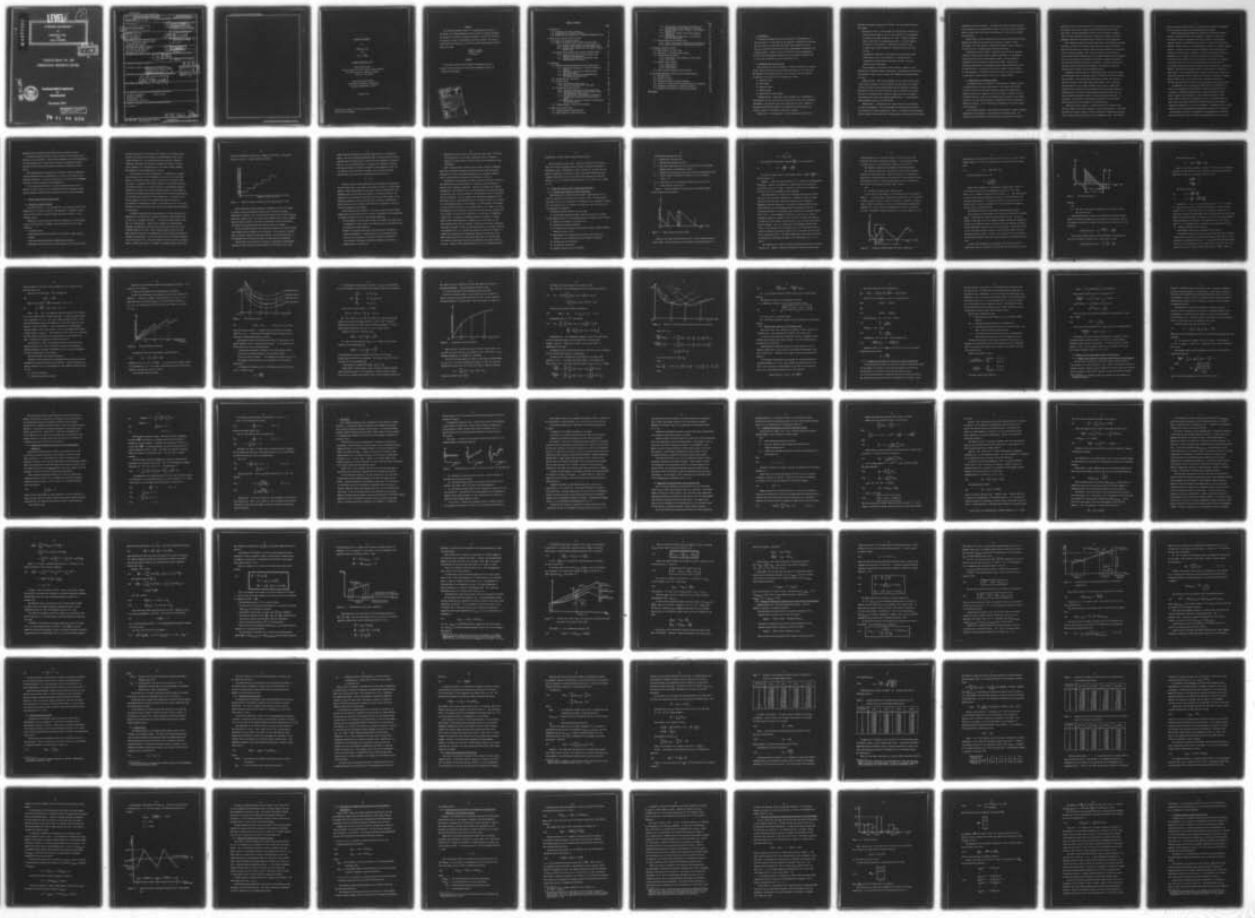
AD-A077 562 MASSACHUSETTS INST OF TECH CAMBRIDGE OPERATIONS RESE--ETC F/G 15/5
INVENTORY MANAGEMENT.(U)
NOV 79 A C HAX , D I CANDEA N00014-75-C-0556
UNCLASSIFIED TR-168 NL

1 OF 3
ADA 077 562



AD A077562

LEVEL II

INVENTORY MANAGEMENT

by

ARNOLDO C. HAX
and
DAN I. CANDEA

DDC
RECEIVED
DEC 3 1979
E

Technical Report No. 168

OPERATIONS RESEARCH CENTER

DDC FILE COPY



MASSACHUSETTS INSTITUTE
OF
TECHNOLOGY

November 1979

This document has been approved for public release and sale; its distribution is unlimited.

79 11 30 032

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 168
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE *(and Subtitle)*: INVENTORY MANAGEMENT

5. TYPE OF REPORT & PERIOD COVERED: Technical Report November 1979

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): Arnoldo C. Hax, Dan I. Candea

8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0556

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
M.I.T. Operations Research Center
77 Massachusetts Avenue
Cambridge, MA 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 347-027

11. CONTROLLING OFFICE NAME AND ADDRESS:
O.R. Branch, ONR Navy Dept.
800 North Quincy Street
Arlington, VA 22217

12. REPORT DATE: November 1979

13. NUMBER OF PAGES: 193 pages

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: 199

15. SECURITY CLASS. *(of this report)*:

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT *(of this Report)*

Releasable without limitation on dissemination.

This document has been approved for public release and sale; its distribution is unlimited.

DDC RECEIVED DEC 3 1979

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*: TR-168

18. SUPPLEMENTARY NOTES:

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Inventory Management
Inventory Systems
Forecasting
Economic Order Quantity
Safety Stocks

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

See page ii.

270 720

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601 |

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

INVENTORY MANAGEMENT

by

ARNOLDO C. HAX

and

DAN I. CANDEA

Technical Report No. 168

Work Performed Under
Contract N00014-75-C-0556, Office of Naval Research
Multilevel Logistics Organization Models
Project No. NR 347-027

Operations Research Center
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

November 1979

Reproduction in whole or in part is permitted for any purpose of the United States Government.

FOREWORD

The Operations Research Center at the Massachusetts Institute of Technology is an interdepartmental activity devoted to graduate education and research in the field of operations research. The work of the Center is supported, in part, by government contracts and grants. The work reported herein was supported by the Office of Naval Research under Contract N00014-75-C-0556.

Richard C. Larson
Jeremy F. Shapiro
Co-Directors

ABSTRACT

This paper surveys the most important mathematical models to support decisions dealing with the Inventory Management process in a production environment.

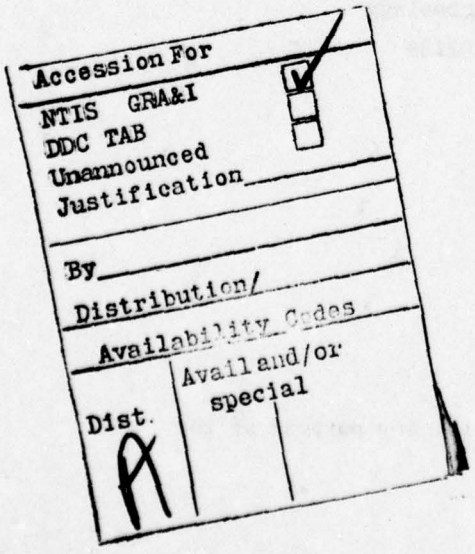


Table of Contents

		Page
1	Introduction	1
1.1	Inventories and Their Functions	1
1.2	Classification of Inventory Systems	3
1.3	Objectives and Structure of an Inventory Management System	6
2	Economic Order Quantity Decision Rules	8
2.1	Costs in an Inventory System	8
2.2	Single Item Economic Order Quantity Decision Rules	13
2.2.1	Economic Order Quantity for Fast Moving Items	13
2.2.2	Economic Order Quantity for Slow Moving Items	29
2.2.3	Economic Order Quantity for Items with a Limited Sales Period	30
2.3	Multiple Items Economic Order Quantity Decision Rules	32
2.3.1	Economic Order Quantity for Multiple Items Sharing the Same Equipment	32
2.3.2	Economic Order Quantity for Multiple Items Under Aggregate Constraints	36
3	Forecasting	39
3.1	Exponential Smoothing Methods for Fast Moving Items	42
3.1.1	Exponential Smoothing for the Base (Constant) Pattern	43
3.1.2	Exponential Smoothing for Demand Patterns with Linear Trend	49
3.1.3	Exponential Smoothing for Demand Patterns with Trend and Seasonalities	57
3.2	Forecasting Over Lead Times	64
3.3	Forecast Errors	65
3.3.1	One Period Forecast Errors	66
3.3.2	Errors in Forecasting Over Lead Times	68
3.4	Forecasting Slow Moving Items	77
3.4.1	Exponential Smoothing for Slow Moving Items	80
3.4.2	Forecasting with Demand Transactions Size and the Time Between Transactions	81
3.4.3	Alternative Ways of Estimating the Parameters of Theoretical Probability Distributions of Demand	82
3.4.4	Empirically Determined Probability Distributions of Lead Time Demand	85
3.5	Tracking Signals to Monitor the Forecasts	89
4	Safety Stock Decision Rules	92
4.1	The ABC Inventory Classification	93
4.2	Safety Stocks for Fast Moving Items	100

