of Bayesian updating scheme to combine current observations of demand with prior estimates of the demand distribution.

5.9 HISTORICAL NOTES AND ADDITIONAL READINGS

Research on stochastic inventory models was sparked by the national war effort and appears to date back to the 1940s, although the first published material appeared in the early 1950s (see Arrow, Harris, and Marschak, 1951; and Dvorestsky, Kiefer, and Wolfowitz, 1952). The important text by Arrow, Karlin, and Scarf (1958) produced considerable interest in the field.

The model we discuss of a lot size–reorder point system with stock-out cost criterion seems to be attributed to Whitin (1957). The extensions to service levels are attributed to Brown (1967). Brown also treats a variety of other topics of practical importance, including the relationship between forecasting models and inventory control. An excellent discussion of exchange curves and ABC analysis, and more in-depth coverage of a number of the topics we discuss, can be found in Silver and Peterson (1985). Love (1979) provides an excellent summary of many of the issues we treat. Hadley and Whitin (1963) also deal with these topics at a more sophisticated mathematical level.

There has been considerable interest in the development of approximate (s, S) policies. Scarf (1960), and later Iglehart (1963), proved the optimality of (s, S) policies, and Veinott and Wagner (1965) considered methods for computing optimal (s, S) policies. A number of approximation techniques have been suggested. Ehrhardt (1979) considers using regression analysis to fit a grid of optimal policies, whereas Freeland and Porteus (1980) adapt dynamic programming methods to the problem. Porteus (1985) numerically compares the effectiveness of various approximation techniques. Some of the issues regarding lead time uncertainty are treated by Kaplan (1970) and Nahmias (1979), among others.

The original formulation of the standard multi-echelon inventory problem came from Clark and Scarf (1960). Extensions have been developed by a variety of researchers, most notably Bessler and Veinott (1966), and, more recently, Federgruen and Zipkin (1984). Retailer-warehouse systems have been studied by Deuermeyer and Schwarz (1981) and Eppen and Schrage (1981). Schmidt and Nahmias (1985) analyzed an assembly system when demand for the final product is random.

The classic work on the type of multi-echelon model that has become the basis for the inventory control systems implemented in the military was done by Sherbrooke (1968). An extension of Sherbrooke's analysis is considered by Graves (1985). Muckstadt and Thomas (1980) discuss the advantages of implementing multi-echelon models in an industrial environment. A comprehensive review of models for managing repairables can be found in Nahmias (1981).

Interest in perishable inventory control models appears to stem from the problem of blood bank inventory control, although food management has a far greater economic impact. Most of the mathematical models for perishables assume that the inventory is issued from stock on an oldest-first basis (although a notable exception is the study by Cohen and Pekelman, 1978). Nahmias (1982) provides a comprehensive review of the research on perishable inventory control.

The style goods problem has been studied by a variety of researchers. Representative work in this area includes that by Murray and Silver (1966), Hartung (1973), and Hausman and Peterson (1972). Those interested in additional readings on stochastic inventory models should refer to the various review articles in the field. Some of the most notable include those of Scarf (1963), Veinott (1966), and Nahmias (1978). A recent collection of up-to-date reviews can be found in Graves et al. (1992).

5.10 Summary

This chapter presented an overview of several inventory control methods when the demand for the item is random. The *newsvendor model* is based on the assumption that the product has a useful life of exactly one planning period. We assumed that the demand during the period is a continuous random variable with cumulative distribution function $F(x)$, and that there are specified overage and underage costs of c_o and c_u charged against the inventory remaining on hand at the end of the period or the excess demand, respectively. The optimal order quantity Q^* solves the equation

$$F(Q^*) = \frac{c_u}{c_u + c_o}.$$

Extensions to discrete demand and multiple planning periods also were considered.

