IE302 CLASS NOTES

To learn and not to do is really not to learn, to know and not to do is really not to know.

Stephen R. Covey

The time is always right to do what is right.

Martin Luther King Jr.

You have to learn the rules of the game. And then you have to play better than anyone.

Albert Einstein

ACKNOWLEDGMENTS

I would like to thank Sena Karadeniz for helping me prepare this LaTeX document.

Table of Contents

Chapter 1

Introduction

Production/Manufacturing is the process of converting *raw materials* (semi-finished products) into finished products that have **value** in the market place.
This process involves the contribution of labor, equipment, energy, and information.
Inventory is both an input and output of the production process. Inventory can be in the form of *raw materials, semi-finished, and finished products.*
Supply Chain Management (SCM) is the set of functions concerned with the *effective utilization of limited resources* and the *management of material flows* through these resources, so as to satisfy customer demands and create **profits.** Next, we discuss decisions to be made in a production system.

1.1 Decisions

A production system refers to the core set of processes, resources, technologies, and strategies employed by an organization to transform inputs into finished goods or services. This system is integral to the organization's overall operations and plays a central role in achieving its production goals and meeting customer demand.
Key decisions of a production system that help an organization achieve its broader objectives and success include determining what should be produced, how much, and when, which involves forecasting. It also involves understanding how much can be produced through capacity planning, assessing current inventory levels and future needs through inventory management and material requirement planning, and deciding when to produce by scheduling and implementing shop floor control.
In general, these key decisions can be categorized based on their effect on the organization and timeline.

1.1.1 *Strategic Level Decisions*

Strategic level decisions in production system analysis refer to the high-level choices and actions taken by an organization to optimize its overall production processes and achieve long-term goals. These decisions involve allocating resources, designing production facilities, selecting technology, and formulating policies that impact the entire production system.

Key aspects of strategic level decisions in production system analysis include strategic network optimization, which involves determining the number, location, and size of distribution centers and facilities. It also encompasses forming strategic partnerships with suppliers, distributors, and customers. Product life cycle management is critical, ensuring new and existing products are optimally integrated into supply chain activities. Information technology operations also play a significant role, as do decisions regarding where to manufacture products and whether to make or buy components.

1.1.2 *Tactical Level Decisions*

Tactical level decisions in the context of production system analysis refer to the mid-range decisions and actions taken by an organization to effectively implement the broader strategic plans. These decisions are more detailed and specific than strategic decisions and are focused on optimizing operations within a shorter time frame, typically spanning weeks to months. Tactical decisions bridge the gap between strategic planning and day-to-day operational activities.

Key aspects of tactical level decisions in production system analysis include sourcing contracts and purchasing decisions, which determine the procurement of materials and services. Production decisions are also critical, encompassing contracting, scheduling, inventory management, and planning to ensure efficient and timely production processes.

1.1.3 *Operational Level Decisions*

Operational level decisions in the realm of production and business refer to the day-to-day choices and actions taken by an organization to manage routine tasks and ensure the smooth functioning of its operational processes. These decisions are specific, short-term in nature, and directly impact the immediate efficiency and effectiveness of ongoing operations.

Key aspects of operational level decisions include routing and scheduling, which ensure the efficient flow of materials and products through the production process, and quality control, which maintains product standards and consistency throughout manufacturing.

Next, we discuss performance measures that are widely used in an organization.

1.2 Performance Measures

Key Performance Indicators (KPIs) are measurable values that help organizations assess their performance and progress toward specific goals. KPIs vary across industries and business functions, but here are some examples across different areas:
Cost: The total expenses incurred in the production process, encompassing both direct and indirect costs, determine if products are being created at a minimum or acceptable cost. It is important to understand the common standard regarding whether the products are produced at a minimum or acceptable cost.
Volume: The number of products that are currently being produced or have the potential to be produced within a given timeframe. It is crucial to ascertain the optimal production capacity, addressing the question of how much can be or is currently being produced.
Variety: The number of different types of products that the production system can manufacture, providing insight into the diversity of the product portfolio. It is crucial to explore the variety of products that are currently produced or have the potential to be manufactured. This involves a comprehensive analysis of market demands, technological capabilities, and diversification strategies. Identifying the range of products ensures businesses are well-positioned to adapt to changing market dynamics and capitalize on opportunities for growth.
Quality: The adherence of products to predefined specifications and standards. It also involves measuring the percentage of shipped products that meet these specifications, indicating the level of quality control. It is essential to define the specifications of products and evaluate the percentage of shipped products that meet these criteria. This entails establishing clear and measurable product standards, monitoring production processes, and implementing quality control measures.
Customer response time: The duration it takes to fulfill a customer order, from the point of order placement to the actual delivery of the product. Understanding the fulfillment time for customer orders is crucial for optimizing customer satisfaction and operational efficiency. This involves analyzing the entire order fulfillment process, from order placement to delivery. By evaluating and streamlining these processes, businesses can reduce fulfillment times, meet customer expectations, and enhance overall service efficiency.
On-time delivery: The percentage of orders delivered on or before the promised date. It is pivotal to assess the consistency of meeting quoted lead times to measure operational efficiency and customer satisfaction. This evaluation involves tracking the accuracy of estimated lead times provided to customers and comparing them with the actual fulfillment

durations.

Flexibility: The speed at which existing resources and processes can be reconfigured or adapted to accommodate the production of new products. It reflects the agility of the production system. Efficiently reconfiguring existing resources to produce a new product is a critical aspect of adaptability and innovation.

Worker satisfaction: The level of contentment and fulfillment experienced by participants in the production process, considering factors such as working conditions, job roles, and overall workplace environment. Assessing the satisfaction levels of participants in the production process is essential for fostering a positive and productive work environment. This includes evaluating the contentment of employees, suppliers, and other stakeholders involved in the production chain.

Safety: The degree to which the production environment ensures the well-being and security of workers and the surrounding community. It involves assessing and mitigating potential hazards. Ensuring a safe production environment for both workers and the surrounding community is a remarkable concern.

Environmental impact: The assessment of how environmentally friendly the production process and the resulting products are. It considers factors such as resource usage, waste generation, and emissions. Evaluating the environmental impact of our production process and products is crucial for sustainable business practices. By prioritizing environmentally friendly practices, businesses can minimize their ecological footprint, comply with regulatory standards, and appeal to environmentally conscious consumers.

These examples illustrate the diversity of KPIs and highlight their role in measuring and monitoring performance in various business functions. Organizations should select KPIs that align with their specific goals and objectives.

The Bottom Line: In the long run, the supply chain must be ***profitable*** by delivering **value** to the end customer.

1.3 Classification of Production Systems

A production system is usually classified based on aspects such as production quantity, product variety, order fulfillment, resource configuration, automation, production processes, and inputs/outputs. Below, we explore each of those in detail.

1. Production Quantity: Production quantity refers to the total number of units or items manufactured within a specified timeframe in a production system. It is a fundamental metric used to measure the output or throughput of the production process and is crucial for assessing the efficiency and productivity of manufacturing operations.

The production quantity is influenced by factors such as production capacity, order demand, and the effectiveness of the production system. This metric is often expressed in terms of the number of finished goods produced, and it can be measured on a daily, weekly, monthly, or annual basis, depending on the organization's reporting and analysis needs. Different types of production systems impact how production quantities are managed and reported, including mass production systems, batch production systems, and job shop systems. Mass production systems focus on high-volume, continuous production of standardized products. Batch production systems produce goods in groups or batches, allowing for flexibility and variation in production. Job shop systems handle custom or small-batch production, catering to specific, often unique, customer requirements.

Understanding and analyzing production quantity is essential for making informed decisions related to capacity planning, resource allocation, and overall production strategy. This metric plays a central role in evaluating the success of a production system in meeting organizational goals and customer expectations.

2. Product Variety: Product variety refers to the range and diversity of different products or variations produced by a manufacturing system. It is a crucial aspect of production planning and control, influencing various aspects of the production process, resource allocation, and overall operational efficiency. The level of product variety within a production system can significantly impact factors such as inventory management, production complexity, and the ability to respond to customer demands. Different approaches to product variety include producing a single product or product line, which simplifies production and inventory management but may limit market responsiveness; a family of similar products, which offers a balance between variety and efficiency by sharing common components or processes; and one-of-a-kind products, which cater to unique customer specifications but require highly flexible production systems and sophisticated planning to manage complexity and variability effectively.

Balancing product variety is a strategic decision for organizations, as it involves trade-offs between meeting diverse customer needs and maintaining operational efficiency. Understanding the implications of product variety is essential for designing production systems that can adapt to market demands while optimizing resources and minimizing costs.

3. Order Fulfillment: Order fulfillment refers to the comprehensive process of receiving, processing, and delivering customer orders efficiently and accurately. It encompasses the entire sequence of activities from the moment an order is placed until the product is delivered to the customer's satisfaction. Order fulfillment is a critical component of the supply chain and production system, directly impacting customer satisfaction, operational efficiency, and overall business success. Different strategies for order fulfillment include make-to-stock systems, where products are produced in advance and stored as inventory until customer orders are received; make-to-order systems, where production begins only after an order is placed, ensuring customization and reducing inventory costs; and hybrid systems, such as make-to-assemble, where components are produced and stocked in advance, but final assembly is done in response to specific customer orders, balancing responsiveness with efficiency.

Efficient order fulfillment is essential for maintaining customer loyalty, a positive brand reputation, and a competitive edge in the market. It requires seamless coordination among various departments within the organization, including production, logistics, and customer service. Analyzing and optimizing the order fulfillment process is a key focus of production system analysis to enhance overall operational effectiveness.

4. Resource Configuration: Resource configuration refers to the strategic arrangement and utilization of various resources within a manufacturing or operational environment to achieve specific production objectives. These resources encompass a wide range of elements, including human resources, machinery, technology, materials, and facilities. The configuration involves the allocation, organization, and coordination of these resources to optimize efficiency, productivity, and overall performance within the production system.

Various types of layouts can be utilized to achieve these goals. For example, a product layout arranges resources in a sequence to produce a specific product, making it ideal for mass production of uniform items. In contrast, a process layout groups similar processes or functions together, which is suitable for customized production with varied workflows. Moreover, group (cellular) layouts create cells of workstations or machines to produce a family of similar products, enhancing flexibility and efficiency. In addition to this, a fixed position layout keeps the product stationary while workers and equipment move to it, which is commonly used for large or heavy products like ships or buildings. Furthermore, hybrid layouts combine different layout types to balance efficiency and flexibility according to specific production needs.

Resource configuration is a dynamic process that requires continuous analysis, adjustment, and optimization to align with changing market conditions, technological advancements, and organizational goals. By strategically configuring resources, organizations aim to enhance their competitiveness, reduce costs, and achieve sustainable growth within their production systems.

5. Automation: Automation refers to the integration and utilization of technology and machinery to perform tasks, processes, or operations with minimal human intervention. The objective of automation is to increase efficiency, enhance productivity, improve accuracy, and reduce labor-intensive efforts within the production system.

Automation can be applied to various stages of manufacturing, from simple, repetitive tasks to complex, intricate processes. For instance, some operations might involve no automation at all, relying entirely on manual operators to perform tasks. On the other hand, dedicated automation is used for specific, repetitive tasks where the same operation is performed continuously without variation. Moreover, programmable automation allows for the reprogramming of equipment to handle different tasks or batches, offering more versatility than dedicated systems. In addition to this, flexible automation can quickly adapt to changes in product design or production schedules, making it suitable for environments with high product variety and frequent changes. Overall, the level of automation implemented in a manufacturing process depends on the specific requirements and goals of the production system.

Automation is a key element in modern production systems, and its application varies across industries and processes. Whether it involves simple programmable logic controllers (PLCs) in basic manufacturing or highly sophisticated robotic systems in advanced industries, the goal remains to enhance efficiency, accuracy, and competitiveness within the production environment. The analysis of automation in production systems considers factors such as initial investment, system integration, training requirements, and the overall impact on operational performance.

6. Production Process: Production process refers to the systematic series of steps and activities involved in transforming raw materials, components, or inputs into finished goods or services within a manufacturing or operational setting.

The analysis of production processes is a fundamental aspect of understanding and optimizing the efficiency, effectiveness, and overall performance of a production system. This analysis encompasses various stages, such as the continuous process of raw material transformation, component fabrication, final assembly, and even re-manufacturing and recycling. Assessing each stage for efficiency involves identifying bottlenecks and implementing strategies to enhance overall performance. Moreover, understanding the intricacies of these processes is essential for production system optimization, cost reduction, and meeting customer expectations for quality and timely delivery. By thoroughly analyzing and optimizing each stage, organizations can ensure a more efficient and effective production system, ultimately leading to better business outcomes.

7. Inputs/Outputs: Inputs and outputs refer to the resources and results involved in the transformation process within a production system. Inputs represent the resources consumed or utilized in the production process, while outputs represent the results or outcomes of the production process.

There are various types of production systems that manage these inputs and outputs differently. Discrete production systems handle discrete inputs and outputs, such as cars, computers, and machine tools. Continuous production systems deal with continuous inputs and outputs, typical in industries like chemicals, textiles, food processing, and pharmaceuticals. On the other hand, hybrid systems can manage discrete inputs with continuous outputs or continuous inputs with discrete outputs, as seen in the production of steel, plastics, and recycling processes.

These terms help describe the flow of materials, energy, information, and labor throughout production. Analyzing inputs and outputs is crucial for understanding efficiency, identifying areas for improvement, and optimizing the overall performance of a production system [Nahmias, 1997].

1.4 Product Cycles

The product cycle is a concept in production system analysis that describes the stages a product goes through from its initial introduction to the market until its eventual decline and discontinuation. This life cycle is typically divided into distinct phases, each with unique characteristics and challenges. The analysis of the product life cycle is essential for making strategic decisions related to production, marketing, and resource allocation.

Introduction/Startup: The product is launched into the market. Sales are low, and there may be high development and marketing costs. (job shop-like, low volume, frequent design changes, manual operation, multiple suppliers, high unit costs)

(Rapid) Growth: The product gains market acceptance, and sales begin to increase. Profits typically improve during this phase. (batch production, larger volumes, partial automation, fewer design changes, fewer suppliers, lower unit costs)

Maturity: Sales stabilize during the maturity phase, and the product reaches its peak market penetration. Competition may increase, leading to price competition. (continuous flow, stable demand, greater automation, periodic design updates, few suppliers with longer-term contracts, unit costs are at their lowest)

Stabilization/Decline: Sales and profitability decline due to factors such as market saturation, changing consumer preferences, or technological advancements. (batch production, resources are shared with the next generation of products, no design changes)

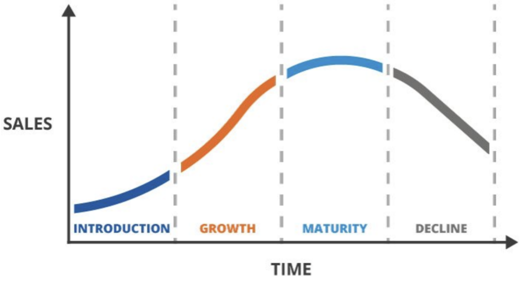


Fig. 1.1 Product life cycle (Credit: TWI)

Question: Think about items that are in one of those phases. Which phase is the electric car in? How about smartphones?

1.5 Process Capabilities and Business Strategy

Product attributes such as price, quality, variety, demand uncertainty, delivery time, and response time must align with process attributes like cost, quality, flexibility, and cycle time. A firm must choose a **business strategy** (attribute values for its portfolio of products) that differentiates it from the competition. A firm must choose **process capabilities** (attribute values for its process) that support its business strategy.

A business strategy can be driven by *market opportunities* or by a *competitive advantage* in process capabilities. In both cases, there must be a fit between process capability and business strategy.

Process Choices

Facility Size

Should we have a few large manufacturing facilities or many smaller ones?

Few large facilities may benefit from economies of scale, but smaller facilities offer flexibility and may reduce transportation costs.

Facility Specialization

Should each facility be dedicated to a few products or shared among many?

Should facilities be specialized or should they have overlapping capabilities?

Specialized facilities can optimize processes, while shared facilities offer versatility. The choice depends on product complexity and market demands.

Production Strategy

Should we produce to stock or make to order?

Producing to stock ensures product availability but may lead to excess inventory. Making to order minimizes inventory but may impact delivery times. The choice depends on customer demand and market expectations.

Facility Location

Where should facilities be located?

Location impacts transportation costs, lead times, and proximity to suppliers and customers. The decision involves considerations of cost, market access, and supply chain efficiency.

Make or Outsource

Should we make our products mostly in-house or should we outsource operations as much as we can?

In-house production provides control but may be costlier. Outsourcing can reduce costs but may impact control and responsiveness. The decision depends on core competencies and strategic goals.

Ownership of Transportation and Distribution

Should we own our transportation and distribution system or should we contract them out?

Owning transportation and distribution provides control but may require substantial investment. Contracting out may offer cost savings. The decision depends on the company's expertise and resources.

Sales Channels

Should we sell through our own dealers, independent retailers, or directly to the customer?

Choosing between dealers, retailers, or direct-to-customer sales depends on market reach, customer preferences, and the desired level of control over the sales process.

Supplier Relationships

Should we have dedicated suppliers or should we always solicit competitive bids?

Dedicated suppliers may offer reliability, but competitive bids may lead to cost savings. The decision involves balancing reliability and cost-effectiveness.

Warehouse Strategy

Should we have multiple regional warehouses of finished goods or should we centralize inventory in one location?

Multiple regional warehouses enhance responsiveness, while centralized inventory reduces costs. The choice depends on distribution efficiency and customer service requirements.

Technology and Globalization

Should we invest in automation technologies or should we offshore manufacturing to countries where labor is cheap? Should we compete locally, nationally, or globally?

Integration of technology and globalization into process capabilities and business strategy is essential for organizations aiming to stay competitive, innovative, and resilient in the dynamic global marketplace. The effective use of technology and a thoughtful approach to globalization can shape the core capabilities of a business and contribute to the formulation of strategic initiatives.

Product Strategy

Should we encourage feature proliferation or should we standardize product offerings?

This strategy encompasses decisions related to product development, differentiation, pricing, distribution, and ongoing product lifecycle management. A well-defined product strategy aligns with overall business goals and leverages the organization's process capabilities to create a competitive advantage.

Criteria for Systems

Optimality: Optimality refers to the condition in which specific performance metrics or criteria are maximized or minimized to achieve the best overall system performance. Assessing the effectiveness of recommended solutions and actions is essential for achieving desired outcomes.

Accuracy: Accuracy as a criterion refers to the precision and correctness with which a production system achieves specific goals, measures, or outcomes. Ensuring the accuracy of produced information is fundamental for informed decision-making and maintaining credibility.

Robustness: Robustness as a criterion refers to the ability of a system to maintain stable and reliable performance in the face of uncertainties, variations, or unexpected changes in conditions. Assessing the consistency of system performance is crucial for reliability and efficiency.

Reconfigurability: Reconfigurability as a criterion refers to the capability of a system to easily and efficiently adapt to changes in its configuration, structure, or operational parameters. Evaluating the adaptability of the system to new situations is vital for staying responsive to evolving needs and challenges. A system that is easy to adapt enables organizations to stay agile and quickly respond to evolving conditions, ensuring sustained relevance and effectiveness in dynamic environments.

Integrability: Integrability as a criterion refers to the ability of a system to seamlessly integrate and interact with various components, technologies, and processes, both within and outside the system. Assessing the ease of integration with other information and decision support systems is crucial for optimizing efficiency and data flow.

Profitability/Cost: Profitability and cost are essential criteria that evaluate the financial performance and efficiency of a production system. These criteria assess the relationship between the costs incurred in the production process and the revenues generated from the sale of goods or services. Evaluating the cost-effectiveness of implementing and operating the system is essential for budgetary considerations and long-term sustainability. Striking a balance between functionality and cost ensures that the system aligns with financial objectives while delivering value to the organization.

Ease of Use/Transparency: Ease of use and transparency are criteria that focus on the accessibility, simplicity, and clarity of the production system's design, operation, and information flow. These criteria are essential for ensuring that the production system is user-friendly, easy to understand, and facilitates efficient decision-making for operators, managers, and other stakeholders. Assessing the user-friendliness and comprehensibility of the system is crucial for maximizing its effectiveness across all participants. A system that is easy to use and comprehend fosters increased user adoption, productivity, and satisfaction, contributing to the overall success of the organization's operations.

Bottomline: Success lies in **value creation.** If there is no value proposed by a company, process, product, etc., you can optimize all you want, and it will not be profitable. By the way, marketing is the clarion of value.

Chapter 2

Forecasting

Forecasting involves the systematic use of historical data, statistical models, and advanced methodologies to predict future trends, demand patterns, and resource needs, empowering businesses to make informed decisions and proactively address challenges in the dynamic landscape of production and manufacturing.

2.1 Need For Forecasting

Forecasting plays a crucial role in various aspects of business and strategic decision-making.

Determining production plans

Optimizing Inventory: Forecasting helps in estimating the demand for products, allowing businesses to optimize inventory levels. This prevents overstocking or stockouts, reducing carrying costs and improving overall operational efficiency.

Resource Allocation: By predicting future demand, businesses can plan their production schedules, ensuring that resources such as raw materials, equipment, and labor are allocated efficiently.

Determining capacity requirements

Resource Planning: Forecasting enables businesses to assess future demand and plan for the required production capacity. This helps in optimizing the use of resources and avoiding underutilization or overloading of production facilities.

Capital Investment: Accurate forecasting aids in making informed decisions regarding the need for expanding or upgrading production facilities. This ensures that capital investments align with future demand.

Determining labor requirements

Workforce Planning: Forecasting helps in predicting the volume of work, allowing businesses to plan their workforce requirements accordingly. This includes hiring, training, and scheduling employees based on anticipated demand.

Cost Management: By aligning labor requirements with production needs, businesses can control labor costs, ensuring that they have the right number of workers at the right time to meet demand.

DETERMINING PRODUCT VIABILITY

Market Research: Forecasting involves analyzing market trends and consumer behavior, providing insights into the potential success of a new product. This helps in assessing the viability of introducing new products or making modifications to existing ones.

Risk Mitigation: Accurate forecasts contribute to risk assessment by identifying potential challenges and uncertainties in the market. Businesses can adjust their strategies or take preventive measures to mitigate risks associated with product viability.

FAMOUS FORECASTING/ PREDICTION ERRORS

"TV won't be able to hold on to any market it captures after the first 6 months. People will soon get tired of staring at a plywood box every night."

- Darryl F. Zanuck, Head of 20th Century Fox, 1946

"I think there is a world market for maybe five computers."

- Thomas Watson, chairman of IBM, 1943.

"There is no reason anyone would want a computer in their home."

- Ken Olson, president, chairman, and founder of Digital Equipment Corp., 1977.

2.2 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 cach.

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, and resource requirements planning.

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.

Long-term production and manufacturing decisions are part of the overall firm's manufacturing strategy. One example is the long-term planning of capacity needs.

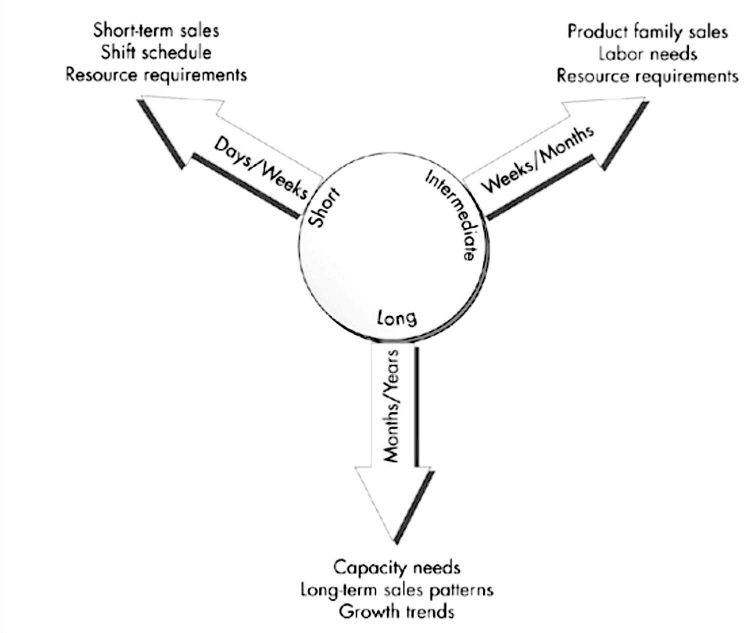


Fig. 2.1 Forecast horizons in operation planning (Credit: [Nahmias, 1997])

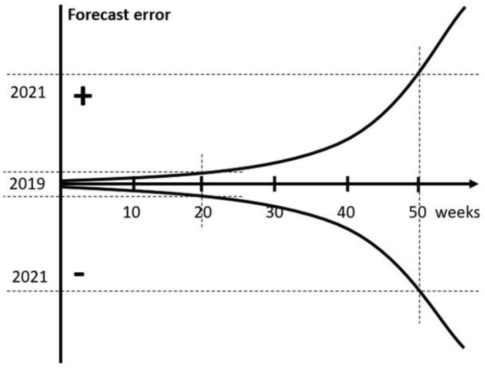


Fig. 2.2 Forecast horizon versus error (Credit: Urbańczyk, DOI: 10.12775/JPM.2022.005)

2.3 Characteristics of Forecasts

(1) *Forecasts 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) *Forecast is more than as single number.* A good forecast also gives some measure of 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 applies to forecasting. 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.

(4) *The longer the horizon, the less accurate forecast will be.* This property is quite intuitive.

(5) *Known (future) information should not be ignored.* 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 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.

2.4 Objective Forecasting

Objective forecasting methods involve deriving forecasts through data analysis, excluding the personal judgment of the forecaster. The primary goals of these forecasts are to predict future trends based on past data, smooth out random variations or "noise", and standardize the forecasting procedure to ensure consistency and reliability. By relying on data-driven techniques, these methods aim to provide accurate and unbiased forecasts essential for effective decision-making and planning.

Forecasting can be categorized into two alternative approaches: (i) **time series** methods that only use past values of the phenomenon we are predicting and (ii) **causal models** that use data from sources other than the series being predicted. These other resources may be other variables with values that are linked in some way to what is being forecasted.

Some of these methods include causal forecasting and time series forecasting. Causal forecasting, such as regression analysis, identifies and models the relationships between different variables to predict future values. Time series forecasting techniques include moving averages, which smooth out fluctuations by averaging data over a specific period; exponential smoothing, which applies decreasing weights to past observations; regression analysis, which

can also be used in a time series context to model trends and patterns; and seasonal models, which account for recurring patterns within the data. These methods help enhance forecasts' accuracy and reliability by systematically analyzing historical data and identifying underlying trends and relationships.
Question: What is the methodology for weather forecasts? Make a guess.

2.4.1 *Causal Models*

Causal models in forecasting involve predicting a dependent variable $f(t)$ based on one or more independent variables x_i with associated lead times l_i. The general form of a causal model is expressed as

$$f(t) = \phi(x_1(t-l_1),\ x_2(t-l_2),\ \ldots,\ x_m(t-l_m))$$

x_i : independent variable (e.g., price, number of hours studied)
l_i : lead time for variable i
$f(t)$: dependent variable (e.g., demand, grade received)

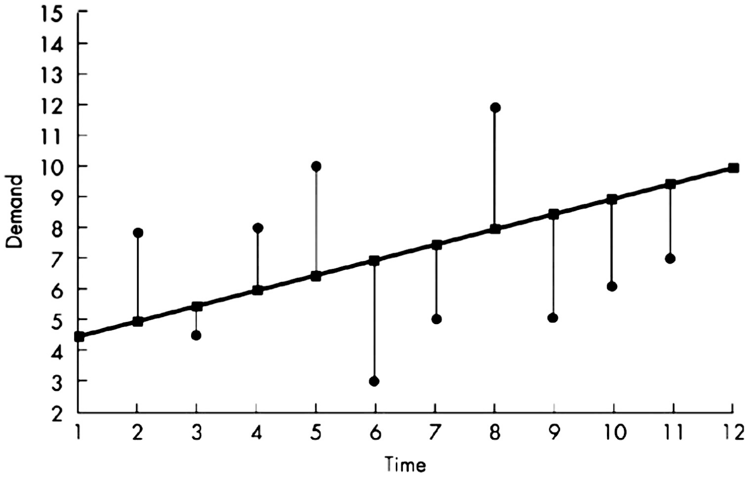


Fig. 2.3 An example of a regression line (Credit: [Nahmias, 1997])

Example:
A specific causal model might be expressed as

$$f(t) : a_0 + a_1 x_1(t-l_1) + a_2 x_2(t-l_2) x_3(t-l_3) + a_2 x_2(t-l_2)^2$$

where $a_0,\ a_1,\ a_2$ are coefficients that quantify the relationship between the dependent and independent variables. This model demonstrates how different independent variables and their interactions can be used to predict the outcome of the dependent variable.

2.4.1.1 *Limitations of Regression Models*

They assume a stable relationship between **dependent** and **independent** variables, which may not always hold in dynamic environments. Additionally, these models require prior knowledge of the values of independent variables, which might necessitate further forecasting efforts for those variables. Furthermore, regression models are data-intensive, necessitating a large database of historical data to produce accurate and reliable forecasts.

2.4.1.2 *Some Derivations for Linear Regression*

Linear regression analysis is a method that fits a straight line to a set of data. It identifies the trend of dependent variables with respect to independent variables.

$$D(t) = a_0 + a_1 t + \epsilon_t$$

$D(t)$: actual demand at period t
$a_0,\ a_1$: intercept and slope of the demand function
ϵ_t: random noise in the process at time t
($E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \sigma^2$)
We would like to build a forecasting model of the form:

$$f(t) = \hat{a_0} + \hat{a_1} t$$

such that error $e(t) = |D(t) - f(t)|$ is as small as possible for any t.
Minimizing the sum of squared errors:

$$\text{SSE} = \sum_{t=1}^{n} (D(t) - f(t))^2$$

we obtain:

$$\hat{a_0} = \frac{\sum_{t=1}^{n} t^2 \sum_{t=1}^{n} D(t) - \sum_{t=1}^{n} t \sum_{t=1}^{n} tD(t)}{n \sum_{t=1}^{n} t^2 - (\sum_{t=1}^{n} t)^2},$$

$$\hat{a_1} = \frac{n \sum_{t=1}^{n} tD(t) - \sum_{t=1}^{n} t \sum_{t=1}^{n} D(t)}{n \sum_{t=1}^{n} t^2 - (\sum_{t=1}^{n} t)^2}$$

The coefficient of determination – some sort of goodness of fit for a regressor relative to the average value:

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (f(t) - D(t))^2}{\sum_{t=1}^{n} (D(t) - \overline{D})^2}$$

Here, $f(t)$ is a forecasted value associated with period t, and $D(t)$ is the actual demand in period t. $\overline{D}$ is the average of all observations, i.e.,

$$\overline{D} = \sum_{t=1}^{n} D(t)/n$$

Example 2.1

Find the R^2 values for each of the following forecasts.

Period	**1**	**2**	**3**	**4**
Demand	90	100	110	100
Forecast 1	95	105	105	105
Forecast 2	100	100	100	100
Forecast 3	90	100	110	100
Forecast 4	100	200	250	275

Example 2.2

Find the linear regressor for the following data and find the R^2 value.

	t	*A(t)*	*t²*	*tA(t)*
	0	100	0	0
	1	115	1	115
	2	116	4	232
	3	125	9	375
	4	135	16	540
Sum	**10**	**591**	**30**	**1262**

$$\hat{a}_0 = \frac{30(591) - 10(1262)}{5(30) - (10)^2} = 102$$

$$\hat{a}_1 = \frac{5(1262) - 10(591)}{5(30) - (10)^2} = 8$$

$$f(t) = 102 + 8t \rightarrow F(5) = 102 + 8(5) = 142$$

$$R^2 = 0.948$$

2.4.1.3 *A Note on Nonlinear Regression*

Here, the only difference is that we aim to fit a nonlinear function of time to represent the dependent variable. We will only present some examples, without any derivation.

Examples:

$$D(t) = BC^t$$

$$D(t) = t/(Bt - C)$$

2.4.2 *Time Series Forecasting*

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.

Notations:

$D(t)$: observation in period t

$f(t+\tau)$: forecast for period $t+\tau$, where t is known to be the current time

Note that we also use a more accurate representation for forecasts using f with two indices. In such cases, we imply the forecast is done from the time denoted with the first index, for the period denoted in the second index. That is

$f(t_1,\ t_2)$: forecast for period t_2 with the data available until t_1

$F(t)$: smoothed estimate as of period t

$T(t)$: smoothed trend as of period t

Historical data ***Forecast***

$D(t),\ t = 1,\ \ldots,\ n \rightarrow \boxed{\text{Time series model}} \rightarrow f(n+\tau),\ \tau = 1, 2,\ \ldots$

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

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.

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

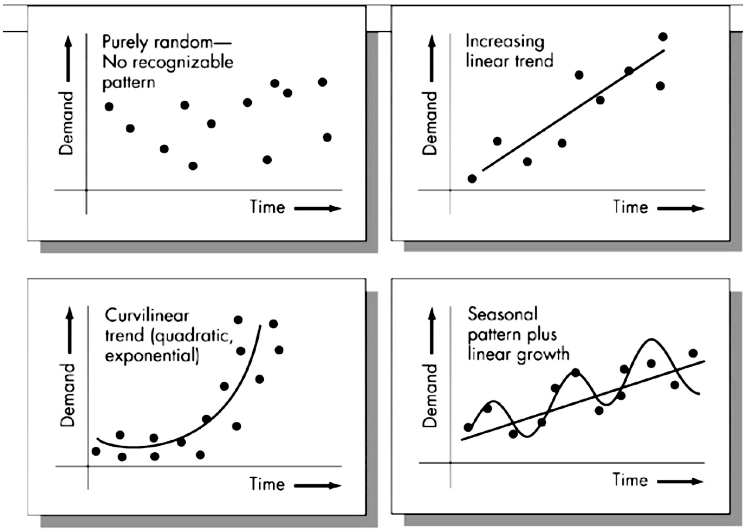


Fig. 2.4 Different time series patterns. (Credit: [Nahmias, 1997])

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.

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.

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.

2.4.2.1 *Data Averaging*

Data averaging assumes that equal weight is given to past observations. The model uses the average of all past observations $D(i)$ at time t, expressed as

$$F(t) = \sum_{i=1}^{t} D(i)$$

$$f(t+\tau) = F(t), \ \tau = 1, 2, \ldots$$

This implies that the forecast for any future period $t + \tau$ is equal to the average of all past observations up to time t. This method is suitable for data that does not exhibit trends or seasonality, as it assumes all past data points contribute equally to the prediction.

2.4.2.2 *Weighted Average*

Weighted averaging assumes that different weights are assigned to each of the past observations. The model calculates the weighted average of all past observations $D(i)$ at time t, expressed as

$$F(t) = \frac{\sum_{i=1}^{t} w_i D(i)}{\sum_{i=1}^{t} w_i}$$

$$f(t+\tau) = F(t), \ \tau = 1, 2, \ldots$$

This method allows more recent observations or those deemed more relevant to have a greater impact on the forecast, making it suitable for data where certain past values are more indicative of future trends.

2.4.2.3 *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 for period t (which is the smoothed estimate in period $t-1$, $F(t-1)$) is given by

$$f(t) = F(t-1) = (1/N) \sum_{i=t-N}^{t-1} D(i) = (1/N)(D(t-1) + D(t-2), + \ldots + D(t-N)).$$

This implies 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.

It may be confusing to understand the difference between forecast (f) and estimate (F) notations. For instance, as of February 2024, the USD exchange rate is 30 TRY and has an increasing trend. We can argue that the F (estimate) is 30 TRY for February and forecast formula is $f(t_\tau) = F(t) + 0.25\tau$.

Do we know the estimate for March? No, we have not realized it yet. Can we come up with a forecast for March? Yes, for a one-month ahead forecast, we need to add one times the trend on top of the estimate for February, as the formula suggests. Likewise, for April it is twice the trend added to the February estimate.

Example 2.3

Quarterly data for the failures of certain aircraft engines at a military base during the last two years (8 quarters) are as follows:

Period	1	2	3	4	5	6	7	8
Demand	200	250	175	186	225	285	305	190

Both three-quarter and six-quarter moving averages are used to forecast the number 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.

$$f(4) = F(3) = (1/3)(200 + 250 + 175) = 208.33$$

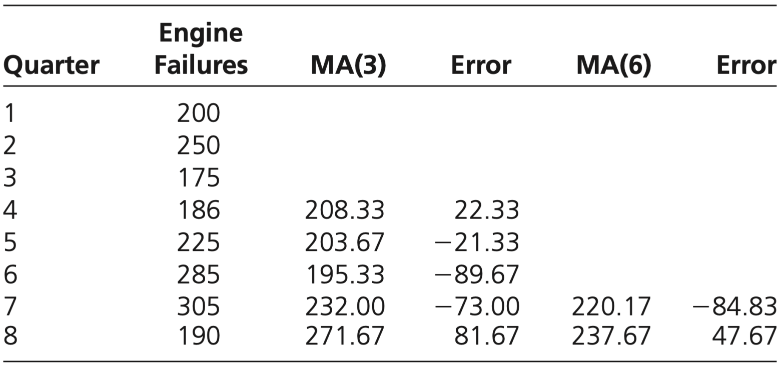
The three-period moving average forecast for period 5 is

$$f(5) = F(4) = (1/3)(250 + 175 + 186) = 203.67$$

Other forecasts are computed similarly. We can also compute six-month moving average forecasts. For instance, the six-period moving average forecast for period 7 is

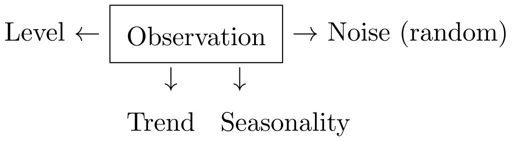
$$f(7) = F(6) = (1/6)(200 + 250 + 175 + 186 + 225 + 285) = 220.17$$

Arranging the forecasts and the associated forecast errors in a table, 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

We will look at smoothing methods from now on. These methods assume that data has the following components:



These methods differ from each other in their use of these components.

2.4.2.4 *Exponential Smoothing (Single)*

Another very popular forecasting method for stationary time series is exponential smoothing. This method assumes that there is no persistent trend and seasonality in data. The method relies on exponentially declining weights assigned to past observations, emphasizing recent data points more.

Model:

$$F(t) = \alpha D(t) + (1-\alpha)F(t-1)$$

$$(\alpha\text{: smoothing constant (given) where } 0 < \alpha \leq 1)$$

$$f(t+\tau) = F(t),\ \tau = 1, 2, \ldots$$

Exponential smoothing is a special case of weighted average.

$$F(t) = \alpha D(t) + (1-\alpha)F(t-1)$$

$$= \alpha D(t) + (1-\alpha)(\alpha D(t-1) + (1-\alpha)F(t-2))$$

$$= \alpha D(t) + (1-\alpha)\alpha D(t-1) + (1-\alpha)^2 \alpha D(t-2) + \ldots + (1-\alpha)^t F(0)$$

$$= \alpha \sum_{i=0}^{t} (1-\alpha)^i D(t-i)$$

A moving average with m periods is equivalent to an exponential smoothing model with $\alpha = 2/(m+1)$ – both methods have the same distribution of forecast errors, but the forecasts are not necessarily the same.

Note that you should choose large α if you want to capture recent data. Conversely, if α is small, then more weight is placed on past data and the forecasts are more stable. In short, as α increases, stability decreases, and closeness to demand increases. Generally, typical values for the parameter α are often found in the [0.1, 0.3] range.

Example 2.4

Consider the example in which moving averages were used to predict aircraft engine failures. The observed number of failures over a two-year period was as follows:

Period	1	2	3	4	5	6	7	8
Demand	200	250	175	186	225	285	305	190

We will now forecast using exponential smoothing to get the method started, let us assume that the forecast for period 1 was 200. Suppose that $\alpha = 0.1$. By assumption, the first-period forecast is assumed to be equal to demand ($f(1) = F(0)$).

Solution

The one-step-ahead forecast for period 2 is

$$f(2) = F(1) = \alpha D(1) + (1-\alpha)F(0) = (0.1)200 + (0.9)200 = 200$$

Similarly,

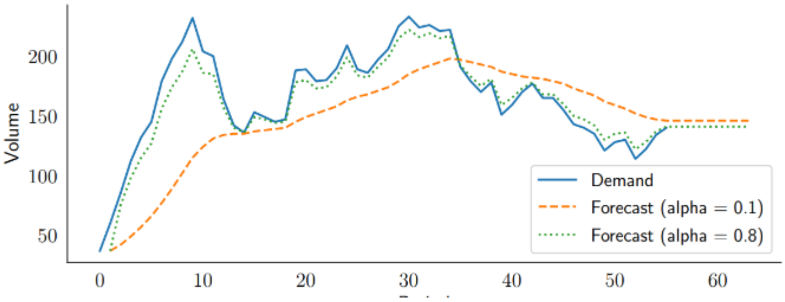
$$f(3) = F(2) = \alpha D(2) + (1-\alpha)F(1) = (0.1)250 + (0.9)200 = 205$$

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 (f)
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 a high variance, the forecasts are quite stable. Repeat the calculations with a value of $\alpha = 0.4$. 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 this example.