Page

4.2.1 Setting Safety Stocks When the Stockout Cost is Proportional to the Number of Units Short 102
4.2.2 Setting Safety Stocks to Achieve a Prespecified Service Level 107
4.2.3 Allocation of Safety Stocks Under Aggregate Constraints 115
4.2.4 Simultaneous Determination of Safety Stocks and Order Quantities 118

4.3 Safety Stocks for Slow Moving Items 123

4.3.1 Safety Stocks Based on Theoretical Distributions 124
4.3.2 Safety Stocks Based on Empirical Distributions 131

5 The System Integrative Module 133

5.1 Made-to-Order vs. Stock Items 133

5.2 Continuous vs. Periodic Review Systems 137

5.3 Continuous Review System 142

5.3.1 Order Point - Order Quantity, (s,Q), Policy 142
5.3.2 (s,S) Policy 144

5.4 Periodic Review Systems 146

5.4.1 (nQ,s,R) Policy 146
5.4.2 (S,R) Policy 146
5.4.3 (s,S,R) Policy 154

5.5 Other Issues in Inventory Control Systems 161

6 The System-Management Interaction and Evaluation Module 171

6.1 Exchange Curves 171

6.2 How Much to Order? 174

6.3 Updating Frequency of System Parameters 175

6.4 Actions to be Taken When the Tracking Signal is Triggered 176

6.5 Management Adjustments to Statistical Forecasts 179

6.6 Production and Inventory Control System Outputs 181

Bibliography 186

1 Introduction

Production planning has among its objectives the determination of inventory levels. In this paper we are primarily concerned with inventories that are involved in inudstrial production, namely inventories of raw materials, purchased and manufactured parts, subassemblies, assemblies, and finished products. However, many of the decision rules presented below are also valid for managing inventories in other kinds of operations such as retailing, distribution, service operations, etc.

1.1 Inventories and their Functions

Since inventories normally represent a sizable investment in a logistic system, legitimate questions can be raised with respect to the causes for the existence of inventories as well as the functions that they perform. In general, one can think of five categories of stocks:

1) Pipeline stocks
2) Cycle stocks
3) Seasonal stocks
4) Safety stocks
5) Stock held for other reasons.

Pipeline stocks. Inventories in this category are a consequence of the finiteness of the production and transportation rates in any industrial environment. The pipeline stock, also called process stock, consists of materials actually being worked on, or moving between work centers, or being in transit to distribution centers and customers.

Cycle stocks. In the majority of cases industrial production and

materials procurement takes place in batches. This can happen because of two reasons:

- Economies of scale: if the average cost of producing, purchasing, or moving stock decreases as the lot size increases it is advantageous to operate with larger quantities at a time. A typical example is when a fixed setup or an administrative cost is incurred whenever an item has to be produced or ordered from an outside vendor. A larger order quantity results in a reduced fixed cost per unit of item.
- Technological requirements: the design of the process may impose certain batch sizes. For instanced, in a chemical reactor processing by tankfuls might be necessary in order to achieve desired reaction parameters.

Cycle inventory, also called lot size inventory, exhibits a time behavior that alternates between a high point, corresponding to the delivery of the batch to stock, and a low point that immediately precedes delivery to stock.

Seasonal stocks. When the requirements for the items vary with time, it may become economical to build inventory during periods of low demand to ease the strain of peak demand periods upon the production facilities. The extent to which this policy should be used is determined by balancing the cost of carrying seasonal inventories against the cost of changing the production rate and of not meeting demand entirely. This problem has been extensively dealt with in Hax [1978].

Safety stocks. Inventories may be carried because of uncertainties of future requirements. Future requirements are estimated by forecasting but forecasts are always accompanied by errors. If planning is done disregarding the possible forecasting errors, shortages may be incurred when materialized

requirements exceed the forecast. To prevent the losses normally associated with shortages, safety stocks have to be held in the form of extra inventories above the level that would result from planning on the basis of the demand forecasts alone.

Safety stocks can offer protection not only against demand uncertainties. Whenever the quantities delivered vary from what is ordered, or when procurement lead times exhibit a probabilistic behavior, safety stocks are an effective tool in hedging against supply uncertainties.

Stock held for other reasons. Inventories can perform the important function of decoupling the various stages of production. By deliberately creating stocking points between adjacent stages, a certain degree of independence can be achieved in operating the stages. Without such decoupling inventories, any disturbance at some stage would shortly affect the entire system. Stocks may be carried for a number of other seasons: to take advantage of a favorable raw material price or quantity discounts, to anticipate an expected rise in price, etc. (Morgan [1963]).

1.2 Classification of Inventory Systems

An inventory system is composed of a large number of elements and has to perform functions of major significance to the company (Hax [1976]). In order to prescribe an inventory system to support the organization's logistics process it is important in the first place to identify the type of system called for, based on the elements that are present in the product structure of the firm and the degree of complexity involved in making the logistics decisions. To this purpose four categories of inventory systems can be identified:

Pure Inventory Systems. These systems are intended to support decisions regarding the replenishment of inventories for individual items. The decision rules associated with these systems are statistically based and

specify for each item an order point (that determines when the item should be ordered), and an order quantity (that determines how much to order). Each item is treated independently from any other item except, perhaps, to allow for joint ordering of families of items, and to account for simple aggregate constraints reflecting storage, financial or other limitations.

Pure inventory systems are normally applicable to raw material purchasing decisions, and retail or wholesale activities, where items are purchased from outside vendors and, possibly, minor production operations are performed such as cutting, packaging. Pure inventory systems can also be used to control production of finished goods in very simple manufacturing environments which are not affected by significant fluctuations in demand requirements and where ample production capacity is available. As these conditions are rarely met in most productions environments, pure inventory systems are basically used to support only purchasing decisions.

Production-Inventory Systems. These systems apply to situations where the firm manufactures the finished products internally rather than procuring them externally. The manufactured items normally compete for production capacity; therefore simple order point-order quantity rules, that ignore item interactions, are no longer effective control tools. Higher level decisions have to be made for the allocation of scarce resources among the competing items. The specific methodologies vary significantly with the type of production process involved in the manufacturing activities. In particular, a fabrication or intermittent process has to be controlled quite differently from an assembly or continuous process.

Obviously, the development of a production-inventory system is much more of a complex task than the design of a pure inventory system. Hax [1978] has given an in-depth coverage of various approaches to modelling the allocation of production capacity and labor at the aggregate level. Hax and Meal

[1975] and Bitran and Hax [1977] have considered the integration of aggregate and detailed level decisions in production-inventory systems.

Distribution-Inventory Systems and *Production-Distribution-Inventory Systems.* The added feature, the distribution, involves the allocation of available inventory (purchased from outside vendors in distribution-inventory systems, or manufactured internally in a production-distribution-inventory system) among a set of stocking points located in a possibly complex network. In practice, sound applicable management science support is not available for these types of decisions. There are two basically different procedures that are used to deal with the inventory allocation aspect. In the first one, referred to as a push system, the allocation of inventory is decided centrally for the whole system taking into account all the distribution requirements and stock availabilities throughout the distribution network. Mathematical programming models are instrumental in supporting push systems. By the second procedure, referred to as the pull system, the individual warehouses generate requests for inventory replenishment independently, based on their own inventory status and demand requirements. Statistically based inventory models have been associated with pull systems.