From a practical standpoint, the newsvendor model has a serious limitation: it does not allow for a positive order setup cost. For that reason, we considered an extension of the EOQ model known as a *lot size–reorder point* model. The key random variable for this case was the demand during the lead time. We showed how optimal values of the decision variables Q and R could be obtained by the iterative solution of two equations. The system operates in the following manner. When the level of on-hand inventory hits R, an order for Q units is placed (which will arrive after the lead time τ). This policy is the basis of many commercial inventory control systems.

Service levels provide an alternative to stock-out costs. Two service levels were considered: Type 1 and Type 2. Type 1 service is the probability of not stocking out in any order cycle, and Type 2 service is the probability of being able to meet a demand when it occurs. Type 2 service, also known as the *fill rate,* is the more natural definition of service for most applications.

Several additional topics for (Q, R) systems also were considered. The *imputed shortage cost* is the effective shortage cost resulting from specification of the service level. Assuming normality, we showed how one transforms the distribution of periodic demand to lead time demand. Finally, we considered the effects of *lead time variability.*

If a setup cost is included in the multiperiod version of the newsvendor problem, the optimal form of the control policy is known as an (s, S) *policy.* This means that if the starting inventory in a period, u, is less than or equal to s, an order for $S - u$ units is placed. An effective approximation for the optimal (s, S) policy can be obtained by solving the problem as if it were continuous review to obtain the corresponding (Q, R) policy, and setting $s = R$ and $S = Q + R$. We also discussed the application of service levels to periodic-review systems.

Most real inventory systems involve the management of more than a single product. We discussed several issues that arise when managing a multiproduct inventory system. One of these is the amount of time and expense that should be allocated to the control of each item. The *ABC system* is a method of classifying items by their annual volume of sales. Another issue concerns the correct choice of the cost parameters used in computing inventory control policies. Because many of the cost parameters used in inventory analysis involve managerial judgment and are not easily measured, it would be useful if management could compare the effects of various parameter settings on the performance of the system. A convenient technique for making these comparisons is via *exchange curves.* We discussed two of the most popular exchange curves: (1) the trade-off between the investment in inventory and the frequency of stock replenishment and (2) the trade-off between the investment in safety stock and service levels.

Additional Problems on Stochastic Inventory Models

28. An artist's supply shop stocks a variety of different items to satisfy the needs of both amateur and professional artists. In each case described, what is the appropriate inventory control model that the store should use to manage the replenishment of the item described? Choose your answer from the following list and be sure to explain your answer in each case:

Simple EOQ	Newsvendor model with service level
Finite production rate	(Q, R) model with stock-out cost
EOQ with quantity discounts	(Q, R) model with Type 1 service level
Resource-constrained EOQ	(Q, R) model with Type 2 service level
Newsvendor model	Other type of model

   a. A highly volatile paint thinner is ordered once every three months. Cans not sold during the three-month period are discarded. The demand for the paint thinner exhibits considerable variation from one three-month period to the next.

   b. A white oil-base paint sells at a fairly regular rate of 600 tubes per month and requires a six-week order lead time. The paint store buys the paint for $1.20 per tube.

   c. Burnt sienna oil paint does not sell as regularly or as heavily as the white. Sales of the burnt sienna vary considerably from one month to the next. The useful lifetime of the paint is about two years, but the store sells almost all the paint prior to the two-year limit.

   d. Synthetic paint brushes are purchased from an East Coast supplier who charges $1.60 for each brush in orders of under 100 and $1.30 for each brush in orders of 100 or greater. The store sells the brushes at a fairly steady rate of 40 per month for $2.80 each.

   e. Camel hair brushes are purchased from the supplier of part (*d*), who offers a discount schedule similar to the one for the synthetic brushes. The camel hair brushes, however, exhibit considerable sales variation from month to month.