Notice that for $\alpha = 0.1$, the predicted value of demand results in a relatively smooth pattern, whereas for $\alpha = 0.8$, 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.

2.4.2.5 *Comparison of ES and MAs*

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

Similarities

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 data.

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. *Both methods will lag behind a trend if it exists.*

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

The ES forecast is a weighted average of *all* past data points (as long as the smoothing constant is strictly less than 1). MA 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 MA forecast after N periods but remains forever in the ES forecast.

To use moving averages, one must save all N past data points. To use ES, one need only save the last forecast. This is the most significant advantage of the ES method and one reason for its popularity in practice. 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 300,000x12= 3,600,000 pieces of information. If ES 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 ES appears to be more popular than MA for production-planning applications.

2.4.2.6 *Forecast Error Analysis*

Procedure

(1) Select model that computes $f(n+\tau)$ from $D(t),\ t = 1, \ldots, n$

(2) Forecast **existing** data and evaluate quality of fit.

(3) Stop if the fit is acceptable. Otherwise, adjust model constants and go to (2) or reject model and go to (1).

Measures of Error

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 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). Note that 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.

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).

They are given by the following formulae:

$$\text{MAD} = \sum_{t=1}^{n} |f(t) - D(t)|/n$$

$$\text{MSE} = \sum_{t=1}^{n} (f(t) - D(t))^2/n$$

$$\text{BIAS} = \sum_{t=1}^{n} (f(t) - D(t))/n$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{|f(t) - D(t)|}{|D(t)|}$$

2.4.2.7 *Double Exponential Smoothing with Holt's Method*

This method assumes that there is a linear trend in the data but not seasonality and weights for past observations or trends decline exponentially.

The method requires the specification of two smoothing constants, α and β, and uses two smoothing equations: one for the value of the series and one for the trend. The model is:

$$F(t) = \alpha D(t) + (1 - \alpha)[F(t-1) + T(t-1)]$$

$$T(t) = \beta[F(t) - F(t-1)] + (1 - \beta)T(t-1)$$

$$f(t + \tau) = F(t) + \tau T(t)$$

Recall what F, T, and f are.

$F(t)$: smoothed estimate for data

$f(t)$: forecast

$T(t)$: trend

Example 2.5

Let us apply Holt's method to the problem of developing one-step-ahead forecasts for the aircraft engine failure data in Example 2.3. Recall that the original series was 200, 250, 175, 186, 225, 285, 305, 190. Assume that both α and β are equal to 0.1. To get the method started, we need estimates for F and T at time zero. Suppose that these are $F(0) = 200$ and $T(0) = 10$.

Solution

$$F(1) = (0.1)(200) + (0.9)(200 + 10) = 209$$

$$T(1) = (0.1)(209 - 200) + (0.9)(10) = 9.9$$

$$F(2) = (0.1)(250) + (0.9)(209 + 9.9) = 222$$

$$T(2) = (0.1)(222 - 209) + (0.9)(9.9) = 10.2$$

$$F(3) = (0.1)(175) + (0.9)(222 + 10.2) = 226.5$$

$$T(3) = (0.1)(226.5 - 222) + (0.9)(10.2) = 9.6$$

and so on.

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

Period	Actual	Forecast (f)	|Error|
4	186	236.1	50.1
5	225	240.3	15.3
6	285	247.7	37.3
7	305	260.8	44.2
8	190	275	85

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.

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.

2.4.3 *Forecasting with Seasonality*

A seasonal series is one that has a pattern that repeats every N period for some value of N (which is at least 3).

We refer to the number of periods before the pattern begins to repeat as the length of the season. Note that this is different from the popular usage of the word *season* as a time of year. 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 multipliers c_t represent 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 $c_5 = 0.6$, 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.

We present a simple method of computing seasonal factors for a time series with seasonal variation and no trend. 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 first period of a season, all the factors corresponding to the 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

Use the following data, and assume it has a cyclic behavior once every 4 observations (e.g., quarterly data with annual cycles). Find the four seasonal factors, and indicate if there is a certain trend using *deseasonalized data.*

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

Solution

First, we compute the average of all the observations. The mean is:

$$\frac{10+20+26+17+12+23+30+22}{8} = 20$$

Then, we divide each observation by the mean value.

$Q_1 : \frac{10}{20} = 0.5$ $Q_2 : \frac{20}{20} = 1$ $Q_3 : \frac{26}{20} = 1.3$ $Q_4 : \frac{17}{20} = 0.85$ $Q_5 : \frac{12}{20} = 0.6$ $Q_6 : \frac{23}{20} = 1.15$ $Q_7 : \frac{30}{20} = 1.5$ $Q_8 : \frac{22}{20} = 1.1$

Finally, we average factors corresponding to the same period of the season. That is, average all factors for period 1, all factors for period 2, and so on. The resulting four seasonal factors are:

$$\frac{0.5+0.6}{2} = 0.550$$

$$\frac{1+1.15}{2} = 1.075$$

$$\frac{1.3+1.5}{2} = 1.400$$

$$\frac{0.85+1.1}{2} = 0.975$$

These factors add up to 4, as expected. Next, we find the deseasonalized data by dividing each observation by the associated seasonal factor.

Quarter	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

Do you see a trend here?

Example 2.7

The data in Table 13 represents the sales for each quarter during four different years. Compute seasonal indices and find the reseasonalized forecast by applying exponential smoothing for each year and each quarter ($\alpha = 0.5$ and $F(0) = 80$).

Table 13: Sales Data

Year	Quarter	Sales
1	1	77
	2	88
	3	92
	4	80
2	1	78
	2	90
	3	94
	4	78
3	1	109
	2	128
	3	130
	4	118
4	1	103
	2	136
	3	141
	4	123

Solution

1. Find the average: $(77 + 88 + 92 + \cdots + 123)/16 = 104.06$
2. Divide each observation by the mean value (Normalization):

$$Q_1 = \frac{77}{104.06} = 0.74$$

$$Q_2 = \frac{88}{104.06} = 0.85 \ldots$$

3. Average factors:

$$\frac{c_1 + c_5 + c_9 + c_{13}}{4} = \frac{0.74 + 0.75 + 1.05 + 0.99}{4} = 0.88 \ldots$$

4. Deseasonalized factors (sales):

For Year 1, 1st Quarter: $\frac{77}{0.88} = 87.33\ldots$

5. Deseasonalized Exponential Smoothing:

$$F(t) = \alpha D(t) + (1-\alpha)F(t-1)$$

$$\alpha = 0.5$$

$$f(1) = F(0) = 80$$

$$f(2) = F(1) = \alpha D(1) + (1-\alpha)F(0)$$

$$f(2) = F(1) = (0.5)87.33 + (1-0.5)80 = 83.66$$

$$\ldots$$

6. Reseasonalized Exponential Smoothing:

$$80(0.88) = 70.53$$

$$83.66(1.06) = 88.84\ldots$$

Year	Quarter	Average Factors	Deseasonalized	Des Exp. Smo. ($f(t)$)	Res Exp. Smo.
1	1	0.8817	87.3331	80	70.5345
	2	1.0619	82.8733	83.6666	88.8423
	3	1.0979	83.7965	83.2699	91.4219
	4	0.9586	83.4586	83.5332	80.0715
2	1	0.8817	88.4673	83.4959	73.6168
	2	1.0619	84.7568	85.9816	91.3006
	3	1.0979	85.6182	85.3692	93.7267
	4	0.9586	81.3722	85.4937	81.9507
3	1	0.8817	123.6274	83.4329	73.5613
	2	1.0619	120.5430	103.5302	109.9347
	3	1.0979	118.4081	112.0366	123.0047
	4	0.9586	123.1015	115.2223	110.4474
4	1	0.8817	116.8222	119.1619	105.0629
	2	1.0619	128.0769	117.9921	125.2913
	3	1.0979	128.4272	123.0345	135.0793
	4	0.9586	128.3177	125.7309	120.5204
		0.8817		**127.0243**	**111.9950**

2.4.3.1 *Winter's Method*

Winter's method is a type of triple exponential smoothing that incorporates components such as trend, seasonality, and level within the data. Notably, it is characterized by its ease of updating as new data becomes available, facilitating dynamic adjustments. Additionally, the model adopts a multiplicative approach, assuming that each season follows a consistent length denoted as N. This combination of features enhances the model's ability to capture and adapt to the underlying patterns in the time series data.

We assume a model of the form

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

μ: base signal or intercept at time zero, excluding seasonality

G_t: trend or slope at time t

c_t: multiplicative seasonal component at time t

ϵ_t: error term associated with time t

(a) Solution Procedure

Three exponential 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

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

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. *Seasonal factors (seasonality).*

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

Finally, the forecast made from period t to period $t+\tau$ is given by

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

(b) Recommended Initialization Procedure

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 as the initial slope estimate.

$$G_0 = (V_2 - V_1)/N$$

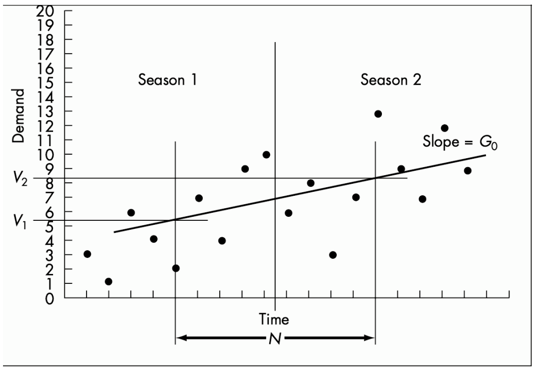


Fig. 2.5 Initialization for Winters's method (Credit: [Nahmias, 1997])

3. Set S_0 as the estimate of time series at time zero. Note that S_0 is the value assumed by the line connecting V_1 and V_2 **at time zero.** Note that this can be computed in several different ways. Below is one of those:

$$S_0 = V_2 + G_0[(N-1)/2]$$

4. Compute seasonal factors as follows.

• Initial seasonal factors are obtained by dividing each of the initial observations by the corresponding point along the line connecting V_1 and V_2. This can be done graphically or by using a formula.

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

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

• Normalize the seasonal factors.

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

Note that the time indices might be confusing depending on the number of periods and what they represent. For forecasting purposes, the time we start forecasting (denoted by t-now or t-zero) is indexed as 0. In the given data that might be period 8, 6, 4, 240 etc. While using the formulae, keep in mind that we will need to re-index the periods so that the most recent data point is associated with 0, and historical data is indexed -1, -2, ... depending on their distance to t-zero. Below is an example that illustrates that.

Example 2.8

Assume that the initial data set is the same as that of Example 2.6. Find V_1 and V_2 to find initial slope estimate (G_0). (Note that $t = 0$ is in fact after 8 periods.)

Time Index For Forecasting	Period	Demand
-7	1	10
-6	2	20
-5	3	26
-4	4	17
-3	5	12
-2	6	23
-1	7	30
0	8	22

Solution

$$V_1 = (10 + 20 + 26 + 17)/4 = 18.25$$

$$V_2 = (12 + 23 + 30 + 22)/4 = 21.75$$

$$G_0 = \frac{V_2 - V_1}{N} = \frac{21.75 - 18.25}{4} = 0.875 \to \textit{slope}!$$

$$S_0 = V_2 + G_0\left(\frac{N-1}{2}\right) = 21.75 + (0.875)(1.5) = 23.06$$

Remember how the initial factors are computed *by dividing each of the initial observations by the corresponding point along the line connecting V_1 and V_2*. In this example, consider the line connecting V_1 and V_2, and the point associated with period -7. This would be $18.25 - (2.5 - 1)(0.875)$ or $21.75 - (6.5 - 1)(0.875)$, because point -7 is 1.5 away from the horizontal of V_1 and 5.5 away from the horizontal of V_2. Note that as the number of points per cycle changes, these distances would change.

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 similarly. They are

$$c_{-5} = 1.391,\ c_{-4} = 0.869,\ c_{-3} = 0.5872$$

$$c_{-2} = 1.079,\ c_{-1} = 1.352,\ c_0 = 0.9539$$

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

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

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

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

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

$$f_{0,1} = (S_0 + G_0)c_{-3} = (23.06 + 0.875)(0.59) = 14.12$$

$$f_{0,2} = (S_0 + 2G_0)c_{-2} = [23.06 + 2(0.875)](1.11) = 27.54$$

$$f_{0,3} = 35.44$$

$$f_{0,4} = 24.38$$

Now, assume that we observe a demand of 16 in period 9 ($t = 1$). We now need to update our equations. Assume that $\alpha = 0.2$, $\beta = 0.1$ and $\gamma = 0.1$. Then

$$S_1 = \alpha(D_1/c_{-3}) + (1-\alpha)(S_0 + G_0) = (0.2)(16/0.59) + (0.8)(23.06 + 0.875) = 24.57$$

$$G_1 = \beta(S_1 - S_0) + (1-\beta)G_0 = (0.1)(24.57 - 23.06) + (0.9)(0.875) = 0.9385$$

$$c_1 = \gamma(D_1/S_1) + (1-\gamma)c_{-3} = (0.1)(16/24.57) + (0.9)(0.59) = 0.5961$$

At this point, we would renormalize c_{-2}, c_{-1}, c_0.

$$c_{-2} = \frac{(1.11)4}{4.0061} = 1.108$$

$$c_{-1} = \frac{(1.38)4}{4.0061} = 1.377$$

$$c_0 = \frac{(0.92)4}{4.0061} = 0.918$$

Forecasting from period 1, we obtain

$$f_{1,2} = (S_1 + G_1)c_{-2} = (24.57 + 0.9385)(1.108) = 28.2634$$

$$f_{1,3} = (S_1 + 2G_1)c_{-1} = [24.57 + 2(0.9385)](1.377) = 36.4175$$

and so on.

Each time a new observation becomes available, the intercept, slope, and most current seasonal factor estimates are updated. Note that their indices shift as you progress.

An important consideration is the choice of the smoothing constants α, β,, and γ to be used in Winter'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.

Example 2.9

As an industrial engineer working for a manufacturing company, your task is to forecast the demand. You have access to the historical data for the past 2 years.

4-month Period	Demand (units)
Jan-Apr 2022	320
May-Aug 2022	360
Sep-Dec 2022	400
Jan-Apr 2023	260
May-Aug 2023	240
Sep-Dec 2023	220

a. Using Winter's method, calculate the one-step and two-step demand forecasts for the upcoming two 4-month periods (i.e., Jan-Apr 2024 and May-Aug 2024). Show each step of your work.

b. Suppose that the demand for Jan-Apr 2024 turns out to be 180 units. Update the forecasting model using the following smoothing parameters: α (level) of 0.4, β (trend) of 0.3, and γ (seasonality) of 0.1. Using the updated model, forecast the demand for May-Aug 2024.

c. Suppose that the demand for May-Aug 2024 turns out to be 220 units. What is the MAPE value for your one-period-ahead forecasts? That means you must use your Jan-Apr 2024 forecast from part a and May-Aug 2024 forecast from part b.

Solution:

Time Index For Forecasting	4-month Period	Demand(units)	
-5	Jan–Apr 2022	320	
-4	May–Aug 2022	360	V_1
-3	Sep–Dec 2022	400	
-2	Jan–Apr 2023	260	
-1	May–Aug 2023	240	V_2
0	Sep–Dec 2023	220	

First, find V_1 and V_2 to find initial slope estimate(G_0)

$$V_1 = \frac{320 + 360 + 400}{3} = 360$$

$$V_2 = \frac{260 + 240 + 220}{3} = 240$$

$$G_0 = \frac{V_2 - V_1}{N} = \frac{240 - 360}{3} = -40 \rightarrow \textit{slope}!$$

Then, you should find S_0

$$S_0 = V_2 + G_0 \left(\frac{N-1}{2} \right) = 240 - 40 \left(\frac{3-1}{2} \right) = 200$$

and find initial seasonal factors

$$c_0 = \frac{220}{V_2 + G_0} = \frac{220}{200} = 1.1$$

$$c_{-1} = \frac{240}{V_2} = \frac{240}{240} = 1$$

$$c_{-2} = \frac{260}{V_2 - G_0} = \frac{260}{280} = 0.93$$

$$c_{-3} = \frac{400}{V_1 + G_0} = \frac{400}{320} = 1.25$$

$$c_{-4} = \frac{360}{V_1} = \frac{360}{360} = 1$$

$$c_{-5} = \frac{320}{V_1 - G_0} = \frac{320}{400} = 0.8$$

Then, you can take the average of seasonal factors and normalize

<u>Normalized Seasonal Factors</u>

$$\frac{c_0 + c_{-3}}{2} = \frac{1.1 + 1.25}{2} = 1.175 \rightarrow c_0 = \frac{(1.175)3}{1 + 1.175 + 0.865} = 1.16$$

$$\frac{c_{-1} + c_{-4}}{2} = \frac{1 + 1}{2} = 1 \rightarrow c_{-1} = \frac{(1)3}{1 + 1.175 + 0.865} = 0.987$$

$$\frac{c_{-2} + c_{-5}}{2} = \frac{0.93 + 0.8}{2} = 0.865 \rightarrow c_{-2} = \frac{(0.865)3}{1 + 1.175 + 0.865} = 0.854$$

Finally, you can forecast

a)

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

$$f_{0,1} = (S_0 + G_0)c_{-2} = 160(0.854) = 136.64$$

$$f_{0,2} = (S_0 + 2G_0)c_{-1} = 120(0.987) = 118.44$$

b)

The series:

$$S_1 = \alpha(D_1/c_{-2}) + (1-\alpha)(S_0 + G_0) = (0.4)(180/0.854) + (0.6)(160) = 180.31$$

The trend:

$$G_1 = \beta(S_1 - S_0) + (1-\beta)G_0 = (0.3)(180.31 - 200) + (0.7)(-40) = -33.907$$

The seasonal factors:

$$c_1 = \gamma(D_1/S_1) + (1-\gamma)c_{-2} = (0.1)(180/180.31) + (0.9)(0.854) = 0.868$$

Renormalize $c_{-1},\ c_0,\ c_1$

$$c_{-1} = \frac{(0.987)3}{0.987 + 1.16 + 0.868} = 0.982$$

$$c_0 = \frac{(1.16)3}{0.987 + 1.16 + 0.868} = 1.54$$

$$c_1 = \frac{(0.868)3}{0.987 + 1.16 + 0.868} = 0.864$$

$$f_{1,2} = (S_1 + G_1)c_{-1} = (180.31 - 33.907)(0.982) = 143.77$$

c)

	Demand	Forecast	ϵ
Jan-Apr 2024	180	136.64	-43.36
May-Aug 2024	220	143.77	-76.23

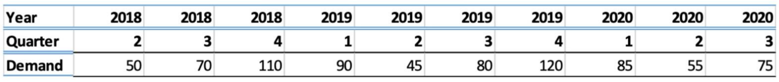
$$|e_1/D_1| = 0.24$$

$$|e_2/D_2| = 0.347$$

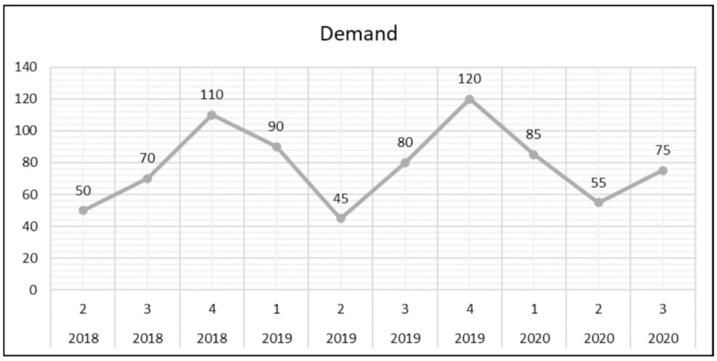
Average MAPE value for one-step ahead forecasts:

$$\text{MAPE} = \frac{1}{2}(0.24 + 0.347)100 = 29.4\%$$

Important 2.1



Year	2018	2018	2018	2019	2019	2019	2019	2020	2020	2020
Quarter	2	3	4	1	2	3	4	1	2	3
Demand	50	70	110	90	45	80	120	85	55	75



A smartphone distributor wants to forecast the demand for a particular brand of smartphones for the fourth quarter of 2020 and the first quarter of 2021 given the demand information for the last 10 quarters shown on the right:

The historical data shows a *seasonal pattern repeating every year* and therefore the planners of the distributor use Winter's Method with $\alpha = 0.2,\ \beta = 0.1,\ \gamma = 0.1$ to forecast the demand for the upcoming quarters.

a. What are your forecasts for the upcoming two quarters?

b. What is the MAD value of this forecasting method if the values in the two quarters are 115 (Q4-2020) and 92 (Q1-2021)?

Important 2.2

Consider a company that sells ice cream, and its sales show a *seasonal pattern repeating every year*. The company collects sales data, and you are tasked with forecasting sales for the first and second halves of the year using Winter's method. A year is split into two as hot and cold weather as follows: (i) Cold season: Fall & Winter, (ii) Hot season: Spring & Summer.

The data is as follows:

Season	**Sales**
Cold Season 2022	150,000
Hot Season 2022	200,000
Cold Season 2023	250,000
Hot Season 2023	300,000

a. Now that the hot season of 2023 is over, the company is trying to estimate sales for the upcoming cold season of 2024. Apply Winter's method to forecast sales for the upcoming cold season of 2024.

b. What would be your MAPE if the cold season 2024 sales is 600,000?

c. Update your Winter's method estimates to forecast sales for the hot season of 2024 if the cold season 2024 sales is 600,000.

For this question, use smoothing constants $\alpha = 0.2,\ \beta = 0.5,\ \gamma = 0.5$. The formulae you need are as follows:

The current level of the deseasonalized series

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

Trend

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

Seasonal factors (seasonality)

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

Forecast

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

Chapter 3

Inventory Control Subject to Known Demand

Effective inventory control is crucial when the demand for a product is known, as it allows businesses to optimize inventory levels, minimize holding costs, and ensure timely fulfillment of customer orders, thereby enhancing operational efficiency and maximizing profitability.

3.1 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.)

Raw Material: Raw materials are the resources required in the production or processing activity of the firm.

Components: Components correspond to items that have not yet reached completion in the production process. Components are sometimes referred to as subassemblies.

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.

Finished goods: Also known as end items, finished goods 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.

Spare parts: Spare parts refer to additional or backup components that are kept on hand to replace or repair any malfunctioning or worn-out parts within a manufacturing or production system. These spare parts are essential for minimizing downtime and ensuring the continuous and efficient operation of the production process.

The appropriate label to place on inventory depends upon the context. For example, components for some operations might be the end products for others.

Inventory locations in production systems analysis are categorized into five key areas:

(1) Manufacturing Location: This refers to the site where items are stored during the production process. It includes raw materials, work-in-progress, and finished goods awaiting further distribution.

(2) In-Transit: In-Transit refers to goods that are currently in transit from one location to another. This stage occurs during the transportation phase of the supply chain.

(3) Warehouse: The warehouse serves as a storage facility for inventory before it is distributed further. These locations are pivotal for managing and organizing goods efficiently.

(4) Retailer: Retailers hold inventory in their stores, ready for purchase by customers. This stage represents a point of sale within the supply chain.

(5) Customer: Once purchased, items are with the customer, representing the final stage in the supply chain. This signifies the completion of the product's journey from manufacturing to the end-user.

3.2 Pros and Cons of Centralized Inventory Management

ADVANTAGES. Some advantageous outcomes result from a well-structured and efficiently managed inventory system within a business. These are some motivations for holding inventories:

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. (This argument assumes that the setup cost is a fixed constant.) In short, it is more economical to produce large quantities and reduce setups.

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, the customer will likely 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 lead time. Lead time is defined as 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.

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.

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.

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 alleviate the disruptions caused by changing production rates and workforce levels.

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 manufacturing; it is virtually impossible to reduce all inventories to zero and expect any continuity in a manufacturing process.

DISADVANTAGES. On the other hand, there are some disadvantages to keeping inventory. These are the following:

Tied up capital. Maintaining inventory can tie up valuable capital that could be used for other strategic investments or business operations.

Storage space/cost (Warehousing cost). The need for storage space and associated warehousing costs can significantly impact overall operational expenses, affecting the bottom line.

Safety hazard. Accumulated inventory may pose safety hazards in the workplace, especially if it involves storing potentially dangerous or flammable materials.

Deterioration. Over time, stored goods may deteriorate, leading to product spoilage or reduced quality, particularly in industries with perishable items.

Obsolescence. Keeping excessive inventory can result in items becoming obsolete, especially in industries where technology or trends evolve rapidly.

Forecasting error. Inaccurate demand predictions may lead to overstock or stockouts, affecting operational efficiency and customer satisfaction.

Design/ manufacturing changes. Frequent changes in product design or manufacturing processes can render existing inventory obsolete, causing financial losses.

Quality problems. Stored goods may face quality issues over time, impacting the overall reputation of the company and customer satisfaction.

Effect on manufacturing lead times. Inventory levels can affect manufacturing lead times, either causing delays due to excess inventory or shortages causing production slowdowns.

Changes in raw material prices. Fluctuations in raw material prices can result in inventory losses or diminished profit margins.

Demand shortfall. An unexpected decrease in demand can lead to excess inventory, tying up resources and affecting profitability.

Changes in product design specifications. Modifications in product design specifications may make existing inventory incompatible, necessitating costly adjustments.

It's important to address these considerations in inventory management to optimize operations and mitigate potential drawbacks.

3.3 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.

- *Constant versus variable.* The simplest inventory models assume that the rate of demand is 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 and materials requirements planning.
- *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 a 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.

Demand	Supply lead time	Review	Inventory quality
✓Constant/Variable ✓Stochastic/Deterministic	✓Deterministic ✓Stochastic ✓Load-dependent	✓Continuous review ✓Periodic review	✓Perishability ✓Obsolescence ✓Imperfect yield

Excess demand	Number of items	Capacity
✓Backordering ✓Lost sales ✓Impatient customers ✓Item substitution ✓Rationing	✓Single item ✓Multiple items	✓Unlimited ✓Limited ✓Deterministic ✓Stochastic

The fundamental problem of inventory management tries to answer the following questions:

- When should an order be placed?
- How much should be ordered?

The complexity of the model depends on the assumptions of demand, physical characteristics of the system, and the form of the cost function.

3.4 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.

3.4.1 *Inventory Holding Costs*

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:

- Storage
- Taxes and insurance
- Breakage, spoilage, deterioration, obsolescence, etc.
- Opportunity cost of alternative investment, etc.

Holding cost can be calculated as:
Inventory Holding Cost = (Annual interest rate)($ value of inventory) + fixed holding cost

$$h = ic + \beta$$

Total inventory holding cost is calculated as the integral of holding cost over the carrying period.

$$\overline{I} = \int_{t_1}^{t_2} hI(t)dt$$

Example 3.1

28% cost of capital

2% taxes and insurance

6% storage cost

1% breakage cost

37% total interest charge (i) per year
$h = ic$ (holding cost in terms of dollars per unit per time)
If $c = \$100$, then $h = \$37$
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 below. 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 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

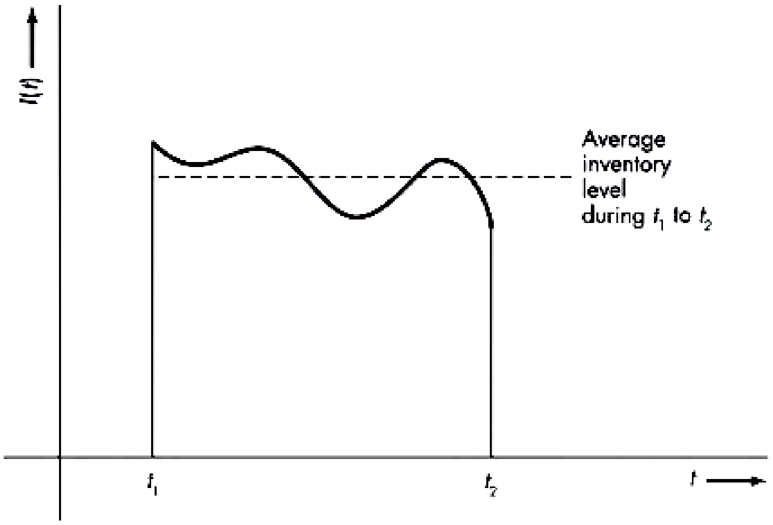


Fig. 3.1 Inventory as a function of time (Credit: [Nahmias, 1997])

area under the inventory level curve. When $I(t)$ is described by a straight line, its average value is obvious. In cases such as 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 $t_2 - t_1$.

3.4.2 *Ordering Cost*

The holding cost includes all those cost 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

$$C(x) = \begin{cases} 0, & x = 0 \\ K + cx, & x > 0 \end{cases}$$

K: Bookkeeping, transportation, etc.

c: Cost of purchasing

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.

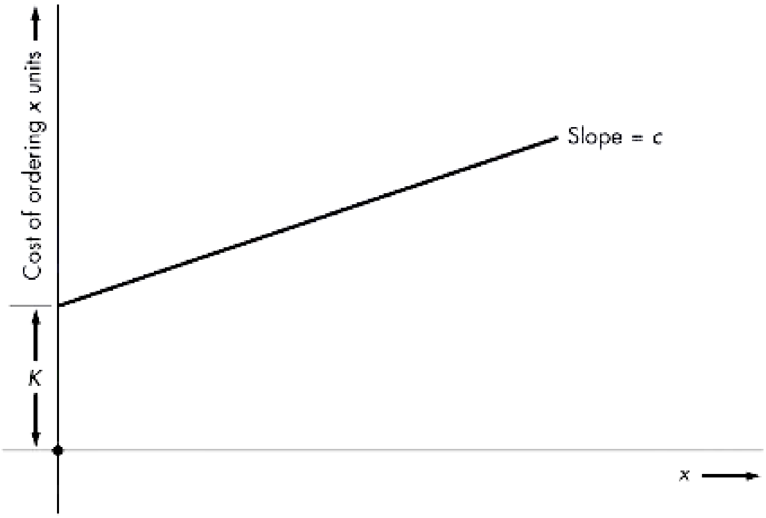


Fig. 3.2 Order cost function (Credit: [Nahmias, 1997])

3.4.3 *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.
In short, penalty cost can be in the form of:

- Backordering cost: bookkeeping and delay costs + loss of goodwill costs.
- Lost sales cost: opportunity + loss of goodwill costs.

p: \$/ unit/time to backorder or \$/unit (independent of time)

3.5 Basic Economic Order Quantity (EOQ) Model

The EOQ 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.
Assumptions:

(1) **Production is instantaneous** – there is no capacity constraint and the entire lot is produced simultaneously.
(2) **Delivery is immediate** – there is no time lag between production and availability to satisfy demand.
(3) **Demand is deterministic** – there is no uncertainty about the quantity or timing of demand.
(4) **Demand is constant over time** – in fact, it can be represented as a straight line, so that if annual demand is 365 units this translates into a daily demand of one unit.
(5) **A production run incurs a fixed setup cost** – regardless of the size of the lot or the status of the factory, the setup cost is constant.
(6) **Products can be analyzed singly** – either there is only a single product or conditions exist that ensure the separability of products.
(7) **Purchasing cost is constant** – unit purchasing/ordering cost is fixed regardless of the quantity purchased.
(8) **Backorders are not allowed** – all demand is satisfied immediately and no customers wait.

Notation
λ: demand rate (units/year)
c: unit production cost, not counting setup or inventory costs (\$/unit)
K: fixed or setup cost to place an order (\$)
h: holding cost (\$/unit/year); if the holding cost consists entirely of the interest on money

tied up in inventory, then

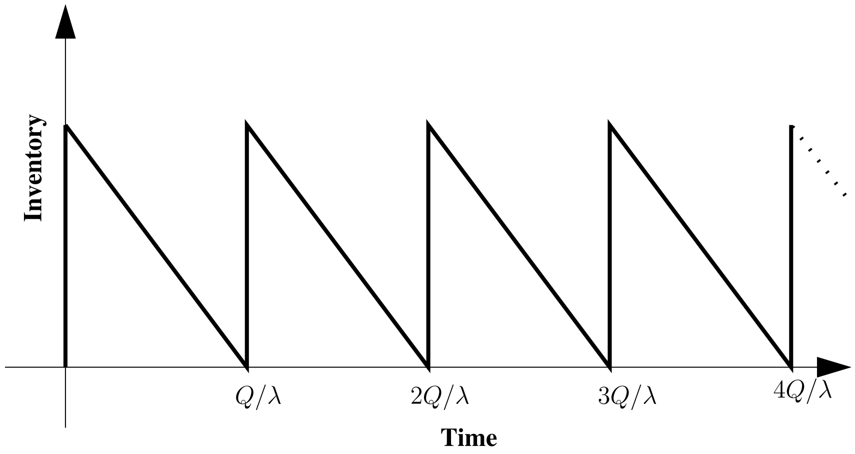
$$h = ic$$

where i is an annual interest rate.

Q: the order size (lot size)

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 the figure below.



3.5.1 *Costs Incurred*

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 $C(Q) = K + cQ$. To obtain the order cost per unit of 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, $-\lambda$, 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 of many cycles is also $Q/2$.

$$\text{Average Inventory} = \frac{Q}{2}$$

Setup Cost: K per lot, therefore

$$\text{Unit Setup Cost} = \frac{K}{Q}$$

Production Cost: c per unit

It follows that the average annual cost function, say $G(Q)$, is given by

$$G(Q) = \frac{hQ}{2} + \frac{K\lambda}{Q} + c\lambda$$

The three terms composing $G(Q)$ are annual holding cost, annual setup cost, and annual purchase cost, respectively.

We now wish to find Q to minimize $G(Q)$. Consider the shape of the curve $G(Q)$. We have that

$$\frac{dG(Q)}{dQ} = \frac{h}{2} - \frac{K\lambda}{Q^2} = 0$$

$$\frac{d^2G(Q)}{dQ^2} = \frac{2K\lambda}{Q^3} > 0 \rightarrow \text{Convex function}$$

Furthermore, since $G'(0) = -\infty$ and $G'(\infty) = h/2$, it follows that $G(Q)$ behaves as pictured in the figure below.

The optimal value of Q occurs where $G'(Q) = 0$. This is true when $Q^2 = 2K\lambda/h$, which gives

$$\boxed{Q^* = \sqrt{\frac{2K\lambda}{h}}} \tag{3.1}$$

Q^* is known as the economic order quantity (EOQ).

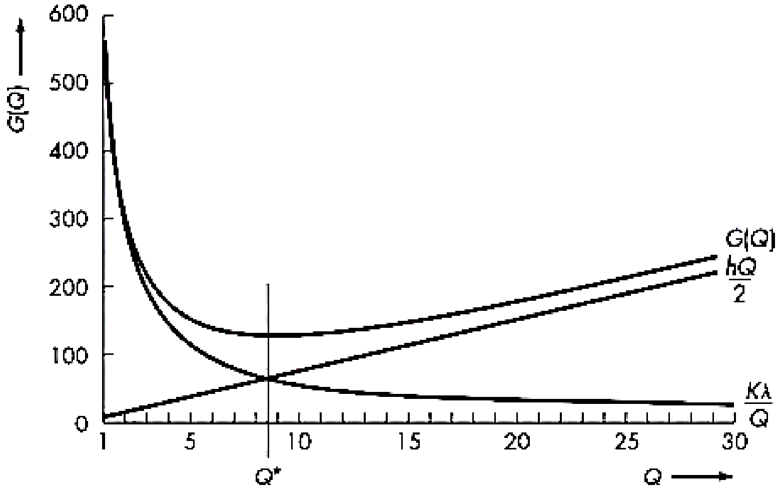


Fig. 3.3 The average annual cost function $G(Q)$ (Credit: [Nahmias, 1997])

To find the optimal cost for the EOQ model, we should replace Q with an economic order quantity Q^* in the $G(Q)$ function.

Optimal average cost per unit time (year):

$$G(Q^*) = \frac{hQ^*}{2} + \frac{K\lambda}{Q^*} + c\lambda$$

$$= \frac{h}{2}\sqrt{2K\lambda/h} + \frac{K\lambda}{\sqrt{2K\lambda/h}} + c\lambda$$

$$\boxed{G(Q^*) = \sqrt{2K\lambda h} + c\lambda} \tag{3.2}$$

We neglect unit cost, c, in the sensitivity analysis since it does not affect Q^*.

3.5.2 *Sensitivity of EOQ Model to Quantity*

In this part, we examine the issue of how sensitive the annual cost function is to errors in the calculation of Q.

Let G^* be the average annual holding and setup cost at the optimal solution. Then

$$G^* = \frac{hQ^*}{2} + \frac{K\lambda}{Q^*}$$

$$= \frac{K\lambda}{\sqrt{2K\lambda/h}} + \frac{h}{2}\sqrt{\frac{2K\lambda}{h}} = \sqrt{2K\lambda h}$$

It follows that for any Q,

$$\frac{G(Q)}{G^*} = \frac{hQ/2 + K\lambda/Q}{\sqrt{2K\lambda h}} = \frac{1}{2}\left[\frac{Q^*}{Q} + \frac{Q}{Q^*}\right]$$

In general, the cost function $G(Q)$ is relatively insensitive to errors in Q.

Example: If $Q = 2Q^*$, then the ratio of the actual to optimal cost is (1/2)[(1/2)+2]= 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.

Bottomline: Large deviations from the optimal order quantity lead to relatively small deviations from the optimal total cost. Considering the variable ordering cost that was left out, the effect is even less dramatic.

As we mentioned before, T represents the time between orders (in years).

$$T = \frac{Q}{\lambda}$$

If we write $T\lambda$ instead of Q in the total cost function:

$$G(Q) = \frac{hQ}{2} + K\frac{\lambda}{Q} + c\lambda = \frac{h\lambda T}{2} + \frac{K}{T} + c\lambda$$

Optimal Order Interval:

$$T^* = \frac{Q^*}{\lambda} = \frac{\sqrt{\frac{2K\lambda}{h}}}{\lambda} = \sqrt{\frac{2K}{h\lambda}}$$

Example 3.2

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 sold for 15 cents each. It costs the bookstore \$12 to initiate an order and holding costs are based on the annual interest rate of 25 percent. Determine the optimal number of pencils for the bookstore to purchase and the time between placement of orders.

Solution

$$\lambda = 60 \text{ items/week}$$

$$c = \$0.02/\text{item}$$

$$i = 25\% \text{ annually}$$

$$K = \$12/\text{order}$$

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:

$$\lambda\text{= (60items/week)(52 weeks/year)= 3120 items/year}$$

The holding cost h is the product of the annual interest rate and the variable cost of the item. Hence,

$$h\text{= (0.25/year)(\$0.02/item)= \$0.005/(item x year)}$$

Substituting into the EOQ formula, we obtain

$$Q^* = \sqrt{\frac{2K\lambda}{h}} = \sqrt{\frac{(2)(12)(3120)}{0.005}} = 3870 \text{ items}$$

Other figures of interest can be found using Q^*.

Time between orders:

$$T^* = \frac{Q^*}{\lambda} = 3870/3120 = 1.24 \text{ years}$$

The average annual holding cost is $h(Q/2) = 0.005(3870/2) = \$9.675$. The average annual setup cost is $K\lambda/Q$, which is also \$9.675.

Annual Inventory Investment (money tied-up):

$$I^* = \frac{cQ^*}{2} = \frac{c\lambda}{2F^*} = \frac{(0.02)(3120)}{2(0.806)} = \$38.7$$

By the way, frequency F is $1/T = \lambda/Q$ (numbers/year).

3.5.3 *EOQ with Non-negative 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.

Assumptions:

(1) Production is instantaneous

(2) ~~Delivery is immediate~~

(3) Demand is deterministic

(4) Demand is constant over time

(5) A production run/order incurs a fixed setup cost

(6) Products can be analyzed singly

(7) Purchasing cost is constant

(8) Backorders are not allowed

Suppose in Example 3.2 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 3.2 is shown in the figure below.

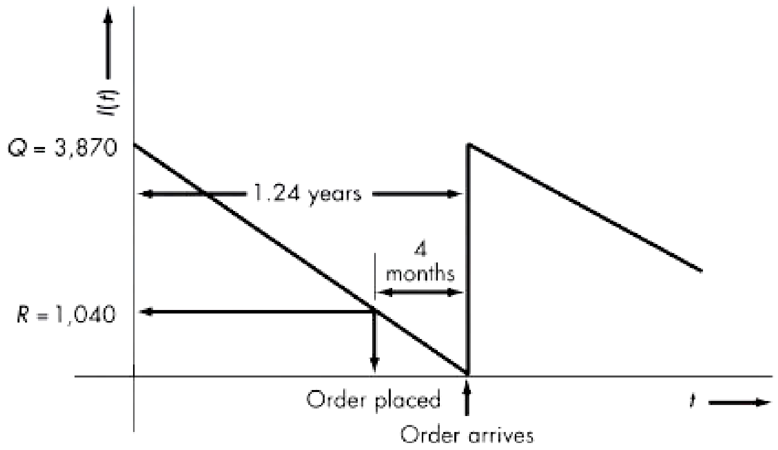


Fig. 3.4 Reorder point calculation for Example 3.2

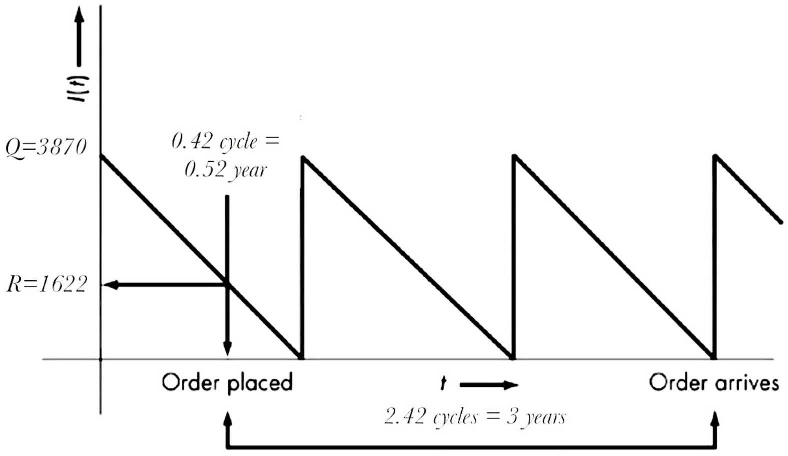


Fig. 3.5 Reorder point calculation for lead times exceeding one cycle

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. Lead time is denoted by L

$$R^* = \begin{cases} \lambda L, & L < T^* \\ \lambda(L - \lfloor \frac{L}{T^*} \rfloor T^*), & L \geq T^* \end{cases}$$

If $L = 4$ months, then $L < T^*$ then $R^* = (3120)(4/12) = 1040$ units. Notice that we converted the lead time to years before multiplying. Always express all relevant variables in the same units of time.
If $L = 3$ years, then $L > T^*$ then $R^* = (3120)(3 - (2)(1.24)) = 1622$ units.