An important issue, that has been hardly tackled by researchers, is that of the integration of distribution, production and scheduling into a coherent, coordinated planning system. Although attempts have been made to approach this problem (e.g.: Hax [1973], Hax and Meal [1975], Krauss [1977]), answers of a more general nature, to questions like: how and at what level should distribution planning interact with production planning and scheduling, are yet to be provided. As Karmarkar [1975, p. 199] points out, while the hierarchical framework holds promise for a solution to the problem, the issues of aggregation and disaggregation of items may prove to be difficult since the bases for aggregation may not be the same

for production and distribution.

In this paper we are primarily concerned with pure inventory systems.

1.3 Objectives and Structure of An Inventory Management System

As Brown [1978a, p. 173] mentioned, there have been two streams of development in the field of inventory operations. One is represented by mathematical abstractions of the inventory system in which the major effort went into modelling the process and searching for optimal policies in terms of minimizing relevant costs. There is a large and still growing literature on inventory theory, in which papers such as Arrow, et al [1953], Dvoretzky, et al. [1952, 1953], and books like Whitin [1953], Arrow, et al. [1958], Scarf, et al. [1963], Hadley and Whitin [1963] have already become classics. The interested reader can find good starting references in any operations management book written with an operations research bias (e.g.: Johnson and Montgomery [1974], Zimmermann and Sovereign [1974]). The other school of thought is primarily concerned with practical issues such as demand and costs measurement, system design, relations among logistics and other industrial management functions, system management, etc. (Magee [1968], Magee and Boodman [1967], Brown [1967, 1977].

In our view, an inventory system can be defined as a coordinated set of rules and procedures that allow for routine decisions on when and how much to order of each item needed in the manufacturing or procurement process to fill customer demand, that call attention to the non-routine situations the rules do not cover, and that provide managers with the necessary information to make these decisions effectively. The objective of a well-designed procedure should be the minimization of the costs incurred in the inventory system, attaining at the same time the customer service level

specified by the company policies. Therefore, the inventory management problem is treated in this chapter from an implementable system perspective. While the theoretical developments have to be given full credit for the insights they bring into the problem and for the establishment of the form of optimal policies, the design of the system has to rely on decision rules that represent some implementable form of the theoretically derived optimal policies.

An inventory management system can be viewed as being structured of subsystems or modules:

- the transactions and file maintenance module,
- the decision rules module,
- the system integrative module,
- the system-management interaction and evaluation module.

The transactions and file maintenance module concerns the bookkeeping of inventory control, namely the entry, auditing, control and processing of inventory transactions, as well as the file maintenance functions. Files have to be continuously updated in order to provide accurate information on available stock (on hand and on order) and opten customer order status. The data base should be posted for any changes that may occur in the cost of items, delivery lead times, source of acquisition, ordering restrictions, etc. "The major systems-design problem is to assure the efficiency and reliability of the system" (Brown [1978, p. 174]).

The decision rules module is concerned with the fundamental components of inventory planning and control procedures. The main decision rules are aimed at answering WHEN and HOW MUCH to order of each item in order to maintain inventories at the "right" level. At the same time, any forward-looking system should also include FORECASTING capabilities, and because of the unavoidable forecast inaccuracies SAFETY STOCKS decision rules are

needed in order to guarantee some desired level of customer service.

The system integrative module brings the decision rules together into distinct inventory policies. Various items, depending on their characteristics as stock-keeping units, require specific degrees of management attention and service levels that can be achieved by using some appropriate stock policy.

The system-management interaction and evaluation module is intended to provide management with such information as to permit an evaluation of the operating performance, to identify problem areas, and to allow for management selection of policy variables (system parameters).

In the sections that follow, we present in some detail each of the modules with the exception of the transactions and file maintenance module whose development concerns almost exclusively the area of data processing.

2 Economic Order Quantity Decision Rules

2.1 Costs in an Inventory System

As stated in section 1.3, the objective of an inventory system is the minimization of the costs involved in the operations. Therefore, at this point we find it useful to briefly discuss the costs relevant to a pure inventory system.

Obviously, the only costs that should be considered are the ones which vary as the stock policy is changed. We can group these costs in three categories:

- procurement costs,
- costs associated with the existence of inventories (supply exceeds demand),
- costs associated with stockouts (demand exceeds supply).

Procurement costs can be seen as being composed of two parts: the cost

that has to be paid by the system to the supplier of the ordered items, and the cost incurred by the system in the procurement process, also called ordering cost. The ordering cost itself can have a number of components: administrative (paper work computer processing, telephone calls, postage, etc.), transportation of the amount ordered, handling and inspection of the shipment when it arrives.

For the purpose of future development, it is important to re-group the procurement costs into two categories: costs per unit of item that depend on the amount ordered, and unit costs that do not depend on the order size. An example of the former category is the case where the cost of ordering an amount Q is A for any value of Q, and therefore the unit ordering cost is A/Q; this is highly relevant because it thus becomes an incentive for purchasing the item in large batches with the obvious purpose of lowering the per unit share of the ordering cost. An instance of the latter category of costs is the constant purchasing unit price (i.e. no economies or diseconomies of scale). Since these kinds of costs cannot influence the decision on order size, they can be dropped from consideration in the analysis.

It should be mentioned that the assumption of a fixed ordering cost, regardless of the order size, is common in inventory systems research and we also use it throughout the paper. We should be aware, however, that this is an approximation. Indeed, if we consider the transportation, handling, and inspection costs per batch, they are invariant only within a limited range of lot sizes. With a radical change in order size one might reconsider the mode of transportation and the handling and inspection equipment and techniques. On the other hand, part of the annual administrative costs does not vary if the total number of orders stays within certain limits. Beyond those limits, the number of purchasing agents, phone lines,

the cost of supporting services, etc. changes, which leads to a cost chart as shown in Figure 1 (Hadley and Whitin [1963]).

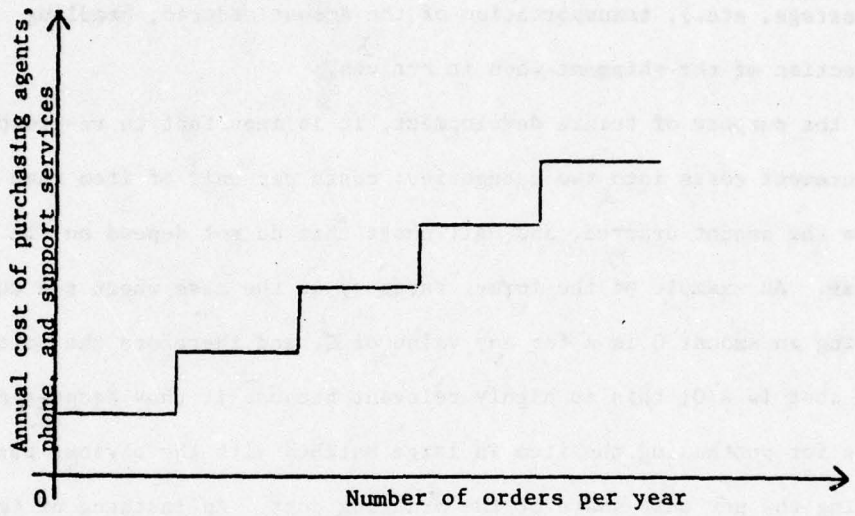


Figure 1: Effect of number of orders upon some administrative costs

Costs associated with the existence of inventories are due to a number of causes: storage and handling, property taxes, insurance, spoilage, obsolescence, pilferage, rent if the inventory system does not own the storing facilities, capital costs. The capital cost represents either direct expenditures for funds (interest) or the rate of return that could be obtained by the system by investing elsewhere the capital tied up in inventory.