29. Annual demand for number 2 pencils at the campus store is normally distributed with mean 1,000 and standard deviation 250. The store purchases the pencils for 6 cents each and sells them for 20 cents each. There is a two-month lead time from the initiation to the receipt of an order. The store accountant estimates that the cost in employee time for performing the necessary paperwork to initiate and receive an order is $20, and recommends a 22 percent annual interest rate for determining holding cost. The cost of a stock-out is the cost of lost profit plus an additional 20 cents per pencil, which represents the cost of loss of goodwill.

   a. Find the optimal value of the reorder point R assuming that the lot size used is the EOQ.

   b. Find the simultaneous optimal values of Q and R.

   c. Compare the average annual holding, setup, and stock-out costs of the policies determined in parts (*a*) and (*b*).

   d. What is the safety stock for this item at the optimal solution?

30. Consider the problem of satisfying the demand for number 2 pencils faced by the campus store mentioned in Problem 29.

    a. Re-solve the problem, substituting a Type 1 service level criterion of 95 percent for the stock-out cost.

    b. Re-solve the problem, substituting a Type 2 service level criterion of 95 percent for the stock-out cost. Assume that Q is given by the EOQ.

    c. Find the simultaneous optimal values of Q and R assuming a Type 2 service level of 95 percent.

31. Answer the following true or false.

    a. The lead time is always less than the cycle time.

    b. The optimal lot size for a Type 1 service objective of X percent is always less than the optimal lot size for a Type 2 service objective of X percent for the same item.

    c. The newsvendor model does not include a fixed order cost.

    d. ABC analysis ranks items according to the annual value of their demand.

    e. For a finite production rate model, the optimal lot size to produce each cycle is equal to the maximum inventory each cycle.

32. One of the products stocked at Weiss's paint store, mentioned in Problem 14, is a certain type of highly volatile paint thinner that, due to chemical changes in the product, has a shelf life of exactly one year. Al Weiss purchases the paint thinner for $20 a gallon can and sells it for $50 a can. The supplier buys back cans not sold during the year for $8 for reprocessing. The demand for this thinner generally varies from 20 to 70 cans a year. Al assumes a holding cost for unsold cans at a 30 percent annual interest rate.

    a. Assuming that all values of the demand from 20 to 70 are equally likely, what is the optimal number of cans of paint thinner for Al to buy each year?

    b. More accurate analysis of the demand shows that a normal distribution gives a better fit of the data. The distribution mean is identical to that used in part (*a*), and the standard deviation estimator turns out to be 7. What policy do you now obtain?

33. Semicon is a start-up company that produces semiconductors for a variety of applications. The process of burning in the circuits requires large amounts of nitric acid, which has a shelf life of only three months. Semicon estimates that it will need between 1,000 and 3,000 gallons of acid for the next three-month period and assumes that all values in this interval are equally likely. The acid costs them $150 per gallon. The company assumes a 30 percent annual interest rate for the money it has invested in inventory, and the acid costs the company $35 a gallon to store. (Assume that all inventory costs are attached to the end of the three-month period.) Acid that is left over at the end of the three-month period costs $75 per gallon to dispose of. If the company runs out of acid during the three-month period, it can purchase emergency supplies quickly at a price of $600 per gallon.

    a. How many gallons of nitric acid should Semicon purchase? Experience with the marketplace later shows that the demand is closer to a normal distribution, with mean 1,800 and standard deviation 480.

b. Suppose that now Semicon switches to a 94 percent fill rate criterion. How many gallons should now be purchased at the start of each three-month period?

34. *Newsvendor simulator.* In order to solve this problem, your spreadsheet program will need to have a function that produces random numbers [@RAND in Lotus 1–2–3 and RAND ( ) in Excel]. The purpose of this exercise is to construct a simulation of a periodic-review inventory system with random demand. We assume that the reader is familiar with the fundamentals of Monte Carlo simulation.

Your spreadsheet should allow for cell locations for storing values of the holding cost, the penalty cost, the proportional order cost, the order-up-to point, the initial inventory, and the mean and the standard deviation of periodic demand.