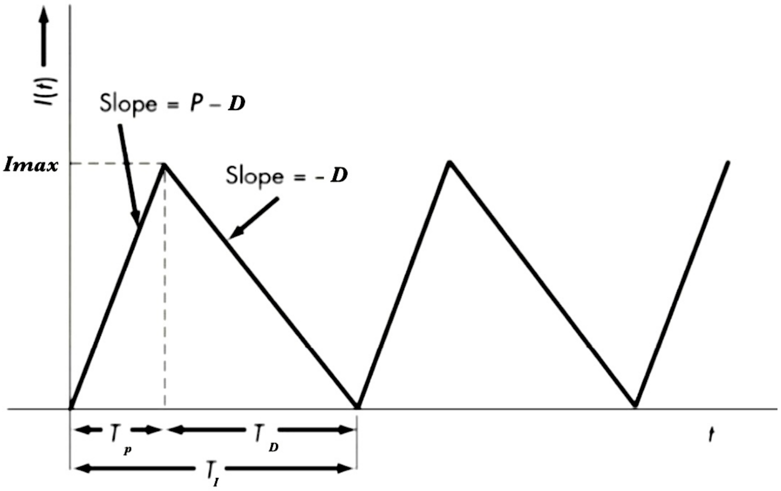
3.5.4 *Economic Production Quantity (EPQ) Model*

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.
Assumptions:

(1) ~~Production is instantaneous~~
(2) Delivery is immediate
(3) Demand is deterministic
(4) Demand is constant over time
(5) A production run/order incurs a fixed setup cost
(6) Products can be analyzed singly
(7) Purchasing cost is constant
(8) Backorders are not allowed

Notation:
P: production rate (number of items/time period)
T_P: production cycle (time facility is producing per order cycle)
T_D: withdrawal cycle (time facility is idle per order cycle)
T_I: total inventory cycle (time between setups)
I_{max}: maximum inventory level (units)
We require that $P > \lambda$ for feasibility. All other assumptions will be identical to those made in the derivation of the simple EOQ.
Let Q be the size of each production run. Let T_I, the cycle length, be the time between suc-

cessive production startups. Write $T_I = T_P + T_D$, where T_P is uptime and T_D is downtime. Note that the maximum level of on-hand inventory during a cycle is not Q.



As items are produced at a rate P for a time T_P, it follows that $Q = PT_P$, or $T_P = Q/P$. From the figure above, we see that $I_{max}/T_P = P - \lambda$. This follows from the definition of the slope as the rise over the run. Substituting $T_P = Q/P$ and solving for I_{max} gives $I_{max} = Q(1 - \lambda/P)$. Another equality we can see from the figure is $T_D = I_{max}/\lambda$.

We now determine an expression for the average annual cost function.

$$\text{Average Inventory} = \frac{I_{max}}{2}$$

$$\text{Total Holding Cost} = \frac{hI_{max}}{2} = \frac{hQ(1-\lambda/P)}{2}$$

$$\text{Total Ordering/Setup Cost} = \frac{K\lambda}{Q}$$

$$\text{Total Production Cost} = c\lambda$$

As a result,

$$\text{Total Cost} = G(Q) = \frac{K\lambda}{Q} + \frac{hQ(1-\lambda/P)}{2} + c\lambda$$

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

$$\frac{dG(Q)}{dQ} = \frac{h(1-\lambda/P)}{2} - \frac{K\lambda}{Q^2} = 0$$

$$Q^* = \sqrt{\frac{2K\lambda}{h'}}$$

$$\boxed{Q^*{=}\sqrt{\frac{2K\lambda}{h(1-\lambda/P)}}} \tag{3.3}$$

When we substitute Q^* in $G(Q)$ function we have

$$G(Q^*) = \sqrt{2K\lambda h'} + c\lambda$$

$$\boxed{G(Q^*){=}\sqrt{2K\lambda h(1-\lambda/P)} + c\lambda} \tag{3.4}$$

Example 3.3

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 $Q^* = 745$.

The time between production runs is $T_I = Q/\lambda = 745/2,500 = 0.298$ year. The uptime each cycle is $T_P = Q/P = 745/10,000 = 0.0745$ year, and the downtime each cycle is $T_D = T_I - T_P = 0.2235$ 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

$$I_{max} = Q^*(1 - \lambda/P) = 559 \text{ units}$$

Note that the EPQ is equivalent to an EOQ model with holding cost $h' = h(1 - \lambda/P)$. Consequently, the optimal cost under the EPQ model is lower than the optimal cost under the EOQ model with the same unit holding cost h.

Question: How would you explain this intuitively?
The formula $U = \lambda/P$, representing capacity utilization, underscores the importance of maintaining operational efficiency within a system. The variable U signifies the utilization ratio, λ represents the arrival rate of tasks or demand, and P denotes the processing capacity. The fundamental guideline is to ensure that the utilization ratio (U) remains equal to or less than 1. Operating above capacity, where $U > 1$, is cautioned against, as it signifies a potential strain on the system, leading to congestion, delays, and decreased overall efficiency. Adhering to $U \leq 1$ ensures that the system operates within its capacity limits, promoting smoother workflows and optimal performance.

EOQ vs. EPQ	
$U < 1$	$Q^*(\text{EPQ}) > Q^*(\text{EOQ})$
$U \to 0$	$Q^*(\text{EPQ}) \cong Q^*(\text{EOQ})$
$U \to 1$ (continuous production)	$Q^*(\text{EPQ}) \to$ infinity $G(Q^*(\text{EPQ})) \to c\lambda$

3.5.5 *EOQ Model with Quantity Discounts*

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.

Assumptions:

(1) Production is instantaneous
(2) Delivery is immediate
(3) Demand is deterministic
(4) Demand is constant over time
(5) A production run/order incurs a fixed setup cost
(6) Products can be analyzed singly
(7) ~~Purchasing cost is constant~~
(8) Backorders are not allowed

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 order (all units), or it is applied only to the additional units beyond the breakpoint (incremental). Note that additional (incremental) unit discounts are more common. Here is an example for all units discount:

$$C(Q) = \begin{cases} 0.3Q, & 0 \leq Q < 500 \\ 0.29Q, & 500 \leq Q < 1000 \\ 0.28Q, & 1000 \leq Q \end{cases}$$

That means, if the item is ordered in quantities less than 500, each unit costs 0.3. Between 500 and 100 it costs 0.29 *each*, and if quantity is 1000 or more, *each* unit costs 0.28.

On the other hand, in the case of incremental discounts, these discounted prices are applied to additional units *only*. Therefore, between 500 and 100 it costs 0.29 *each unit that exceeds 500*, whereas the first 500 units are still 0.30 each. Likewise, if the quantity is 1000 or more, the first 500 is 0.30 each, the second 500 is 0.29 each, and *each unit that exceeds 1000* costs 0.28. Mathematically, that is

$$C(Q) = \begin{cases} 0.3Q, & 0 \leq Q < 500 \\ 150 + 0.29(Q - 500) = 5 + 0.29Q, & 500 \leq Q < 1000 \\ 295 + 0.28(Q - 1000) = 15 + 0.28Q, & 1000 \leq Q \end{cases}$$

3.5.5.1 *All Units Discounts*

The solution technique for all unit discounts is as follows:

(1) Determine the largest *realizable* EOQ value[1]. 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.

Below is an example:

Example 3.4

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 1000 bags; it charges 29 percent per bag; and for orders of 1000 or more, it charges 28 cents per bag. In this case, the breakpoints occur at 500 and 1000.

[1] An EOQ value is realizable if it falls within the interval that corresponds to the unit cost used to compute it.

The discount schedule is for all units because the discount is applied to all of the units in an order. The order cost function $C(Q)$ is defined as

$$C(Q) = \begin{cases} 0.3Q, & 0 \leq Q < 500 \\ 0.29Q, & 500 \leq Q < 1000 \\ 0.28Q, & 1000 \leq Q \end{cases}$$

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. $c_0 = 0.3,\ c_1 = 0.29,\ c_2 = 0.28$ are the respective unit costs.
The function $C(Q)$ is pictured in the figure below.

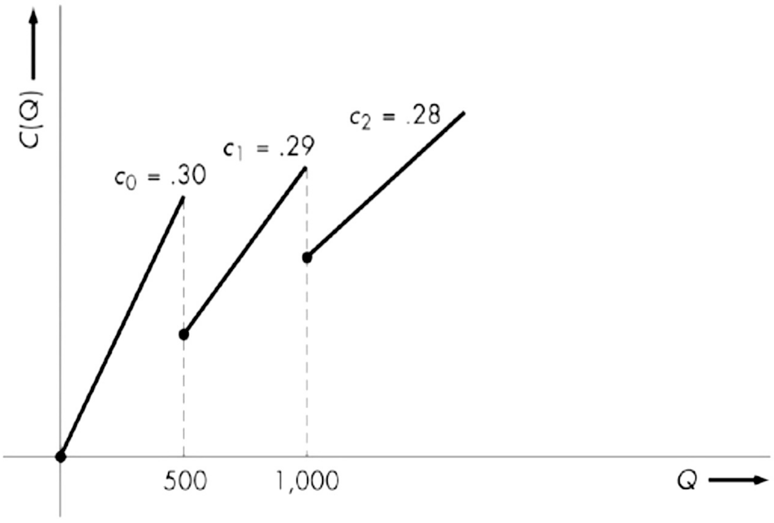


Fig. 3.6 All units discount order cost function (Credit: [Nahmias, 1997])

Solution

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.

$$Q^{(0)} = \sqrt{\frac{2K\lambda}{ic_0}} = \sqrt{\frac{(2)(8)(600)}{(0.2)(0.3)}} = 400 \quad \textbf{Realizable}$$

$$Q^{(1)} = \sqrt{\frac{2K\lambda}{ic_1}} = \sqrt{\frac{(2)(8)(600)}{(0.2)(0.29)}} = 406 \quad \textbf{Non-Realizable}$$

$$Q^{(2)} = \sqrt{\frac{2K\lambda}{ic_2}} = \sqrt{\frac{(2)(8)(600)}{(0.2)(0.28)}} = 414 \quad \textbf{Non-Realizable}$$

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 1000, and $Q^{(2)}$ would have to have been 1000 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 the figure below. 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.

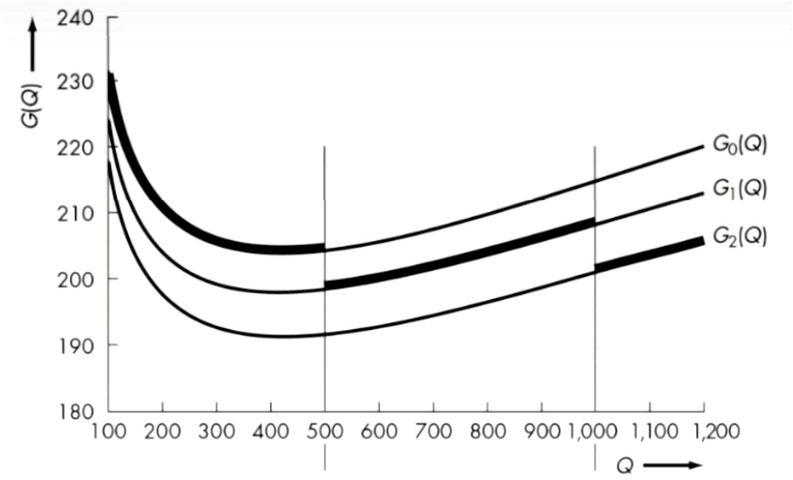


Fig. 3.7 All units discount annual cost function(Credit: [Nahmias, 1997])

There are three candidates for the optimal solution: 400, 500, and 1000. 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.

The average annual cost functions are given by

$$G_j(Q) = \lambda c_j + \lambda K/Q + ic_jQ/2 \quad \text{for } j = 0, 1, \text{and } 2.$$

The broken curve pictured in the figure above, $G(Q)$, is defined as,

$$G(Q) = \begin{cases} G_0(Q), & 0 \leq Q < 500 \\ G_1(Q), & 500 \leq Q < 1000 \\ G_2(Q), & 1000 \leq Q \end{cases}$$

Substituting Q equals 400, 500, and 1000, and using the appropriate values of c_j, we obtain

$$G(400) = G_0(400) = (600)(0.3) + (600)(8)/400 + (0.2)(0.3)(400)/2 = \$204.00$$

$$G(500) = G_1(500) = (600)(0.29) + (600)(8)/500 + (0.2)(0.29)(500)/2 = \$198.10 \leftarrow$$

$$G(1000) = G_2(1000) = (600)(0.28) + (600)(8)/1000 + (0.2)(0.28)(1000)/2 = \$200.80$$

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.

3.5.5.2 *Incremental Quantity Discounts*

The solution technique for incremental discounts is as follows:

(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.

Below is an example.

Example 3.5 Consider Example 3.4, 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 1000, the first 500 costs 30 cents each, and the remaining amount costs 29 cents each; for quantities of 1000 and over the first 500 costs 30 cents each, the next 500 costs 29 cents each, and the remaining amount costs 28 cents each. We need to determine a mathematical expression for the function $C(Q)$ pictured below.

From the figure, we can see that the first price break corresponds to $C(Q) = (500)(0.3) =$ \$150, and the second price break corresponds to $C(Q) = 150+(0.29)(500) =$ \$295. It follows that

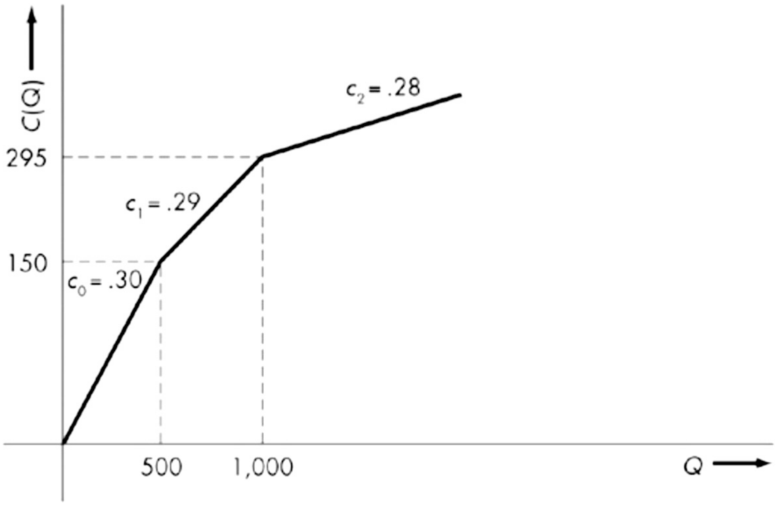


Fig. 3.8 Incremental discount order cost function (Credit: [Nahmias, 1997])

$$C(Q) = \begin{cases} 0.3Q, & 0 \leq Q < 500 \\ 150 + 0.29(Q - 500) = 5 + 0.29Q, & 500 \leq Q < 1000 \\ 295 + 0.28(Q - 1000) = 15 + 0.28Q, & 1000 \leq Q \end{cases}$$

so that

$$C(Q)/Q = \begin{cases} 0.3, & 0 \leq Q < 500 \\ 0.29 + 5/Q, & 500 \leq Q < 1000 \\ 0.28 + 15/Q, & 1000 \leq Q \end{cases}$$

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 $[G_0(Q),\ G_1(Q),\text{ and } G_2(Q)]$ depending upon into which interval Q falls. Because $C(Q)$ is continuous, $G(Q)$ also will be continuous. The function $G(Q)$ appears in the figure below. 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

$$G_0(Q) = (600)(0.3) + (600)(8)/Q + (0.2)(0.3)Q/2$$

which is minimized at

$$Q^{(0)} = \sqrt{\frac{2K_0\lambda}{ic_0}} = \sqrt{\frac{(2)(8)(600)}{(0.2)(0.3)}} = 400 \quad \textbf{Realizable}$$

The next interval to check is between 500 and 1000, where

$$G_1(Q) = (600)(0.29 + 5/Q) + (600)(8)/Q + (0.2)(0.29 + 5/Q)(Q/2)$$

$$= (600)(0.29) + (600)(13)/Q + (0.2)(0.29)(Q/2) + (0.2)(5)/2$$

This function is minimized at

$$Q^{(1)} = \sqrt{\frac{2K_1\lambda}{ic_1}} = \sqrt{\frac{(2)(13)(600)}{(0.2)(0.29)}} = 519 \quad \textbf{Realizable}$$

Notice that, while computing the EOQ, the fixed cost is not 8 as given in the question, but instead 13. This is because there is an additional fixed cost of 5 for this interval, as per the $C(Q)$ equation above.

The next interval is above 1000.

$$G_2(Q) = (600)(0.28 + 15/Q) + (600)(8)/Q + (0.2)(0.28 + 15/Q)(Q/2)$$

$$= (600)(0.28) + (600)(23)/Q + (0.2)(0.28)(Q/2) + (0.2)(15)/2$$

which is minimized at

$$Q^{(2)} = \sqrt{\frac{2K_2\lambda}{ic_2}} = \sqrt{\frac{(2)(23)(600)}{(0.2)(0.28)}} = 702 \quad \textbf{Non-Realizable}$$

It should be noted that the fixed cost is 8+15, hence 23 in the EOQ formula. This 15 comes from the $C(Q)$ equation above for the order quantities above 1000.

Both $Q^{(0)}$ and $Q^{(1)}$ are realizable. $Q^{(2)}$ is not realizable because $Q^{(2)} < 1000$. 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

$$G_0(Q^{(0)}) = \$204.00$$

$$G_1(Q^{(1)}) = \$204.58$$

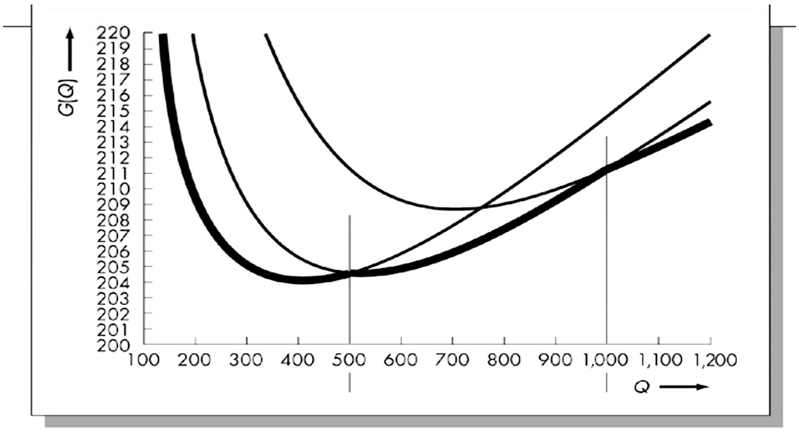


Fig. 3.9 Additional/incremental units discount annual cost function(Credit: [Nahmias, 1997])

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.

☛ Pay attention to the fact that in case of incremental discounts, the variable per unit cost is not a constant, but a function of quantity. This leads to an additional fixed cost component, affecting EOQ calculations. Moreover, while choosing among the realizable quantities, we have to compute the total cost function. During this cost computation, the fixed cost is the original fixed cost, but the variable cost is a function of quantity.

In summary; EOQ computation takes the discounted (constant) per unit cost and artificial fixed cost into account.

On the other hand, total cost function takes the discounted per unit cost (variable as a function of quantity) and original fixed cost into account.

3.5.6 *EOQ Model with Backorders/Lost Sales*

The EOQ model with backorders extends the traditional EOQ model by incorporating the impact of backorders and lost sales. In this modified approach, the model aims to find the optimal order quantity that minimizes the total cost, taking into account the costs associated with holding inventory, ordering, backordering, and potential revenue loss from unmet customer demand.

Assumptions:

(1) Production is instantaneous

(2) ~~Delivery is immediate~~

(3) Demand is deterministic

(4) Demand is constant over time

(5) A production run/order incurs a fixed setup cost

(6) Products can be analyzed singly

(7) Purchasing cost is constant

(8) ~~Backorders are not allowed~~

The assumptions for this system encompasses several key aspects. Firstly, it acknowledges that demand does not require immediate satisfaction from on-hand inventory, indicating flexibility in meeting customer needs over time. Customers are assumed to be willing to wait for the fulfillment of their orders. Additionally, a penalty cost denoted as b is incurred for each unit back-ordered per unit of time, emphasizing the cost implication of delayed fulfillment. Lastly, the system accounts for the fact that orders are received L units of time after they are placed, introducing a delay factor in the ordering process. These assumptions collectively provide a basis for modeling and analyzing a system where customer demand, order fulfillment, and associated costs are intricately interconnected over time.

Notation:

$I(t)$: inventory level at time t

I: average inventory level

$B(t)$: number of backorders at time t

B: average backorder level

$N(t)$: net inventory at time t, $N(t) = I(t) - B(t)$

$P_B(t)$: stock out indicator at time t

$(P_B(t) = 1 \text{ if } \ N(t) < 0)$

P_B: average fraction of time that a stockout occurs (sometimes referred to as stockout probability)

The goal is to minimize the costs of purchasing, ordering, holding, and backordering. By collectively minimizing these costs, the objective is to optimize the overall efficiency and economic performance of the inventory management system.

Let $s = r - \lambda L$ represent the safety stock, where r is the reorder level, λ is the demand rate, and L is the lead time.

The given conditions outline the inventory and backordering scenarios based on the relationship with the safety stock:

(1) If $s > 0$, indicating that the inventory position is above the safety stock, then $I(t) > 0$

(inventory is positive), and $B(t) = 0$ (no backorder).

(2) If $s < -Q$, where Q is the order quantity, implying that the inventory position is significantly below the safety stock, then $I(t) = 0$ (no inventory), and $B(t) > 0$ (backorder occurs).

(3) If $-Q \le s \le 0$, suggesting the inventory position is within an acceptable range around the safety stock, then both $I(t)$ and $B(t)$ can be positive. This is considered the most common and least costly scenario in practice.

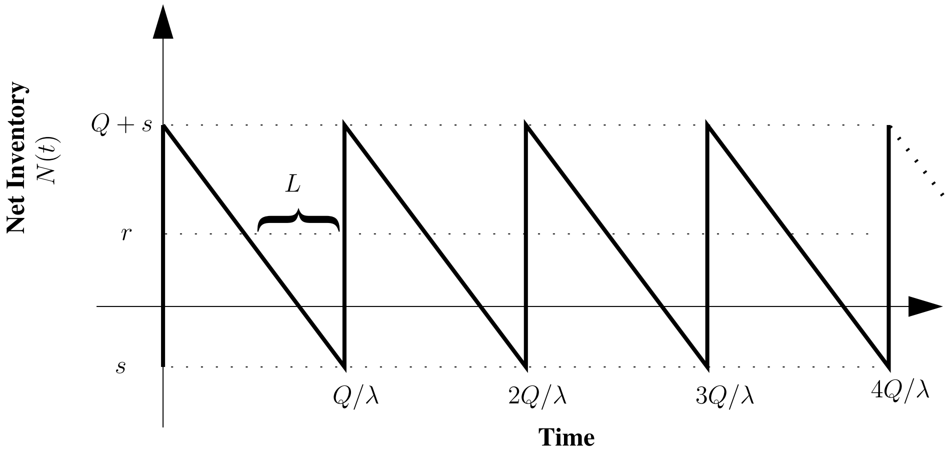


Fig. 3.10 Net inventory vs. time

Let us provide some derivations. T represents the time between orders and is calculated as Q/λ, where Q is the order quantity and λ is the demand rate.

T_1 signifies the time interval within T during which positive inventory is maintained and is calculated as $(Q+s)/\lambda$, where s is the safety stock.

T_2 represents the time interval within T during which backorders are positive and is calculated as $-s/\lambda$.

P_B is the fraction of time with backorders and is calculated as $T_2/T = -s/Q$.

I denotes the average inventory level over T and is calculated using $(1-P_B)(Q+s)/2 = (Q+s)^2/2Q$

B represents the average backorder level over T and is calculated as $P_B(-s/2) = s^2/2Q$.

These equations provide a quantitative understanding of the timing and quantities systematically associated with inventory and backorders.

The expression $G(Q,\ s)$ represents a total cost function.

$$G(Q,\ s) = c\lambda + K\lambda/Q + \frac{h(Q+s)^2}{2Q} + b\frac{s^2}{2Q}$$

To find the optimal order quantity, Q, and safety stock, s, we are equating the partial derivative of $G(Q,\ s)$ with respect to Q to zero.

$$\frac{\partial G(Q,\ s)}{\partial Q} = \frac{-K\lambda}{Q^2} + \frac{h(Q^2 - s^2)}{2Q^2} - \frac{bs^2}{2Q^2} = 0$$

The partial derivative of $G(Q,\ s)$ with respect to s is expressed as:

$$\frac{\partial G(Q,\ s)}{\partial s} = \frac{h(Q+s)}{Q} + \frac{bs}{Q}$$

Setting this equation to zero and solving for s results in

$$\frac{h(Q+s)}{Q} + \frac{bs}{Q} = 0 \rightarrow s = -\frac{hQ}{h+b}$$

Let $\alpha = b/(b+h)$, then the partial derivative of $G(Q,\ s)$ with respect to Q is given by

$$\frac{\partial G(Q,\ s)}{\partial Q} = \frac{-K\lambda}{Q^2} + \frac{h(Q^2 - s^2)}{2Q^2} - \frac{bs^2}{2Q^2}$$

$$= \frac{-K\lambda}{Q^2} + \frac{h(Q^2 - (1-\alpha)^2 Q^2)}{2Q^2} - \frac{b(1-\alpha)^2 Q^2}{2Q^2}$$

Setting this derivative to zero and solving for Q provides the optimal order quantity that minimizes the cost function

$$\frac{-K\lambda}{Q^2} + \frac{h(Q^2 - (1-\alpha)^2 Q^2)}{2Q^2} - \frac{b(1-\alpha)^2 Q^2}{2Q^2} = 0$$

These values only ensure that the point found is a critical point. Hessian must be checked for positive semi-definiteness to ensure the critical point found is a local minimizer, which is omitted for brevity.

<u>Conclusions:</u>

$$\alpha = b/(b+h)$$

$$\boxed{Q^* = \sqrt{\frac{2K\lambda}{h\alpha}}} \tag{3.5}$$

The optimal safety stock is calculated as

$$\boxed{s^* = -(1-\alpha)Q^*} \tag{3.6}$$

The optimal reorder level is expressed as

$$r^* = \lambda L + s^*$$

The total cost at the optimal values is given by

$$\boxed{G(Q^*,\ s^*) = \sqrt{2K\lambda h\alpha} + c\lambda} \tag{3.7}$$

Finally, the optimal fraction of time with stockouts is

$$\boxed{P_B(Q^*,\ s^*) = \frac{-s^*}{Q^*} = 1 - \alpha = \frac{h}{b+h}} \tag{3.8}$$

Important 3.1

A large automobile repair shop installs about 12,500 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 $12.60 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.
(Consider each part below independently, unless otherwise mentioned)

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?
d. If the mufflers are discounted to $10 each for orders larger than or equal to 250, how would you update your answer in part a?
e. If an unsatisfied customer costs $7 annually to a backlog, what percent of the time do you expect stock-outs at optimality? What would be the optimal order quantity, reorder point (when lead time is 10 weeks), and length of the backlog at most?

3.5.7 *Optional Reading: EOQ Model with Multiple Products*

In this section, we consider an extension of the EOQ model with a finite production rate, to the problem of producing N products on a single machine.

Assumptions:

(1) ~~Production is instantaneous~~
(2) Delivery is immediate
(3) Demand is deterministic
(4) Demand is constant over time
(5) A production run/order incurs a fixed setup cost
(6) ~~Products can be analyzed singly~~

(7) Purchasing cost is constant

(8) Backorders are not allowed

Notation:

N: number of products

λ_i: demand rate for product i

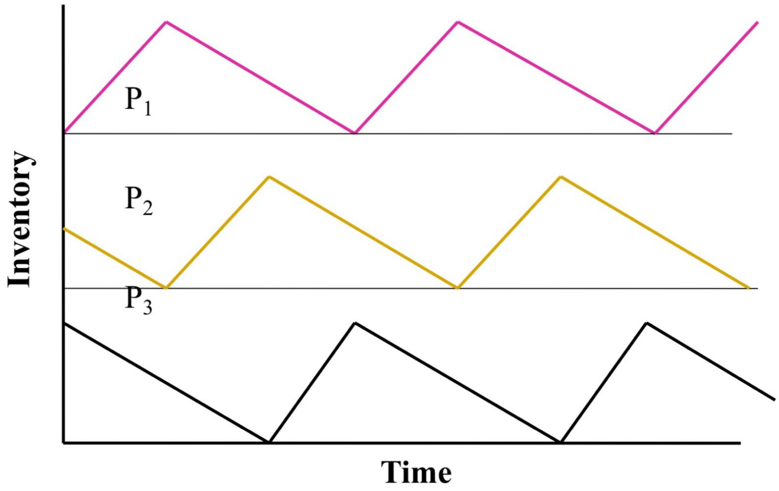
P_i: production rate for product i

h_i: holding cost per unit per unit time for product i

K_i: Ordering/setup cost for product i

c_i: production cost for product i

Our objectives are minimizing total cost while guaranteeing that no stock-outs occur for any product and producing only one product at a time.



To ensure feasibility, we require the assumption that

$$\sum_{i=1}^{N} \frac{\lambda_i}{P_i} \leq 1$$

This assumption is needed to ensure that the facility has sufficient capacity to satisfy the demand for all products.

The economic order quantity (EOQ) for each product i formula is given by

$$Q_i^* = \sqrt{\frac{2K_i\lambda_i}{h_i(1-\lambda i/P_i)}}$$

can lead to stockouts. Why would that lead to stockouts?

The fraction λ_i/P_i in the denominator represents the fraction of demand that is satisfied by the production rate. If this fraction is close to or exceeds 1, it indicates that the production rate is not sufficient to meet the demand, and stock-outs may occur.

In summary, the EOQ expression, along with the feasibility constraint, helps in determining the optimal order quantity for each product while ensuring that the overall production rates are sufficient to meet the total demand. Violating the feasibility constraint may result in insufficient production and, consequently, stock-outs.

We also will assume that a strictly cyclic policy is used. That means that in each cycle, there is exactly one setup per product, and products are produced in the same sequence in each production cycle.

Let cycle time, T, is the time between two consecutive setups for any given product and during T, a quantity Q_i of each product i is produced and consumed ($Q_i = \lambda_i T$)

So, each product undergoes a regular setup, and during the cycle time, a specific quantity is produced and consumed to meet the demand. This cyclic approach is often employed to balance the costs associated with setup and holding, optimizing the overall inventory system.

The average annual cost associated with product i can be written in the form:

$$G(Q_i) = \frac{h_i Q_i (1 - \lambda_i / P_i)}{2} + \frac{K_i \lambda_i}{Q_i} + c_i \lambda_i$$

The average annual cost for all products is the sum:

$$G(Q_1, \ \ldots, \ Q_N) = \sum_{i=1}^{N} \left\{ \frac{h_i Q_i (1 - \lambda_i / P_i)}{2} + \frac{K_i \lambda_i}{Q_i} + c_i \lambda_i \right\}$$

If we write T instead of Q_i/λ_i and h'_i instead of $h_i(1 - \lambda_i/P_i)$ in the total cost function we have,

$$G(T) = \sum_{i=1}^{N} \left\{ \frac{h'_i \lambda_i T}{2} + \frac{K_i}{T} + c_i \lambda_i \right\}$$

The goal is to find T to minimize $G(T)$. The necessary condition for an optimal T is

$$\frac{dG(T)}{dT} = 0$$

Solving for T, we obtain the optimal cycle time T^* as:

$$T^* = \sqrt{\frac{2 \sum_{i=1}^{N} K_i}{\sum_{i=1}^{N} h'_i \lambda_i}}$$

The optimal quantity for each product i is calculated as:

$$Q_i^* = \lambda_i T^*$$

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_i be the setup time for product i. Ensuring that the total time required for setups and production each cycle does not exceed T leads to the constraint

$$\sum_{i=1}^{N}(s_i + \frac{Q_i}{P_i}) \leq T$$

Using the fact that $Q_i = \lambda_i T$, this condition translates to

$$\sum_{i=1}^{N}(s_i + \frac{\lambda_i T}{P_i}) \leq T$$

which gives, after rearranging terms,

$$T \geq \frac{\sum_{i=1}^{N} s_i}{1 - \sum_{i=1}^{N}(\lambda_i / P_i)} = T_{min}$$

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}.

Chapter 4

Stochastic Inventory Models

In this chapter, we are deep-diving into a crucial aspect of businesses: real-world unpredictability. Stochastic inventory models, unlike simpler models, take into account the uncertainty that comes with factors such as how many people want something, how much they want, how long it takes to receive the items and unexpected market changes. Think of this section as a guide to handling the unpredictability of today's business world. We will break down the basics, discuss how to use them, and show how these models can help businesses stay flexible and ready for whatever comes their way.

The primary source of uncertainty in this context is attributed to demand variations. Despite the inherent unpredictability in demand, deterministic models remain crucial for understanding the trade-offs involved in inventory management. Depending on the degree of uncertainty, deterministic models can serve as effective approximations, especially when the relative variance is small and a significant portion of the variation is predictable. In cases where the system is too complex to incorporate randomness, deterministic models provide a valuable and practical approach to navigating inventory management challenges. These models help businesses strike a balance between the need for accurate predictions and the complexities posed by uncertain demand patterns.

In handling uncertainty within inventory management, different modeling philosophies are employed, each addressing uncertainty in its own way.

(1) *Deterministic Model with Adjusted Solution*

EOQ (Economic Order Quantity): Utilize the EOQ model to calculate the optimal order quantity, then incorporate safety stock to account for demand variability.

Deterministic Scheduling Algorithm: Employ deterministic scheduling algorithms and subsequently introduce safety lead time to accommodate uncertainties in delivery or processing times.

(2) *Stochastic Models*

Newsvendor Model: A stochastic model that considers the trade-off between ordering too much or too little when faced with uncertain demand and shortage costs.

Base Stock and $(Q,\ R)$ Models: Stochastic inventory models that use base stock levels or reorder points to manage inventory in the face of demand variability.

These approaches reflect a spectrum of strategies, from deterministic models with adjustments to explicitly stochastic models, allowing businesses to choose models that align with the nature and level of uncertainty present in their specific inventory management scenarios. In general, optimization of inventory models is about finding a control policy to minimize cost (maximize profit). 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 models minimize the expected cost.

Empirical distributions, while reflective of actual demand variability, may pose practical challenges in inventory management. Storing past demand data could be resource-intensive and may not always be feasible. Additionally, expressing the empirical distribution in a precise mathematical form might be complex, making it difficult to compute optimal inventory policies.

To address these challenges, a common strategy is to approximate the empirical distribution with a continuous distribution, often modeled using a Normal distribution. This simplification facilitates the application of well-established statistical methods and allows for more straightforward computations of optimal inventory policies. While this approximation might not capture all nuances of the empirical distribution, it provides a practical and manageable approach for handling uncertainty in demand and optimizing inventory management strategies.

4.1 Trade-offs

Demand is modeled as a random variable X with $E(X) = \lambda$ and standard deviation σ. The inherent variability in demand introduces the possibility of both overstocking and understocking situations, each associated with specific costs.

(1) *Overstocking*

Excess inventory can lead to overage costs, including capital costs for holding surplus stock and potential disposal costs. To mitigate the risk of overstocking, businesses may need to carefully manage inventory levels, considering the associated holding costs.

(2) *Understocking*

Understocking results in shortages, incurring costs such as backordering expenses. To address the risk of understocking, safety stock is introduced to provide a buffer, helping meet unexpected demand and minimizing the likelihood of shortages.

Balancing the trade-off between overage costs and shortage costs is a key challenge in inventory management. Strategies involve optimizing the level of safety stock to maintain a balance between the costs associated with excess and insufficient inventory, ensuring efficient and cost-effective operations in the face of demand uncertainty.

Stochastic models in inventory management can be classified into single-period models and multiple-period models. Single-period models are suitable for fashion goods, perishable items, and those with short lifecycles or seasonal demand. These models involve a one-time decision, determining the optimal order quantity for a specific period, and typically do not consider backordering.

On the other hand, multiple-period models are applicable when dealing with goods characterized by recurring demand, where demand varies from period to period. These models allow for decisions over multiple periods, addressing questions of how much to order in each period. Unlike single-period models, multiple-period models often consider the possibility of backordering, providing a more flexible approach to managing inventory levels. The choice between these models depends on factors such as the nature of the goods, demand patterns, and the desired level of flexibility in decision-making.

Safety stock is crucial in inventory management to address uncertainties in demand and supply. It serves to compensate for variations in customer demand and uncertainties in the replenishment lead time, providing a buffer against potential stockouts. Safety stock contributes to improving service levels by ensuring more consistent availability of products, reducing the risk of customer dissatisfaction. In terms of the reorder point, r, where L is the order replenishment lead time and λ is the average demand, the reorder point is calculated as $r = \lambda L$. This ensures that there is sufficient inventory to cover the average demand during the lead time. In an example where demand during the lead time follows a Normal distribution with a mean of $r = \lambda L$, $P(\text{demand during lead time} \leq r) = 0.5 \rightarrow 50\%$ probability of running out of products during the lead time, reflecting a balance between holding excess inventory and the cost of potential stockouts.

4.2 The Newsvendor Model

The newsvendor model is particularly applicable in situations that involve **one-time** production or purchasing **decisions**, where the disposal of unused or unsold inventory is a crucial consideration. This model is commonly used in industries such as magazines, newspapers, perishable products like food items, fashion goods, and seasonal products like snow shovels. The key assumptions underlying this model include a single-period focus, random demand with a known distribution, linear overage and shortage costs, and a minimum expected cost criterion.

The objective is to determine the optimal order quantity that minimizes the expected total cost, factoring in the trade-off between costs associated with holding excess inventory and costs associated with shortages. If the order quantity is too low, resulting in insufficient inventory to meet demand, a shortage cost is incurred. On the other hand, if the order quantity is too high, leading to excess inventory, a disposal or overage cost is incurred. Essentially, the decision on how much to order hinges on carefully assessing and managing the associated shortage and overage costs. The newsvendor model gives us a plan for making this decision.

Notation:

X : demand (in units), a random variable

$F(x) = P(X \leq x)$: cumulative distribution function for demand (assumed continuous)

$f(x) = \frac{dF(x)}{dx}$: probability density function for demand

c_o : cost (in dollars) per unit left over after demand is realized (i.e., *overage cost*)

c_u : cost (in dollars) per unit of shortage (i.e., *underage cost*) (c_u is sometimes called c_s)

Q : production/order quantity (in units); this is the decision variable

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 X 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 $Y(Q,\ X)$ 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 X is the demand. If Q units are purchased and X is the demand, $Q - X$ units are left at the end of the period as long as $Q \geq X$. If $Q < X$, then $Q - X$ is negative and the number of units remaining on hand at the end of the period is 0. Notice that

$$\max\{Q - X,\ 0\} = \begin{cases} Q - X, & Q \geq X \\ 0, & \text{otherwise} \end{cases}$$

In the same way, $\max\{X - Q,\ 0\}$ 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 X, either one or the other of these terms will be zero.

N_O : Number of units over

$$N_O = \begin{cases} Q - X, & Q \geq X \\ 0, & \text{otherwise} \end{cases} = \max(Q - X,\ 0)$$

N_S : Number of units short

$$N_S = \begin{cases} X - Q, & Q \leq X \\ 0, & \text{otherwise} \end{cases} = \max(X - Q,\ 0)$$

It follows that,

$$Y(Q,\ X) = c_o\ \max\{Q - X,\ 0\} + c_u\ \max\{X - Q,\ 0\}$$

Next, we derive the expected cost function. Define

$$Y(Q) = E[Y(Q,\ X)]$$

We obtain

$$Y(Q) = c_o\ E[N_O] + c_u\ E[N_S]$$

$$= c_o \int_0^\infty \max\{Q - x,\ 0\}\ f(x)dx + c_u \int_0^\infty \max\{x - Q,\ 0\}\ f(x)dx$$

$$= c_o \int_0^Q (Q - x) f(x)dx + c_u \int_Q^\infty (x - Q) f(x)dx$$

Why do the bounds start from 0 instead of $-\infty$? Because we deal with a demand function, which *ideally* should not have a probability density function defined sub-zero.