Of all the above components only those which change as the level of inventory changes should be brought into the analysis. For instance, the amounts spent on heating, lighting, and security services for the warehouse tend to be invariable with the stock level and if so they should be disregarded.

The diversity in the inventory carrying cost components is undoubtedly reflected in their functional relationships to the inventory level, which

creates serious difficulties in modelling these costs in a satisfactory manner. The usual simplifying assumption made in inventory theory is that carrying costs are proportional to the size of the investment in inventory. Thus, if r is the carrying charge (or holding cost) expressed in dollars per year per dollar of inventory investment (or percentage per year), and C is the unit cost of the item in dollars, then the annual inventory carrying cost H for that item, in dollars per year per unit is:

$$H = rC$$

A special class of costs associated with the existence of inventories are the salvage costs. Such costs occur when,at the end of a limited sales season or after some operation is shut down, there is inventory left over. Depending on the action taken with respect to the excess inventory, the salvage cost can be either the carrying cost until the next season, or the difference between the cost of the item to the inventory system and the price at which it can be disposed (this price may be negative if there is a cost to disposal of the surplus).

Costs associated with stockouts. A stockout situation arises whenever demand occurs and the system is out of stock. Depending on the circumstances, a stockout may result in one of the following conditions:

- To meet demand a priority special order is released; the stockout cost is represented in this case by the additional cost of the special order as compared to normal operation.
- Demand is backordered and filled when stock becomes available by routine replenishment. Stockout costs of a less tangible nature occur in this situation such as loss of customer's goodwill, or the lowering of the degree of military readiness if a military supply system is involved, etc.

- Demand which cannot be met is lost (the lost sales case). The stock-out costs would have to account, besides the loss of customer's goodwill, for lost profit on the units which were requested but were unavailable.

Other costs may apply to either of the stockout situations described above such as penalties stemming from failure to meet legal contractual obligations, the cost of getting information on the customer's belated order, contacting him and relaying the information to him, etc.

The problem of quantifying the stockout costs has long been a difficult and unsatisfactorily resolved question in inventory theory, especially because of the intangible components. There are two aspects that require consideration: the functional form of the mathematical expression describing the stockout costs, and the estimation of the parameters once the functional form is established. The most largely used simplifying assumption is that the stockout cost is proportional either to the number of units out of stock, or to the maximum duration of the shortage, or to the product of the number of units times the duration of the stockout (Holt, et al. [1960, ch. 12-2]). The simplest but not necessarily a good way to assess the unit stockout cost is to consider it equal to the lost profit, or to the cost of expenditure a rush order. If, however, a stockout influences customers to transfer some of their ensuing business to competitors, the lost profit on the later business should also be captured. Along this line and by considering a decision tree model in which all possible outcomes of a stockout situation are included, Oral, et al. [1972] have conducted a statistical study for a company with the objective of determining expected unit shortage costs. Their results show that, for the company under study, the unit shortage cost bears an exponential functional relationship to the gross profit for the item. Schwartz [1966,1970] approaches the loss of customer's goodwill

by modelling its effect upon the future demand pattern.

For the purpose of this paper costs are considered time invariant. Also, given that the planning horizon relevant to inventory problems is sufficiently short, discounting of costs to account for the time value of money is unimportant, and hence not used. The reader interested in more extensive discussions on costs can consult such references as: McGarrah [1963, ch. 1.4, 5.4], Hadley and Whitin [1963, ch. 1], Arrow, et al. [1958, ch. 2].

2.2 Single Item Economic Order Quantity Decision Rules

2.2.1 Economic Order Quantity for Fast Moving Items

When dealing with stocked items the question regarding how much to order is answered by the *economic order quantity* (EOQ). The EOQ provides the proper lot size for purchasing by minimizing the cost components involved, i.e. the ordering cost, the inventory carrying cost, and the stockout cost (if shortages are permitted).

The Simple Classical Economic Lot Size Model

If the following assumptions apply, then what has come to be known as the standard Wilson lot size formula provides the economic order quantity:

a) Demand is continuous at a constant rate.

b) The process continues infinitely.

c) No constraints are imposed (on quantities ordered, storage capacity, available capital, etc.).

d) Replenishment is instantaneous (the entire order quantity is received all at one time as soon as the order is released).

e) All costs are time invariant.

f) No shortages are allowed.

g) Quantity discounts are not available.

The following notations are used:

D = demand rate, units per year

A = ordering cost, dollars per order

C = unit cost, dollars per unit of item (the value of the item immediately after it is delivered to stock)

r = inventory carrying charge, dollars per dollar of inventory per year

H = annual inventory holding cost, dollars per unit of item per year; H = rC

TC = total annual cost of operating the system, dollars per year

Q = order quantity, units per lot.

Figure 2 shows the behavior of an inventory system which operates by the assumptions listed above.

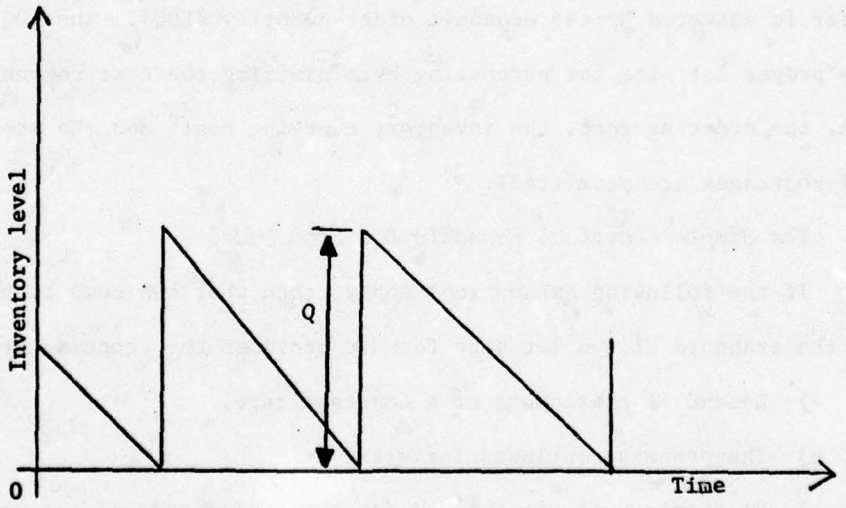


Figure 2: Simple classical inventory model

There are D/Q orders placed during one year, and the average inventory is Q/2; hence, the total annual cost of ordering and carrying inventory is:

$$TC = A\frac{D}{Q} + rC\frac{Q}{2}$$

The optimum Q^* is obtained by setting $\frac{d(TC)}{dQ} = 0$; the solution is

$$(1) \qquad Q^* = \sqrt{\frac{2AD}{rC}} = \sqrt{\frac{2AD}{H}}$$

It is easy to check that Q^* is the global minimum. Indeed, $\frac{d^2(TC)}{dQ^2} > 0$ for any finite $Q > 0$.

Formula 1 was first derived in 1915 by F. W. Harris of the Westinghouse Corporation. However, the expression is widely known as the Wilson lot size formula since R. H. Wilson, a consultant, has used such a formula in his work on inventory management in many companies.

The annual demand D is obtained by forecasting based upon the demand history for the item under consideration. Parameters A and H are either derived by some statistical estimation technique or are inferred by management from past experience. It is shown by Solomon [1959] that the total inventory cost in the neighborhood of the optimum lot size is relatively insensitive to moderately small variations in the amount ordered. Brown [1977, p. 212] recommends as practical to round off lot sizes to reasonable values; his argument is that if the lot is within the range 70 to 140% of the true optimum, the total annual costs rise less than 6% above the true minimum. Also, by a sensitivity test Zimmermann and Sovereign [1974, p. 360] conclude that the sensitivity of total cost with respect to errors in ordering and inventory holding costs is very small if the errors are in the same direction. This makes unnecessary a great deal of accuracy in estimating the parameters involved in the calculations (i.e. D, A, r, and C).

The ordering cost A used in the EOQ calculations should be the marginal ordering cost. However, inventory carrying costs should be based on the

total purchasing cost of the item, since it is this total cost that determines the capital invested in inventory. In the American economy a typical value of r used in practice ranges from 0.20 to 0.25.