An efficient means of generating an observation from a standard normal variate is the formula

$$Z = [-2 \ln(U_1)]^{0.5} \cos(2\pi U_2),$$

where U_1 and U_2 are two independent draws from a (0, 1) uniform distribution. (See, for example, Fishman, 1973.) Note that two independent calls of @RAND are required. Since Z is approximately standard normal, the demand X is given by

$$X = \sigma Z + \mu$$

where μ and σ are the mean and the standard deviation of one period's demand.

A suggested layout of the spreadsheet is

NEWSVENDOR SIMULATOR

Holding cost = **Mean demand =**
Order cost = **Std. dev. demand =**
Penalty cost = **Initial inventory =**
Order-up-to point =

Period	Starting Inventory	Order Quantity	Demand	Ending Inventory	Holding Cost	Penalty Cost	Order Cost
1							
2							
3							
.							
.							
.							
20							
Totals							

Each time you recalculate, the simulator will generate a different sequence of demands. A set of suggested parameters is

$$h = 2,$$
$$c = 5,$$
$$p = 20,$$
$$\mu = 100,$$
$$\sigma = 20,$$
$$I_0 = 50,$$

Order-up-to point = 150.

35. *Using the simulator for optimization.* You will be able to do this problem only if the program you are using does not restrict the size of the spreadsheet. Assume the parameters given in Problem 34 but use $\sigma = 10$. Extend the spreadsheet from Problem 34 to 1,000 rows or more and compute the average cost per period. If the cost changes substantially as you recalculate the spreadsheet, you need to add additional rows.

    Now, experiment with different values of the order-up-to point to find the one that minimizes the average cost per period. Compare your results to the theoretical optimal solution.

36. *Newsvendor calculator.* Design a spreadsheet to calculate the optimal order-up-to point for a newsvendor problem with normally distributed demands. In order to avoid table look-ups, use the approximation formula

$$Z = 5.0633[F^{0.135} - (1 - F)^{0.135}]$$

    for the inverse of the standard normal distribution function. (This formula is from Ramberg and Schmieser, 1972.) The optimal order-up-to point, y^*, is of the form $y^* = \sigma Z + \mu$.

    a. Assume that all parameters are as given in Problem 34. Graph y^* as a function of p, the penalty cost, for $p = 10, 15, \ldots, 100$. By what percentage does y^* increase if p increases from 10 to 20? from 50 to 100?

    b. Repeat part (*a*) with $\sigma = 35$. Comment on the effect that the variance in demand has on the sensitivity of y^* to p.

37. A large national producer of canned foods plans to purchase 100 combines that are to be customized for its needs. One of the parts used in the combine is a replaceable blade for harvesting corn. Spare blades can be purchased at the time the order is placed for \$100 each, but will cost \$1,000 each if purchased at a later time because a special production run will be required.

    It is estimated that the number of replacement blades required by a combine over its useful lifetime can be closely approximated by a normal distribution with mean 18 and standard deviation 5.2. The combine maker agrees to buy back unused blades for \$20 each. How many spare blades should the company purchase with the combines?

38. Crazy Charlie's, a discount stereo shop, uses simple exponential smoothing with a smoothing constant of $\alpha = .2$ to track the mean and the MAD of monthly item demand. One particular item, a stereo receiver, has experienced the following sales over the last three months: 126, 138, 94.

    Three months ago the computer had stored values of mean = 135 and MAD = 18.5.

    a. Using the exponential smoothing equations given in Section 5.1, determine the current values for the mean and the MAD of monthly demand. (Assume that the stored values were computed *prior* to observing the demand of 126.)

    b. Suppose that the order lead time for this particular item is 10 weeks (2.5 months). Assuming a normal distribution for monthly demand, determine the current estimates for the mean and the standard deviation of lead time demand.

    c. This particular receiver is ordered directly from Japan, and as a result there has been considerable variation in the replenishment lead time from one order