Note: For any given season, we will be either over or short, not both. But in expectation, overage and shortage can both be positive.

Leibnitz's Rule

In the analysis of the newsvendor model, Leibnitz's Rule is used to determine the derivative of $Y(Q)$. According to Leibnitz's Rule:

$$\frac{d}{dQ} \int_{a_1(Q)}^{a_2(Q)} g(x,\ Q)dx = \int_{a_1(Q)}^{a_2(Q)} \frac{\partial[g(x,\ Q)]}{\partial Q} dx + g(a_2(Q),\ Q)\frac{da_2(Q)}{dQ} - g(a_1(Q),\ Q)\frac{da_1(Q)}{dQ}$$

Application of Leibnitz's Rule

For instance, the partial derivative of the first term above with respect to Q is as follows:

$$\frac{d}{dQ}\int_0^Q (Q-x)f(x)dx + c_u = \int_0^Q \frac{(Q-x)f(x)}{\partial Q}dx + (Q-Q)f(x)\frac{dQ}{dQ} - (Q-0)f(x)\frac{d0}{dQ}$$

$$= \int_0^Q \frac{(Q-x)f(x)}{\partial Q}dx = \int_0^Q f(x)dx$$

Determining the Optimal Policy

We would like to determine the value of Q that minimizes the expected cost $Y(Q)$. To do so, it is necessary to obtain an accurate description of the function $Y(Q)$. We have that

$$\frac{\partial Y(Q)}{\partial Q} = c_o\int_0^Q f(x)dx + c_u\int_Q^\infty -f(x)dx$$

$$= c_oF(Q) - c_u(1-F(Q))$$

This is a result of Leibniz's rule, which indicates how one differentiates integrals.
It follows that

$$\frac{d^2Y(Q)}{dQ^2} = (c_o+c_u)f(Q) \geq 0 \quad \text{for all} \quad Q \geq 0$$

Because the second derivative is nonnegative, the function $Y(Q)$ is said to be *convex.* The function $Y(Q)$ is pictured in the figure below.

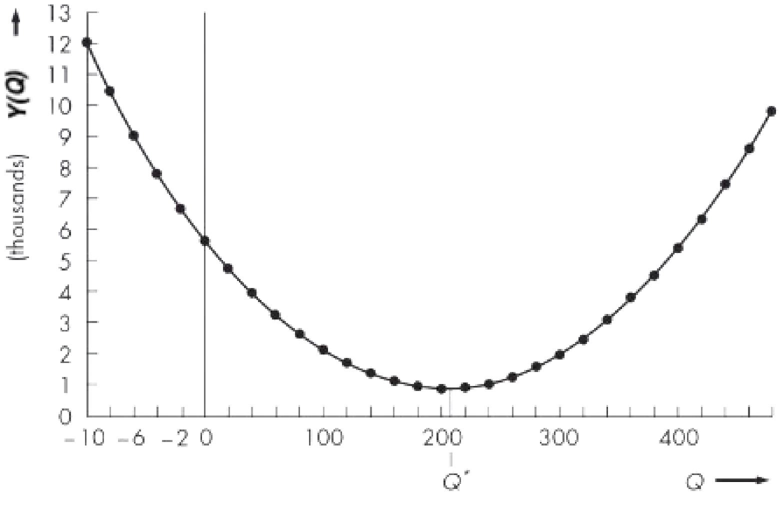


Fig. 4.1 Expected cost function for newsvendor problem (Credit: [Nahmias, 1997])

It follows that the optimal solution, say Q^*, occurs where the first derivative of $Y(Q)$ equals zero. That is,

$$Y'(Q^*) = (c_o+c_u)F(Q^*) - c_u = 0$$

Rearranging terms gives

$$\boxed{F(Q^*)=\frac{c_u}{c_o+c_u}} \tag{4.1}$$

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.

The optimal solution satisfies

$$F(Q^*) = P(X \leq Q^*) = \frac{c_u}{c_o+c_u}$$

where $F(Q^*)$ is defined as the probability that the demand does not exceed Q^*. This means that the optimal solution satisfies the equality of the critical ratio and the probability of satisfying all the demand during the period if Q^* units are purchased at the start of the period.

Note that Q^* decreases as c_o increases. In contrast, as c_u increases Q^* also increases.

If c_o and c_u are difficult to estimate, we determine the desired *service level* α (i.e., with probability α, all demand is met) and choose Q^* accordingly. Specifically, we select Q^* in a way that ensures the cumulative distribution function of the demand up to Q^* equals the desired service level α. This is expressed as $F(Q^*) = P(X \leq Q^*) = \alpha$.

Short Review of The Exponential Distribution

For the exponential distribution with parameters λ, the cumulative distribution function $F(x)$ is given by

$$F(x) = 1 - e^{-\lambda x}$$

The probability density function $f(x)$ is defined as follows

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The expected value is calculated as

$$E(X) = \frac{1}{\lambda}$$

Finally, the variance is computed as

$$Var(X) = \frac{1}{\lambda^2}$$

If we write Q in the cumulative distribution function we have

$$F(Q) = 1 - e^{-\lambda Q}$$

and if we write Q^* in the $F(Q)$ function we obtain

$$F(Q^*) = 1 - e^{-\lambda Q^*} = \frac{c_u}{c_u + c_o} = \alpha$$

The resulting equation can be solved for the optimal order quantity Q^* in terms of the desired service level α

$$Q^* = \frac{-\ln(1-\alpha)}{\lambda}$$

Example 4.1

In a unique scenario where the demand for T-shirts follows an exponential distribution with a mean of 1000, represented by the cumulative distribution function $F(x) = P(X \leq x) = 1 - e^{-x/1000}$, the cost of each shirt is \$10 and the selling price is \$15. Additionally, any unsold shirts can be sold off at \$8. Given this information, calculate the optimal order quantity.

Note that exponential distribution is not a widely used demand distribution, whereas Poisson or Normal are more common. Exponential distribution is commonly used for interarrival times.

Solution

c_u= opportunity lost = price − cost + loss of goodwill = \$15 − \$10= \$5

c_o= cost of redundant purchase = cost + disposal/handling − salvage= \$10 − \$8= \$2

We can compute $F(Q^*)$ by using the formula given in the question

$$F(Q^*) = 1 - e^{-Q^*/1000} = \frac{c_u}{c_u + c_o} = \frac{5}{5+2} = 0.714 \text{ (which is also } \alpha)$$

Solving for Q^*

$$Q^* = \frac{-\ln(1-\alpha)}{\lambda} = \frac{-\ln(1-0.714)}{1000} = 1,253 \text{ units}$$

Note that if $c_o = \$10$ (i.e., shirts must be discarded) then

$$F(Q^*) = 1 - e^{-Q^*/1000} = \frac{c_u}{c_u + c_o} = \frac{5}{5+10} = 0.333$$

The optimal order quantity is recalculated as

$$Q^* = 405 \text{ units}$$

The change in c_o has a notable impact on the optimal order quantity. In practical terms, this means a more conservative approach to ordering, aligning with the increased expense associated with discarding unsold shirts.

Short Review of The Normal Distribution

The probability density function of the Normal distribution with parameters μ and σ, denoted by $N(\mu,\ \sigma)$ is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$$

The expected value (mean) of the distribution is

$$E(x) = \mu$$

and the variance is

$$Var(x) = \sigma^2$$

Additionally, if a random variable X has the Normal distribution $N(\mu,\ \sigma)$, then the standardized variable $(X-\mu)/\sigma$ has the Standard normal distribution $N(0,\ 1)$.
The cumulative distributive function of the Standard normal distribution is denoted by ϕ. We defined the α as

$$F(Q^*) = \alpha$$

So,

$$P(X \le Q^*) = \alpha$$

If we subtract the mean from all sides and divide each side by the standard deviation we have

$$P\left(\frac{X-\mu}{\sigma} \le \frac{Q^*-\mu}{\sigma}\right) = \alpha$$

Let $Y = (X-\mu)/\sigma$, then Y has the standard Normal distribution

$$P\left(Y \le \frac{Q^*-\mu}{\sigma}\right) = \phi\left(\frac{Q^*-\mu}{\sigma}\right) = \alpha$$

If we take the inverse of the function

$$\frac{Q^*-\mu}{\sigma} = \phi^{-1}(\alpha)$$

After rearranging terms we obtain

$$Q^* = \mu + \phi^{-1}(\alpha)\sigma$$

Let $z_\alpha = \phi^{-1}(\alpha)$, then

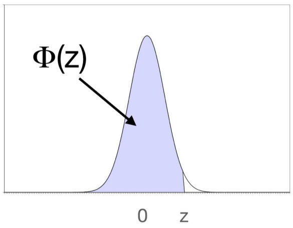
$$Q^* = \mu + z_\alpha\sigma$$

Suppose demand is normally distributed with mean μ and standard deviation σ. Then the critical ratio formula reduces to

$$Q^* = \mu + z_\alpha \sigma$$

where

$$\alpha = \frac{c_u}{c_u + c_o}$$



Note: Q* increases in μ,
it also increases in σ if z is positive
(i.e., if ratio is greater than 0.5).

Example 4.2

On consecutive Sundays, Mac, the owner of a local newsstand, purchases several copies of *The Computer Journal*, a popular weekly magazine. From past experience, we saw that weekly demand for the *Journal* is approximately normally distributed with mean $\mu = 10,000$ and standard deviation $\sigma = 1,000$. He pays 75 cents for each copy and sells each for 175 cents. Copies he has not sold during the week can be returned to his supplier for 25 cents each. The supplier can salvage the paper for printing future issues. How many copies should he purchase every week?

Solution

Because Mac purchases the magazines for 75 cents and can salvage unsold copies for 25 cents, his overage cost is

$$c_o = 75 - 25 = 50 \text{ cents}$$

His underage cost is the profit on each sale, so that

$$c_u = 175 - 75 = 100 \text{ cents}$$

The critical ratio is

$$\frac{c_u}{c_u + c_o} = \frac{1}{1.5} = 0.67 \ \rightarrow \alpha = 0.67$$

Hence, he should purchase enough copies to satisfy all the weekly demand with a probability of 0.67. The optimal Q^* is the 67th percentile of the demand distribution.
From a standard normal table, we find that

$$z_{0.67} = 0.44$$

The optimal Q is

$$Q^* = \mu + z_\alpha \sigma = 10,000 + 0.44(1,000) = 10,440 \text{ units}$$

Hence, he should purchase 10,440 copies every week.

4.3 Service Levels

Service levels are crucial metrics in inventory management, providing insights into different aspects of performance. A common substitute for a stock-out cost is a service level. Although there are many different definitions of service, it generally refers to the probability that a demand or a collection of demands is met. Service levels for continuous-review systems are considered here. Two types of service are considered, labeled Type 1 and Type 2, respectively.

α service level (Type 1)

The α service level, often referred to as a Type 1 service level, focuses on event-oriented criteria. It quantifies the probability that all customer orders received within a specific time frame will be fulfilled entirely from available stock, without any delay.

β service level (Type 2)

In contrast, the β service level, categorized as a Type 2 service level, concentrates on quantity-oriented performance measurements. It represents the proportion of total demand that can be met directly from existing inventory, without any delays in fulfillment.

The probability of no-stockout (Type 1 Service), denoted by $F(Q^*)$, is calculated as

$$F(Q^*) = P(X \leq Q^*) = \frac{c_u}{c_o + c_u}$$

The fill rate (Type 2 Service) is calculated as

$$\frac{E[\min(Q, X)]}{E[X]} = \frac{Q - E[\max(Q - X, 0)]}{E[X]}$$

$$= \frac{E[X] - E[\max(X - Q, 0)]}{E[X]} = 1 - \frac{E[N_S]}{E[X]}$$

The fill rate is always greater than or equal to the no-stockout probability.

Question: How can you prove this?

4.4 Results for the Discrete Case

X represents a discrete random variable (demand), and $Y(Q)$ is a function representing the total cost, which includes the holding or overage cost (c_o) and the underage or shortage cost (c_u).

$$Y(Q) = c_o\ E[N_O] + c_u\ E[N_S]$$

$$= c_o \sum_{x=0}^{\infty} \max\{Q - x,\ 0\} P(X = x) + c_u \sum_{x=0}^{\infty} \max\{x - Q,\ 0\} P(X = x)$$

$$= c_o \sum_{x=0}^{Q} (Q - x) P(X = x) + c_u \sum_{x=Q}^{\infty} (x - Q) P(X = x)$$

To find the optimal value of Q, which minimizes the total cost $Y(Q)$, we need to identify the smallest integer Q that satisfies the condition:

$$Y(Q + 1) - Y(Q) \geq 0$$

This condition ensures that increasing the order quantity by one does not lead to a decrease in total cost. In other words, it ensures that the current order quantity Q is optimal. Mathematically, this condition can be expressed as

$$\sum_{x=1}^{Q} P(X = x) \geq \frac{c_u}{c_u + c_o}$$

This inequality indicates that the cumulative probability of demand up to Q should be greater than or equal to the ratio of underage or shortage cost to the sum of costs.

or equivalently

$$P(X \leq Q) \geq \frac{c_u}{c_u + c_o}$$

Short Review of The Geometric Distribution

Geometric distribution has two interpretations. We often come across the shifted geometric distribution, where the probability distribution of the number of trials needed to get the first success. Here we use the less common interpretation: the probability distribution of the number of successes before the first failure, where success probability is ρ. A common example is the death of a circuit, where success is considered the survival of a circuit in a period. The random variable denotes the lifetime of that circuit (the number of success periods before the first period it fails).

The geometric distribution with parameter ρ, where $0 \leq \rho \leq 1$ is characterized by the probability mass function and related statistics:

Probability mass function

$$P(X = x) = \rho^x(1 - \rho)$$

The expected value (mean) is calculated as

$$E(X) = \frac{\rho}{1 - \rho}$$

The probability of at least x occurrences is given by

$$P(X \geq x) = \rho^x$$

and the probability of at most x occurrences is

$$P(X \leq x) = 1 - \rho^{x+1}$$

These formulas describe the behavior of the geometric distribution, which models the number of trials needed to achieve the first success in a sequence of independent Bernoulli trials, where each trial has a success probability ρ.

Remember that the optimal order quantity Q^* is the smallest integer that satisfies

$$P(X \leq Q^*) \geq \frac{c_u}{c_u + c_o}$$

If we write Q^* instead of x in the $P(X \leq x)$ function we have

$$P(X \leq Q^*) = 1 - \rho^{Q^*+1} \geq \frac{c_u}{c_u + c_o}$$

Rearranging terms gives

$$Q^* \geq \frac{\ln(\frac{c_o}{c_u+c_o})}{\ln(\rho)} - 1$$

So,

$$Q^* = \left\lfloor \frac{\ln(\frac{c_o}{c_u+c_o})}{\ln(\rho)} \right\rfloor$$

4.5 Multi-period Newsvendor Problem

The extension of the newsvendor model to a multi-period problem involves dealing with periodic demands, typically on a monthly basis, which are independent and identically distributed (i.i.d.) according to a given distribution $F(x)$. In this setup, orders placed are either backordered and fulfilled in subsequent periods or lost entirely if demand exceeds the available inventory, with no setup costs associated with the order. The objective remains to minimize expected costs or maximize expected profits over the entire planning horizon, considering factors such as demand uncertainty, inventory holding costs, and costs associated with backorders or lost sales.

In this scenario, the interpretation of both c_o and c_u will be different. Here, c_o represents the overage cost, that is incurred to hold one unit of inventory in stock *for a single period.* On the other hand, c_u represents the underage cost, that is either the cost of backordering one unit *for one period* or the cost of a stockout/lost sale.

In handling starting inventory and backorders within the newsvendor model, several key terms and concepts are employed. First, S_0 represents the starting inventory position, while S denotes the order-up-to level, with $S - S_0$ representing the order quantity.

The function $Y(S)$ calculates the total expected cost, incorporating costs of holding excess inventory (c_o) and costs associated with stockouts or backorders (c_u).

$$Y(S) = c_o E[(S - X)^+] + c_u E[(X - S)^+]$$

Note that, in general, $E[(Y)^+]$ denotes the expectation of the positive values of a random variable Y. That is $E[(Y)^+] = E[\max(0, Y)]$.

The derivation of the optimal order-up-to level, denoted by S^*, is omitted here for brevity. It is determined by differentiation, and at optimality, the probability that demand in a period (X) does not exceed the order-up-to level has to be equal to the ratio of the cost of stockouts to the sum of the stockout and holding costs.

$$P(X \leq S^*) = \frac{c_u}{c_u + c_o}$$

The optimal policy dictates ordering nothing if the starting inventory position, S_0, is greater than or equal to S^*; otherwise, the order quantity is set to $S^* - S_0$, ensuring efficient inventory management and cost minimization.

Insights derived from the newsvendor model include the following:

Order size tends to increase as shortage or underage costs rise. This is because higher costs associated with stockouts incentivize larger orders to minimize the likelihood of running out of inventory.

Conversely, order size tends to decrease as overage costs increase. Elevated overage costs discourage excessive inventory holding, leading to smaller orders to avoid unnecessary excess inventory.

Order size typically rises with demand variability. Increased demand variability necessitates larger stock levels to buffer against uncertainty, resulting in larger order quantities to meet potential fluctuations in demand effectively.

Inventory is a hedge against demand uncertainty. By maintaining inventory, businesses can better manage fluctuations in customer demand and ensure product availability. The amount of protection depends on "overage" and "underage" costs, as well as the distribution of demand.

If shortage cost exceeds the overage cost, it suggests that the business faces higher penalties for stockouts than for maintaining excess inventory. In this scenario, the optimal order quantity increases in both the mean and standard deviation of demand.

Important 4.1

A retailer buys sunbeds at the beginning of each summer for sales during summer. The demand for sunbeds is distributed uniformly between 100 and 200. The purchasing cost of each sunbed is \$30. Any sunbed can be sold for \$50 each. By the end of the selling season, leftover sunbeds can be returned to their suppliers for a discounted price. This discounted selling price is \$20. There is also a \$2 handling cost for each leftover sunbed.

a. Compute the optimal number of sunbeds that a retailer should buy at the beginning of summer.
b. What is the probability of satisfying the entire demand with the quantity in part a?
c. What is the number of expected shortages and expected leftovers when the optimal quantity is ordered?

4.6 The Base-Stock Model

In the realm of inventory management, the base stock model stands as a cornerstone strategy for businesses aiming to maintain optimal stock levels while meeting customer demands efficiently. This section delves into the fundamental principles and practical applications of the base stock model, a concept that revolves around replenishing inventory to a predetermined level whenever stock falls below a certain threshold. Unlike other inventory models, the base stock model allows for flexibility in responding to fluctuations in demand and lead times, ensuring that businesses can fulfill orders promptly while minimizing excess inventory costs.

The base-stock model in inventory management operates under several key assumptions to facilitate analysis and decision-making. Firstly, it assumes that demand for the product occurs continuously over time, rather than in discrete intervals. Additionally, the times between consecutive orders, known as inter-arrival times, are stochastic, meaning they follow a probabilistic distribution and are independent and identically distributed (***i.i.d.***). Inventory levels are continuously monitored, and orders are placed whenever the inventory level reaches a predetermined threshold, known as the base stock level. The model also assumes a fixed supply lead time, representing the time it takes for an order to be fulfilled from placement until inventory replenishment, simplifying the timing of order placement. Furthermore, there are no fixed costs associated with placing an order, streamlining the analysis to focus solely on inventory holding and backordering costs. Lastly, the model allows for backorders, ensuring unfilled customer demand is satisfied once inventory is replenished, contributing to customer satisfaction and mitigating the risk of lost sales. These assumptions collectively form the foundation of the base-stock model, providing a structured framework for optimizing inventory management decisions.

4.6.1 *The Base-Stock Policy*

The base-stock policy in inventory management involves initiating operations with an initial inventory level denoted by R. Whenever a new demand occurs, a replenishment order is promptly placed with the supplier. However, due to the stochastic nature of demand, multiple orders (referred to as inventory on-order) may exist at any given time, awaiting delivery. These orders are fulfilled by the supplier after a fixed lead time of L units. The demand that arises during this lead time is termed lead time demand. In the base-stock policy, the lead time demand and inventory on-order are considered equivalent. If the lead time demand (or inventory on-order) surpasses the initial inventory level R, backorders occur, indicating unfulfilled customer demand awaiting inventory replenishment. This policy ensures that the inventory level is maintained at or above a predetermined threshold (the base stock level), facilitating efficient order fulfillment while minimizing the risk of stockouts.

Notation:

I: inventory level, a random variable

B: number of backorders, a random variable

X: lead time demand (inventory on-order), a random variable

I^P: inventory position

$E[I]$: expected inventory level

$E[B]$: expected backorder level

$E[X]$: expected lead time demand

$E[D]$: average demand per unit time (demand rate)

The inventory balance equation serves as a fundamental principle in inventory management, particularly under a base-stock policy where maintaining the inventory position at a predetermined level is crucial. The inventory position at any given time is calculated by summing the on-hand inventory with the inventory on order and subtracting the backorder level. Here is the representation:

Inventory position= on-hand inventory + inventory on order − backorder level

Under a base-stock policy with base-stock level R, inventory position is always kept at R, hence $I^P = R$. This equation can be represented as:

$$I^P = I + X - B = R$$

Therefore,

$$I + X - B = R$$

Taking the expectations of both sides, we have

$$E[I] + E[X] - E[B] = R$$

Under a base-stock policy, the lead time demand X is independent of the base stock level R and depends solely on the lead time L and the demand distribution D with the expected value of X given by $E[X] = E[D]L$. The distribution of X depends on the distribution of D.

The objective is to select an optimal value of R that minimizes the sum of the expected inventory holding cost and expected backorder cost, denoted by $Y(R) = hE[I] + bE[B]$, where h is the unit holding cost per unit time and b is the backorder cost per unit per unit time.

The expected inventory holding cost ($hE[I]$) is determined by the average level of inventory held during the cycle, while the expected backorder cost ($bE[B]$) is influenced by the average number of units backordered during the cycle.

Expanding the cost function, $Y(R)$, expression further:

$$= h(R - E[X] + E[B]) + bE[B]$$

$$= h(R - E[X]) + (h + b)E[B]$$

$$= h(R - E[D]L) + (h + b)E([X - R]^+)$$

$$= h(R - E[D]L) + (h+b)\sum_{x=R}^{\infty}(x-R)P(X=x)$$

By selecting an appropriate value of R, the company aims to strike a balance between holding enough inventory to meet demand without incurring excessive holding costs and minimizing the occurrence of backorders to reduce associated costs. This optimization process allows for efficient management of inventory levels while optimizing costs, contributing to improved operational efficiency and profitability.

The Optimal Base-Stock Level

The optimal value of R is the smallest integer that satisfies the condition:

$$Y(R+1) - Y(R) \geq 0$$

This condition ensures that increasing the base stock level by one unit does not lead to a decrease in the total cost $Y(R)$.

$$Y(R+1) = h(R+1-E[D]L) + (h+b)\sum_{x=R+1}^{\infty}(x-R-1)P(X=x)$$

$$Y(R) = h(R - E[D]L) + (h+b)\sum_{x=R}^{\infty}(x-R)P(X=x)$$

If we calculate the difference between $Y(R+1)$ and $Y(R)$ we have

$$Y(R+1) - Y(R) = h + (h+b)\sum_{x=R+1}^{\infty}[(x-R-1)-(x-R)]P(X=x)$$

$$= h - (h+b)\sum_{x=R+1}^{\infty}P(X=x)$$

$$= h - (h+b)P(X \geq R+1)$$

$$= h - (h+b)[1 - P(X \leq R)]$$

$$= -b + (h+b)P(X \leq R)$$

Remember that we have the following condition

$$Y(R+1) - Y(R) \geq 0$$

So,

$$-b + (h+b)P(X \leq R) \geq 0$$

Rearranging terms gives

$$P(X \leq R) \geq \frac{b}{b+h}$$

As a result, choosing the smallest integer R that satisfies $Y(R+1)-Y(R) \geq 0$ is equivalent to choosing the smallest integer R that satisfies

$$\boxed{P(X \leq R) \geq \frac{b}{b+h}} \tag{4.2}$$

Computing Expected Backorders

To compute the expected backorders $E[B]$ for a specified base stock level R, it is sometimes easier to first compute (for a given R),

$$E[I] = \sum_{x=0}^{R}(R-x)P(X=x)$$

and then obtain $E[B] = E[I] + E[X] - R$

For the case where lead time demand has the Poisson distribution (with mean $q = E[D]L$), the following relationship (for a fixed R) applies

$$E[B] = qP(X=R) + (q-R)[1 - P(X \leq R)]$$

Here, $P(X = R)$ denotes the probability of lead time demand being exactly R units, and $[1-P(X \leq R)]$ signifies the probability of lead time demand exceeding R units. By utilizing these equations, businesses can efficiently compute the expected backorders for a given base stock level, aiding in strategic decision-making regarding inventory management to balance costs and customer satisfaction.

Example 1:

Demand arrives one unit at a time according to a Poisson process with mean λ. If $D(t)$ denotes the amount of demand that arrives in the interval of time of length t, then

$$P(D(t) = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!},\ x \geq 0$$

Lead time demand, X, can be shown in this case to also have the Poisson distribution with a mean of λL and variance λL. We have

$$P(X = x) = \frac{(\lambda L)^x e^{-\lambda L}}{x!}$$

If X can be approximated by a normal distribution, the optimal base stock level R^* can be computed as

$$R^* = E[D]L + z_{b/(b+h)}\sqrt{Var(X)}$$

The total cost function can then be expressed as

$$Y(r^*) = (h+b)\sqrt{Var(X)}\ \phi(z_{b/(b+h)})$$

where ϕ represents the cumulative distribution function of the standard normal distribution. In the case where X has the Poisson distribution with mean λL, R^* can be computed as

$$R^* = \lambda L + z_{b/(b+h)}\sqrt{\lambda L}$$

and $Y(r^*)$ is given by

$$Y(r^*) = (h+b)\sqrt{\lambda L}\ \phi(z_{b/(b+h)})$$

These expressions provide a means to determine the optimal base stock level and the associated total cost under the normal approximation for the given demand distribution.

Example 2:

If X has the geometric distribution with parameter r, where $0 \le r \le 1$, the probability mass function for X is given by

$$P(X = x) = \rho^x(1-\rho)$$

The expected value of X is calculated as

$$E[X] = \frac{\rho}{1-\rho}$$

Additionally, the probability of observing a lead time demand greater than or equal to x is

$$P(X \ge x) = \rho^x$$

and the probability of observing a lead time demand less than or equal to x is

$$P(X \le x) = 1 - \rho^{x+1}$$

Recall from Equation 4.2 that the optimal base-stock level is the smallest integer R^* that satisfies

$$P(X \le R^*) \ge \frac{b}{b+h}$$

So,

$$P(X \le R^*) = 1 - \rho^{R^*+1} \ge \frac{b}{b+h}$$

Rearranging terms gives

$$R^* \ge \frac{\ln[\frac{b}{b+h}]}{\ln[\rho]} - 1$$

As a result,

$$R^* = \left\lfloor \frac{\ln[\frac{b}{b+h}]}{\ln[\rho]} \right\rfloor$$

Insights derived from the base-stock model include the following:

Firstly, reorder points play a pivotal role in managing the probability of stock-outs by establishing safety stock levels. By setting appropriate reorder points, businesses can ensure that inventory is replenished before running out, thereby minimizing the risk of stockouts and maintaining customer satisfaction.

Secondly, the required base stock level tends to increase with both the mean and variance of demand during replenishment lead time, particularly in scenarios where the unit backordering cost exceeds the unit holding cost. This adjustment ensures sufficient inventory is held to accommodate demand fluctuations and reduce the likelihood of stockouts or backorders.

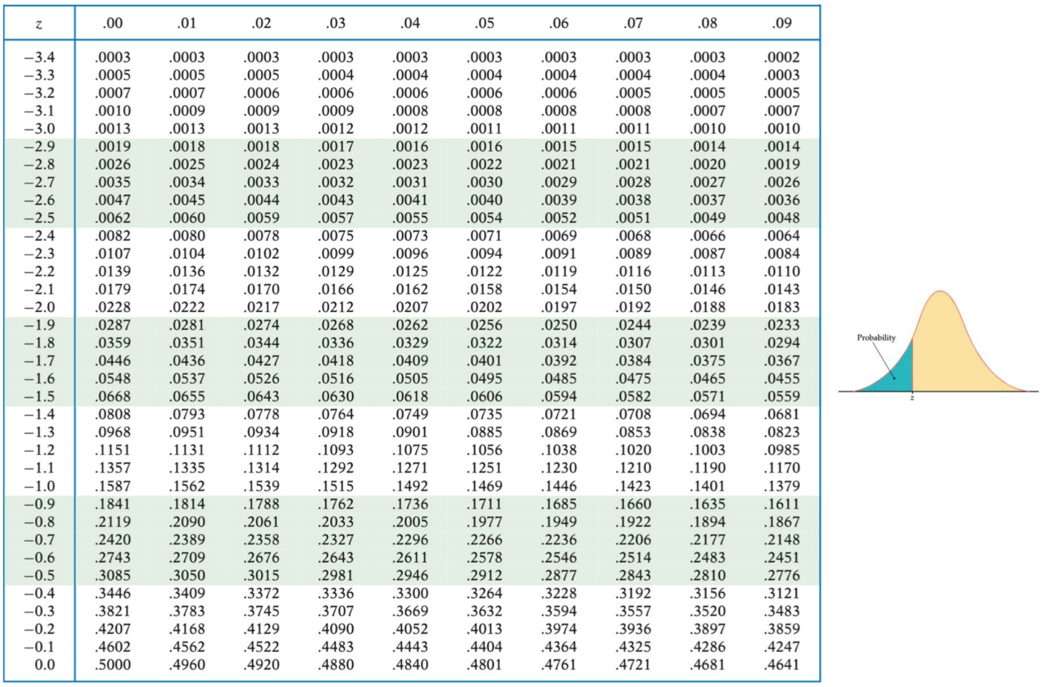
Lastly, base-stock levels in multi-stage production systems exhibit similarities to the Kanban system, a lean manufacturing technique that regulates inventory levels to synchronize production with demand. Implementing base-stock levels in such systems helps streamline production processes, optimize inventory flow, and enhance overall operational efficiency. These insights provide valuable guidelines for organizations seeking to optimize inventory management practices and improve supply chain performance.

Example 4.3

A retailer sells beds *throughout the year*. The demand for beds is Poisson with a mean of 365/year. There is no fixed ordering cost, and the lead time is one week. The holding cost for sunbeds is \$6, and backordering cost is \$10. What would be the best ordering policy? (You can use normal approximation with Figure 4.2 and the fact that the mean and variance of Poisson are the same.)

4.7 The $(Q,\ r)$ Model

Within this chapter, we embark on an exploration of the $(Q,\ r)$ model, a foundational framework in inventory management that offers a dynamic approach to maintaining optimal stock levels. We delve into the principles, methodologies, and practical applications of the model, which revolves around the periodic review of inventory levels and the replenishment of stock to predetermined thresholds. Unlike static models, the $(Q,\ r)$ model allows businesses to adapt to fluctuating demand patterns and lead times, striking a balance between minimizing stockouts and excess inventory costs. Through a comprehensive examination of its theoretical foundations, calculation techniques, and strategic implications, this chapter aims to



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Fig. 4.2 Standard normal table

provide readers with a thorough understanding of how to effectively utilize this model to enhance supply chain efficiency and responsiveness.

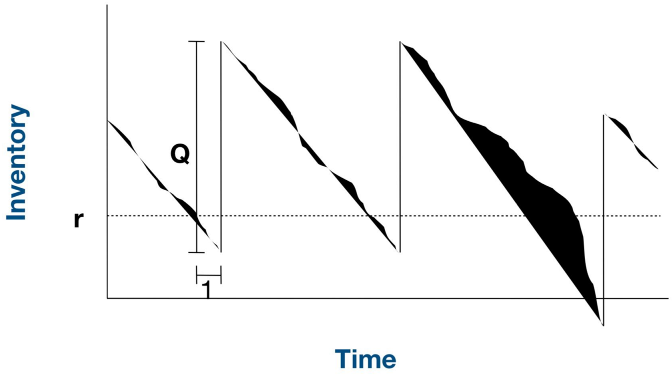
The $(Q,\ r)$ inventory model operates within a framework of specific assumptions to facilitate analysis and decision-making in inventory management. Firstly, it assumes that demand for the product occurs continuously over time, rather than in discrete intervals.

Additionally, the times between consecutive orders (inter-arrival times) are stochastic, meaning they follow a probabilistic distribution, and are assumed to be independent and identically distributed ($\boldsymbol{i.i.d.}$). Inventory levels are continuously monitored, and orders are placed whenever the inventory level reaches a predetermined reorder point (r), ensuring prompt replenishment to meet demand while minimizing excess stock. The lead time for order fulfillment is fixed at a constant value of L, aiding in determining when to place orders to maintain inventory availability. Furthermore, there is a fixed cost associated with placing an order. Finally, the model allows for backorders if an order cannot be immediately fulfilled from on-hand inventory, contributing to customer satisfaction and mitigating the risk of lost sales. These assumptions collectively provide a structured foundation for analyzing and optimizing inventory management decisions within the $(Q,\ r)$ framework.

4.7.1 *The (Q, r) Policy*

The $(Q,\ r)$ policy involves starting with an initial inventory amount $Q+r$ and continuously monitoring the inventory position. When the inventory position falls to a predetermined reorder point r, an order is placed in the quantity Q to replenish the inventory back to the initial level $Q+r$. Subsequently, whenever the inventory position drops to the reorder point r, another order of the same size Q is placed. This policy ensures that the inventory remains within desired levels, preventing stockouts while minimizing excess inventory.

Note that the base-stock policy is the special case of the $(Q,\ r)$ policy where $Q = 1$.



Notation:

I: inventory level, a random variable

B: number of backorders, a random variable

X: lead time demand (inventory on-order), a random variable

I^P: inventory position, net inventory and open orders

N: net inventory, which is actual inventory (I) minus the backorders (B)

$E[I]$: expected inventory level

$E[B]$: expected backorder level

$E[X]$: expected lead time demand

$E[D]$: average demand per unit time (demand rate)

The inventory position (I^P) is indeed determined by subtracting the backorder level from the sum of the on-hand inventory and the inventory on order. The formulation is as follows:

Inventory position = net inventory + inventory on-order

(Inventory position = on-hand inventory − backorder level + inventory on-order)

Under the $(Q,\ r)$ policy I^P typically takes on values ranging from $r+1$ to $r+Q$, where r

represents the reorder point and Q denotes the order quantity.

The time I^P remains at any specific value is the time between consecutive demand arrivals. Since the times between consecutive arrivals are independent and identically distributed, the long-run fraction of time I^P remains at any value is the same for all values.

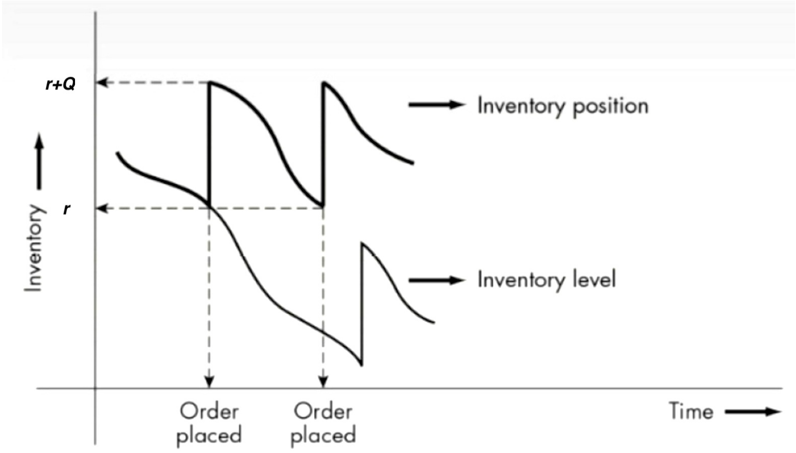


Fig. 4.3 Inventory Position vs. Inventory Level(Credit: [Nahmias, 1997])

Expected Backorders and Inventory

To highlight the dependency of inventory and backorder levels on the choice of order quantity Q and r, let's denote the inventory level as $I(Q,\ r)$ and the backorder level as $B(Q,\ r)$. The inventory level $I(Q, r)$ represents the number of items available on hand when the $(Q,\ r)$ policy is implemented, and it depends on the specific values chosen for Q and r. Similarly, the backorder level $B(Q,\ r)$ represents the quantity of unfulfilled customer demand when inventory levels are insufficient, also contingent on the selected values of Q and r.

By introducing these notation conventions, it becomes clearer that the inventory and backorder levels are not fixed but rather influenced by the decisions made regarding the order quantity and reorder point in the $(Q,\ r)$ policy.

Since $N = I^P - X$, we have

$$E[N] = E[I^P] - E[X] = r + (Q + 1)/2 - E[D]L$$

Similarly, given that $I - B = N$, we have

$$E[I(Q,\ r)] = r + (Q + 1)/2 - E[D]L + E[B(Q,\ r)]$$

This equation provides insights into the expected inventory level, accounting for backorders and other relevant inventory management factors.

The Expected Total Cost

The expected total cost $Y(Q,\ r)$ in the $(Q,\ r)$ inventory policy incorporates various cost components, including the ordering cost, inventory holding cost, and backorder cost. Let's break down the expression:

h: inventory holding cost per unit per unit time

b: backorder cost per unit per unit time

K: ordering cost per order

$$Y(Q,\ r) = KE[D]/Q + hE[I(Q,\ r)] + bE[B(Q,\ r)]$$

$$= KE[D]/Q + h[r + (Q+1)/2 - E[D]L + E[B(Q,\ r)]] + bE[B(Q,\ r)]$$

$$= KE[D]/Q + h[r + (Q+1)/2 - E[D]L] + (h+b)E[B(Q,\ r)]]$$

We want to choose r and Q so that the expected total cost (the sum of expected ordering cost, inventory holding cost, and backorder cost per unit time) is minimized.

This equation offers insights into the factors influencing total inventory-related expenses and aids in optimizing inventory management decisions by considering the trade-offs between these cost components.

4.7.1.1 *Approximate solution*

To find the optimal values of Q and r denoted by Q^* and r^* respectively, we employ an efficient computational search method. Since $Y(Q,\ r)$ is jointly convex in Q and r, this implies that the cost function has a single minimum point. Therefore, we can implement an optimization algorithm to systematically search for the values of Q and r that minimize $Y(Q,\ r)$.

The approximate solution approach involves simplifying the calculation of the expected total cost under the $(Q,\ r)$ policy by making certain assumptions and approximations. Here's a breakdown of the approach:

(1) Approximate $E[B(Q,\ r)]$ by $E[B(r)]$
Instead of explicitly calculating the expected backorder level $E[B(Q,\ r)]$, it is approximated by $E[B(r)]$, assuming that the backorder level primarily depends on the reorder point r rather than the order quantity Q. This simplification streamlines the optimization process.

(2) Assume demand is continuous

Demand is treated as continuous, which allows for easier mathematical analysis and approximation of inventory management metrics.

(3) Treat Q and r as continuos variables
Q and r are treated as continuous variables rather than discrete values, enabling the use of optimization techniques that operate on continuous domains.

The expected total cost function is then formulated based on these approximations:

$$Y(Q,\ r) = KE[D]/Q + h[r + (Q+1)/2 - E[D]L] + (h+b)E[B(r)]]$$

Using this formulation, the optimal order quantity is approximated by:

$$\boxed{Q^*{=}\sqrt{\frac{2KE[D]}{h}}} \tag{4.3}$$

$$\boxed{F(r^*){=}\frac{b}{b+h}} \tag{4.4}$$

where F denotes the cumulative distribution function of demand during lead time. If the distribution of lead time demand is approximated by a normal distribution, then the optimal reorder point can be approximated by

$$r^* \approx E[D]L + z_{b/b+h}\sqrt{Var(X)}$$

$$= E[D]L + z_{b/b+h}\sqrt{Var(D)L}$$

$$\boxed{{=}E[D]L + z_{b/b+h}\sigma_D\sqrt{L}} \tag{4.5}$$

$$\boxed{Var(X){=}L\ Var(D)} \tag{4.6}$$

Insights derived from the $(Q,\ r)$ model include the following:

Firstly, if we increase the reorder point r, it means we keep more safety stock on hand. This helps us avoid running out of stock when demand unexpectedly rises. Additionally, raising the order quantity Q means that we hold more inventory in each batch, which can reduce the number of orders we need to place and the associated setup costs. Other insights include that the longer lead times tend to mean we need to set higher reorder points, as we need more buffer stock to cover delays. In addition to this, more unpredictable demand patterns often call for higher reorder points to ensure that we have enough stock to meet varying demands. Lastly, if holding costs are high, it's usually best to order smaller quantities less frequently to minimize the expense of holding excess inventory. These insights help managers find the right balance between inventory costs and ensuring products are available when customers need them.