The Classical Economic Lot Size Model with Finite Supply Rate

This instance represents a relaxation of assumption (d) mentioned earlier, all others being kept. Thus, for instance, the item might be supplied by a packaging machine that operates at a finite production rate. The supplying process is continuous and takes place at a constant rate until Q units are delivered to stock, then it stops.

Let:

P = production (supply) rate, units per year.

In this case the time behavior of the inventory level is depicted in Figure 3. In this picture T represents the cycle time; t_s is the period of time over which a batch of size Q is supplied to inventory. During t_s stock is input to inventory at a rate P and simultaneously removed from inventory at a rate D; thus stock accumulates at a rate of (P-D) units per year. The highest inventory position is $I = t_s(P-D)$, where t_s is the

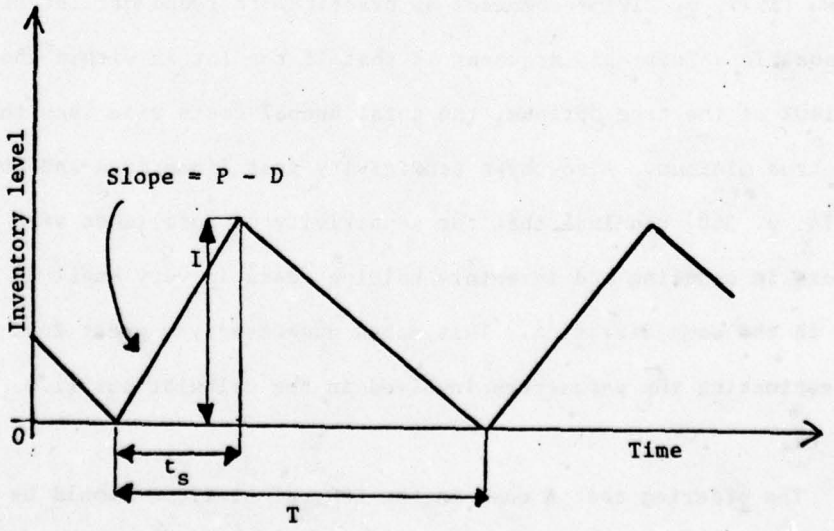


Figure 3: Classical inventory model with finite supply rate

time necessary to process the Q units at a rate of P, i.e. $t_s = Q/P$. As the average inventory is I/2 it is a simple matter to express the total annual cost:

$$TC = A\frac{D}{Q} + H\frac{Q}{2}\left(1 - \frac{D}{P}\right) \tag{2}$$

The global minimum of TC is Q^*:

$$Q^* = \sqrt{\frac{2AD}{H\left(1 - \frac{D}{P}\right)}}$$

Notice that by making $P = \infty$ (instantaneous replenishment formula (1) of the simple classical case is immediately recovered. On the other hand, if D = P (demand and supply rates are equal) $Q \to \infty$. The interpretation is that in this case the supplying source has to be fully devoted to the item under consideration. No inventory builds up, and there is just one initial setup or ordering; consequently, the cost minimization requires the lot size to be extremely large.

The Classical Economic Lot Size Model with Backlogging Allowed

For this case we relax assumption (f) and keep the other ones unchanged. Figure 4 describes the inventory level variation in time. The stock positions range from a low of -B (amount of demand deliberately unsatisfied and put on the backorder list) to a high of Q-B which represents the amount on hand immediately after a lot of size Q is delivered. Notice that B units out of Q are never carried in stock; as soon as delivery takes place the backlog of orders is filled. Clearly, a low stockout cost is an incentive for backordering demand, since this yields savings on the inventory holding cost.

Assume the backordering cost incurred by the inventory system is proportional both to the number B of units short and the duration t_b of the

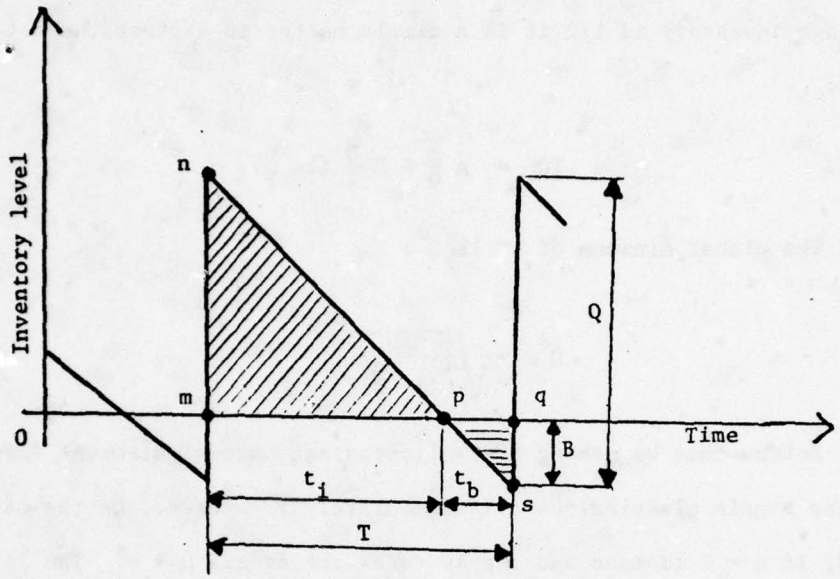


Figure 4: The backordering case

shortage.

Let:

b = the cost to have one unit backordered for one year, dollars per unit year year.

The average inventory on hand can be computed by dividing the area under the inventory triangle mnp by the duration T of the cycle. Time t_i of inventory availability is given by (Q-B)/D, and the cycle time T = Q/D. Therefore:

$$\text{Average inventory} = \frac{1}{T} \cdot \frac{(Q-B)t_i}{2} = \frac{(Q-B)^2}{2Q}$$

The average backorder level is obtained similarly by dividing the area of the backorder triangle pqs to T; note that $t_b = B/D$:

$$\text{Average backorder level} = \frac{1}{T} \cdot \frac{Bt_b}{2} = \frac{B^2}{2Q}$$

The total annual cost is

$$TC = A\frac{D}{Q} + H\frac{(Q-B)}{2Q} + b\frac{B}{2Q}$$

There are two policy parameters: Q and B. The first order (necessary) conditions for the optimality of Q^*, T^* are given below (for second-order sufficient conditions see Luenberger [1973, p. 114]):

$$\begin{cases} \frac{\partial(TC)}{\partial Q} = 0 \\ \frac{\partial(TC)}{\partial B} = 0 \end{cases}$$

The optimum solution is:

$$Q^* = \sqrt{\frac{2AD}{H}}\sqrt{\frac{H+b}{b}} \tag{3}$$

$$B^* = \frac{HQ^*}{H+b} = \sqrt{\frac{2AD}{b}}\sqrt{\frac{H}{H+b}} \tag{4}$$

Expressions (3) and (4) are defined only for $b > 0$. Of course, $b = 0$ does not make sense if stockouts are permitted. Indeed, in such a case $B^* = Q^* \to \infty$. This means that, with no charge for backorders, one would keep piling up unfilled demand until the backlog gets infinitely large. Then, one single order would be released to satisfy all accumulated demand, thus driving the per unit share of the ordering cost to zero. Notice that if $b \to \infty$ no backorders may be carried in the optimum solution, and (3) becomes the Wilson lot size formula.