Service Level Approximations

Service level approximations offer an understanding of the effectiveness of inventory management policies in meeting customer demand. In Type 1 Service, denoted by $S(Q,\ r)$, the approximation $G(r)$ represents the probability of not experiencing a stockout, which is crucial for ensuring products are available when customers need them. (Type 1 Service: $S(Q,\ r) \approx G(r)$)

In Type 2 Service, the approximation $1 - E[B(r)]/Q$ indicates the proportion of time the inventory level remains above the safety stock level s, reflecting the reliability of inventory levels in meeting demand. (Type 2 Service: $S(Q,\ r) \approx 1 - E[B(r)]/Q$)

Example 4.4

A retailer sells beds *throughout the year*. The demand for beds is Poisson with a mean of 365/year. There is a fixed cost of $200 per order, and the lead time is one week. The holding cost for sunbeds is $6, and backordering cost is $10. What would be an approximate solution for the best ordering policy? (You can use normal approximation with Figure 4.2 and the fact that the mean and variance of Poisson are the same.)

4.7.1.2 *Exact Solution*

There is no closed form solution for the exact approach. The proof is omitted for brevity, but there are several equations that need to be solved recursively until convergence. The list of equations is as follows:

(1) Start with approximate $Q = \sqrt{\frac{2KE[D]}{h}}$

(2) Using this Q find the associated r using the following Optimality Condition I:

$$\boxed{1 - F(r^*) = \frac{Q^* h}{bE(D)}} \tag{4.7}$$

Note that α denotes the Type I Service Level, and we usually use normal approximations for the demand during lead time.

$$\boxed{F(r) = \alpha} \tag{4.8}$$

(3) Using the above r^* value, find the loss function and $n(r)$ as follows:

$$\boxed{n(r) = \sigma \mathrm{L}\left(\frac{r - \mu}{\sigma}\right) = \sigma \mathrm{L}(z)} \tag{4.9}$$

L($\cdot$) denotes the loss function, which can be read from Figure 4.4. μ and σ are the mean and standard deviation of *demand during lead time*. Thus, under normal assumptions: $\mu = E[D]L$ and $\sigma = \sqrt{Var[D]L}$.

The associated Type II Service Level (β) can be calculated via

$$\boxed{n(r)/Q = 1 - \beta} \tag{4.10}$$

(4) Finally, use the Optimality Condition II to calculate the associated Q value:

$$\boxed{Q^* = \sqrt{\frac{2E(D)[K + bn(r^*)]}{h}}} \tag{4.11}$$

If this is reasonably close to the previous Q value, stop. Otherwise, proceed with step 2 above.

During this procedure, we need not only the standard normal z-chart, but the loss function values as well. Figure 4.4 is an example, where $F(Z)$ is the probability that a variable from a standard normal distribution will be less than or equal to Z, or alternately, the service level for a quantity ordered with a z-value of Z.

Recall that L(Z) is the standard loss function, i.e., the expected number of lost sales as a fraction of the standard deviation. Hence, the lost sales is equal to the L(Z) value multiplied by the standard deviation of demand σ. Figure 4.4 includes this loss function values for the standard normal variate.

Solve 4.4 using the exact method, and compare your answer with the approximate solution.

Important 4.2

A chemical compounds manufacturer buys a certain raw material. The annual demand for this raw material is normally distributed with a mean of 5000 tons and a standard deviation of 120 tons. The raw material costs \$75 per ton and the annual interest rate is %25. The shop pays \$1500 for each order placed and the order arrives in 6 weeks. If the raw material stock is out, then the production is disrupted. The annual penalty cost of disrupting the production is estimated to be \$45 per ton of demanded raw material.

a. Find the optimal order quantity, reorder point, and safety stock[1] for this item.

b. What are the Type 1 and Type 2 service levels obtained by the policy in part a?

Solution

Let's summarize the given quantities:

[1] Safety stock is introduced to provide a buffer, helping meet unexpected demand. Thus, it is the items in excess of the expected demand during lead time.

Z	F(Z)	L(Z)	Z	F(Z)	L(Z)	Z	F(Z)	L(Z)	Z	F(Z)	L(Z)
-3.00	0.0013	3.000	-1.48	0.0694	1.511	0.04	0.5160	0.379	1.56	0.9406	0.026
-2.96	0.0015	2.960	-1.44	0.0749	1.474	0.08	0.5319	0.360	1.60	0.9452	0.023
-2.92	0.0018	2.921	-1.40	0.0808	1.437	0.12	0.5478	0.342	1.64	0.9495	0.021
-2.88	0.0020	2.881	-1.36	0.0869	1.400	0.16	0.5636	0.324	1.68	0.9535	0.019
-2.84	0.0023	2.841	-1.32	0.0934	1.364	0.20	0.5793	0.307	1.72	0.9573	0.017
-2.80	0.0026	2.801	-1.28	0.1003	1.327	0.24	0.5948	0.290	1.76	0.9608	0.016
-2.76	0.0029	2.761	-1.24	0.1075	1.292	0.28	0.6103	0.274	1.80	0.9641	0.014
-2.72	0.0033	2.721	-1.20	0.1151	1.256	0.32	0.6255	0.259	1.84	0.9671	0.013
-2.68	0.0037	2.681	-1.16	0.1230	1.221	0.36	0.6406	0.245	1.88	0.9699	0.012
-2.64	0.0041	2.641	-1.12	0.1314	1.186	0.40	0.6554	0.230	1.92	0.9726	0.010
-2.60	0.0047	2.601	-1.08	0.1401	1.151	0.44	0.6700	0.217	1.96	0.9750	0.009
-2.56	0.0052	2.562	-1.04	0.1492	1.117	0.48	0.6844	0.204	2.00	0.9772	0.008
-2.52	0.0059	2.522	-1.00	0.1587	1.083	0.52	0.6985	0.192	2.04	0.9793	0.008
-2.48	0.0066	2.482	-0.96	0.1685	1.050	0.56	0.7123	0.180	2.08	0.9812	0.007
-2.44	0.0073	2.442	-0.92	0.1788	1.017	0.60	0.7257	0.169	2.12	0.9830	0.006
-2.40	0.0082	2.403	-0.88	0.1894	0.984	0.64	0.7389	0.158	2.16	0.9846	0.005
-2.36	0.0091	2.363	-0.84	0.2005	0.952	0.68	0.7517	0.148	2.20	0.9861	0.005
-2.32	0.0102	2.323	-0.80	0.2119	0.920	0.72	0.7642	0.138	2.24	0.9875	0.004
-2.28	0.0113	2.284	-0.76	0.2236	0.889	0.76	0.7764	0.129	2.28	0.9887	0.004
-2.24	0.0125	2.244	-0.72	0.2358	0.858	0.80	0.7881	0.120	2.32	0.9898	0.003
-2.20	0.0139	2.205	-0.68	0.2483	0.828	0.84	0.7995	0.112	2.36	0.9909	0.003
-2.16	0.0154	2.165	-0.64	0.2611	0.798	0.88	0.8106	0.104	2.40	0.9918	0.003
-2.12	0.0170	2.126	-0.60	0.2743	0.769	0.92	0.8212	0.097	2.44	0.9927	0.002
-2.08	0.0188	2.087	-0.56	0.2877	0.740	0.96	0.8315	0.090	2.48	0.9934	0.002
-2.04	0.0207	2.048	-0.52	0.3015	0.712	1.00	0.8413	0.083	2.52	0.9941	0.002
-2.00	0.0228	2.008	-0.48	0.3156	0.684	1.04	0.8508	0.077	2.56	0.9948	0.002
-1.96	0.0250	1.969	-0.44	0.3300	0.657	1.08	0.8599	0.071	2.60	0.9953	0.001
-1.92	0.0274	1.930	-0.40	0.3446	0.630	1.12	0.8686	0.066	2.64	0.9959	0.001
-1.88	0.0301	1.892	-0.36	0.3594	0.605	1.16	0.8770	0.061	2.68	0.9963	0.001
-1.84	0.0329	1.853	-0.32	0.3745	0.579	1.20	0.8849	0.056	2.72	0.9967	0.001
-1.80	0.0359	1.814	-0.28	0.3897	0.554	1.24	0.8925	0.052	2.76	0.9971	0.001
-1.76	0.0392	1.776	-0.24	0.4052	0.530	1.28	0.8997	0.047	2.80	0.9974	0.001
-1.72	0.0427	1.737	-0.20	0.4207	0.507	1.32	0.9066	0.044	2.84	0.9977	0.001
-1.68	0.0465	1.699	-0.16	0.4364	0.484	1.36	0.9131	0.040	2.88	0.9980	0.001
-1.64	0.0505	1.661	-0.12	0.4522	0.462	1.40	0.9192	0.037	2.92	0.9982	0.001
-1.60	0.0548	1.623	-0.08	0.4681	0.440	1.44	0.9251	0.034	2.96	0.9985	0.000
-1.56	0.0594	1.586	-0.04	0.4840	0.419	1.48	0.9306	0.031	3.00	0.9987	0.000
-1.52	0.0643	1.548	0.00	0.5000	0.399	1.52	0.9357	0.028			

Fig. 4.4 Cumulative (F) and Loss (L) Function Values for Standard Normal Variate

$K = 1500,\ h = ic = 75 \cdot 0.25 = 18.75, b = 45,\ D \sim N(5000, \sigma_D = 120), L = 6$ weeks $= \frac{6}{52}$ years

Start using EOQ with expected annual demand.

$Q = \sqrt{\frac{2 \cdot 1500 \cdot 5000}{18.75}} \approx 894$

<u>Iteration 1:</u>

$1 - F(r) = \frac{894 \cdot 18.75}{45 \cdot 5000} = 0.0745 \Rightarrow F(r) = 0.9255$

Using Figure 4.4, we observe that $z = 1.44$ and L(z) = 0.034.

Notice that for the demand during lead time, the distribution has mean $\mu = 5000 \cdot \frac{6}{52} = 577$ and standard deviation $\sigma = 120 \cdot \sqrt{\frac{6}{52}} \approx 40$

That leads to $r = 577 + 1.44 \cdot 40 \approx 635$ and the associated expected lost customers is $n(r) = 40 \cdot 0.034 = 1.36$

Plugging these back in the optimality condition gives

$Q = \sqrt{\frac{2 \cdot 5000[1500 + 45 \cdot 1.36]}{18.75}} \approx 912$

<u>Iteration 2:</u>

$1 - F(r) = \frac{912 \cdot 18.75}{45 \cdot 5000} = 0.076 \Rightarrow F(r) = 0.924$

Using Figure 4.4 and interpolating between two intervals there, we observe that $z = 1.43$

and L(z) = 0.035.

That leads to $r = 577 + 1.43 \cdot 40 = 634$ and the associated expected lost customers is $n(r) = 40 \cdot 0.035 = 1.4$

Plugging these back in the optimality condition gives

$Q = \sqrt{\frac{2 \cdot 5000[1500+45 \cdot 1.4]}{18.75}} \approx 913$

Iteration 3:

$1 - F(r) = \frac{913 \cdot 18.75}{45 \cdot 5000} \approx 0.076 \Rightarrow F(r) = 0.924$

Thus, we will obtain the same z, $L(z)$, $n(r)$, and Q values. Convergence is achieved and we stop.

$Q^* = 913$, $r^* = 634$

Type I service level: $\alpha = F(r^*) = 92.4\%$

Type II service level: $\beta = 1 - n(r^*)/Q = 1 - 1.4/913 = 99.85\%$.

The $(S,\ s)$ Model

Transitioning to the $(S,\ s)$ model, which is also known as the $(R,\ r)$ model, introduces a periodic approach to inventory management. In this model, each demand order can consist of multiple units, and demand orders occur stochastically. When the inventory level is **less than or equal to** s, a replenishment order of size $S - s$ is placed. Here, s serves as the safety stock level, ensuring a buffer against stockouts, while S, commonly referred to as the order-up-to level or par level, represents the desired inventory level to meet demand effectively. This periodic version of the $(Q,\ r)$ model provides a structured framework for inventory replenishment, balancing inventory levels to ensure optimal service levels while minimizing the risk of stockouts.

Dealing with Lead Time Variability

Lead time refers to the time between placing an order and receiving it. Variability in lead time occurs due to factors such as supplier reliability, transportation delays, and production issues. Variability in lead time can lead to stockouts or excess inventory if not managed properly. It can be characterized by its mean and variance. Similar to lead time variability, demand variability is also a critical factor in inventory management. Here are the notations:

L: replenishment lead time (a random variable)

$E[L]$: expected replenishment lead time

$Var(L)$: variance of lead time

D: demand per period

$E[D]$: expected demand per period

$Var(D)$: variance of demand

Lead time demand, X, is the demand during the lead time. The expected lead time demand is given by:

$$\boxed{E[X]{=}E[L]E[D]} \tag{4.12}$$

The variance of the lead time demand is affected by both lead time variability and demand variability, which is stated as:

$$\boxed{Var(X){=}E[L]Var(D) + E[D]^2 Var(L)} \tag{4.13}$$

The proofs are omitted for brevity.

Example 4.5

Jack, the maintenance manager, has collected historical data that indicate one of the replacement parts he stocks has an annual demand (D) of 14 units per year. The unit cost (c) of the part is \$150, and since the firm uses an interest rate of 20 percent, the annual holding cost (h) has been set at 0.2(\$150) = \$30 per year. It takes 45 days to receive a replenishment order, so the average demand during a replenishment lead time is

$$\theta = \frac{14}{365}(45) = 1.726$$

The part is purchased from an outside supplier, and Jack estimates that the cost of time and materials required to place a purchase order (K) is about \$15. The one remaining cost required by our model is the backorder cost. Although he was very uncomfortable trying to estimate this, when pressed, Jack guessed that the annualized cost of backorder is about b = \$100 per year. Finally, Jack has decided that demand is Poisson distributed, which means the standard deviation is equal to the square root of the mean.
The order quantity is computed as follows:

$$Q^* = \sqrt{\frac{2KE[D]}{h}} = \sqrt{\frac{2(15)(14)}{30}} = 3.7 \approx 4$$

To compute the reorder point, we approximate the Poisson by the normal, with the following mean and standard deviation

$$E[D]L = 1.726$$

$$\sigma = \sqrt{1.726} = 1.314$$

The critical fractile is given by

$$\frac{b}{b+h} = \frac{100}{100+30} = 0.769$$

Using a standard normal table,

$$z = \phi(0.736) = 0.769$$

$$r^* = E[D]L + z\sigma = 1.726 + 0.736(1.314) = 2.693 \approx 3$$

where $z\sigma$ is safety stock

Important 4.3

An electronic device retail shop buys and sells Bluetooth speakers. The annual demand for speakers is normally distributed with a mean of 1500 and a standard deviation of 100. A speaker costs \$250 to the shop and the annual interest rate is 25%. The shop pays \$2500 for each order placed and an order arrives in 8 weeks. Any demand that is not satisfied on time is fully backordered and the penalty cost incurred by not satisfying a customer demand on time is estimated to be \$40. (Assume a year has 52 weeks.)

The electronic device retailer shop applies $(Q,\ r)$ control policy for Bluetooth speakers where $Q = 500$ and $r = 300$. Calculate Type 1 and Type 2 service levels of this inventory control policy.

Find the optimal order quantity, reorder point, and safety stock for Bluetooth speakers. What is the expected number of shortages when the optimal policy is applied?

Chapter 5

Aggregate Production Planning

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 changes the course of one's life: where to live, what to major in, which job to take, and whom to marry. A company also must make both micro and macro decisions every day. In this chapter, we explore decisions made at the macro level, such as planning companywide workforce and production levels.

Aggregate planning, which might also be called macro production planning, 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. So, the goal of aggregate planning is to determine aggregate production quantities and the levels of resources required to achieve production goals.

The scope encompasses various aspects of operational planning and decision-making to optimize production processes and resource utilization while minimizing costs and maximizing profits. Key areas of focus include:

Production Scheduling: Efficiently organizing production activities over time to maximize profitability while ensuring that production levels do not exceed available capacity. This involves balancing production rates with demand fluctuations and capacity constraints to avoid underutilization or overloading of resources.

Production Smoothing: Implementing strategies to build inventory ahead of demand fluctuations, enabling smoother production levels and minimizing disruptions caused by variations in demand or supply.

Product Mix Planning: Determining the optimal combination of products to manufacture based on resource availability, demand forecasts, and profitability considerations. This involves allocating resources to produce the most profitable product mix while ensuring the efficient use of available resources.

Staffing: Managing workforce requirements through hiring, firing, and training activities to match production needs and ensure optimal utilization of labor resources. This includes aligning staffing levels with production schedules and adjusting workforce size and skill sets based on demand fluctuations and production requirements.

Procurement: Negotiating supplier contracts for materials and components to secure favorable terms and ensure a reliable supply chain. This involves optimizing procurement strategies to minimize costs, reduce lead times, and mitigate supply chain risks while maintaining quality standards.

Sub-Contracting: Leveraging external capacity through subcontracting arrangements to meet production demands during peak periods or capacity constraints. This may involve outsourcing certain production processes or tasks to specialized vendors to optimize resource utilization and improve production flexibility.

Marketing: Integrating promotional activities and marketing strategies into production planning to align production levels with anticipated demand and market trends. This includes coordinating production schedules with marketing campaigns to ensure timely availability of products and capitalize on sales opportunities.

By addressing these aspects comprehensively, businesses can optimize their production operations, enhance resource efficiency, and maintain competitiveness in dynamic market environments while achieving cost-effective production outcomes and maximizing profitability.

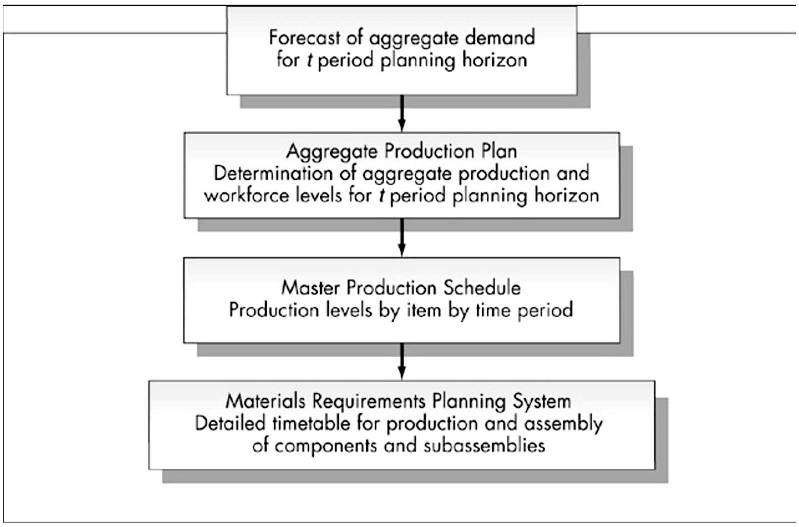

Forecast of aggregate demand for *t* period planning horizon
Aggregate Production Plan Determination of aggregate production and workforce levels for *t* period planning horizon
Master Production Schedule Production levels by item by time period
Materials Requirements Planning System Detailed timetable for production and assembly of components and subassemblies

Fig. 5.1 The hierarchy of production planning decisions (Credit: [Nahmias, 1997])

Issues in APP

In aggregate production planning, several critical issues must be addressed to optimize production processes and maximize profitability. These issues include:

Limited Capacity: Limited production capacity poses a significant challenge in meeting aggregate demand while ensuring efficient resource utilization. Balancing production levels with available capacity requires strategic capacity planning, considering factors such as workforce availability, equipment capacity, and facility constraints.

Varying Demand: Fluctuations in demand introduce uncertainty and complexity into aggregate production planning. Managing varying demands requires robust forecasting methods and flexibility in production scheduling to adjust output levels in response to changing demand patterns while minimizing production costs and meeting customer service requirements.

Inventory Holding Costs vs. Lost Revenue: Finding the right balance between inventory holding costs and potential lost revenue due to stockouts is critical. Excessive inventory ties up capital and incurs holding costs, while insufficient inventory levels can lead to lost sales and dissatisfied customers. Effective inventory management strategies aim to optimize inventory levels to minimize holding costs while ensuring product availability to meet customer demand.

Multiple Products with Varying Characteristics: Managing multiple products with diverse demand patterns, prices, production costs, and capacity requirements adds complexity to aggregate production planning. It requires the segmentation of products based on demand characteristics and strategic allocation of resources to maximize overall profitability. Techniques such as product mix optimization and capacity leveling can help in efficiently utilizing resources across different product lines.

By addressing these critical issues, companies can achieve better alignment between production resources and market requirements.

Tools for Decision Making

In the realm of decision-making, various tools and methods are available to guide organizations in making informed choices and optimizing outcomes. These include:

Trial and Error: This approach involves testing different strategies or solutions iteratively until the desired outcome is achieved. While simple, trial and error can be time-consuming and may not always yield optimal results.

Heuristics: Heuristics are proven rules or strategies that provide shortcuts for decision-making in complex situations. They are based on past experiences, best practices, or common sense, allowing decision-makers to quickly arrive at satisfactory solutions. However, heuristics may not always guarantee the best possible outcome and can be susceptible to biases.

Optimization Methods: Optimization methods aim to find the best solution from a set of feasible alternatives, considering specific objectives, constraints, and variables. Some common optimization methods include:

- ***Linear Programming (LP):*** LP is a mathematical technique used to optimize a linear objective function subject to linear equality and inequality constraints. It is widely used in resource allocation, production planning, and supply chain optimization.
- ***Mixed Integer Programming (MIP):*** MIP extends LP by allowing some decision variables to take integer values, enabling the modeling of discrete decision variables. MIP is useful for solving problems with both continuous and discrete decision variables, such as production scheduling and network optimization.
- ***Nonlinear Programming (NLP):*** NLP deals with optimization problems where the objective function or constraints are nonlinear. It is used to solve complex optimization problems with non-convex objectives or constraints, such as portfolio optimization, process optimization, and engineering design.

These decision-making tools provide organizations with systematic approaches to analyze complex problems, identify optimal solutions, and make data-driven decisions that drive efficiency, productivity, and competitive advantage. By leveraging these tools effectively, businesses can enhance their decision-making processes, mitigate risks, and achieve their strategic objectives more effectively.

5.1 Basic Aggregate Planning

In basic aggregate planning, the goal is to project the production of a single product over a defined planning horizon. This study is motivated by the exploration of the mechanics and value of linear programming (LP) as a tool for optimizing production decisions, as well as understanding the concept of production smoothing. The inputs to this planning process include the demand forecast over the planning horizon, capacity constraints of the production facilities, unit profit from each unit produced, and the inventory carrying cost rate. The objectives of basic aggregate planning encompass minimizing costs and maximizing profits, while also enabling a quick response to changes in demand or market conditions, maximizing customer service, minimizing inventory investment, minimizing changes in production rates, minimizing changes in workforce levels, and maximizing the utilization of plant and equipment. By addressing these objectives, businesses can achieve greater efficiency, profitability, and responsiveness in their production processes.

5.1.1 *Relevant Costs*

As with most of the optimization problems considered in production management, the goal of the analysis is 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 workforce means that workers must be laid off. Severance pay is thus one cost of decreasing the size of the workforce. Other costs, somewhat harder to measure, are (i) the costs of a decline in worker morale that may result and (ii) the potential for decreasing the size of the labor pool in the future, as workers who are laid off acquire jobs with other firms or in other industries.

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 of 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 the figure below.

(2) ***Holding costs.*** Holding costs are the costs that accrue as a result of having capital tied up in inventory. They 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 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 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

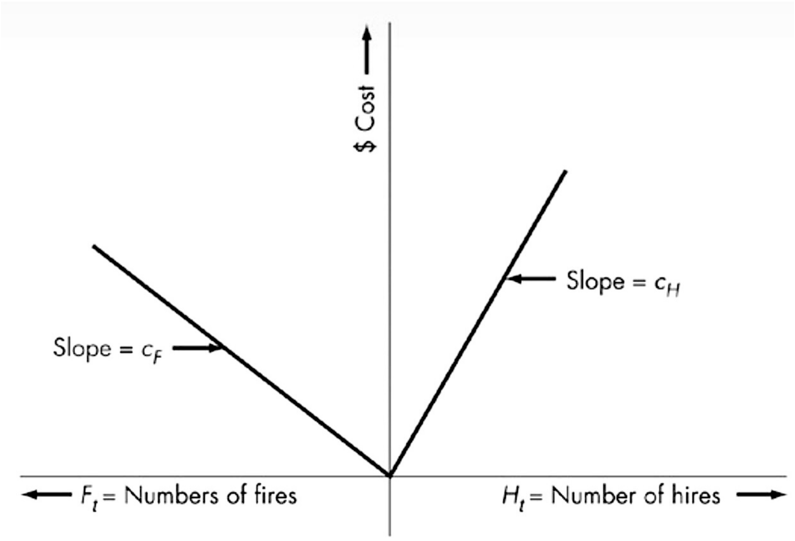


Fig. 5.2 Cost of changing the size of the workforce (Credit: [Nahmias, 1997])

than anticipated. For 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 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. The figure below shows a typical holding/shortage cost function.

(4) ***Other costs.*** Basic aggregate planning also considers other costs such as payroll, overtime, and subcontracting expenses. These costs are essential components of the overall production expenditure and need to be factored into the planning process. Payroll costs involve the expenses associated with hiring and compensating the workforce, including wages, salaries, benefits, and other related expenses. Overtime costs arise when employees work beyond their regular hours, typically at a higher rate of pay, to meet increased demand or production requirements. Subcontracting expenses refer to the costs incurred when outsourcing certain production activities or tasks to external vendors or subcontractors.

5.1.2 *Important Issues*

The primary issues related to the aggregate planning problem include

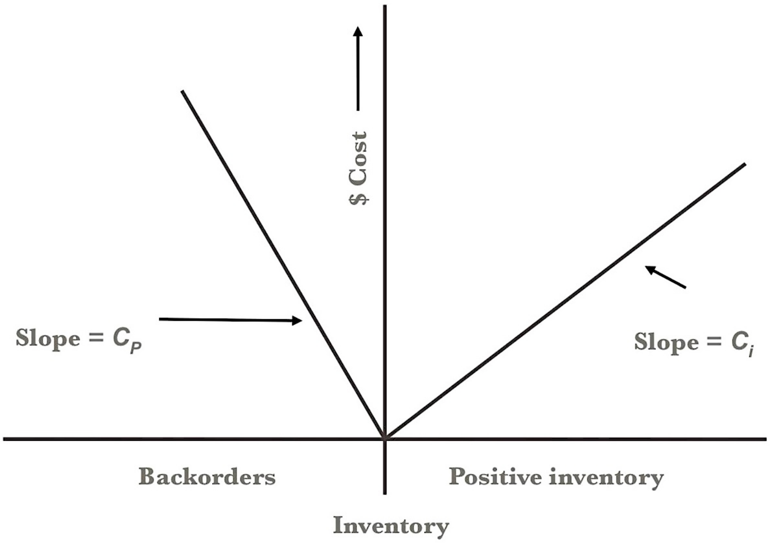


Fig. 5.3 Holding and backorder costs (Credit: [Nahmias, 1997])

(1) ***Smoothing.*** Smoothing refers to costs that result from changing production and workforce levels from one period to the next. Two of the key components of smoothing costs are the costs that result from hiring and firing workers. Aggregate planning methodology requires the specification of these costs, which may be difficult to estimate. Firing workers could have far-reaching consequences and costs that may be difficult to evaluate. Firms that hire and fire frequently develop a poor public image. This could adversely affect sales and discourage potential employees from joining the company. Furthermore, workers who 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. However, it is still valuable for management to be aware of the cost trade-offs associated with varying workforce levels and the attendant savings in inventory costs.

(2) ***Bottleneck problems.*** We use the term *bottleneck* to refer to the inability of the system to respond to sudden changes in demand as a result of capacity restrictions. For example, a bottleneck could arise when the forecast for demand in one month is unusually high, and the plant does not have sufficient capacity to meet that demand. A breakdown of a vital piece of equipment also could result in a bottleneck.

(3) ***Planning horizon.*** The number of periods for which the demand is to be forecasted,

and hence the number of periods for which workforce and inventory levels are to be determined, must be specified in advance. The choice of the value of the forecast horizon, T, can be significant in determining the usefulness of the aggregate plan. If T is too small, then current production levels might not be adequate for meeting the demand beyond the horizon length. If T is too large, the forecasts far into the future will likely prove inaccurate. If future demands turn out to be very different from the forecasts, then current decisions indicated by the aggregate plan could be incorrect. Another issue involving the planning horizon is the *end-of-horizon* effect. For example, the aggregate plan might recommend that the inventory at the end of the horizon be drawn to zero to minimize holding costs. This could be a poor strategy, especially if demand increases at that time. (However, this particular problem can be avoided by adding a constraint specifying minimum ending inventory levels.)

In practice, rolling schedules are almost always used. This means that at the time of the next decision, a new forecast of demand is appended to the former forecasts and old forecasts might be revised to reflect new information. The new aggregate plan may recommend different production and workforce levels for the current period than were recommended one period ago. When only the decisions for the current planning period need to be implemented immediately, the schedule should be viewed as dynamic rather than static.

Although rolling schedules are common, it is possible that because of production lead times, the schedule must be frozen for a certain number of planning periods. This means that decisions over some collection of future periods cannot be altered. The most direct means of dealing with frozen horizons is simply to label as period 1 the first period in which decisions are not frozen.

(4) ***Treatment of demand.*** As noted above, 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 ignores the possibility (and, in fact, likelihood) of forecast errors. As noted in the discussion of forecasting techniques in Chapter 2, it is virtually a certainty that demand forecasts are wrong. Aggregate planning 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.

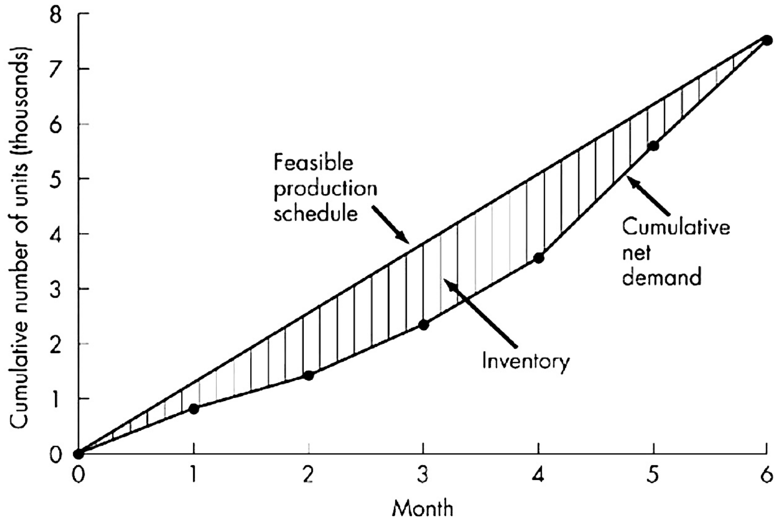


Fig. 5.4 Feasible aggregate plan (Credit: [Nahmias, 1997])

5.1.3 *Aggregate Units*

The aggregate planning approach is predicated on the existence of an aggregate unit of production. When the types of items produced are similar, an aggregate production unit can correspond to an "average" item, but if many different types of items are produced, it would be more appropriate to consider aggregate units in terms of weight (tons of steel), volume (gallons of gasoline), amount of work required (workers-years of programming time), or dollar value (value of inventory in dollars). What the appropriate aggregating scheme should be is not always obvious. It depends on the context of the particular planning problem and the level of aggregation required.

Example 5.1

A plant manager working for a large national appliance firm is considering implementing an aggregate planning system to determine the workforce and production levels in his plant. This particular plant produces six models of washing machines. The characteristics of the machines are

Model Number	Number of Worker-Hours	Selling Price ($)	Sales (%)
A5532	4.2	285	32
K4242	4.9	345	21
L9898	5.1	395	17
L3800	5.2	425	14
M2624	5.4	525	10
M3880	5.8	725	6

The plant manager must decide on the particular aggregation scheme to use. One possibility is to define an aggregate unit as one dollar of output. Unfortunately, the selling prices of the various models of washing machines are not consistent with the number of worker hours required to produce them. The ratio of the selling price divided by the worker hours is $67.86 for A5532 and $125.00 for M3880. (The company bases its pricing on the fact that the less expensive models have a higher sales volume). The manager notices that the percentages of the total number of sales for these six models have been fairly constant, with values of 32 percent for A5532, 21 percent for K4242, 17 percent for L9898, 14 percent for L3800, 10 percent for M2624, and 6 percent for M3880. He decides to define an aggregate unit of production as a fictitious washing machine requiring $(0.32)(4.2)+(0.21)(4.9)+(0.17)(5.1)+(0.14)(5.2)+(0.10)(5.4)+(0.06)(5.8) = 4.856$ hours of labor time. He can obtain sales forecasts for aggregate production units in essentially the same way by multiplying the appropriate fractions by the forecasts for unit sales of each type of machine.

The approach used by the plant manager in Example 5.1 was possible because of the relative similarity of the products produced. However, defining an aggregate unit of production at a higher level of the firm is more difficult. In cases in which the firm produces a large variety of products, a natural aggregate unit is sales dollars. Although, as we saw in the example, this will not necessarily translate to the same number of units of production for each item, it will generally provide a good approximation for planning at the highest level of a firm that produces a diverse product line.

Solution Approaches:

In addressing basic aggregate planning problems, several solution approaches can be employed to determine optimal production strategies. Graphical solutions offer an intuitive understanding of the trade-offs involved in production decisions, enabling approximate solutions through visual analysis of cost and demand curves. The constant workforce strategy entails maintaining a steady level of employees without hiring or firing, with production adjustments made through inventory holding. The zero inventory, or chase strategy, aims to align production with demand fluctuations, minimizing inventory levels through workforce adjustments.

Linear programming provides exact solutions by optimizing linear objective functions subject to constraints, but challenges arise when dealing with integer variables or complex constraints, leading to computational difficulties.

Each approach offers unique advantages and trade-offs, and the choice of method depends on factors such as the linearity of costs, the complexity of constraints, and the available computational resources. By leveraging these solution approaches effectively, businesses can develop robust aggregate production plans that meet their objectives while balancing cost considerations and operational constraints.

Example 5.2

The washing machine plant is interested in determining workforce and production levels for the next 8 months. Forecasted demands for Jan-Aug. are: 420, 280, 460, 190, 310, 145, 110, 125. Starting inventory at the end of December is 200 and the firm would like to have 100 units on hand at the end of August. Find monthly production levels.

Solution

Step 1: Determine "net" demand by subtracting starting inventory from the period 1 forecast and add ending inventory to the period 8 forecast.

Month	Net Demand	Cumulative Demand
Jan	220	220
Feb	280	500
Mar	460	960
Apr	190	1150
May	310	1460
June	145	1605
July	110	1715
Aug	225	1940

Step 2: Graph cumulative net demand to find plans graphically

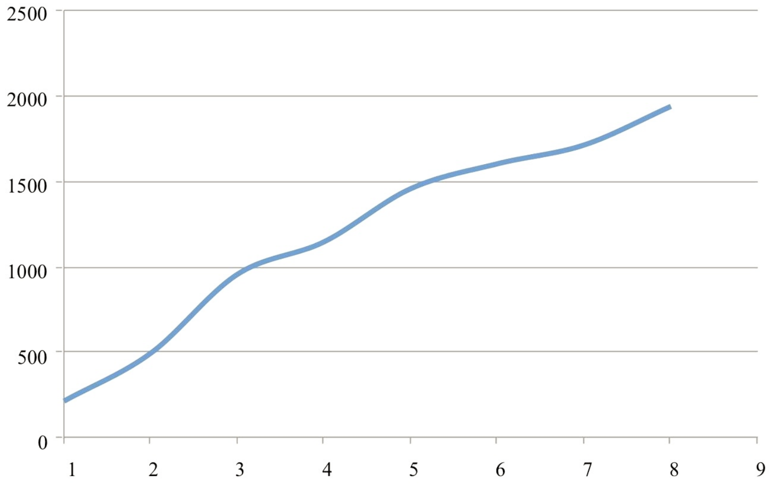


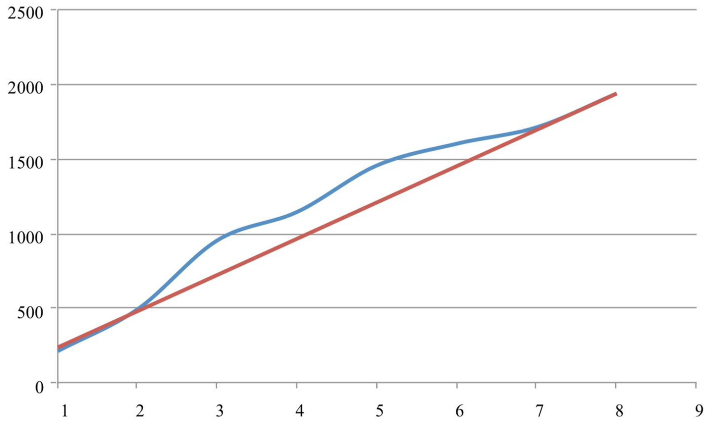
Fig. 5.5 Cumulative demand over time

We will evaluate two alternative plans for managing the workforce that represent two essentially opposite management strategies. Plan 1 is to maintain the minimum constant workforce necessary to satisfy the net demand. This is known as the *constant workforce plan.* Plan 2 is to change the workforce each month to produce enough units to most closely match the demand pattern. This is known as a *zero inventory plan.*

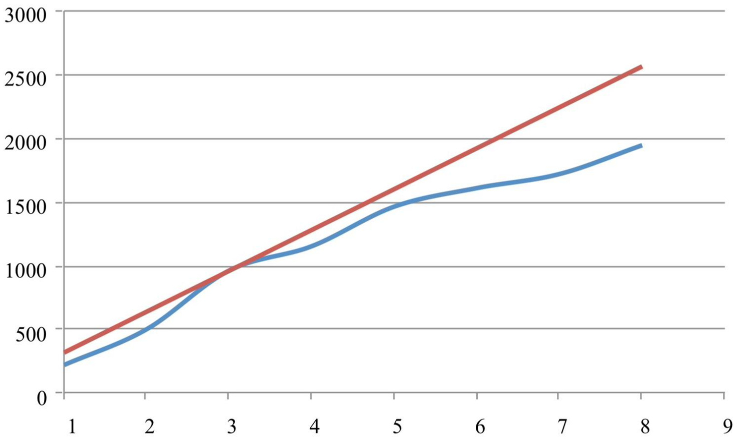
5.2 Constant Workforce Plan

Now assume that the goal is to eliminate the need for hiring and firing during the planning horizon. So, we are interested in determining a production plan that doesn't change the size of the workforce over the planning horizon.

Graphical Method: In the previous picture, draw a straight line from the origin to 1940 units in month 8. The slope of the line is the number of units produced each month.



Monthly production is calculated as $1940/8 = 242.2$ or rounded to 243/month. However, there are stockouts in this case because the cumulative demand curve is above the line representing the constant workforce.



We should consider the plan pictured above. Here there are no stockouts, because the cumulative demand curve is below the line representing the constant workforce. Now, we can calculate the monthly production rate of such a plan. We can calculate the production rate by dividing the cumulative demand by the number of months that have passed so far.

Month	Net Demand	Cumulative Demand	Production Rate
Jan	220	220	220
Feb	280	500	250
Mar	460	960	320
Apr	190	1150	287.5
May	310	1460	292
June	145	1605	267.5
July	110	1715	245
Aug	225	1940	242.5

From the graph, we see that the cumulative net demand curve is crossed at period 3 so that monthly production is $960/3 = 320$. Ending inventory each month is found from:

Month	Net Demand	Cumulative Demand	Ending Inventory
Jan	220	220	$320 - 220 = 100$
Feb	280	500	$2(320) - 500 = 140$
Mar	460	960	$3(320) - 960 = 0$
Apr	190	1150	130
May	310	1460	140
June	145	1605	315
July	110	1715	525
Aug	225	1940	620

The constant workforce plan with no stockouts, while appealing in theory, may encounter practical challenges that render it less feasible in real-world scenarios. Firstly, maintaining a large inventory level to buffer against fluctuations in demand may not be cost-effective, as it ties up capital and incurs holding costs.

Additionally, achieving a production level of 320 units per month with an integer number of workers may not be feasible due to workforce constraints and production capacity limitations. Furthermore, the assumption of a constant production level each month overlooks the variability in workdays across different months, which may necessitate adjustments in workforce levels to maintain consistent production output.

To address these shortcomings, modifications to the plan can be made. For instance, the number of workdays per month can be explicitly defined, allowing for more accurate workforce planning. Additionally, a K factor, representing the number of aggregate units produced by one worker in one day, can be computed or provided, enabling more precise calculations and better alignment between workforce levels and production targets.

Suppose that we are told that over 40 days, the plant had 38 workers who produced 520 units. It follows that:

$$\boxed{K = \frac{520}{38(40)} = 0.3421 \Rightarrow \text{ average number of units produced by one worker in one day}} \tag{5.1}$$

Also, assume we are given the following number of working days per month: 22, 16, 23, 20, 21, 22, 21, 22. Considering March as the critical month in the scenario described, a corner case arises when analyzing the cumulative net demand and the cumulative number of working days up to that point.