The Classical Economic Lot Size Model with Lost Sales

As different from the previous case, we consider here that if demand occurs when the system is out of stock the potential sales are lost. This situation is illustrated in Figure 5. by t_ℓ we have denoted the duration of the out of stock situation. The in-stock time is $t_i = Q/D$. The total sales per cycle are L units, and the cycle time is $T = \frac{Q+L}{D}$. There are

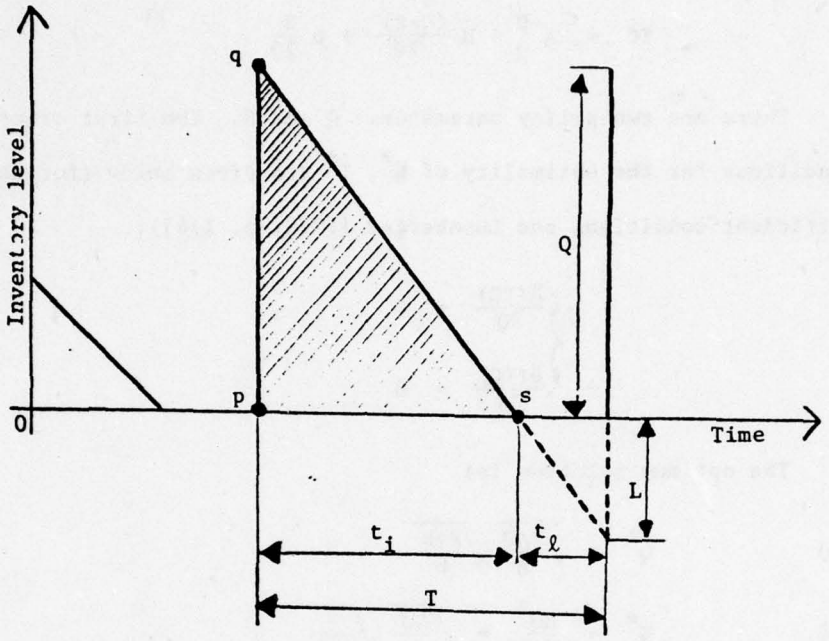


Figure 5: The lost sales case

$\frac{D}{Q+L}$ cycles per year.

To calculate the average inventory we use the inventory triangle pqs:

$$\text{Average inventory} = \frac{1}{T} \cdot \frac{Q\, t_i}{2} = \frac{Q^2}{2(Q+L)}$$

Assume that for each unit of demand which occurs during the stockout situation a cost c_ℓ is incurred by the system. Then:

$$\text{Annual shortage cost} = c_\ell L \cdot \frac{1}{T} = c_\ell \frac{LD}{Q+L}$$

The total annual cost is:

$$TC = A \frac{D}{Q+L} + H \frac{Q^2}{2(Q+L)} + c_\ell \frac{LD}{Q+L}$$

The necessary conditions for Q^*, L^* to be optimal are:

$$\begin{cases} \frac{\partial(TC)}{\partial Q} = 0 \\ \frac{\partial(TC)}{\partial Q} = 0 \end{cases}$$

or:

$$\begin{cases} \frac{1}{2} HQ^2 + HQL - c_\ell DL - AD = 0 & (5) \\ \frac{1}{2} HQ^2 - c_\ell DQ + AD = 0 & (6) \end{cases}$$

Equation (6) yields:

$$Q^* = \frac{c_\ell D \pm \sqrt{(c_\ell D)^2 - 2HAD}}{H} \qquad (7)$$

For Q^* to take on real values the expression under the square root has to be nonnegative:

$$(c_\ell D)^2 \geq 2HAD \qquad (8)$$

$$c_\ell D > \sqrt{(c_\ell D)^2 - 2HAD} \qquad (9)$$

It appears that when (8) is satisfied Q^* is always positive. Q^* takes on one positive value when (8) holds at equality, or two distinct positive values when (8) is an inequality.

To calculate the optimum L consider first the case where (8) holds as strict inequality: $(c_\ell D)^2 > 2HAD$. Then, by substituting (7) into (5), the optimum lost sales position is:

$$L^* = -\frac{c_s D \pm \sqrt{(c_\ell D)^2 - 2HAD}}{H} \qquad (10)$$

From (7) and (10) it follows that $L^* = -Q^* < 0$, i.e. L^* is not contained in the interval $0 < L^* < \infty$. Therefore, the optimal operating value is $L^* = 0$ since nothing can be gained from running the system with any

positive amount of lost sales, and the optimal lot size is given by the Wilson's formula (1).

Consider now the case when (8) is an equality:

$$(c_\ell D) = 2HAD \tag{11}$$

Under this setting $Q^* = \frac{c_\ell D}{H}$; substitute it into (5)

$$\frac{I}{2} \cdot \frac{(c_\ell D)^2}{H} + c_\ell DL - c_\ell DL - AD = 0 \tag{12}$$

Given (11), (12) is an identity, hence any value of L^* is optimal. It is also a simple matter to check that, when (11) holds, the total cost TC is independent of L, i.e., $TC = c_\ell D$. Hadley and Whitin [1963, p. 50] give the following intuitive interpretation to this last result: the time sequence of events in Figure 5 can be rearranged by consolidating an arbitrary number of lost sales periods. Thus, the new graph would present a region similar to Figure 2 where there are no lost sales for a long time, and then another region where for a long time there is nothing but lost sales. This does not change the average cost per year.

The conclusion is that, for the given stockout cost structure and because constraint (8) has to be observed, nothing can be gained by allowing shortages to develop, and therefore it is at least as economical to run the system without any stockouts.

The Lot Size Model with Quantity Discounts

Frequently when deciding upon a purchasing quantity vendor's discount schedules are available and have to be considered. Quantity discounts are usually offered in one of the following two forms (Hadley and Whitin [1963, ch. 2]):

1) all units discounts,

2) incremental quantity discounts.

Consider that all assumptions made at the beginning of section 2.2.1, except g), are in effect.

1) *The Lot Size Model with All Units Quantity Discounts*

In this form the discount price applies to all units purchased (see Figure 6). There are a number of price breaks $b_0 = 0, b_1, b_2, \ldots, b_i, \ldots$ given such that if the ordered amount Q is within discount interval i, $b_{i-1} \leq Q < b_i$, then the unit price for each of the Q units is c_i, where $c_i < c_{i-1}$.

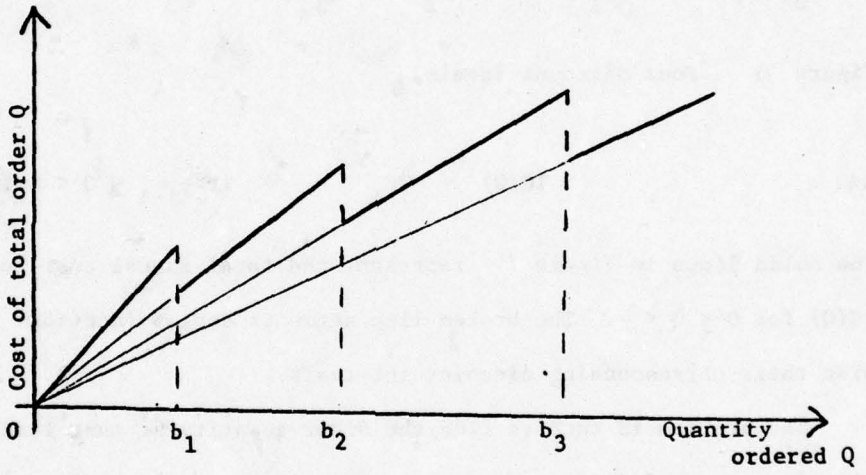


Figure 6: All units quantity discounts

For discount level i the total annual cost function is:

$$TC_i = c_i D + A \frac{D}{Q} + r\, C_i \frac{Q}{2} \tag{13}$$

defined for $b_{i-1} \leq Q < b_i$. In Figure 7 a case with four discount levels is illustrated ($b_4 = \infty$). It is easy to show that these cost curves do not intersect and that $TC_i < TC_{i-1}$ for all Q.

Let us define TC(Q) as follows:

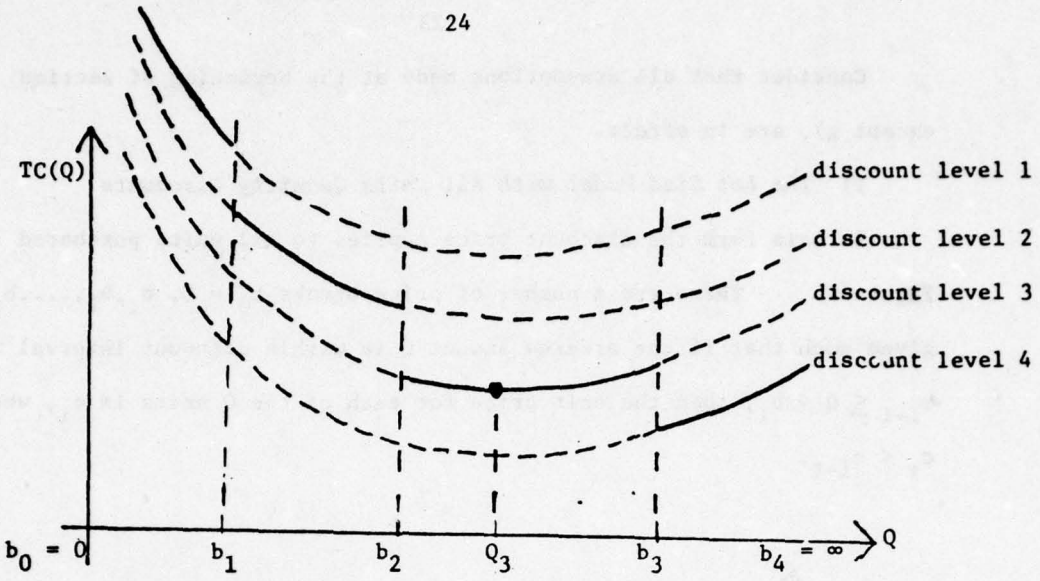


Figure 7: Four discount levels

(14) $$TC(Q) = TC_i \quad \text{if } b_{i-1} \leq Q < b_i,\ i=1,2,\ldots$$

The solid lines in Figure 7 represent the total annual cost function TC(Q) for $0 \leq Q < \infty$. The broken line segments depict functions TC_i outside their corresponding discount intervals.

The problem is then to find the order quantity Q^* that leads to the global optimum of TC(Q). From inspecting Figure 7 it is apparent that Q^* can take on one of the following values:

- the EOQ which minimizes the annual cost for some discount level;
- one of the two extreme points of some discount interval.

In the case shown in Figure 7 the optimum order quantity is $Q^* = b_3$.

The optimum purchase quantity can be determined by the procedure given below:

I - Determine the minimum point of the mathematical function TC_i for $0 \leq Q < \infty$; denote it Q_i:

$$Q_i = \sqrt{\frac{2AD}{r\,C_i}}$$

II - Determine for each discount interval i, $[b_{i-1}, b_i]$, the value Q_i^* of Q which minimizes the total annual cost function TC(Q) in that interval:

$$Q_i^* = \begin{cases} b_i - 1 & \text{if } Q_i < b_{i-1} \\ Q_i & \text{if } b_{i-1} \leq Q_i \leq b_i \\ b_i & \text{if } b_i < Q_i \end{cases}$$

Thus, for the case of Figure 7 we have:

$$Q_1^* = b_1; \quad Q_2^* = b_2; \quad Q_3^* = Q_3; \quad Q_4^* = b_3.$$

III - Out of the set of all $Q_i^* = Q_i$ (there is at least one such Q_i^*) choose the one with the largest subscript i; call it Q_k^*. The global optimum order quantity Q^* cannot lie in a discount interval $i < K$; this follows from the property that for $i < K$, $TC_i > TC_k$ for all Q.

Compute the total cost function at Q_K:

$$TC(Q_K) = C_K D + A \frac{D}{Q_K} + r C_K \frac{Q_K}{2}$$

IV - Test all other discount levels $j > K$. Compute the total cost function at the corresponding Q_j^*:

$$TC(Q_j^*) = C_j D + A \frac{D}{Q_j^*} + r C_j \frac{Q_j^*}{2}$$

The optimum discount level is given by that value of j for which $TC(Q_K) - TC(Q_j^*)$ is positive and a maximum. The optimal overall Q^* is then set equal to the corresponding Q_j^*.

If all $TC(Q_j)$ are larger than $TC(Q_K)$, then $Q^* = Q_K$.

Brown [1978, p. 184] advances a rule of thumb for making decisions under quantity discounts: it is economical to order M months of supply over the usual economic order quantity only if the percentage reduction in

unit cost that can be obtained is at least 2M% (based on the typical case where management's carrying charge is 20-24% per year).

2) *The Lot Size Model with Incremental Quantity Discounts*

This form of discount offers a lower price only for the number of units in the particular discount interval; for units in other intervals, although belonging to the same order, other prices apply (see Figure 8):

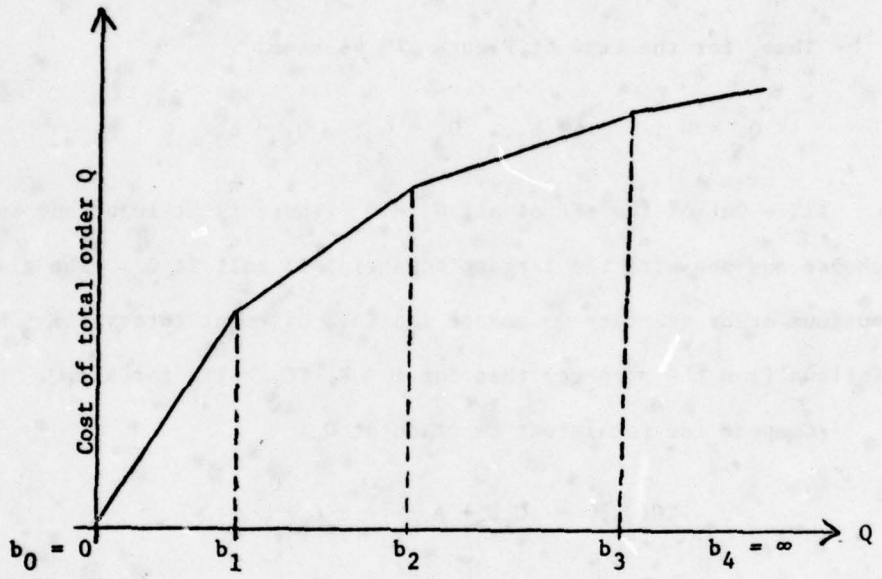


Figure 8: Incremental quantity discounts

Thus, the first b_1 units cost C_1 each, the next (b_2-b_1) units cost C_2 each; the units in the $[b_{i-1},b_i]$ interval cost C_i each, with $C_i < C_{i-1}$. Consider assumptions (a) through (f) are holding.

To derive the total annual cost TC(Q) of operating the system assume $b_{i-1} \leq Q < b_i$. The yearly ordering cost amounts to $A\frac{D}{Q}$. Let P_i be the purchase price (cost) of a lot of size Q in the i-th discount interval:

$$P_i = \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) + C_i(Q - b_{i-1})$$

The annual purchase cost is $\frac{D}{Q} P_i$.

The annual inventory holding cost is given by $r \frac{P_i}{2}$.

Then, the annual cost corresponding to the i-th discount interval is:

$$(15) \qquad TC_i = A \frac{D}{Q} + \frac{D}{Q} \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) + \frac{D}{Q} C_i(Q - b_{i-1}) +$$

$$+ \frac{r}{2} \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) + \frac{r}{2} C_i(Q - b_{i-1})$$

The annual cost function TC(Q) is defined as:

$$(16) \qquad TC(Q) = TC_i \qquad \text{if} \quad b_{i-1} \leq Q < b_i, \qquad i=1,2,\ldots$$

By grouping terms in (15) one obtains:

$$(17) \qquad TC_i = \left[A + \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) - C_i b_{i-1}\right] \frac{D}{Q} + r\, C_i \frac{Q}{2} +$$

$$+ \left[C_i D + \frac{r}{2} \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) - \frac{r}{2} C_i b_{i-1}\right]$$

Functional form (17) immediately suggests a curve of the same shape as in the classical economic lot size model, with a unique global minimum. Figure 9 illustrates the fact; cost function TC(Q) is represented by the solid.

As opposed to the all units discount case, the TC(Q) curve is continuous [it can be shown that $TC_i(Q=b_i) = TC_{i+1}(Q=b_i)$].