If the cumulative net demand through March is 960 units and the cumulative number of working days across January, February, and March is 61 days (22 days in January, 16 days in February, and 23 days in March), we can calculate the average daily demand as 960 units divided by 61 days, resulting in approximately 15.7377 units per day.

$$\boxed{\frac{960}{61} = 15.7377 \text{ units/day}} \tag{5.2}$$

To ensure that production meets this demand without stockouts, the number of required workers can be determined by dividing the average daily demand by the productivity factor, 0.3421 units per worker per day, yielding approximately 46 workers required for March.

$$\boxed{\frac{15.7377}{0.3421} = 46 \text{ workers required}} \tag{5.3}$$

Month	# Day	Prod Level	Cuml. Prod	Cuml. Demand	Ending Inventory
Jan	22	22(15.7377) = 346	346	220	126
Feb	16	252	598	500	98
Mar	23	362	960	960	0
Apr	20	315	1275	1150	125
May	21	330	1605	1460	145
June	22	346	1951	1605	346
July	21	330	2281	1715	566
Aug	22	346	2627	1940	687

If we assume that additional costs such as the \$8.50 holding cost per unit per month, \$800 hiring cost per worker, \$1,250 firing cost per worker, and \$75 payroll cost per worker per day are to be incurred by the company. Let's think about how would this affect our constant workforce plan.

Evaluation: Considering the additional costs incurred by the company, we assess the impact on our constant workforce plan. Beginning with the assumption of 40 workers at the end of December, the cost to hire 6 additional workers amounts to

$$6(800) = \ \$4,800$$

In terms of inventory cost, we accumulate the ending inventory across months, including an adjustment for 100 units netted out in August, resulting in a total inventory cost of

$$\text{Accumulate ending inventory: } (126 + 98 + 0 + \ldots + 687) = 2,093$$

$$\text{Add in 100 units netted out in Aug} = 2,193$$

$$\text{Hence inventory cost} = 2,193(8.5) = \$18,640.50$$

and the payroll cost is calculated as

$$(\$75/\text{worker/day})(46 \text{ workers})(167 \text{ days}) = \$576,150$$

Summing these costs, the total cost of the plan amounts to \$599,590.50.

5.2.1 *Modification of CWF Plan*

We may achieve some cost reduction in the constant workforce plan by modifying the labor usage. In the original cumulative net demand curve, consider making reductions in the workforce one or more times over the planning horizon to decrease inventory investment. Suppose we are interested in determining a production plan that doesn't use unnecessary workforce over the planning horizon and *we are not allowed to change our workforce in two consecutive months.* Now the question is the following: how do we determine the workforce requirements?

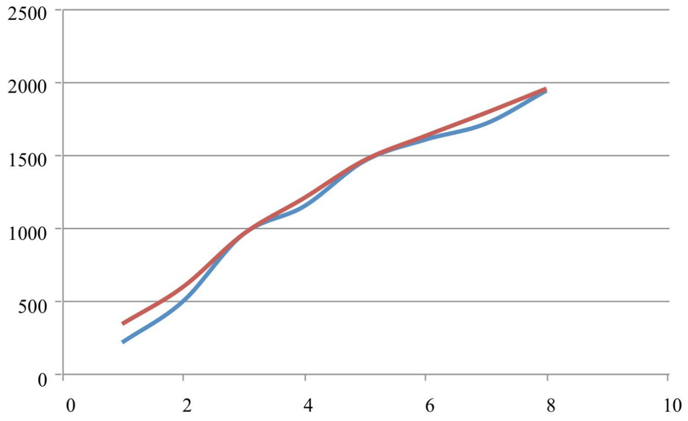


Fig. 5.6 Modification of CWF for a piecewise linear form.

We first determine the maximum cumulative workforce requirement in a way that the production rate allows us to complete a subset of periods without any shortages.

We compare 220/22 versus 500/38 versus 960/61, which are cumulative demand divided by a cumulative number of workdays. The rate keeps increasing until the fourth month which is 1150/81. So if we aim for month 4, stockouts are inevitable. Thus we stop at month 3, use the rate of $960/61 \approx 15.74$, and start things over as of the beginning of month 4. Using the production rate of 15.74, we can produce the following table for the first 3 months.
[1]

[1] **Note that** depending on how many days there are in each month, it is no longer guaranteed that March will be critical. This might change if the number of workdays varies a lot from month to month. It needs to be recalculated, but this is not the case in this question.

Month	# Day	Prod Level	Cuml. Prod	Cuml. Demand	Ending Inventory
Jan	22	346	346	220	126
Feb	16	252	598	500	98
Mar	23	362	960	960	0

Next, we continue in the same manner from month 4 forward.

Month	# Day	# Cum Days	Prod Per Day	Cuml. Demand
Apr	20	20	9.5	190
May	21	41	12.195122	500
June	22	63	10.2380952	645

We stop in month May, and we can now compute the inventory quantities for April and May.

Month	# Day	Prod Level	Cuml. Prod	Cuml. Demand	Old End. Inv.	Ending Inv.
Apr	20	246	246	190	125	56
May	21	259	505	500	145	5

Note that due to rounding there is a leftover inventory of 5, instead of zero. We will treat that as zero, but alternatively, this could have been deducted from the following cumulative demand computations. Finally, we check the remaining months starting with June.

Month	# Day	Prod Per Day	Cuml. Demand
June	22	6.590909	145
July	21	5.930233	255
Aug	22	7.384615	480

Notice that the rate initially decreases but we cannot stop there because we are not allowed to change our workforce in two consecutive months.

Month	# Day	Prod Level	Cuml. Prod	Cuml. Demand	Old End. Inv.	Ending Inv.
June	22	165	165	145	131	25
July	21	158	323	255	279	73
Aug	22	165	488	480	325	13

Final Plan:

Month	# Day	Prod Level	Cuml. Prod	Cuml. Demand	Ending Inventory
Jan	22	346	346	220	126
Feb	16	252	598	500	98
Mar	23	362	960	960	0
Apr	20	246	1206	1150	56
May	21	259	1465	1460	5
June	22	165	1630	1605	25
July	21	158	1788	1715	73
Aug	22	165	1953	1940	13

Cost of the modified plan: The modified plan calls for reducing the workforce to 36 at the start of April and making another reduction to 22 at the start of June. The cost to hire 6 workers amounts to

$$6(800) = \$4,800$$

In terms of inventory cost, we accumulate the ending inventory across months, including an adjustment for 100 units netted out in August, resulting in a total inventory cost of

$$\text{Inventory cost: Accumulate ending inventory: } (126 + 98 + 0 + \ldots + 13) = 396$$

$$\text{Add in 100 units netted out in Aug} = 496$$

$$\text{Hence inventory cost} = 496(8.5) = \$4,216$$

Regarding payroll cost, the calculation involves multiplying the daily wage ($75 per worker per day) by the number of workers and the total number of days worked across different workforce levels. This yields a total payroll cost of

$$(\$75/\text{worker/day})[(46 \text{ workers})(61 \text{ days})+(36 \text{ workers})(41 \text{ days})+(22 \text{ workers})(65 \text{ days})] = \$428,400$$

Additionally, the modified plan incurs an additional cost of $30,000 for layoffs, but this is offset by reduced holding costs of only $4,216. Also, the total payroll costs are reduced to $428,400.

As a result, the total cost of the modified plan is obtained by summing these costs

$$\$428,400 + \$4,216 + \$4,800 + \$30,000 = \$467,416$$

5.3 Zero Inventory Plan

The idea is to change the workforce each month to reduce ending inventory to nearly zero by matching the workforce with monthly demand as closely as possible. By adjusting the workforce levels based on the anticipated demand for each month, the company can optimize its production plan to meet customer needs efficiently while minimizing inventory holding costs.

Determining the production plan involves calculating the number of units produced by one worker each month, which is computed by multiplying the productivity factor K by the number of days per month. Next, the net demand for each month is divided by this quantity to obtain a ratio representing the required workforce level. This ratio is then rounded up to ensure sufficient capacity and possibly adjusted downward to account for factors such as efficiency improvements or seasonal fluctuations in demand. In addition to these, production level is obtained by multiplying the number of workers by monthly production per worker. By following this method, the company can develop a production plan that optimally utilizes its workforce while meeting customer demand effectively and minimizing inventory costs.

Month	# Day	Monthly Prod.	Net	# Workers	Cumulative	Cumulative	Ending
		Per Worker	Demand		Production	Demand	Inventory
Jan	22	7.5262	220	30	225	220	5
Feb	16	5.4736	275	51	505	500	5
Mar	23	7.8683	455	58	961	960	1
Apr	20	6.842	189	28	1152	1150	2
May	21	7.1841	308	43	1461	1460	1
June	22	7.5262	144	20	1612	1605	7
July	21	7.1841	103	15	1720	1715	5
Aug	22	7.5262	220	30	1945	1940	5

The number of hired and fired workers for each month is given below.

Month	# Hired	# Fired
Jan	–	10
Feb	21	–
Mar	7	–
Apr	–	30
May	15	–
June	–	23
July	–	5
Aug	15	–

Cost of the ZI Plan: In evaluating the production plan, the cost analysis reveals several key expenditures. The cost to hire $21 + 7 + 15 + 15 = 58$ workers in successive months totals

$$58(800) = \$46,400$$

While cost to fire $10 + 30 + 23 + 5 = 68$ workers amounts to

$$68(1250) = \$85,000$$

Calculating the inventory cost involves accumulating ending inventory across months, which totals 31 units, with an additional 100 units netted out in August, bringing the total to 131 units. Consequently, the inventory cost is computed as

$$\text{Inventory cost: Accumulate ending inventory: } 5 + 5 + 1 + \ldots + 5 = 31$$

$$\text{Add in 100 units netted out in Aug} = 131$$

$$\text{Hence inventory cost} = 131(8.5) = \$1,113.50$$

Payroll costs for the plan are calculated at $426,600. Combining these costs yields a total plan cost of

$$\$46,400 + \$85,000 + \$1,113.50 + \$426,600 = \$559,113.5$$

This analysis provides insights into the financial implications of workforce adjustments and inventory management, essential for optimizing production strategies and maintaining cost-effectiveness.

Implementing a zero inventory plan presents challenges that may render it impractical in certain contexts. Foremost, without maintaining inventory, the company risks being unable to fulfill future demand if production capacity is insufficient or inflexible to adapt to fluctuations in market requirements. Additionally, the frequent hiring and firing of workers necessary for zero inventory management could disrupt workforce stability, leading to discontent among employees and potentially damaging the company's reputation, making it harder to attract new talent. Moreover, existing labor agreements may impose constraints that hinder our ability to implement the necessary changes for a zero inventory plan to be effective. Considering these factors is crucial in assessing the feasibility and potential drawbacks of adopting a zero inventory approach in production planning.

5.4 The Optimization Framework

To determine the optimal quantity to produce, we need to balance revenue against the cost of holding inventory, taking into account both demand and production capacity constraints. Our objective is to either minimize costs or maximize profits while ensuring that production meets demand and does not exceed available capacity. By solving this problem, we can identify the most efficient production quantity that balances supply and demand while optimizing financial performance.

Notation:

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

D_t: forecasted demand in period t

Decision Variables:

W_t: workforce level in period t

Note that W_0 is the initial workforce at the start of the planning horizon.

P_t: production level in period t

I_t: inventory level in period t

Note that I_0 is given as the initial inventory on hand at the start of the planning horizon.

H_t: number of workers hired in period t

F_t: number of workers fired in period t

O_t: overtime production in units in period t

U_t: worker idle time in units in period t

S_t: number of units subcontracted from outside in period t

The objective is to minimize the total cost over the T periods, which includes costs associated with hiring, firing, holding inventory, regular production, overtime production, worker idle time, and subcontracting. The optimization problem can be formulated as follows:

$$\min \sum_{t=1}^{T} (c_H H_t + c_F F_t + c_I I_t + c_R P_t + c_O O_t + c_U U_t + c_S S_t)$$

subject to

$$\begin{aligned}
W_t &= W_{t-1} + H_t - F_t && \forall t \in [1,\ T] && \text{(conservation of workforce)}\\
P_t &= K n_t W_t + O_t - U_t && \forall t \in [1,\ T] && \text{(production and workforce)}\\
I_t &= I_{t-1} + P_t + S_t - D_t && \forall t \in [1,\ T] && \text{(inventory balance)}\\
& \text{all variables} \geq 0
\end{aligned}$$

Here is the graphical representation of the conservation of workforce and inventory balance.

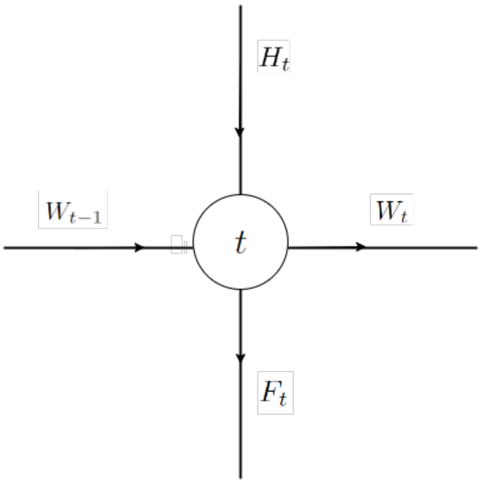


Fig. 5.7 Workforce balance

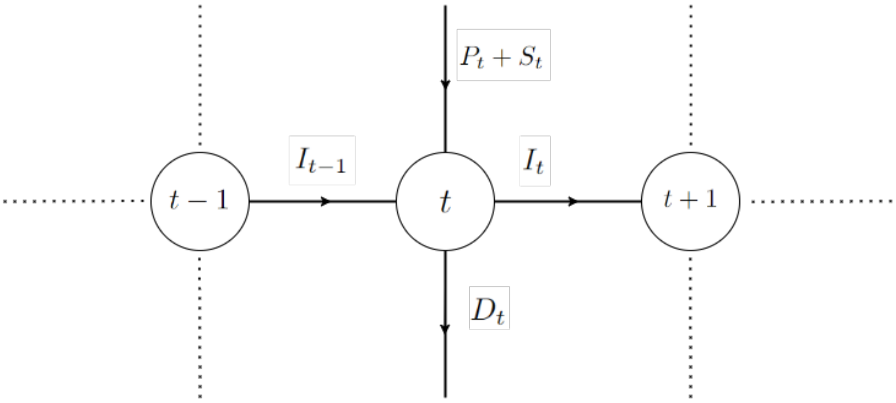


Fig. 5.8 Inventory balance

5.4.1 *A Simple APP Model*

Notation:

t: an index of the time periods, $t = 1, \ldots, \bar{t}$

d_t: demand in period t

c_t: capacity (number of items) in period t

r: unit profit (not including holding cost)

h: cost to hold one unit of inventory for one period

X_t: quantity produced during period t

S_t: quantity sold during period t

I_t: inventory at the end of period t

A company must manage its inventory over a series of periods indexed by $t = 1, 2, \ldots, T$. In each period t, the company faces demand d_t, has a production capacity c_t, and incurs a unit profit of r and a holding cost of h per unit of inventory per period.

The objective is to maximize the total profit over the T periods by determining the optimal production quantities while meeting demand and capacity constraints and minimizing holding costs.

The constraints are as follows: The quantity sold S_t in each period t must satisfy the demand d_t. The total production quantity X_t in each period t must not exceed the production capacity c_t. The optimization problem can be formulated as follows:

$$\max \sum_{t=1}^{\bar{t}} rS_t - hI_t \quad \text{(sales revenue - holding cost)}$$

where S_t is summed over planning horizon subject to

$$
\begin{aligned}
S_t &\leq d_t \quad t = 1,\ \ldots,\ \bar{t} \quad \text{(demand)}\\
X_t &\leq c_t \quad t = 1,\ \ldots,\ \bar{t} \quad \text{(capacity)}\\
I_t &= I_{t-1} + X_t - S_t \quad t = 1,\ \ldots,\ \bar{t} \quad \text{(inventory balance)}\\
X_t,\ S_t,\ I_t &\geq 0 \quad t = 1,\ \ldots,\ \bar{t} \quad \text{(non-negativity)}
\end{aligned}
$$

Example 5.3

A company operates over a planning horizon of six periods, indexed by $t = 1,\ 2,\ \ldots,\ 6$. The company produces a certain product and must decide on production quantities to meet demand while minimizing holding costs. The unit profit earned from selling the product is \$10, and the cost to hold one unit of inventory for one period is \$1. Initially, the company had no inventory. The production capacity varies over time, with the following values: $c_t = 100$ for periods $t = 1,\ 2,\ 3$ and 120 for periods $t = 4,\ 5,\ 6$. The demand for the product in each period is as follows: $d_t = 80,\ 100,\ 120,\ 120,\ 90,\ 140$. The company aims to determine the optimal production quantities for each period to meet demand while minimizing costs.

Solution

$$r = \$10$$

$$h = \$1$$

$$I_0 = 0$$

$$c_t = \begin{cases} 100, & t = 1,\ 2,\ 3 \\ 120, & t = 4,\ 5,\ 6 \end{cases}$$

$$d_t = 80,\ 100,\ 120,\ 120,\ 90,\ 140$$

The optimal solution suggests the following production quantities, sales quantities, and inventory levels for each period:

t	1	2	3	4	5	6
X_t	100	100	100	120	110	120
S_t	80	100	120	120	90	140
I_t	20	20	0	0	20	0

Question: What is the total profit for this scenario?

5.4.2 *Product Mix Planning*

The product mix planning problem aims to ascertain the optimal mix of products across a designated planning horizon. Its motivation lies in the integration of marketing and promotional strategies with logistical operations, as well as in identifying bottlenecks within the production process. Key inputs include demand forecasts for individual products or product families, often provided in the form of ranges to accommodate uncertainty, as well as data on the time required to produce one unit of each product. Additionally, capacity constraints, unit profit margins for each product, and holding costs are essential inputs for this problem.

Basic Pseudo-formulation

$$\begin{gathered}\text{maximize profit}\\ \text{subject to:}\\ \text{production} \leq \text{capacity, at all workstations in all periods}\\ \text{sales} \leq \text{demand, for all products in all periods}\end{gathered}$$

Note that we will need some technical constraints to ensure that variables represent reality.

Notation:

i: product index, $i = 1, \ \ldots, \ m$

j: workstation index, $j = 1, \ \ldots, \ n$

t: time period index, $t = 1, \ \ldots, \ \bar{t}$

$\overline{d}_{it}$: maximum demand for product i in period t

$\underline{d}_{it}$: minimum sales allowed of product i in period t

a_{ij}: time required on workstation j to produce one unit of product i

c_{jt}: time capacity of workstation j in period t

r_i: net profit from one unit of product i

h_i: cost to hold one unit of i for one period

Decision Variables:

X_{it}: amount of product i produced in period t

S_{it}: amount of product i sold in period t

I_{it}: inventory of product i at the end of t

The objective of the model is to maximize the overall sales revenue while considering holding costs. Constraints ensure that demand requirements are met, production capacities are not exceeded, and inventory balances are maintained, while non-negativity constraints enforce that production, sales, and inventory levels remain non-negative. The formulation is:

$$\max \sum_{t=1}^{\bar{t}} \sum_{i=1}^{m} r_i S_{it} - h_i I_{it} \qquad \text{(sales revenue - holding cost)}$$

subject to

$$\begin{aligned} \underline{d}_{it} \leq S_{it} \leq \overline{d}_{it} \quad &\forall i,t \quad \text{(demand)} \\ \sum_{i=1}^{m} a_{ij} X_{it} \leq c_{jt} \quad &\forall j,t \quad \text{(capacity)} \\ I_{it} = I_{it-1} + X_{it} - S_{it} \quad &\forall i,t \quad \text{(inventory balance)} \\ X_{it},\ S_{it},\ I_{it} \geq 0 \quad &\forall i,t \quad \text{(non-negativity)} \end{aligned}$$

5.4.2.1 *Extensions to Product Mix Model*

Extensions to the Product Mix Model involve incorporating additional constraints such as resource constraints, utilization matching, backorders, and overtime. They provide a more realistic representation of the production environment and can help optimize production planning and scheduling.

Notation:

b_{ij}: units of resource j required per unit of product i

k_{jt}: number of units of resource j available in period t

Decision Variables:

X_{it}: amount of product i produced in period t

Introduce constraints for other resources besides production capacity. For each resource j, ensure that the sum of resource usage across all products does not exceed the available capacity k_{jt}.

$$\sum_{i=1}^{m} b_{ij} X_{it} \leq k_{jt}$$

Implement utilization matching, where the utilization of each resource j is limited to a certain fraction q of its rated capacity. This constraint ensures that resources are utilized efficiently without exceeding their capacity limits.

$$\sum_{i=1}^{m} a_{ij} X_{it} \leq q c_{jt} \qquad \forall j,t$$

We can incorporate backorders by allowing the inventory level I_{it} to become unrestricted. Backordering costs can be penalized differently in the objective function. Thus, we substitute

$I_{it} = I_{it}^+ - I_{it}^-$ and penalize I_{it}^+, I_{it}^- differently in objective if desired where I_{it}^+ represents positive inventory (excess supply) while I_{it}^- represents negative inventory (excess demand). Overtime considerations involve defining O_{jt} as hours of overtime used on resource j in period t and incorporating it into the capacity constraint (add it to c_{jt} in capacity constraint) and objective function as needed.

5.4.3 *Workforce Planning*

The problem at hand involves determining the most profitable production and hiring/firing policy over a planning horizon. This study is motivated by the need to balance hiring/firing with overtime and inventory buildup, highlighting the tradeoffs involved. Additionally, the iterative nature of optimization modeling is recognized as a key aspect of the investigation. Inputs for this analysis include a demand forecast (assuming a single product for simplicity), unit hour data, labor content data, capacity constraints, hiring/firing costs, overtime costs, holding costs, and unit profit. These inputs will guide the decision-making process toward achieving optimal production and employment strategies while maximizing profitability.

Notation:

j: an index of workstation, $j = 1, \ \dots, \ n$

t: an index of period, $t = 1, \ \dots, \ \bar{t}$

$\overline{d}_t$: maximum demand in period t

$\underline{d}_t$: minimum sales allowed in period t

a_j: unit hours on workstation j

b: number of man hours required to produce one unit

c_{jt}: capacity of work center j in period t

r: net profit from one unit

h: cost to hold one unit for one period

l: cost of regular time in dollars/man-hour

l': cost of overtime in dollars/man-hour

e: cost to increase workforce by one man-hour

e': cost to decrease workforce by one man-hour

Decision Variables:

X_t: amount produced in period t

S_t: amount sold in period t

I_t: inventory at the end of t

W_t: workforce in man-hours of regular time in period t

H_t: increase (hires) in workforce from period $t-1$ to t in man-hours

F_t: decrease (fires) in workforce from period $t-1$ to t in man-hours

O_t: overtime in period t in hours

The objective function aims to maximize the total profit over all periods, considering sales revenue, holding costs, labor costs, overtime costs, and workforce adjustment costs. Constraints ensure that demand is met, production does not exceed workstation capacity, inventory balances, workforce is maintained, and overtime usage is within capacity limits. The model is:

$$\max \sum_{t=1}^{\bar{t}} rS_t - hI_t - lW_t - l'O_t - eH_t - e'F_t$$

subject to

$$\begin{aligned} \underline{d}_t \leq S_t \leq \overline{d}_t & \quad \forall t \\ a_j X_t \leq c_{jt} & \quad \forall t \\ I_t = I_{t-1} + X_t - S_t & \quad \forall t \\ W_t = W_{t-1} + H_t - F_t & \quad \forall t \\ bX_t \leq W_t + O_t & \quad \forall t \\ X_t,\ S_t,\ I_t,\ O_t,\ W_t,\ H_t,\ F_t \geq 0 & \quad \forall t \end{aligned}$$

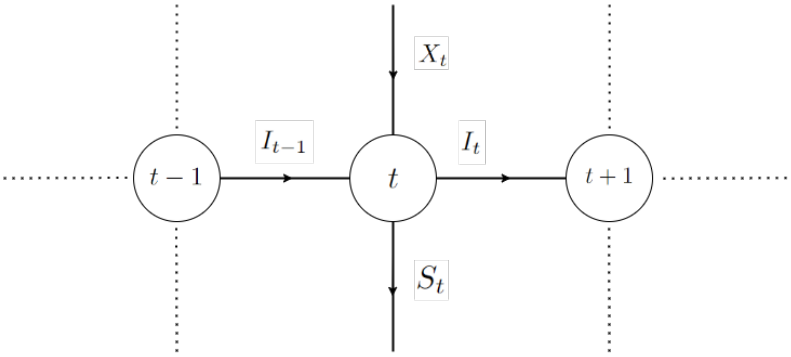


Fig. 5.9 Inventory balance

Conclusions:

In conclusion, it's important to recognize that there isn't a one-size-fits-all solution when it comes to Aggregate Planning (AP) models. Instead, the choice of model should be tailored to fit the specific characteristics and requirements of each unique situation. Embracing simplicity in model design promotes better understanding and accessibility, allowing for easier implementation and interpretation by stakeholders. Linear programming stands out as a valuable tool in AP, offering structured optimization techniques to address complex planning challenges. However, it's crucial to prioritize robustness over precision, as real-world scenarios often involve uncertainties and variations that cannot be perfectly captured in models. Lastly, it's essential to view formulation and solution as interconnected activities, emphasizing the iterative nature of the planning process and the need for continuous refinement and adaptation to changing conditions.

Example 5.4

A TV company produces Smart TVs. The company wants to plan production and workforce levels for the next 6 months. The table below shows the number of workdays and Smart TV demand (in thousands) for each month. Note that if the demand is exceeded, the leftovers can be carried to the next month. If the demand is not satisfied, the items are backordered. The annual inventory holding cost per Smart TV is \$120. If the company cannot satisfy a demand on time, a backorder cost of \$30 is incurred. Additionally, past data shows that 25 workers can produce 60 thousand Smart TVs in 20 days. The company currently has 35 workers and 3000 Smart TVs in stock. The company incurs a hiring cost of \$800 and a firing cost of \$1200 per worker. The company cannot hire or fire more than 10% of its workforce in any month (i.e., if there are 30 workers at the end of a month, they can hire or fire a maximum of 3 people starting next month). The payroll is \$60 per worker per day. Suppose that the company can make an initial adjustment to the number of workers, but then they cannot change the number of workers for the first 3 months. The company also aims to have at least 25 workers and 10 Smart TVs in stock at the end of the 6 month planning horizon. Write a Linear Programming (LP) model that identifies the workforce and production plan, aiming to minimize the total cost.

Month	Number of workdays	Demand (in thousands)
1	20	80
2	22	100
3	21	90
4	20	85
5	22	105
6	21	95

Solution

Table 1: Parameters

D_t	Demand at month t
d_t	Number of workdays in month t
K	average number of units produced by one worker in one day

Table 2: Decision Variables

I_t	Smart TVs in stock at the end of month t
B_t	Smart TVs backordered in month t
H_t	Number of workers hired in month t
F_t	Number of workers fired in month t
W_t	Workforce at the end of month t

$$K = \frac{60,000}{25(20)} = 120 \text{ units}$$

Note that the inventory holding cost per Smart TV is given to us annually. When determining the inventory cost, it's necessary to convert it monthly, as our calculations will be based on the inventory at the end of each month ($120/yr=$10/mo).

Model:

$$\min \sum_{t=1}^{6} 10I_t + 800H_t + 1200F_t + 30B_t + 60d_tW_t$$

s.t.

$$I_0 = 3000$$

$$B_0 = 0$$

$$I_{t-1} - B_{t-1} + 120d_tW_t = I_t - B_t + D_t \qquad \forall t = 1,\ 2,\ \ldots,\ 6$$

The above constraint originally is

$$N_{t-1} + 120d_tW_t = N_t + D_t \qquad \forall t = 1,\ 2,\ \ldots,\ 6$$

where N_t denotes the net inventory at the end of period t. Next, we substitute N_t with $I_t - B_t$. You will see that with the given objective either one of the inventory or backorder quantity will be zero.

$$W_{t-1} + H_t - F_t = W_t \quad \forall t = 1,\ 2,\ \ldots,\ 6$$

$$H_t + F_t \leq 0.1 W_{t-1} \quad \forall t = 1,\ 2,\ \ldots,\ 6$$

$$W_0 = 35$$

$$H_2 = F_2 = H_3 = F_3 = 0$$

The company cannot change the number of workers they have for the first 3 months. They can hire/fire in the beginning of month 1, but not in 2 or 3.

$$W_6 \geq 25$$

$$I_6 \geq 10$$

$$I_t,\ W_t,\ B_t,\ H_t,\ F_t \geq 0 \quad \forall t = 1,\ 2,\ \ldots,\ 6$$

Important 5.1

Month	Number of workdays	Demand
1	20	90
2	18	75
3	20	130
4	23	120
5	22	80

A TV company produces Smart TVs. The company wants to plan production and workforce levels for the next 5 months. The above table shows the number of workdays and Smart TV demand (in thousands) for each month.

The inventory holding cost per Smart TV per month is calculated as \$15. If the company cannot satisfy a demand on time, a backorder cost of \$25 is incurred. Additionally, past data shows that 25 workers can produce 60 thousand Smart TVs in 20 days. The company currently has 35 workers. The company incurs a hiring cost of \$750 and a firing cost of \$1300 per worker. The payroll is \$50 per worker per day.

a. What is the number of workers needed if the company wants to apply a constant workforce plan where stock-outs are not allowed?

b. What is the total inventory holding cost for the constant workforce plan with the number of workers found in part a?

c. Suppose that the company cannot change the number of workers that they have for the first 3 months. In month 4 or 5, they can change the number of employees so that they can satisfy the total demand. Propose a production and workforce plan under these circumstances. Calculate the cost of this plan.

d. Write an LP that solves this problem.

Chapter 6

Transportation Problem in Supply Chain Management

In the vast landscape of supply chain management, an array of interconnected processes and strategies come into play to ensure the seamless flow of goods and services from suppliers to end consumers. While the scope of supply chain management encompasses procurement, production, inventory management, logistics, and customer service, this chapter will specifically focus on a critical component: transportation.

The transportation problem is a mathematical model for optimally scheduling the flow of goods from production facilities to distribution centers. Assume that a fixed amount of product must be transported from a group of sources (plants) to a group of sinks (warehouses). The unit cost of transporting from each source to each sink is assumed to be known. The goal is to find the optimal flow paths and the amounts to be shipped on those paths to minimize the total cost of all shipments.

The transportation problem can be viewed as a prototype supply chain problem. Although most real-world problems involving the shipment of goods are more complex, the model illustrates the issues and methods one would encounter in practice.

The Greedy Heuristic

The transportation problem can be formulated as a linear program and thus can be solved using any linear programming code, such as Solver in Excel. To gain some intuition about the structure of the problem, however, we will consider a simple heuristic that usually gives good, but possibly suboptimal, solutions. To implement the heuristic, we construct a transportation tableau.

Solving Transportation Problems with Linear Programming

Several heuristics for solving transportation problems have been proposed, such as greedy heuristics. However, it is unlikely that anyone with a real problem would use a heuristic since optimal solutions can be found efficiently by linear programming. In fact, because of the special structure of the transportation problem, today's specialized codes can solve problems with millions of variables. Let us introduce the decision variable x_{ij}.

x_{ij}: flow from source i to sink j for $1 \leq i \leq m$ and $1 \leq j \leq n$.

And define c_{ij} as the cost of shipping one unit from i to j. It follows that the total cost of making all shipments is

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}$$

Since many routes are not economical, many of the decision variables will likely equal zero at the optimal solution.

The constraints are designed to ensure that the total amount shipped out of each source equals the amount available at that source, and the amount shipped into any sink equals the amount required at that sink. Since there are m sources and n sinks, there are a total of $m + n$ functional constraints (excluding nonnegativity constraints). Let a_i be the total amount to be shipped out of source i and b_j the total amount to be shipped into sink j. The linear programming constraints may be written:

$$\begin{aligned} \sum_{j=1}^{n} x_{ij} = a_i \quad &\text{for} \quad 1 \leq i \leq m \\ \sum_{i=1}^{m} x_{ij} = b_j \quad &\text{for} \quad 1 \leq j \leq n \\ x_{ij} \geq 0 \quad \text{for} \quad 1 \leq i \leq m \quad &\text{and} \quad 1 \leq j \leq n \end{aligned}$$

Example 6.1

Bose is a manufacturer of home theater and sound systems for households. Bose has four manufacturing plants located in Atlanta, Detroit, Memphis, and Portland. Periodically, shipments are made from these four plants to three warehouses located in New York, Chicago, and Los Angeles. The production capacity at the factories is 150, 250, 180, and 110 respectively (in thousands). The forecasted demand for the warehouses is 300, 120, and 270 (in thousands). The shipping costs of a unit from the factories to the warehouses are given in the table below.

Shipping costs	NY	CHI	LA
ATL	180	150	260
DET	200	160	220
MEM	300	510	420
POR	250	440	380

a. What is the solution obtained using the greedy heuristic assuming that ATL/CHI and MEM/NY routes are eliminated?

b. Write an LP that solves this problem assuming that there is a transshipment point at Denver, which has no demand of its own. Assume that the unit costs of shipping from Atlanta, Detroit, Memphis, and Portland to Denver are 90, 110, 220, and 200 respectively. Also, assume that the unit costs of shipping from Denver to New York, Chicago, and Los Angeles are 200, 75, and 100 respectively.

c. What is the solution obtained using the greedy heuristic for the version with the transshipment point?

Solution

a)

	NY	CHI	LA	Supply
ATL	180	M	260	150
DET	200	160	220	250
MEM	M	510	420	180
POR	250	440	380	110
Demand	300	120	270	690

Rows correspond to supply sources (plants) and columns to sinks (warehouses). The numbers we will place in the cells of the tableau will be the value of the flow of product from each source to each sink. To implement the greedy heuristic, we search the tableau for the minimum unit cost, which is \$160, and corresponds to the DET/CHI cell. Into this cell, we place the minimum of the following two quantities: the availability at DET (250) and the requirement at CHI (120). Hence, we place a 120 into this cell. At this point, we've saturated the second column (meaning no more can be shipped to CHI). Since 120 is already allocated, the supply capacity from DET will drop to $250 - 120 = 130$. The tableau now becomes

	NY	CHI	LA	Supply
ATL	180		260	150
DET	200	120	220	130
MEM	M		420	180
POR	250		380	110
Demand	300	0	270	570

Of the uncovered cells, the least-cost cell is ATL/NY at \$180 [i.e., cell (1,1)]. In this cell, we can assign a maximum flow of 150 units (the minimum of 150 and 300), which we do. This saturates the first row. The demand capacity for NY will drop to $300 - 150 = 150$. The tableau becomes

	NY	CHI	LA	Supply
ATL	150			0
DET	200	120	220	130
MEM	M		420	180
POR	250		380	110
Demand	150	0	270	420

The next least-cost cell is DET/NY at \$200. The capacity of this cell is the minimum of 130 and 150 which is 130. Hence, we allocate 130 to this cell, which saturates the second row. The demand capacity for NY will drop to $150 - 130 = 20$. The tableau at this stage is

	NY	CHI	LA	Supply
ATL	150			0
DET	130	120		0
MEM	M		420	180
POR	250		380	110
Demand	20	0	270	290

The cell with the minimum cost in the remaining cells is POR/NY with $250. Into this cell, we place a minimum of 20 and 110. Hence, we allocate 20 to this cell, which saturates the first column. The supply capacity from POR will drop to $110 - 20 = 90$. The tableau transforms

	NY	CHI	LA	Supply
ATL	150			0
DET	130	120		0
MEM			420	180
POR	20		380	90
Demand	0	0	270	270

Finally, the least-cost cell is POR/LA at $380. In this cell, we can assign a maximum flow of 90 units (the minimum of 90 and 270). This saturates the fourth row. The remaining capacity from LA is 270 − 90 = 180. The tableau now looks like this

	NY	CHI	LA	Supply
ATL	150			0
DET	130	120		0
MEM			420	180
POR	20		90	0
Demand	0	0	180	180

180 units are sent to the last remaining cell, MEM/LA, for 420 dollars. The final solution is

	NY	CHI	LA	Supply
ATL	150			0
DET	130	120		0
MEM			180	0
POR	20		90	0
Demand	0	0	0	0

If we let x_{ij} be the amount of flow from source i to sink j, then the solution shown in the last tableau is

$$x_{11} = 150,\ x_{21} = 130,\ x_{22} = 120,\ x_{33} = 180,\ x_{41} = 20,\ \text{and}\ \ x_{43} = 90$$

and all other $x_{ij} = 0$. The total cost of this solution is $150(\$180) + 130(\$200) + 120(\$160) + \ldots + 90(\$380) = \$187,000$.

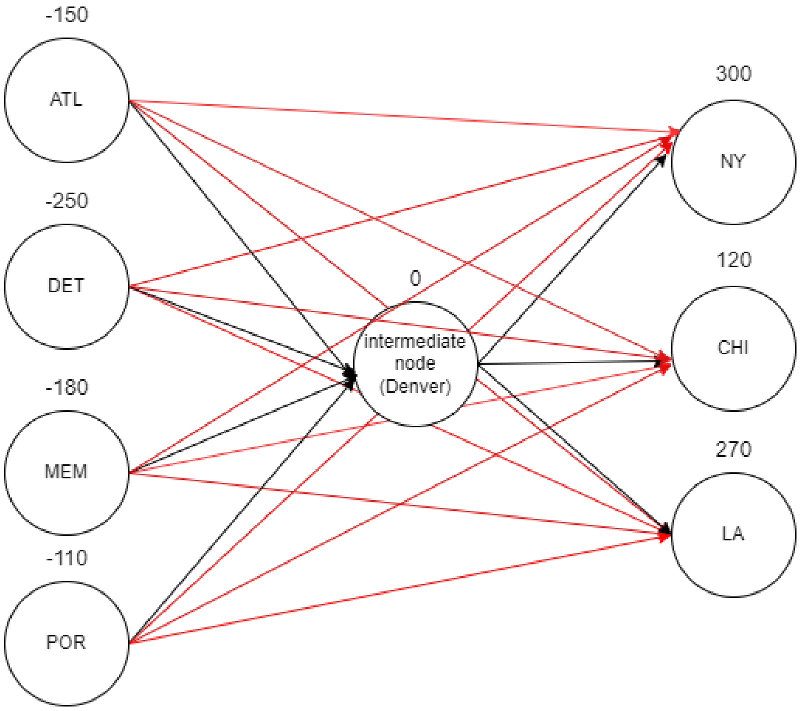
b)

Table 1: Parameters

c_{ij}	the cost of shipping one unit from i to j

Table 2: Decision Variables

x_{ij}	amount of flow from city $i = 1,\ 2,\ 3,\ 4$ (ATL, DET, MEM, POR) to city $j = 1,\ 2,\ 3$ (NY, CHI, LA)
z_i	amount of flow from city $i = 1,\ 2,\ 3,\ 4$ to Denver
y_j	amount of flow from Denver to city $j = 1,\ 2,\ 3$



Model:

$$\min \sum_{i=1}^{4} \sum_{j=1}^{3} c_{ij} x_{ij}$$

subject to

$$x_{11} + x_{12} + x_{13} + z_1 = 150 \quad \text{(shipments out of ATL)}$$

$$x_{21} + x_{22} + x_{23} + z_2 = 250 \quad \text{(shipments out of DET)}$$

$$x_{31} + x_{32} + x_{33} + z_3 = 180 \quad \text{(shipments out of MEM)}$$

$$x_{41} + x_{42} + x_{43} + z_4 = 110 \quad \text{(shipments out of POR)}$$

$$x_{11} + x_{21} + x_{31} + x_{41} + y_1 = 300 \quad \text{(shipments into NY)}$$

$$x_{12} + x_{22} + x_{32} + x_{42} + y_2 = 120 \quad \text{(shipments into CHI)}$$

$$x_{13} + x_{23} + x_{33} + x_{43} + y_3 = 270 \quad \text{(shipments into LA)}$$

$$\sum_{i=1}^{4} z_i = \sum_{j=1}^{3} y_j$$

$$x_{ij} \geq 0 \ \text{ for } \ 1 \leq i \leq 4 \ \text{ and } \ 1 \leq j \leq 3 \quad \text{(non-negativity)}$$

$$z_i \geq 0 \ \text{ for } \ 1 \leq i \leq 4 \ \text{ and } \ y_j \geq 0 \ \text{ for } \ 1 \leq j \leq 3 \quad \text{(non-negativity)}$$

c)

You have two options: you can send units either directly or through the via point. You should send items through one of those cheaper compared to the other alternative. You will always choose the cheaper route between any supply and demand node. For example, we will send units from ATL to NY directly because the via point costs us \$290 (ATL to Denver costs \$90, Denver to NY costs \$200) while direct is \$180. We will choose via point option when we send items from ATL to LA (\$260 vs \$190), DET to LA (\$220 vs \$210), MEM to CHI (\$510 vs \$295), MEM to LA (\$420 vs \$320), POR to CHI (\$440 vs \$275) and POR to LA (\$380 vs \$300).

	NY	CHI	LA	Supply
ATL	180	150	190	150
DET	200	160	210	250
MEM	300	295	320	180
POR	250	275	300	110
Demand	300	120	270	690

We search the tableau for the minimum unit cost, which is $150, and corresponds to the ATL/CHI cell. Into this cell, we place a minimum of 120 and 150. Hence, we place a 120 into this cell. At this point, we've saturated the second column. Since 120 is already allocated, the supply capacity from ATL will drop to 150 − 120 = 30. The tableau now becomes

	NY	CHI	LA	Supply
ATL	180	120	190	30
DET	200		210	250
MEM	300		320	180
POR	250		300	110
Demand	300	0	270	570

Of the uncovered cells, the least-cost cell is ATL/NY at $180. In this cell, we can assign a maximum flow of 30 units (the minimum of 30 and 300), which we do. This saturates the first row. The demand capacity for NY will drop to 300 − 30 = 270. The tableau becomes

	NY	CHI	LA	Supply
ATL	30	120		0
DET	200		210	250
MEM	300		320	180
POR	250		300	110
Demand	270	0	270	540

The next least-cost cell is DET/NY at $200. The capacity of this cell is the minimum of 270 and 250 which is 250. Hence, we allocate 250 to this cell, which saturates the second row. The demand capacity for NY will drop to 270 − 250 = 20. The tableau at this stage is

	NY	CHI	LA	Supply
ATL	30	120		0
DET	250			0
MEM	300		320	180
POR	250		300	110
Demand	20	0	270	290

The greedy heuristic continues in this manner until all rows and columns are saturated. The final solution is

	NY	CHI	LA	Supply
ATL	30	120		0
DET	250			0
MEM			180	0
POR	20		90	0
Demand	0	0	0	0

The total cost of this solution is $30(\$180)+120(\$150)+250(\$200)+\ldots+90(\$300) = \$163,000$.

Supply Chain Management Issues

Supply chain management involves grappling with a variety of complex issues that impact the flow of goods and information across the network. One critical concern is the role of information, particularly in mitigating the Bullwhip effect, where small fluctuations in demand at the consumer level can amplify as they move up the supply chain, leading to inefficiencies and excess inventory. Another significant challenge is the transportation problem, which involves optimizing the movement of goods from suppliers to manufacturers to distributors and ultimately to customers, considering factors like cost, time, and capacity constraints. Additionally, determining the most efficient delivery routes is essential for ensuring timely and cost-effective distribution while balancing factors such as distance, traffic conditions, and service level requirements. These issues underscore the complexity of managing supply chains and the importance of strategic planning and coordination to achieve optimal performance.

Important 6.1

Simbo is a manufacturer of home theater and sound systems for households. Simbo has three manufacturing plants located in Kocaeli, Giresun, and Gaziantep. The production capacity at the manufacturing plants is 750, 250, 200, respectively (in thousands). Periodically, shipments are made from these three plants to four warehouses located in Istanbul, Antalya, Izmir, and Mardin. The forecasted demand for the warehouses are 800, 150, 170, and 80, respectively (in thousands).

The shipping costs of a unit from the factories to the warehouses are given in the table below.

Shipping Cost	Istanbul	Antalya	Izmir	Mardin
Kocaeli	50	300	220	600
Giresun	450	500	590	200
Gaziantep	350	440	460	80

- What is the solution obtained using the greedy heuristic?
- Assume that there is a transshipment point in Ankara, which has no demand or supply of its own. Assume that the unit costs of shipping from Kocaeli, Giresun, and Gaziantep to Ankara are 200, 250, and 50, respectively. Also assume that the unit costs of shipping from Ankara to Antalya and Mardin are 75 and 100, respectively. Suppose we cannot ship from Ankara to Istanbul and Izmir. What is the solution obtained using the greedy heuristic for the version with the transshipment point?

Food for thought How would you model this? How would you update the model for the following condition: "Giresun cannot ship anything to Ankara if Gaziantep sends items to Mardin."

Chapter 7

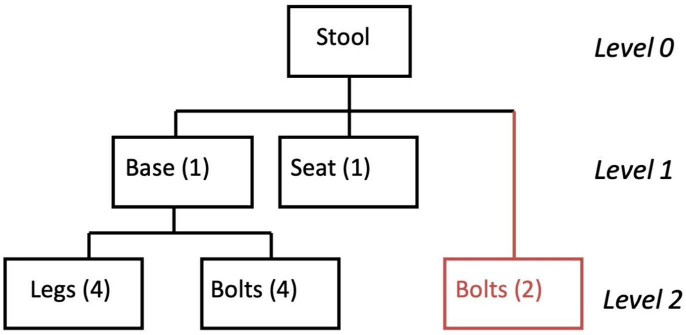
MRP, JIT, and Lot Sizing

This chapter is dedicated to exploring Material Requirements Planning (MRP), a fundamental concept in production and inventory management. MRP serves as a pivotal tool for businesses aiming to optimize their production processes by effectively planning and controlling the flow of materials required for manufacturing. We delve into the principles, methodologies, and practical applications of MRP. This includes a systematic analysis of production requirements, inventory levels, and lead times to ensure the timely availability of materials while minimizing excess inventory costs.

Example 7.1

Relax is a furniture manufacturer that produces several kinds of furniture including stools. Each stool requires one base, one seat, and two bolts to assemble. Each base requires four legs and four bolts to assemble. It takes one week to produce a stool from the base and seats and one week to produce a base from the legs. Legs have 2 week lead time. Relax currently has 20 stools in the finished goods inventory. Also, there is an existing order for legs of size 200, which will arrive in the second week. If customers place an order of 120, which is to be satisfied five weeks from now, determine the demand for base and legs for each week. How should the company plan the production/ordering process?

Graphical Bill of Materials



Bolts are treated at the lowest level in which they occur for MRP calculations. They might be left off BOM altogether in practice.

Netting

The netting table provides a structured overview of the inventory management process for two items: stools and bases. Each table contains columns representing different weeks, from week 0 to week 6, and rows representing various aspects of inventory management.

Let's explain what the rows in the table mean to make it more understandable. Gross requirements represent the total demand for the item each week. This includes both existing demand and any new orders or requirements. Scheduled receipts indicate any incoming inventory or orders that are scheduled to be received each week. This includes both existing inventory and any planned orders that are expected to arrive.

Project inventory reflects the projected inventory levels for each week. It combines the existing inventory from previous weeks, scheduled receipts, and any adjustments made based on net requirements. Net requirements represent the actual inventory needs after accounting for existing inventory and scheduled receipts. It is calculated by subtracting project inventory from gross requirements. Planned orders indicate any new orders that need to be placed to meet the net requirements. These orders are planned based on the difference between gross requirements and project inventory.

In summary, the netting table helps in tracking inventory levels, determining actual inventory needs, and planning orders to ensure that sufficient inventory is available to meet demand while minimizing excess inventory.

Item: Stool (Leadtime = 1 week)

Week	0	1	2	3	4	5	6
Gross Requirements						120	
Scheduled Receipts							
Project Inventory	20	20	20	20	20	-100	-100
Net Requirements						100	
Planned Orders					100		

To produce 1 stool, 1 base is needed, so they have a one-to-one relationship. If our requirement is 120 (*remember that the company currently has 20 stools in the finished goods inventory*), we specify our requirement as 100 and place our order 1 week in advance because the lead time is 1 week and it should arrive when we want.

Item: Base (Leadtime = 1 week)

Week	0	1	2	3	4	5	6
Gross Requirements					100		
Scheduled Receipts							
Project Inventory	0	0	0	0	-100	-100	-100
Net Requirements					100		
Planned Orders				100			

We create the base from the legs, the legs are one layer down, and the lead time is 2 weeks. We need 4 legs to produce 1 base so our demand is 400 legs in the third week to meet our requirement (*remember also that an existing order for legs of size 200 will arrive in the second week*).

Item: Legs (Leadtime = 2 weeks)

Week	0	1	2	3	4	5	6
Gross Requirements				400			
Scheduled Receipts			200				
Project Inventory	0	0	200	-200	-200	-200	-200
Net Requirements				200			
Planned Orders		200					

In the realm of manufacturing planning and control, the synergy between Bill of Materials (BOM) explosion and Material Requirements Planning (MRP) tables is paramount in optimizing lot sizing strategies. BOM explosion, the process of breaking down finished products into their constituent parts, lays the groundwork by delineating the intricate relationships between components and assemblies. Concurrently, MRP tables meticulously analyze inventory levels, demand forecasts, and lead times to determine the replenishment needs of each item. When integrated seamlessly, these systems provide invaluable insights into the dynamic interplay of supply and demand. Lot sizing, a pivotal aspect of production planning, is intricately linked to this harmonious blend. By leveraging the granular details unearthed through BOM explosion and MRP tables, manufacturers can judiciously tailor lot sizes to minimize costs, optimize resources, and synchronize production with market demands. Thus, the fusion of BOM explosion and MRP tables serves as the bedrock for informed decision-making in lot sizing, fostering efficiency and agility within manufacturing operations. Thus, next, we discuss lot sizing problems with different solution approaches.

7.1 Lot Sizing Schemes

The problem of finding the best (or near best) production plan can be characterized as follows: we have a known set of time-varying demands and costs of setup and holding. What production quantities will minimize the total holding and setup costs over the planning horizon?

In this section, we discuss several popular heuristics (i.e., approximate) lot-sizing methods that easily can be incorporated into the MRP calculus as well as one exact approach. The methods will be demonstrated through the example below.

Example 7.2

For the next 10 weeks, Relax legs assembly department has the following demand for legs. Suppose that the fixed cost of production/ordering is \$100 and the holding cost of an item per week is \$1. What is the optimal production/ordering policy for the legs assembly department that minimizes the total fixed cost plus the total inventory holding costs?

t	1	2	3	4	5	6	7	8	9	10
D_t	20	50	10	50	50	10	20	40	20	30

7.1.1 *Lot for Lot Method*

The simplest lot sizing scheme for MRP systems is lot-for-lot, which leads to **zero inventory**. That is, the number of units scheduled for production each period was the same as the net requirements for that period. This policy is assumed for convenience and ease of use only. It is, in general, not optimal.

t	1	2	3	4	5	6	7	8	9	10	Total
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	20	50	10	50	50	10	20	40	20	30	300
Setup	100	100	100	100	100	100	100	100	100	100	\$1000
Holding											\$0
Total											\$1000

7.1.2 *EOQ Lot Sizing Method*

To apply the EOQ formula, we need three inputs: the average demand rate, λ; the holding cost rate, h; and the setup cost, K. Using the average demand as the demand rate, EOQ can be calculated, which is to be used as a fixed order size. This method tries to balance the total fixed and total inventory holding costs. It will not work well if the demand is highly variable.

$$Q = \sqrt{\frac{2K\overline{\lambda}}{h}} = \sqrt{\frac{2(100)(30)}{1}} = 77$$

t	1	2	3	4	5	6	7	8	9	10	Total
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	77		77		77			69			300
Setup	100		100		100			100			\$400
Holding		57	7	74	24	51	41	21	50	30	\$355
Total											\$755

Note that the last order is artificially decreased to 69 from 77 to avoid excessive inventory within this planning horizon.

Also note that we don't use annual demand, but weekly instead. This is not an issue because the holding is weekly as well.

7.1.3 *Fixed Order Period Method*

Another simple heuristic for lot sizing is the fixed order period method. It involves determining a fixed order period and then calculating order quantities using that fixed order interval.

If the fixed order period is three weeks, then the tableau is

t	1	2	3	4	5	6	7	8	9	10	Total
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	80			110			80			30	300
Setup	100			100			100			100	\$400
Holding		60	10	0	60	10	0	60	20	0	\$220
Total											\$620

7.1.4 *Silver Meal Method*

The Silver Meal heuristic is a forward method that requires determining the average cost per period as a function of the number of periods the current order spans, and stopping the computation when this function first increases.

Define $C(T)$ as the average holding and setup cost per period if the current order spans the next T periods. Let $(r_1, \ldots, r_n)$ be the requirements over the n-period horizon. Consider period 1. If we produce just enough in period 1 to meet the demand in period 1, then we just incur the order cost of K. Hence,

$$C(1) = K$$

If we order enough in period 1 to satisfy the demand in both periods 1 and 2, then we must hold r_2 for one period. Hence,

$$C(2) = (K + hr_2)/2$$

Similarly,

$$C(3) = (K + hr_2 + 2hr_3)/3$$

In general,

$$C(j) = (K + hr_2 + 2hr_3 + \ldots + (j - 1)hr_j)/j$$

Once $C(j) > C(j-1)$, we stop and set $y_1 = r_1 + r_2 + \ldots + r_{j-1}$, and begin the process again starting at period j.

In our example,

$$C(1) = 100,$$

$$C(2) = \frac{100 + 1(50)}{2} = 75,$$

$$C(3) = \frac{100 + 1(50) + 2(10)}{3} = 56.66,$$

$$C(4) = \frac{100 + 1(50) + 2(10) + 3(50)}{4} = 80$$

Stop because $C(4) > C(3)$. Set $y_1 = r_1 + r_2 + r_3 = 20 + 50 + 10 = 80$.

Starting in period 4:

$$C(1) = 100,$$

$$C(2) = \frac{100 + 1(50)}{2} = 75,$$

$$C(3) = \frac{100 + 1(50) + 2(10)}{3} = 56.66,$$

$$C(4) = \frac{100 + 1(50) + 2(10) + 3(20)}{4} = 57.5$$

Stop because $C(4) > C(3)$. Set $y_4 = r_4 + r_5 + r_6 = 50 + 50 + 10 = 110$.

and finally starting in period 7:

$$C(1) = 100,$$

$$C(2) = \frac{100 + 1(40)}{2} = 70,$$

$$C(3) = \frac{100 + 1(40) + 2(20)}{3} = 60,$$

$$C(4) = \frac{100 + 1(40) + 2(20) + 3(30)}{4} = 67.5$$

Stop because $C(4) > C(3)$. Set $y_7 = r_7 + r_8 + r_9 = 20 + 40 + 20 = 80$.

t	1	2	3	4	5	6	7	8	9	10	Total
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	80			110			80			30	300
Setup	100			100			100			100	\$400
Holding		60	10	0	60	10	0	60	20	0	\$220
Total											\$620

7.1.5 *Least Unit Cost Method*

The least unit cost heuristic is similar to the Silver Meal method except that instead of dividing the cost over j periods by the number of periods, j, we divide it by the total number of units demanded through period j, $r_1 + r_2 + \ldots + r_j$. We choose the order horizon that minimizes the cost per unit of demand rather than the cost per period.

Define $C(T)$ as the average holding and setup cost per unit for a T period order horizon. Then,

$$C(1) = K/r_1$$

$$C(2) = (K + hr_2)/(r_1 + r_2)$$

In general,

$$C(j) = [K + hr_2 + 2hr_3 + \ldots + (j-1)hr_j]/(r_1 + r_2 + \ldots + r_j)$$

As with the Silver Meal heuristic, this computation is stopped when $C(j) > C(j-1)$, and the production level is set equal to $r_1 + r_2 + \ldots + r_{j-1}$. The process is then repeated, starting at period j and continuing until the end of the planning horizon is reached.

Starting in period 1:

$$C(1) = \frac{100}{20} = 5,$$

$$C(2) = \frac{100 + 1(50)}{20 + 50} = 2.14,$$

$$C(3) = \frac{100 + 1(50) + 2(10)}{20 + 50 + 10} = 2.13,$$

$$C(4) = \frac{100 + 1(50) + 2(10) + 3(50)}{20 + 50 + 10 + 50} = 2.46$$

Because $C(4) > C(3)$, we stop and set $y_1 = r_1 + r_2 + r_3 = 20 + 50 + 10 = 80$.

Starting in period 4:

$$C(1) = \frac{100}{50} = 2,$$

$$C(2) = \frac{100 + 1(50)}{50 + 50} = 1.5,$$

$$C(3) = \frac{100 + 1(50) + 2(10)}{50 + 50 + 10} = 1.55$$

Because $C(3) > C(2)$, we stop and set $y_4 = r_4 + r_5 = 50 + 50 = 100$.

Starting in period 6:

$$C(1) = \frac{100}{10} = 10,$$

$$C(2) = \frac{100 + 1(20)}{10 + 20} = 4,$$

$$C(3) = \frac{100 + 1(20) + 2(40)}{10 + 20 + 40} = 2.86,$$

$$C(4) = \frac{100 + 1(20) + 2(40) + 3(20)}{10 + 20 + 40 + 20} = 2.89$$

Because $C(4) > C(3)$, we stop and set $y_6 = r_6 + r_7 + r_8 = 10 + 20 + 40 = 70$.

Finally, starting in period 9:

$$C(1) = \frac{100}{20} = 5,$$

$$C(2) = \frac{100 + 1(30)}{20 + 30} = 2.6,$$

As we have reached the end of the horizon, we set $y_9 = r_9 + r_{10} = 20 + 30 = 50$.

t	1	2	3	4	5	6	7	8	9	10	Total
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	80			100		70			50		300
Setup	100			100		100			100		$400
Holding	0	60	10	0	50	0	60	40	0	30	$250
Total											$650

It is interesting to note that the policy obtained by this method is different from that for the Silver Meal heuristic. It turns out that the Silver Meal method gives the optimal policy, with a cost \$620, whereas the LUC gives a suboptimal policy, with a cost \$650.

7.1.6 *Part Period Balancing Method*

Another approximate method for solving this problem is part period balancing. Although the Silver Meal technique seems to give better results in a greater number of cases, part period balancing seems to be more popular in practice.

The method is to set the order horizon equal to the number of periods that most closely matches the total holding cost with the setup cost over that period. The order horizon that exactly equates holding and setup costs will rarely be an integer number of periods (hence the origin of the name of the method).

Again consider our example. Starting in period 1, we find

$$C^{\text{Holding}}(1) = 0,$$

$$C^{\text{Holding}}(2) = 1(50) = 50,$$

$$C^{\text{Holding}}(3) = 1(50) + 2(10) = 70,$$

$$C^{\text{Holding}}(4) = 1(50) + 2(10) + 3(50) = 220$$

Because 220 exceeds the setup cost of 100, we stop. As 100 is closer to 70 than 220, the first order horizon is three periods. That is, $y_1 = r_1 + r_2 + r_3 = 20 + 50 + 10 = 80$.

We start the process again in period 4.

$$C^{\text{Holding}}(1) = 0,$$

$$C^{\text{Holding}}(2) = 1(50) = 50,$$

$$C^{\text{Holding}}(3) = 1(50) + 2(10) = 70,$$

$$C^{\text{Holding}}(4) = 1(50) + 2(10) + 3(20) = 130$$

Because 130 exceeds the setup cost of 100, we stop. Since 100 is equally distant to 70 and 130, we can arbitrarily choose the order horizon as four periods. That is, $y_4 = r_4 + r_5 + r_6 + r_7 = 50 + 50 + 10 + 20 = 130$.

We start the process again in period 8.

$$C^{\text{Holding}}(1) = 0,$$

$$C^{\text{Holding}}(2) = 1(20) = 20,$$

$$C^{\text{Holding}}(3) = 1(20) + 2(30) = 80$$

As we have reached the end of the horizon, we set $y_8 = r_8 + r_9 + r_{10} = 40 + 20 + 30 = 90$.

t	1	2	3	4	5	6	7	8	9	10	Total
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	80			130				90			300
Setup	100			100				100			\$300
Holding	0	60	10	0	80	30	20	0	50	30	\$280
Total											\$580

7.1.7 *Wagner Whitin Method*

The Wagner Whitin algorithm relies on a fundamental insight: the production in any period must account for the total demand over subsequent periods. Therefore, in an optimal lot-sizing strategy, either the inventory carried forward from a previous period will be depleted entirely, or production in that period will suffice to meet demand. This algorithm employs backward dynamic programming, working from the end of the planning horizon toward the beginning, to determine the most efficient ordering policy. By iteratively considering future demands and minimizing costs over the entire planning horizon, the Wagner Whitin algorithm helps companies optimize their production and inventory management strategies.

Notation:

c_{tj}: cost of producing enough items in period t for periods $t,\ t+1,\ \ldots,\ t+j$

f_t: minimum cost incurred during periods $t,\ t+1,\ \ldots,\ T$, given that at the beginning of period t, the inventory level is zero

$$\boxed{f_t = \min_{j=0,\ 1,\ \ldots,\ T-t} (c_{tj} + f_{t+j+1})} \tag{7.1}$$

The initial condition is $f_{T+1} = 0$. Now, we will solve our example by dynamic programming to illustrate the technique. One starts with the initial condition and works backward from period $T+1$ to period 1. In each period one determines the value of f_t that achieves the minimum.

$$f_{11} = 0$$

$$f_{10} = 100 \text{ (there is no holding cost)}$$

$$f_9 = \begin{cases} 100 + f_{10} = 200 \\ 100 + 1(30) + f_{11} = 130^* \end{cases}$$

$$f_8 = \begin{cases} 100 + f_9 = 230 \\ 100 + 1(20) + f_{10} = 220 \\ 100 + 1(20) + 2(30) + f_{11} = 180^* \end{cases}$$

$$f_7 = \begin{cases} 100 + f_8 = 280 \\ 100 + 1(40) + f_9 = 270^* \\ 100 + 1(40) + 2(20) + f_{10} = 280 \\ 100 + 1(40) + 2(20) + 3(30) + f_{11} = 270^* \end{cases}$$

$$f_6 = \begin{cases} 100 + f_7 = 370 \\ 100 + 1(20) + f_8 = 300^* \\ 100 + 1(20) + 2(40) + f_9 = 330 \\ 100 + 1(20) + 2(40) + 3(20) + f_{10} = 360 \\ 100 + 1(20) + 2(40) + 3(20) + 4(30) + f_{11} = 380 \end{cases}$$

$$f_5 = \begin{cases} 100 + f_6 = 400 \\ 100 + 1(10) + f_7 = 380 \\ 100 + 1(10) + 2(20) + f_8 = 330^* \\ 100 + 1(10) + 2(20) + 3(40) + f_9 = 400 \\ 100 + 1(10) + 2(20) + 3(40) + 4(20) + f_{10} = 450 \\ 100 + 1(10) + 2(20) + 3(40) + 4(20) + 5(30) + f_{11} = 500 \end{cases}$$

$$f_4 = \begin{cases} 100 + f_5 = 430 \\ 100 + 1(50) + f_6 = 450 \\ 100 + 1(50) + 2(10) + f_7 = 440 \\ 100 + 1(50) + 2(10) + 3(20) + f_8 = 410^* \\ 100 + 1(50) + 2(10) + 3(20) + 4(40) + f_9 = 520 \\ 100 + 1(50) + 2(10) + 3(20) + 4(40) + 5(20) + f_{10} = 590 \\ 100 + 1(50) + 2(10) + 3(20) + 4(40) + 5(20) + 6(30) + f_{11} = 670 \end{cases}$$

$$f_3 = \begin{cases} 100 + f_4 = 510 \\ 100 + 1(50) + f_5 = 480^* \\ 100 + 1(50) + 2(50) + f_6 = 550 \\ 100 + 1(50) + 2(50) + 3(10) + f_7 = 550 \\ 100 + 1(50) + 2(50) + 3(10) + 4(20) + f_8 = 440 \\ 100 + 1(50) + 2(50) + 3(10) + 4(20) + 5(40) + f_9 = 690 \\ 100 + 1(50) + 2(50) + 3(10) + 4(20) + 5(40) + 6(20) + f_{10} = 780 \\ 100 + 1(50) + 2(50) + 3(10) + 4(20) + 5(40) + 6(20) + 7(30) + f_{11} = 890 \end{cases}$$

$$f_2 = \begin{cases} 100 + f_3 = 580 \\ 100 + 1(10) + f_4 = 520^* \\ 100 + 1(10) + 2(50) + f_5 = 540 \\ 100 + 1(10) + 2(50) + 3(50) + f_6 = 660 \\ 100 + 1(10) + 2(50) + 3(50) + 4(10) + f_7 = 670 \\ 100 + 1(10) + 2(50) + 3(50) + 4(10) + 5(20) + f_8 = 680 \\ 100 + 1(10) + 2(50) + 3(50) + 4(10) + 5(20) + 6(40) + f_9 = 870 \\ 100 + 1(10) + 2(50) + 3(50) + 4(10) + 5(20) + 6(40) + 7(20) + f_{10} = 980 \\ 100 + 1(10) + 2(50) + 3(50) + 4(10) + 5(20) + 6(40) + 7(20) + 8(30) + f_{11} = 1120 \end{cases}$$

$$f_1 = \begin{cases} 100 + f_2 = 620 \\ 100 + 1(50) + f_3 = 630 \\ 100 + 1(50) + 2(10) + f_4 = 580^* \\ 100 + 1(50) + 2(10) + 3(50) + f_5 = 650 \\ 100 + 1(50) + 2(10) + 3(50) + 4(50) + f_6 = 820 \\ 100 + 1(50) + 2(10) + 3(50) + 4(50) + 5(10) + f_7 = 840 \\ 100 + 1(50) + 2(10) + 3(50) + 4(50) + 5(10) + 6(20) + f_8 = 870 \\ 100 + 1(50) + 2(10) + 3(50) + 4(50) + 5(10) + 6(20) + 7(40) + f_9 = 1100 \\ 100 + 1(50) + 2(10) + 3(50) + 4(50) + 5(10) + 6(20) + 7(40) + 8(20) + f_{10} = 1230 \\ 100 + 1(50) + 2(10) + 3(50) + 4(50) + 5(10) + 6(20) + 7(40) + 8(20) + 9(30) + f_{11} = 1400 \end{cases}$$

To determine the optimal order policy, we retrace the solution back from the beginning. In period 1 the optimal value of f_1 is 580. This means that the production level in period 1 is equal to the sum of demands in periods 1, 2, and 3 so that $y_1 = r_1 + r_2 + r_3 = 20 + 50 + 10 = 80$. The next order period is period 4. The optimal value of f_4 is 410, which implies that the production quantity in period 4 is equal to the sum of demands in periods 4, 5, 6, and 7, or $y_4 = r_4 + r_5 + r_6 + r_7 = 50 + 50 + 10 + 20 = 130$. The next period of ordering is period 8. The optimal value of f_8 is 180. This gives $y_8 = r_8 + r_9 + r_{10} = 40 + 20 + 30 = 90$.

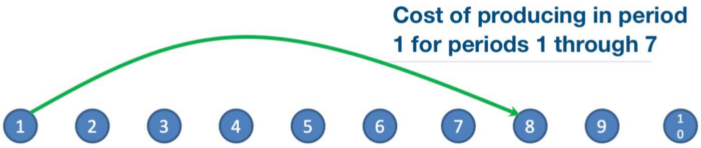
t	1	2	3	4	5	6	7	8	9	10	Total
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	80			130				90			300
Setup	100			100				100			\$300
Holding	0	60	10	0	80	30	20	0	50	30	\$280
Total											\$580

7.1.8 *Optimal Lot Sizing*

It shows how dynamic programming can be used to find the shortest path. In the context of the lot sizing problem, an alternative approach to the Wagner Whitin Algorithm involves modeling it as a shortest path problem. Here, each vertex in the graph represents an order moment, while an arc $(i,\ j)$ signifies an order placed in period i to fulfill demand from periods i to $j - 1$. This method will also yield the optimal solution under the given assumptions.

Shortcomings of MRP

MRP is a valuable tool for production scheduling, but it also has its limitations and chal-



lenges. One significant issue is uncertainty, particularly in forecasting future sales and estimating production lead times accurately. Lead times can be independent of lot sizes, leading to inaccuracies in planning. Additionally, MRP assumes infinite production capacity, which doesn't always align with real-world constraints. The static nature of MRP, with fixed planning horizons, can make it challenging to adapt to dynamic production environments, leading to system nervousness and unanticipated changes in the Master Production Schedule (MPS). Moreover, there can be an incentive for stakeholders to inflate lead times to compensate for uncertainties, potentially creating a planning loop and further complicating the process. These shortcomings highlight the need for complementary strategies and systems to address the limitations of MRP effectively.

The Planning Loop

The planning loop can be a frustrating cycle within production management, often arising from efforts to enhance due-date performance. It typically begins with fixed lead times, which can result in a poor performance against due dates. In response, management may opt to increase lead times to allow for more flexibility. However, longer lead times necessitate extending the forecasting horizon to anticipate demand accurately. Unfortunately, a longer forecasting horizon often introduces errors in demand estimation, ultimately contributing to continued poor performance against due dates. This situation prompts management to once again consider increasing lead times, perpetuating the loop. Breaking this cycle requires careful consideration of lead time adjustments alongside improvements in forecasting accuracy to achieve better due-date performance without exacerbating the planning loop.

Question: What causes long delays in processing departments: limited capacity, demand variability, or process variability?

Nervousness Example

Item A (Leadtime = 2 weeks, Order Interval = 5 weeks)

Week	0	1	2	3	4	5	6	7	8
Gross Requirements		2	24	3	5	1	3	4	50
Scheduled Receipts									
Project Inventory	28	26	2	-1	-6	-7	-10	-14	-64
Net Requirements				1	5	1	3	4	50
Planned Orders		14					50		

Component B (Leadtime = 4 weeks, Order Interval = 5 weeks)

Week	0	1	2	3	4	5	6	7	8
Gross Requirements		14					50		
Scheduled Receipts		14							
Project Inventory	2	2	2	2	2	2	-48		
Net Requirements							48		
Planned Orders			48						

Note that we are using fixed order period lot sizing rule.

Item A (Leadtime = 2 weeks, Order Interval = 5 weeks)

Week	0	1	2	3	4	5	6	7	8
Gross Requirements		2	23	3	5	1	3	4	50
Scheduled Receipts									
Project Inventory	28	26	3	0	-5	-6	-9	-13	-63
Net Requirements					5	1	3	4	50
Planned Orders			63						

Component B (Leadtime = 4 weeks, Order Interval = 5 weeks)

Week	0	1	2	3	4	5	6	7	8
Gross Requirements			63						
Scheduled Receipts		14							
Project Inventory	2	16	-47						
Net Requirements			47						
Planned Orders	47*								

*: Past due

Note that small reduction in requirements caused a large change in orders and made the schedule infeasible.

To diminish nervousness in production planning, several strategies can be implemented to mitigate the triggers for plan changes. One approach involves stabilizing the Master Production Schedule (MPS) by incorporating tactics like frozen zones and time fences. Additionally, integrating spare parts forecasts into gross requirements can help reduce unplanned demands. Employing discipline in adhering to the MRP plan for releases and controlling changes in safety stocks or lead times can also foster stability. Adjusting lot sizing procedures is another effective tactic, such as using fixed order quantities at the top level, lot-for-lot at intermediate levels, and fixed order intervals at the bottom level. Furthermore, adopting firm planned orders that require managerial intervention to adjust can help maintain consistency in production plans, thereby reducing nervousness and promoting smoother operations.

Handling Change

In response to various causes of change in production planning, different strategies can be employed to manage them effectively. These changes may include new orders in the Master Production Schedule (MPS), delays in order completion, scrap losses, or engineering changes in the Bill of Materials (BOM). One approach is regenerative MRP, which involves completely redoing MRP calculations, starting from the MPS and cascading through the BOMs to reflect the updated situation comprehensively. Another method is Net Change MRP, where the material requirements plan is stored, and only the parts impacted by the change are altered, offering a more targeted and efficient adjustment process without the need for a full recalculation. These responses help streamline the adaptation to changes while maintaining operational continuity and efficiency in production planning.

Rescheduling

Rescheduling in production planning can follow two primary strategies: top-down planning and bottom-up replanning. In top-down planning, the Material Requirements Planning system is utilized alongside any modifications, such as adjustments to the Master Production Schedule (MPS) or scheduled receipts, to recalculate the overall plan. However, this approach may encounter infeasibilities, often indicated by exception codes. To address this, Joseph Orlicky suggested incorporating minimum lead times to enhance feasibility. Conversely, bottom-up replanning involves leveraging pegging and firm planned orders to direct the rescheduling process. Pegging facilitates tracing releases back to their sources in the MPS, while fixed order periods aid in securing releases essential for fulfilling firm customer orders. Additionally, the use of compressed lead times, known as expediting, is common to accelerate the process and ensure timely order fulfillment. These methods enable efficient rescheduling while maintaining operational integrity in response to changes in production plans.

Safety Stocks and Safety Lead Time

Safety stocks serve as a buffer against inventory uncertainties by ensuring a minimum level of inventory is maintained at all times. They are essential for mitigating risks associated with quantity uncertainties, such as yield loss or unexpected fluctuations in demand. Safety lead times, on the other hand, involve inflating production lead times recorded for parts. These inflated lead times act as a safeguard against time uncertainties, such as delays in delivery. By incorporating safety lead times into the planning process, organizations can better anticipate and mitigate the impact of potential delays, ensuring smoother operations and enhanced reliability in meeting customer demands.

Item A (Leadtime = 2 weeks, Order Quantity = 50)

Week	0	1	2	3	4	5
Gross Requirements		20	40	20	0	30
Scheduled Receipts			50			
Project Inventory	40	20	30	10	10	-20
Net Requirements						20
Planned Orders				50		

Safety Stock = 20 units

Week	0	1	2	3	4	5
Gross Requirements		20	40	20	0	30
Scheduled Receipts			50			
Project Inventory	40	20	30	10	10	-20
Net Requirements				10		30
Planned Orders		50				

Safety Leadtime = 1 week

Week	0	1	2	3	4	5
Gross Requirements		20	40	20	0	30
Scheduled Receipts			50			
Project Inventory	40	20	30	10	10	-20
Net Requirements						20
Planned Orders			50			

7.2 MRP II: Manufacturing Resource Planning

MRP II, or Manufacturing Resource Planning, represents an evolution from traditional Material Requirements Planning (MRP) systems. More modern implementations have evolved into Enterprise Resource Planning (ERP) systems, which encompass a broader range of functions beyond manufacturing. MRP II extends the capabilities of MRP by incorporating additional modules such as Master Production Scheduling (MPS), Rough Cut Capacity Planning (RCCP), Capacity Requirements Planning (CRP), and Production Activity Control (PAC). These modules enable organizations to effectively plan and manage their manufacturing operations by integrating various aspects such as production scheduling, capacity planning, and production control into a comprehensive system.

MRP is a closed production planning system that converts an MPS into planned order releases. Manufacturing resource planning (MRP II) is a philosophy that attempts to incorporate the other relevant activities of the firm into the production planning process. In particular, the financial, accounting, and marketing functions of the firm are tied to the operations function. As an example of the difference between the perspectives offered by MRP and MRP II, consider the role of the master production schedule. In MRP, the MPS is treated as input information. In MRP II, the MPS would be considered a part of the system and, as such, would be considered a decision variable as well. Hence, the production control manager would work with the marketing manager to determine when the production schedule should be altered to incorporate revisions in the forecast and new order commitments. Ultimately, all divisions of the company would work together to find a production schedule consistent with the overall business plan and long term financial strategy of the firm.

Another important aspect of MRP II is the incorporation of capacity resource planning (CRP). Capacity considerations are not explicitly accounted for in MRP. MRP II is a closed-loop cycle in which lot sizing and the associated shop floor schedules are compared to capacities and recalculated to meet capacity restrictions. However, capacity issues continue to be an important issue in both MRP and MRP II operating systems.

Such a global approach to the production scheduling problem is quite ambitious. Whether such a philosophy can be converted to a workable system in a particular operating environment remains to be seen.

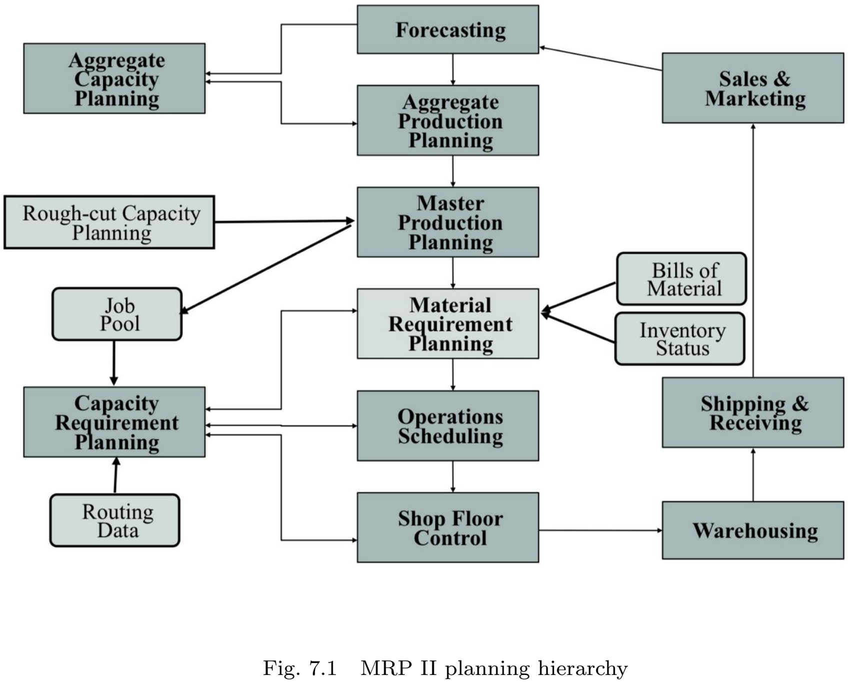

Forecasting
Aggregate Capacity Planning
Aggregate Production Planning
Sales & Marketing
Rough-cut Capacity Planning
Master Production Planning
Job Pool
Material Requirement Planning
Bills of Material
Inventory Status
Capacity Requirement Planning
Operations Scheduling
Shipping & Receiving
Routing Data
Shop Floor Control
Warehousing

Fig. 7.1 MRP II planning hierarchy

7.2.1 *Master Production Scheduling (MPS)*

Master Production Scheduling (MPS) serves as a crucial component within the broader framework of Manufacturing Resource Planning systems. It acts as a primary driver for MRP by providing a detailed plan for production activities over a specific time horizon. While MPS is expected to be highly accurate in the near term, particularly for firm orders, its accuracy may diminish when forecasting for the long term. Software tools supporting MPS typically incorporate functionalities such as forecasting, order entry, and netting against existing inventory levels. To maintain stability and reliability in production planning, MPS often establishes a "frozen zone", ensuring that certain aspects of the schedule remain unchanged within defined timeframes.

7.2.2 *Rough Cut Capacity Planning (RCCP)*

Rough Cut Capacity Planning (RCCP) serves as a preliminary assessment of the capacity requirements for key resources within the production process. It provides a rapid evaluation of whether the available capacity aligns with the projected demands outlined in the Master Production Schedule (MPS). RCCP utilizes a bill of resource (BOR) approach for each item listed in the MPS, allowing it to calculate the resource utilization by systematically exploding the MPS against the BOR while considering lead times. In situations where capacity constraints are identified, RCCP offers recommendations for addressing these constraints, such as adjusting the MPS or increasing capacity through measures like overtime production.

7.2.3 *Capacity Requirements Planning (CRP)*

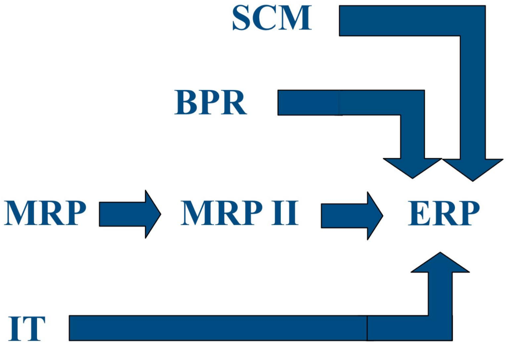
Capacity Requirements Planning (CRP) represents a step further in detail compared to Rough Cut Capacity Planning (RCCP). It incorporates comprehensive routing data, including work centers and associated processing times, for all items in the production process. By exploring the orders against this routing information, CRP generates a usage profile for each work center, allowing for a thorough analysis of capacity utilization across the entire production system. CRP is particularly valuable for identifying potential overload conditions within work centers. However, it is important to note that it does not provide mechanisms for directly addressing capacity-related issues. Additionally, CRP maintains fixed lead times despite potential queueing effects, which may impact the accuracy of capacity assessments in dynamic production environments.

7.2.4 *Production Activity Control (PAC)*

Production Activity Control (PAC), also known as "shop floor control", serves as a critical link between planning and execution within manufacturing operations. It involves the coordination of production activities by providing detailed routing and standard time information for each operation. PAC sets planned start times for tasks, allowing for effective prioritization and expediting when necessary. Additionally, PAC facilitates input-output control by comparing planned production throughput with actual performance on the shop floor. In modern manufacturing contexts, similar functionalities are encompassed within Manufacturing Execution Systems (MES), which bridge the gap between planning activities and real-time production control. MES systems play a vital role in optimizing production processes and ensuring efficient utilization of resources within manufacturing facilities.

7.2.5 *Enterprise Resource Planning (ERP)*

Enterprise Resource Planning (ERP) systems aim to integrate and streamline information flow across various functional areas within an organization. The primary goal of ERP is to unify disparate business processes and data sources, enabling seamless coordination and collaboration across the entire enterprise. This integration typically spans key areas such as manufacturing, distribution, accounting, financial management, and personnel administration. By centralizing data and processes, ERP systems facilitate real-time visibility, improved decision-making, and enhanced operational efficiency throughout the organization.



7.2.5.1 *"Integrated" ERP Approach*

The "integrated" ERP approach offers several advantages, including integrated functionality, consistent user interfaces, and a unified database. Having a single vendor and contract simplifies management, while a unified architecture and product support enhance operational efficiency. However, this approach also presents disadvantages, such as incompatibility with existing systems and management practices, as well as long and expensive implementation processes. Additionally, it may limit flexibility in using tactical point systems and result in long product development and implementation cycles, leading to a lengthy payback period. Furthermore, there may be a lack of technological innovation compared to more agile solutions.

Challenges and Limitations of Material Requirements Planning

MRP can fail due to several reasons. Firstly, inadequate commitment from top management can hinder the successful implementation and operation of the system. Secondly, a lack of proper education and training for those using the system can lead to misunderstandings and misuse of the MRP software. Additionally, an unrealistic master production schedule (MPS), which serves as the foundation of the MRP, can result in inefficient planning and resource allocation. Moreover, inaccurate data, including bills of material (BOM) and inventory records, can undermine the accuracy and effectiveness of the MRP process, leading to suboptimal decision-making and planning outcomes.

7.3 Just In Time

Just In Time (JIT) principles have deep roots in Japanese culture and history, emerging from the country's post-World War II efforts to revitalize its economy and compete globally. In 1949, Japanese domestic production figures revealed a stark reality: while 25,622 trucks and 1,008 cars were manufactured in Japan, the productivity ratio between American and Japanese industries stood at 9:1. Facing scarcity of resources and striving to overcome the vast productivity gap with America, Japanese firms, notably Toyota, focused on cost reduction, quality improvement, and responsiveness to customer demands. This led to the development of the Toyota Production System by Taiichi Ohno and Shigeo Shingo, emphasizing efficient resource utilization and continuous improvement. JIT gained significant traction in the 1980s and 90s in the United States, becoming a cornerstone of the agile and lean manufacturing movement, characterized by streamlined processes and minimal waste. JIT challenges fundamental aspects of Western manufacturing practices by questioning established norms. It questions the necessity of fixed setups, long delivery times, high ordering costs, and the extensive time spent on material handling. Moreover, it challenges the traditional reliance on inventory as a buffer against uncertainty, advocating instead for streamlined processes that minimize waste and maximize efficiency.

The environment within which manufacturing operations take place serves as both constraints and controls. Factors such as machine setup times, vendor deliveries, quality levels including scrap and rework rates, production schedules aligned with customer due dates, and even product designs influence the efficiency and effectiveness of production processes. By optimizing these environmental factors, the manufacturing system can be significantly streamlined and made easier to manage, leading to improved overall performance and productivity.

Toyota Production System

The Toyota Production System (TPS) is built upon two fundamental pillars: Just in Time (JIT) and Autonomation, known as jidoka, which embodies automation with human intervention. Jidoka employs limit switches or mechanisms to halt a process under specific conditions such as completing the required number of pieces, detecting defective parts, or encountering equipment jams. TPS incorporates various practices including setup reduction (SMED), extensive worker training, fostering strong vendor relations, stringent quality control measures, and implementing foolproofing techniques (baka-yoke) to prevent errors. These elements collectively enable TPS to achieve high efficiency, quality, and flexibility in manufacturing operations.

The Seven Zeros

The concept of the "Seven Zeros" in lean manufacturing encapsulates a set of principles aimed at streamlining production processes and eliminating waste. These principles include achieving zero defects by focusing on quality at the source to prevent delays caused by errors. Additionally, it emphasizes the importance of minimizing or eliminating excess lot sizes to prevent delays associated with waiting for inventory, often targeting a lot size of one to facilitate efficient production. Another key aspect is achieving zero setups to reduce setup delays and support smaller lot sizes without sacrificing efficiency. The goal of zero breakdowns aims to maintain uninterrupted production flow while minimizing excess handling, which helps promote smoother material flow within the production system. Ensuring zero lead time is crucial for rapid replenishment of parts, aligning closely with the overarching objective of achieving zero inventories. Finally, zero surging is essential, particularly in systems without work-in-process (WIP) buffers, to maintain consistent production rates and avoid disruptions. These principles collectively drive lean manufacturing efforts toward greater efficiency and productivity.

JIT Strategies

JIT strategies encompass a range of practices aimed at enhancing efficiency and minimizing waste in manufacturing processes. These strategies include efforts to reduce setup times and batch sizes to enable quicker changeovers and more frequent production runs. Variability reduction aims to stabilize processes and minimize disruptions, while efforts to reduce material handling streamline workflow and minimize unnecessary movement of materials. Strategies to reduce defects and rework enhance product quality and minimize delays, while initiatives to reduce breakdowns ensure continuous operation of machinery and equipment. Increasing capacity helps meet demand fluctuations while smoothing production schedules promotes consistent workflow and resource utilization. Maintaining constant work-in-process (WIP) and limiting inventory levels of finished goods and raw materials minimize excess inventory and associated costs. Synchronizing operations within the factory and coordinating material delivery with suppliers and customers further optimize supply chain efficiency. Additionally, empowering workers to make improvements and simplifying workflow promote employee engagement and operational effectiveness. Overall, these JIT strategies collectively contribute to leaner, more agile manufacturing systems.

Reducing inventory across various stages of production, including raw materials, work in process (WIP), and finished goods, is a core objective of JIT methodologies. Efforts to minimize raw material inventory involve closely monitoring and optimizing procurement processes to maintain adequate supplies without excess stockpiling. Similarly, reducing WIP inventory requires streamlining production processes, minimizing queue times, and implementing just-in-time manufacturing practices to ensure smooth workflow and minimize idle inventory. Strategies to decrease finished goods inventory involve aligning production with customer demand to prevent overproduction and excess stock.

Concurrently, reducing variability throughout the supply chain is essential for achieving JIT goals. This involves stabilizing demand through accurate forecasting, synchronizing production schedules with suppliers and customers, and implementing measures to regulate work release and minimize process variability. Additionally, efforts to stabilize WIP involve optimizing workflow, reducing breakdowns through preventive maintenance, and minimizing scrap and rework through improved quality control measures. Lastly, assembling products to order instead of maintaining pre-built inventory levels helps minimize excess inventory and align production with customer demand more effectively.

A popular analogy is to compare a production process with a river and the level of inventory with the water level in the river. When the water level is high, the water will cover the rocks. Likewise, when inventory levels are high, problems are masked. However, when the water level (inventory) is low, the rocks (problems) are evident (see the figure below).

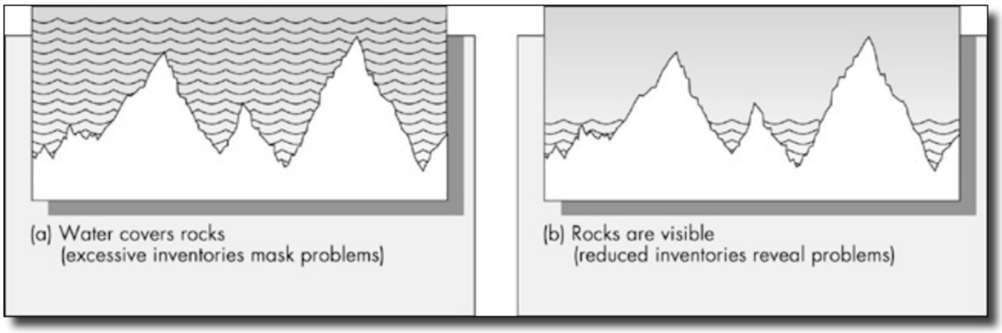

(a) Water covers rocks
(excessive inventories mask problems)
(b) Rocks are visible
(reduced inventories reveal problems)

Fig. 7.2 River/inventory analogy illustrating the advantages of just in time (Credit: [Nahmias, 1997])

Production Smoothing

Production smoothing aims to maintain relatively constant production volumes and product mix over time, facilitating efficient operations and resource utilization. For instance, in a scenario where 10,000 units are produced monthly across 20 working days with two shifts, and each shift produces 250 units within 480 minutes, the goal is to achieve a consistent production rate of one unit every 1.92 minutes. To accomplish this, a balanced production sequence is implemented, ensuring that the daily output of 500 units comprises a mix of products A, B, and C in the ratio of 50%, 25%, and 25%, respectively. This results in a consistent pattern of production, alternating between products A, B, and C, thereby smoothing the overall production process and enhancing efficiency.

$0.5(500) = 250$ units of A

$0.25(500) = 125$ units of B

$0.25(500) = 125$ units of C

A-B-A-C—A-B-A-C—A-B-A-C—A-B-A-C...

This pattern ensures production of equal B and C's and twice as many A's.

Inherent Inflexibility of JIT

The JIT system, while highly efficient, faces inherent inflexibility due to several factors. These include the need for stable production volumes, a consistent product mix, precise production sequences, and rapid replenishment processes. To address these challenges and promote flexibility within the JIT framework, several measures can be implemented. These include maintaining capacity buffers to accommodate fluctuations in demand, reducing setup times to enable quick changeovers between products, implementing cross-training programs to enhance workforce versatility, and optimizing plant layout to facilitate smooth and adaptable operations. By adopting these strategies, organizations can mitigate the constraints associated with JIT and enhance their ability to respond effectively to changing market conditions and customer demands.

Increasing production capacity entails several strategies aimed at optimizing resource utilization and efficiency within a manufacturing environment. One approach involves reducing resource loading by balancing workloads across available resources to minimize idle time and maximize productivity. Another strategy is to enhance processing speed through technological upgrades or process optimizations to shorten production cycles. Additionally, increasing the number of resources, whether through investments in additional equipment or workforce expansion, can help meet growing demand and alleviate bottlenecks. Resource sharing involves maximizing the utilization of existing resources across multiple processes or shifts to optimize overall capacity utilization. Improving resource availability by reducing mean time to repair (MTTR) and increasing mean time to failure (MTTF) enhances operational uptime. To streamline operations further, reducing internal setups and eliminating bottlenecks related to auxiliary resources such as operators, tools, and fixtures are essential. Finally, implementing two-shift operations can effectively extend production hours to meet increased demand without significant infrastructure investments.

Setup reduction is a key strategy aimed at minimizing the time and effort required to switch between different production runs, thereby enabling more frequent and efficient production of smaller lot sizes. The motivation behind this approach lies in the recognition that large setups can hinder the feasibility of producing small lot sequences. Internal setups, which occur while the machine is offline, contrast with external setups, which are performed while the machine continues running. The approach to setup reduction involves several steps: firstly, separating internal setups from external setups to minimize downtime; secondly, converting internal setups into external setups wherever possible to keep machines operational; thirdly, eliminating adjustment processes to streamline setup procedures; and finally, abolishing setups altogether through strategies like uniform product design, combined production, or parallel machines.

Other techniques to reduce setup times include minimizing online setups, standardizing products to simplify changeovers, designing flexible fixtures, and standardizing tooling and fixturing. Additionally, adopting practices such as group technology and cellular manufacturing can facilitate setup reduction by organizing similar parts into production cells and streamlining workflows. Investing in flexible and programmable automation technologies further enhances setup flexibility and efficiency, while implementing group release and scheduling methodologies helps coordinate production activities to optimize setup utilization across multiple production runs. By systematically implementing these setup reduction techniques, manufacturers can achieve faster changeovers, increased production flexibility, and improved overall operational efficiency.

Worker cross-training, a practice where employees are trained to perform tasks outside their primary roles, offers numerous benefits to manufacturing operations. Firstly, it enhances flexibility by enabling workers to adapt to changing production demands and fill in for absent or overloaded colleagues. This flexibility allows capacity to float across different workstations, smoothing the flow of production and minimizing bottlenecks. Additionally, cross-training reduces boredom among workers by providing variety in their tasks, which can lead to higher morale and job satisfaction. Moreover, exposing employees to different aspects of the production process fosters an appreciation for the overall picture of manufacturing operations, encouraging a deeper understanding of how individual tasks contribute to the larger workflow. Finally, cross-training increases the potential for idea generation and innovation as employees gain insights from diverse experiences and perspectives, leading to more creative problem-solving and process improvements. Overall, worker cross-training serves as a valuable strategy for enhancing operational flexibility, efficiency, and employee engagement in manufacturing environments.

U-Shaped Cells

U-shaped cells in manufacturing layout design offer several advantages that contribute to improved workflow efficiency and productivity. Firstly, they promote smooth flow with minimal work-in-process (WIP), as the U-shaped configuration allows for easy movement of materials and components between workstations, reducing the need for excess inventory. Additionally, U-shaped cells facilitate workers staffing multiple machines within the cell, enabling multitasking and efficient utilization of labor resources. The layout also provides maximum visibility, as workers have a clear line of sight to all machines and processes within the cell, enhancing monitoring and quality control efforts. Furthermore, the compact U-shaped design minimizes walking distances for workers, reducing time wasted on unnecessary movement and improving overall productivity. The flexible nature of U-shaped cells allows for varying numbers of workers to be deployed based on production demands, ensuring optimal resource allocation and responsiveness to changing needs. Moreover, the layout facilitates collaboration among workers within the cell, enabling them to cooperate seamlessly to smooth flow and address any issues or bottlenecks that may arise during production. Overall, U-shaped cells offer a versatile and efficient layout solution that enhances workflow visibility, flexibility, and collaboration in manufacturing environments.

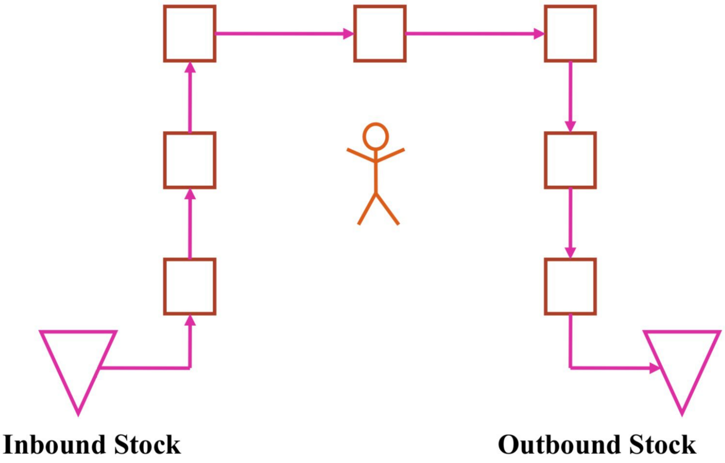


Fig. 7.3 U-shaped manufacturing cell

Reducing transfer batches in manufacturing involves several strategies aimed at minimizing the amount of material transferred between different stages of production. One approach is to implement a relayout of the production floor, reorganizing workstations and processes to minimize the distance and time required for material movement. Another strategy is to establish continuous transfer processes, where materials flow seamlessly between workstations without interruptions or delays, reducing the need for large transfer batches. Synchronizing production schedules across different stages of the manufacturing process can also help minimize transfer batches by ensuring that materials are available exactly when needed, avoiding unnecessary accumulation or waiting times. Additionally, adopting cellular manufacturing techniques, which group together machines and processes needed to produce a specific product or product family, can further reduce transfer batches by streamlining production within dedicated work cells.

Total Quality Management

Total Quality Management (TQM) emerged from the work of American quality experts such as Shewhart, Deming, Juran, and Feigenbaum. However, its principles found fertile ground in Japan due to cultural factors such as the Japanese aversion to wasting resources and their resistance to specialists, including quality assurance personnel. TQM became integral to Just-In-Time manufacturing because JIT relies on high-quality inputs and processes. TQM promotes high quality by emphasizing the identification and rapid detection of problems, creating pressure to continually improve quality throughout the production process.

Improving quality entails implementing various measures aimed at enhancing product consistency and reliability. These include Statistical Process Control (SPC) to monitor and maintain process stability, emphasizing visual indicators for quality assessment, prioritizing compliance with quality standards over output quantity, employing line stop mechanisms to prevent defective products from proceeding, encouraging self-correction of errors to eliminate rework loops, conducting 100 percent inspections rather than relying on statistical sampling, fostering a culture of continual improvement, adopting small lot sizes to detect and address issues early, certifying vendors to ensure quality inputs, and implementing total preventive maintenance to minimize equipment downtime and defects. By integrating these strategies, organizations can achieve higher levels of quality assurance and customer satisfaction.

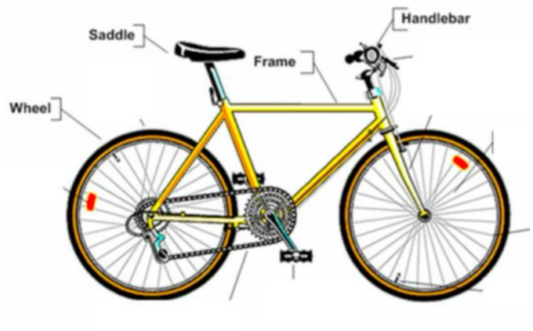
JIT Implementation

Implementing Just-in-Time involves a series of strategies aimed at streamlining production processes and minimizing waste. Key elements of JIT implementation include scheduling frequent deliveries of raw materials and components to match production needs, limiting Work-in-Process (WIP) inventory to reduce storage costs and lead times, coordinating activities across the supply chain to ensure timely delivery and availability of materials, smoothing production volumes to minimize fluctuations and maintain consistent workflow, implementing pull production control methods such as Kanban to signal production needs based on actual demand, maintaining an excess capacity to accommodate fluctuations in demand or unexpected disruptions, enabling rapid changeovers to switch between production tasks efficiently, cultivating a flexible and cross-trained workforce capable of adapting to changing production requirements, fostering a culture of continuous improvement to identify and eliminate inefficiencies, and enforcing strict quality control measures to ensure product reliability and consistency.

As a result, Just-in-Time philosophy offers several valuable lessons for efficient production management. Firstly, it emphasizes that the production environment serves as a natural control mechanism, highlighting the significance of optimizing operational details to enhance overall performance. Secondly, JIT underscores the importance of controlling Work-in-Process (WIP) inventory to minimize waste and improve workflow efficiency. Additionally, JIT teaches that speed and flexibility are essential assets in meeting customer demands and adapting to changing market conditions. Moreover, JIT demonstrates that prioritizing quality can lead to improved productivity and customer satisfaction. Finally, JIT advocates for a culture of continual improvement, recognizing that ongoing enhancements are necessary for long-term success and competitiveness in today's dynamic business environment.

By embracing these lessons, organizations can enhance their operational effectiveness and achieve sustainable growth.

Important 7.1



A bike is made of one unit of the saddle, one unit of the frame assembly, and one unit of the handlebar. Each frame assembly needs one frame and 2 wheels. The lead times for each item are given in the table below.

Item	Purchased or Produced	Lead time (in weeks)
Bike	Produced	1
Frame Assembly	Produced	1
Saddle	Purchased	2
Handlebar	Purchased	2
Frame	Purchased	3
Wheel	Purchased	1

The company has two customer orders for bikes: 100 in week 3 and 20 in week 5. The company has 50 bikes in the inventory and 40 additional bikes will be ready to be delivered in the first week. There are 5 frame assemblies, 40 saddles, 10 handlebars, 70 frames, and 10 wheels in the inventory. The company will receive 5 handlebars in week 1. There are also three open purchase orders for wheels: the company will receive 10 wheels on week 1, 20 wheels on week 2, and 10 wheels on week 4.

Determine the component and subassembly requirements and when to plan orders for these items. (Note: You need to create MRP tables.)

Important 7.2

Comfortable Furniture Co. produces a type of chairs. The demand for the chairs for the next 5 weeks is given in the table below.

	1	2	3	4	5
Demand	40	50	30	70	20

The setup cost of production is \$200 and the holding cost per item per week is equal to \$10. Determine a production plan using the Silver-Meal method.

Chapter 8

Operations Scheduling

Operation scheduling is a critical component in operational management, providing the framework for coordinating and optimizing organizational tasks, resources, and timelines. In essence, scheduling entails strategically allocating available resources to various activities to meet production goals, deadlines, and customer demands efficiently. From manufacturing plants to service industries, the effective implementation of scheduling techniques can enhance productivity, minimize delays, and streamline operations, ultimately contributing to the success and competitiveness of organizations across diverse sectors.

It may be helpful to first mention several key terminologies. Firstly, there's the concept of a flow shop, where n jobs are processed through m machines in the same sequence. In contrast, a job shop allows for different sequencing of jobs through machines, and some machines may handle multiple operations. The distinction between parallel processing, where machines are identical and work concurrently, and sequential processing is also crucial. The flow time of a job refers to the duration from its initiation to completion, while makespan denotes the flow time of the last completed job. Moreover, tardiness represents the positive difference between a job's completion time and its due date. At the same time, lateness indicates the difference between the completion time and the due date, which may be negative if completed before the due date. Understanding these terms is fundamental for effective scheduling and optimization in machine environments.

Some KPIs that the manufacturing is focused on are makespan, average flow time, tardiness, lateness, etc. What a company focuses on depends on the company goals, and perspective is valid. One such example is presented in Section 8.4.

8.1 Single Machine Scheduling

In single-machine scheduling, the focus lies on scenarios where a set of jobs awaits processing on a single machine, each job characterized by its **available time**, **processing time**, and

due date. The primary objective is to determine the optimal processing sequence for these jobs to achieve predefined goals or objectives.

In single-machine scheduling, different sequencing rules dictate the order in which jobs are processed. The purpose of this section is to illustrate how these sequencing rules affect various measures of system performance. We compare the following four sequence rules:

(1) ***First come first served (FCFS).*** Jobs are processed in the sequence in which they entered the shop.

(2) ***Shortest processing time (SPT).*** Jobs are sequenced in increasing order of their processing times. The job with the shortest processing time is first, the job with the next shortest processing time is second, and so on.

(3) ***Earliest due date (EDD).*** Jobs are sequenced in increasing order of their due dates. The job with the earliest due date is first, the job with the next earliest due date is second, and so on.

(4) ***Critical ratio (CR).*** Critical ratio scheduling requires forming the ratio of the remaining time until the due date, divided by the processing time of the job, and scheduling the job with the smallest ratio next.

There are several **KPIs** that we can assess, including makespan, mean flow time, average/maximum tardiness, average lateness, etc. The selection of the KPI(s) must meet the strategic goals of the company. Here, only for illustration purposes, we compare the performance of the above four rules for a specific case based on (i) mean flow time, (ii) average tardiness, and (iii) the number of tardy jobs. The purpose of the next example is to help the reader develop an intuition for the mechanics of scheduling before presenting formal results.

Example 8.1

A machining center has five unprocessed jobs remaining at a particular point in time. All the jobs are available at this point but suppose they were made available and placed here in the order of 1-2-3-4-5. The jobs, their processing times, and their due dates are given in the table below. Determine the best sequence of jobs that minimizes mean flow time, average tardiness, and number of tardy jobs.

Job Number	Processing Time	Due Date
1	11	61
2	29	45
3	31	31
4	1	33
5	2	32

8.1.1 *First Come First Served*

Because the jobs have entered the shop in the sequence that they are numbered, FCFS scheduling means that the jobs are scheduled in the order 1, 2, 3, 4, and 5.

Sequence	Job Number	Processing Time	Completion Time	Due Date	Tardiness
1	1	11	11	61	0
2	2	29	40	45	0
3	3	31	71	31	40
4	4	1	72	33	39
5	5	2	74	32	42
Total			**268**		**121**

$$\text{Mean flow time} = 268/5 = 53.6$$
$$\text{Average tardiness} = 121/5 = 24.2$$
$$\text{Number of tardy jobs} = 3$$

The tardiness of a job is equal to zero if the job is completed before its due date and is equal to the number of days late if the job is completed after its due date.

8.1.2 *Shortest Processing Time*

In this rule, jobs are processed in the ascending order of their process time. Hence the sequence is 4, 5, 1, 2, and 3.

Sequence	Job Number	Processing Time	Completion Time	Due Date	Tardiness
1	4	1	1	33	0
2	5	2	3	32	0
3	1	11	14	61	0
4	2	29	43	45	0
5	3	31	74	31	43
Total			**135**		**43**

Mean flow time = 135/5 = 27.0

Average tardiness = 43/5 = 8.6

Number of tardy jobs = 1

8.1.3 *Earliest Due Date*

In this rule, jobs are processed in the ascending order of their due date. Hence the sequence is 3, 5, 4, 2, and 1.

Sequence	Job Number	Processing Time	Completion Time	Due Date	Tardiness
1	3	31	31	31	0
2	5	2	33	32	1
3	4	1	34	33	1
4	2	29	63	45	18
5	1	11	74	61	13
Total			**235**		**33**

Mean flow time = 235/5 = 47.0

Average tardiness = 33/5 = 6.6

Number of tardy jobs = 4

8.1.4 *Critical Ratio Scheduling*

After each job has been processed, we compute

$$(\text{Due date} - \text{Current time})/\text{Processing time}$$

which is known as the critical ratio, and schedule the next job to minimize the value of the critical ratio. The idea behind critical ratio scheduling is to provide a balance between SPT, which only considers processing time, and EED, which only considers due dates. The ratio will grow smaller as the current time approaches the due date, and more priority will be given to those jobs with longer processing times. One disadvantage of the method is that the critical ratios need to be recalculated each time a job is scheduled.

The numerator may be negative for some or all of the remaining jobs. When that occurs it means that the job is late, and we will assume that late jobs are automatically scheduled next. If there is more than one late job, then the late jobs are scheduled in the SPT sequence.

First, we compute the critical ratios starting at time $t = 0$.

Job Number	Processing Time	Due Date	Critical Ratio
1	11	61	$61/11 = 5.545$
2	29	45	$45/29 = 1.552$
3	31	31	$31/31 = 1.000$
4	1	33	$33/1 = 33.000$
5	2	32	$32/2 = 16.000$

The minimum value corresponds to job 3, so job 3 is performed first. As job 3 requires 31 units of time to process, we must update all the critical ratios to determine the next job to process. We move the clock to time $t = 31$ and recompute the critical ratios.

Job Number	Processing Time	Due Date	Critical Ratio
1	11	61	$30/11 = 2.727$
2	29	45	$14/29 = 0.483$
4	1	33	$2/1 = 2.000$
5	2	32	$1/2 = 0.500$

The minimum is 0.483, which corresponds to job 2. Hence, job 2 is scheduled next. Since job 2 has a processing time of 29, we update the clock time to $t = 31 + 29 = 60$

Job Number	Processing Time	Due Date	Critical Ratio
1	11	61	$1/11 = 0.091$
4	1	33	$-27/1 = -27$
5	2	32	$-28/2 = -14$

The minimum is -27, thus job 4 is scheduled next, and we update the clock time to $t = 61$

Job Number	Processing Time	Due Date	Critical Ratio
1	11	61	$0/11 = 0$
5	2	32	$-29/2 = -14.5$

The minimum is -14.5 and we schedule 5 next, and finally, job 1 is scheduled last.

Summary of the results for critical ratio scheduling

Sequence	Job Number	Processing Time	Completion Time	Due Date	Tardiness
1	3	31	31	31	0
2	2	29	60	45	15
3	4	1	61	33	28
4	5	2	63	32	31
5	1	11	74	61	13
Total			**289**		**87**

$$\text{Mean flow time} = 289/5 = 57.8$$
$$\text{Average tardiness} = 87/5 = 17.4$$
$$\text{Number of tardy jobs} = 4$$

We summarize the results of this section for all four scheduling rules:

Rule	Mean Flow Time	Average Tardiness	Number of Tardy Jobs
FCFS	53.6	24.2	3
SPT	27.0	8.6	1
EDD	47.0	6.6	4
CR	57.8	17.4	4

As a result, the Shortest Processing Time (SPT) rule emerges as the optimal choice for minimizing the mean flow time of all jobs. Additionally, several criteria, including mean flow time, mean waiting time, and mean lateness, are deemed equivalent, implying that optimizing one of these metrics inherently optimizes the others as well.

8.2 Multiple Machine Sequencing

We now extend the analysis of Section 8.1 to the case in which several jobs must be processed on more than one machine. Assume that n jobs are to be processed through m machines. For each machine, there are $n!$ different ordering of the jobs. If the jobs may be processed on the machines in any order, there are $(n!)^m$ possible schedules.

Tailored solution methods are developed to address the challenges posed by multiple-machine sequencing. Depending on the problem's characteristics, various techniques exist that can produce either optimal or near-optimal results. Evaluation of these scheduling methods typically focuses on two primary criteria: the optimality gap, which measures how closely the solution approximates the optimal one, and the computation (running) time required to generate the schedule. These metrics help assess the effectiveness and efficiency of the different approaches in solving the multiple-machine sequencing problem.

8.2.1 *Scheduling n Jobs on Two Machines*

Assume that n jobs must be processed through two machines and that each job must be processed in the order of machine 1 and then machine 2. Furthermore, assume that the optimization criterion is to minimize the makespan. The problem of scheduling on two machines turns out to have a relatively simple solution.

Theorem: The optimal solution for scheduling n jobs on two machines is always a permutation schedule. That is, the sequences of jobs on both machines will be the same.

Question: Does this theorem help in solving the two-machine problem? Why/Why not?

Because the total number of permutation schedules is exactly $n!$, which is still quite large, determining optimal schedules for two machines is roughly of the same level of difficulty as determining optimal schedules for one machine. Hence we need an efficient algorithm to solve this two-machine problem.

A very efficient algorithm for solving the two-machine problem was discovered by Johnson (1954). Following Johnson's notation, denote the machines by A (first machine) and B (second machine). Suppose that the jobs are labeled i, for $1 \leq i \leq n$. Let A_i be the processing time of job i on machine A and B_i be the processing time of job i on machine B. Johnson's Algorithm to compute the optimal schedule is as follows:

(1) List the values of A_i and B_i in two columns.

(2) Find the smallest remaining element in the two columns. If it appears in column A, then schedule that job next. If it appears in column B, then schedule that job last.

(3) Cross off the jobs as they are scheduled. Stop when all jobs have been scheduled.

Example 8.2

In a job shop, five jobs are waiting to be processed on two machines (in order A-B). The processing times are given in the table below. What is the optimal schedule of jobs to minimize the makespan?

Job	Machine A	Machine B
1	5	2
2	1	6
3	9	7
4	3	8
5	10	4

Solution

The first step is to identify the minimum job time. It is 1, for job 2 on machine A. Because it appears in column A, job 2 is scheduled first and row 2 is crossed out.

The next smallest processing time is 2, for job 1 on machine B. This appears in the B column, so job 1 is scheduled last. The next smallest processing time is 3, corresponding to job 4 in column A, so job 4 is scheduled next (it comes after job 2). Next, we schedule job 5 before job 1 as it has the lowest processing time and is on machine B. Finally, the sequence is 2-4-5-1 and the makespan is equal to 30.

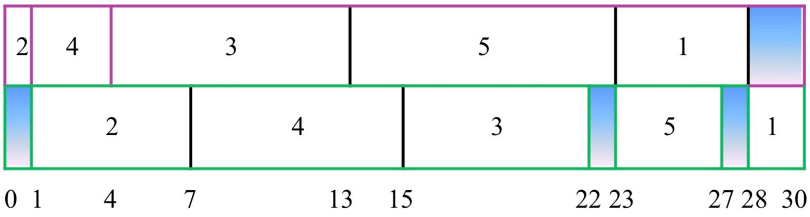


Fig. 8.1 The Gantt chart for the optimal schedule

8.2.2 *Extension to Three Machines*

In the problem setting where n jobs must be processed through three machines, with each job sequentially processed on machine 1, then machine 2, and finally machine 3, the aim is to minimize the makespan. According to the theorem, the optimal solution for scheduling n jobs on three machines will always be a permutation schedule if the objective is to minimize the makespan or total flow time. This means that the sequences of jobs processed on each machine will be identical. However, it's important to note that while permutation schedules are optimal for minimizing makespan or total flow time, they may not necessarily be optimal when considering average flow time as a criterion.

Denote the machines by A (first machine), B (second machine), and C (third machine). Let A_i be the processing time of job i on machine A, B_i be the processing time on machine B, and C_i be the processing time on machine C. We can apply modified Johnson's algorithm which reduces the three-machine problems to (essentially) a two-machine problem if the following condition is satisfied:

$$\min A_i \geq \max B_i \quad \text{or} \quad \min C_i \geq \max B_i$$

It is only necessary that *either one* of these conditions be satisfied. If that is the case, then the problem is reduced to a two-machine problem in the following way.

Define $A'_i = A_i + B_i$, and define $B'_i = B_i + C_i$. Now solve the problem using the rules described for two machines, treating A'_i and B'_i as the processing times.

Important 8.1

In a job shop, five jobs are waiting to be processed on three machines (in order A-B-C). The processing times are given in the table below. What is the optimal schedule of jobs to minimize the makespan?

Job	Machine A	Machine B	Machine C
1	4	5	8
2	9	6	10
3	8	2	6
4	6	3	7
5	5	4	11

Solution

Checking the conditions, we find

$$\min A_i = 4, \ \max B_i = 6, \ \min C_i = 6$$

so that required condition is satisfied. We now form the two columns A' and B'.

Job	Machine A'	Machine B'
1	9	13
2	15	16
3	10	8
4	9	10
5	9	15

The problem is now solved using the two-machine algorithm. The optimal solution is 1-4-5-2-3. Note that because of ties in column A, the optimal solution is not unique. In other words, we have alternative optimal solutions.

8.3 Assembly Line Balancing

Assembly line balancing involves managing a set of n tasks to be completed during each cycle. These tasks are assigned to stations, and they must be sequenced properly, with certain tasks potentially excluded from certain stations. Additionally, tasks may have precedence relationships, meaning that specific tasks must be completed before others can begin. The primary objective of assembly line balancing is to assign tasks to stations in a way that minimizes the cycle time, denoted as C. While solving the general problem optimally is challenging, effective heuristics are available to address it.

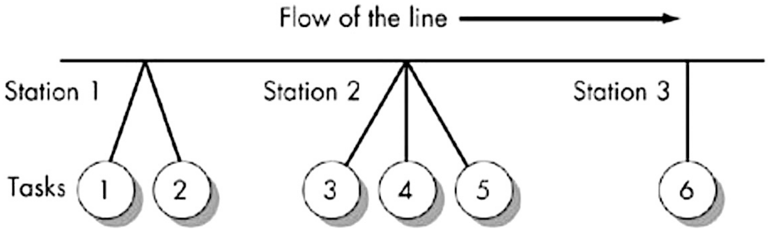


Fig. 8.2 Schematic of a typical assembly line (Credit: [Nahmias, 1997])

Let $t_1,\ t_2,\ \ldots,\ t_n$ be the time required to complete the respective tasks. The total work content associated with the production of an item, say T, is given by

$$T = \sum_{i=1}^{n} t_i$$

For a cycle time of C, the minimum number of workstations possible is $[T/C]$, where the brackets indicate that the value of T/C is to be rounded to the next larger integer. Because of the discrete and indivisible nature of the tasks and the precedence constraints, it is often true that more stations are required than this ideal minimum value. If there is leeway in the choice of the cycle time, it is advisable to experiment with different values of C to see if a more efficient balance can be obtained.

We will present one heuristic method known as the ***ranked positional weight technique***. The method places a weight on each task based on the total time required by all the succeeding tasks. Tasks are assigned sequentially to stations based on these weights. We illustrate the method by example.

Important 8.2

The final assembly of Noname personal computers, a generic mail-order PC clone, requires a total of 12 tasks. The assembly is done at the Lubbock, Texas, plant using various components imported from the Far East. The tasks required for the assembly operations are

(1) Drill holes in the metal casing and mount the brackets to hold disk drives.
(2) Attach the motherboard to the casing.
(3) Mount the power supply and attach it to the motherboard.
(4) Place the main processor and memory chips on the motherboard.
(5) Plug in the graphics card.
(6) Mount the DVD burner. Attach the controller and the power supply.
(7) Mount the hard disk drive. Attach the hard disk controller and the power supply to the hard drive.
(8) Set switch settings on the motherboard for the specific configuration of the system.
(9) Attach the monitor to the graphics board before running system diagnostics.
(10) Run the system diagnostics.
(11) Seal the casing.
(12) Attach the company logo and pack the system for shipping.

The job times and precedence relationships for this problem are summarized in the following table. The network representation of this particular problem is given in the figure below.

Task	Predecessor	Time
1	–	12
2	1	6
3	2	6
4	2	2
5	2	2
6	2	12
7	3, 4	7
8	7	5
9	5	1
10	6, 9	4
11	8, 10	6
12	11	7

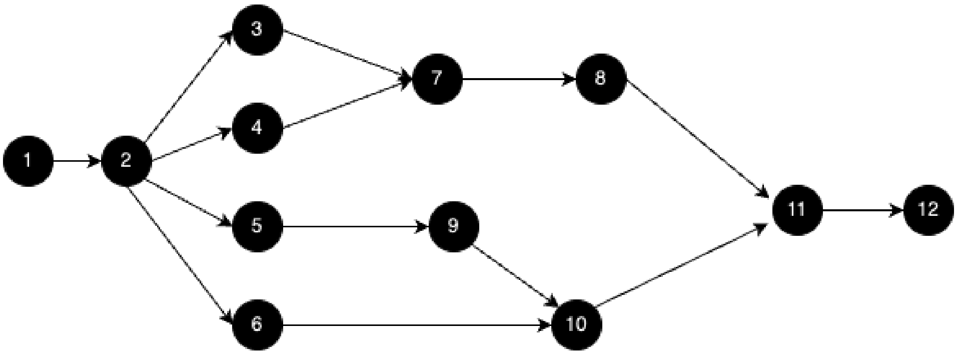


Fig. 8.3 Precedence constraints for Noname computer

Solution

Suppose that the company is willing to hire enough workers to produce one assembled machine every 15 minutes. The sum of the task times is 70, which means that the minimum number of workstations is the ratio $70/15 = 4.67$ rounded to the next larger integer, which is 5. This does not mean that a five-station balance necessarily exists.

The solution procedure requires determining the positional weight of each task. The positional weight of task i is defined as the time required to perform task i plus the times required to perform all tasks having task i as a predecessor. Remember that we include the times of the task of all the successors not only the immediate successor tasks.

As task 1 must precede all other tasks, its positional weight is simply the sum of the task times, which is 70. Task 2 has positional weight 58[1]. From Figure 8.3 we see that task 3 must precede tasks 7, 8, 11, and 12 so that the positional weight of task 3 is $t_3 + t_7 + t_8 + t_{11} + t_{12} = 31$. The other positional weights are listed in the table below.

Task	Predecessor	Time	Weight
1	–	12	70
2	1	6	58
3	2	6	31
4	2	2	27
5	2	2	20
6	2	12	29
7	3, 4	7	25
8	7	5	18
9	5	1	18
10	6, 9	4	17
11	8, 10	6	13
12	11	7	7

The next step is to rank the tasks in order of decreasing positional weight. The ranked tasks are given in the table below. Finally, the tasks are assigned sequentially to stations in the ranking order, and assignments are made only as long as the precedence constraints are not violated.

[1] Pay attention to the nodes that appear in multiple paths. The processing time on those nodes should not be added more than once.

Task	Predecessor	Time	Weight
1	–	12	70
2	1	6	58
3	2	6	31
6	2	12	29
4	2	2	27
7	3, 4	7	25
5	2	2	20
8	7	5	18
9	5	1	18
10	6, 9	4	17
11	8, 10	6	13
12	11	7	7

Let us now consider the balance obtained using this technique assuming a cycle time of 15 minutes. Task 1 is assigned to station 1. That leaves a slack of three minutes at this station. However, because task 2 must be assigned next, to avoid violating the precedence constraints, and the sum $t_1 + t_2$ exceeds 15, we close station 1. Tasks 2, 3, and 4 are then assigned to station 2, resulting in an idle time of only one minute. Continuing in this manner we obtain the following balance for this problem:

Station	Tasks	Time	Idle Time
1	1	12	3
2	2, 3, 4	14	1
3	5, 6, 9	15	0
4	7, 8	12	3
5	10, 11	10	5
6	12	7	8

Notice that although the minimum possible number of stations for this problem is five, the ranked positional weight technique results in a six-station balance. As the method is only a heuristic, there may be a solution with 5 stations. In this case, however, the optimal balance requires six stations when $C = 15$ minutes.

The head of the firm assembling Noname computers is interested in determining the minimum cycle time that would result in a five-station balance. If we increase the cycle time from $C = 15$ to $C = 16$, then the balance obtained is

Station	Tasks	Time	Idle Time
1	1	12	4
2	2, 3, 4, 5	16	0
3	6, 9	13	3
4	7, 8, 10	16	0
5	11, 12	13	3

There is a much more efficient balance. The total idle time has been cut from 20 minutes per unit to only 10 minutes per unit. The number of stations decreases by 16 percent, while the cycle time increases by only about 7 percent.

8.4 Average Flow Time

Here we outline the rationale for prioritizing the minimization of average flow time over other KPIs such as makespan or tardiness within a company context. While various approaches are valid and there is no definitive right or wrong, we delve into the justification for this particular focus.

The significance of average flow time lies in its direct correlation with maximizing throughput. Consequently, in companies aiming to optimize throughput, efforts are directed towards reducing average flow time. This principle finds its roots in queueing theory, a branch of stochastic modeling that examines waiting lines or queues.

Little's Law is a fundamental principle at the core of queueing theory. This law establishes a relationship between three essential metrics in queueing systems: $L,\ \lambda,\ W$.

Formally, it is expressed as:

$$L = \lambda \times W$$

Here, L represents the average number of customers in the system, λ denotes the average arrival rate of customers (i.e., the average number of arrivals per unit time), and W signifies the average time a customer spends in the system, also referred to as the average waiting time or residence time.

It essentially posits that the average number of customers in a queuing system equals the product of the average arrival rate of customers and the average time each customer spends in the system.

Little's Law also finds application across diverse queuing systems, spanning customer service centers, telecommunications networks, computer systems, and manufacturing processes. It furnishes valuable insights into system dynamics and performance, empowering analysts to make data-driven decisions regarding capacity planning, resource allocation, and process enhancement.

Bibliography

S. Nahmias. *Production and Operations Analysis.* Richard D. Irwin, Burr Ridge, Illinois, 3rd edition, 1997.