An aspect we are interested in is the nature of the price break points b_i, i=1,2,... As Hadley and Whitin [1963, p. 67] mention, the slope of TC_i at $Q = b_i$ is larger than the slope of TC_{i+1} at $Q = b_i$. Indeed:

$$\frac{d(TC_i)}{dQ} = -\left[A + \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) - C_i b_{i-1}\right] \frac{D}{Q^2} + \frac{1}{2} r\, C_i$$

$$\frac{d(TC_{i+1})}{dQ} = -\left[A + \sum_{k=1}^{i} C_k(b_k - b_{k-1}) - C_{i+1} b_i\right] \frac{D}{Q^2} + \frac{1}{2} r\, C_{i+1}$$

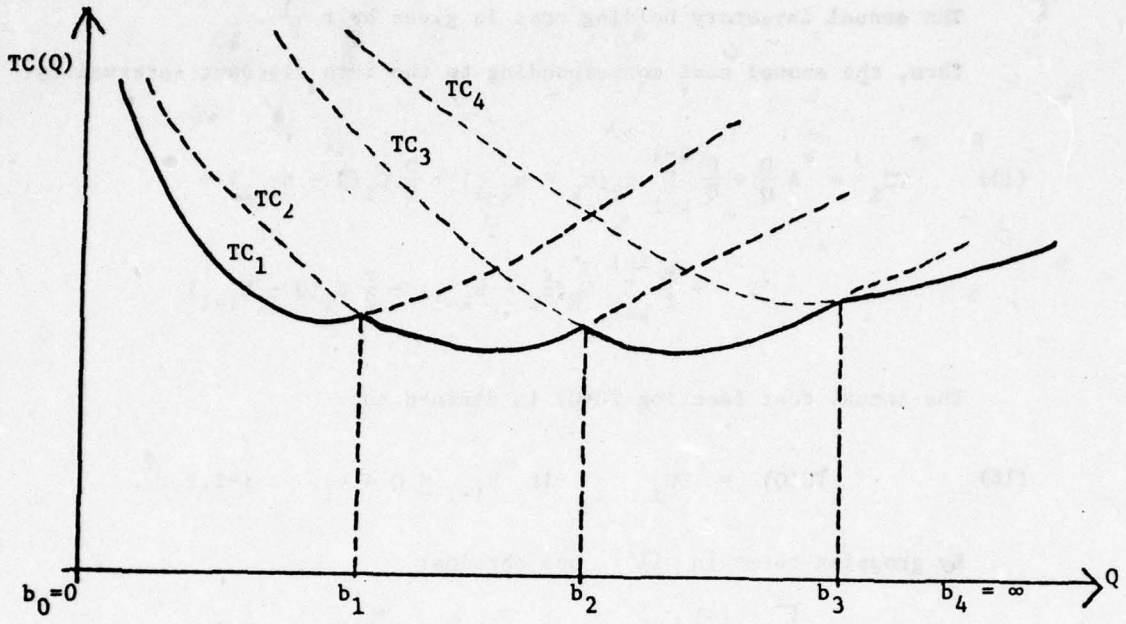


Figure 9: Annual cost functions under incremental quantity discounts

Then, for $Q = b_i$:

$$\frac{d(TC_i)}{dQ}(Q=b_i) = -A + \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) \frac{D}{b_i^2} - C_i b_{i-1} \frac{D}{b_i^2} + \frac{1}{2} r C_i$$

$$\frac{d(TC_{i+1})}{dQ}(Q=b_i) = -A + \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) \frac{D}{b_i^2} - C_i(b_i - b_{i-1}) \frac{D}{b_i^2} +$$

$$+ C_{i+1} \frac{D}{b_i} + \frac{1}{2} r C_{i+1}$$

It is clear that $\frac{1}{2} r C_i > \frac{1}{2} r C_{i+1}$.

Also:

$$C_i b_{i-1} \frac{D}{b_i^2} > -C_i(b_i - b_{i-1}) \frac{D}{b_i^2} + C_{i+1} \frac{D}{b_i} = C_i b_{i-1} \frac{D}{b_i^2} - (C_i - C_{i+1}) \frac{D}{b_i}$$

Hence:

$$\frac{d(TC_i)}{dQ}\,(Q=b_i) \;>\; \frac{d(TC_{i+1})}{dQ}\,(Q=b_i) \tag{17}$$

By (17) the global minimum of TC(Q) cannot occur at a price break point b_i.

The optimum of TC(Q) can be determined as follows:

I - Compute the optimal lot size for every discount level:

$$Q_i = \sqrt{\frac{2D\left[A + \sum_{k=1}^{i-1} C_k(b_k - b_{k-1}) - C_i b_{i-1}\right]}{r\,C_i}} \tag{18}$$

II - If $b_{i-1} \leq Q_i < b_i$ compute $TC_i(Q_i)$.

III - The overall optimum Q^* equals the Q_i corresponding to the smallest $TC_i(Q_i)$.

2.2.2 Economic Order Quantity for Slow Moving Items

With slow moving items, demand can no longer be assumed continuous at a constant rate, and consequently the classical EOQ formula fails.

In order to study this case, assume that one unit of demand occurs at constant time intervals known with certainty. All other assumptions may be maintained as previously.

Cost minimization obviously requires the order quantity Q to be a positive integer and to arrive exactly at the moment when demand for one unit of item occurs. Therefore, the inventory level varies between zero and Q-1.

As the annual requirement is D units/year the interval between two consecutive demands is t = 1/S. Then, during time interval [0,t] there are Q-1 units in stock, in time interval [t,2t] there are Q-2 units in stock, etc. The inventory holding cost for one cycle is:

$$Ht[(Q-1)+(Q-2)+\ldots+1+0] = \frac{H}{D} \cdot \frac{Q(Q-1)}{2}$$

The total annual cost can be expressed as:

$$(19) \qquad TC(Q) = A\frac{D}{Q} + \frac{D}{Q} \cdot \frac{H}{D} \cdot \frac{Q(Q-1)}{2} = A\frac{D}{Q} + \frac{H}{2}(Q-1)$$

For Q^* to be the optimum lot size it is necessary that:

$$(20) \qquad TC(Q^*) < TC(Q^*-1)$$

and

$$(21) \qquad TC(Q^*) < TC(Q^*+1)$$

By substituting (19) into (20) one gets

$$(22) \qquad \frac{H}{2} < \frac{AD}{Q^*(Q^*-1)}$$

Similarly, (21) yields

$$(23) \qquad \frac{H}{2} > \frac{AD}{Q^*(Q^*+1)}$$

Inequalities (22) and (23) can be combined into:

$$(24) \qquad \frac{Q^*(Q^*-1)}{2}\frac{H}{A} < S < \frac{Q^*(Q^*+1)}{2}\frac{H}{A}$$

Hanssmann [1962, p. 21] suggests that Q^* is one of the two integers bracketing the real number

$$Q_0 = \sqrt{\frac{2AD}{H}}$$

2.2.3 Economic Order Quantity for Items with a Limited Sales Period

A basic difference between this case and the models presented previously is that demand is stochastic, with some known probability distribution. The statement of the problem is then: the item whose inventory is to be controlled presents a demand pattern with a limited sales period. The item can be procured only once at the beginning of the period. After

the sales period is over there is a cost associated with being left with the item in stock: it either has to be discarded because of spoilage or obsolescence (e.g. newspapers), or has to be sold at a reduced price (e.g. fashions), or has to be stored until the next season (e.g. Christmas decorations, snow tires, etc.). At the same time there is a cost for running out of stock while the sales season is still on. The problem is known in the operations research literature as the "newsboy problem" or as the "single-period inventory model with stochastic demand".

The following notations are used:

D = demand during the sales period; it is a continuous random variable

$f(d)$ = probability density function of D, assumed known

c_o = cost associated with having one unit of item in stock at the end of the sales period, dollars per unit

c_u = cost incurred by the system for each unit of demand which occurs when the system is out of stock, dollars per unit.

To derive the decision rules we seek the amount Q^* that has to be purchased at the beginning of the season (assume no initial inventories) such as to minimize the expected cost incurred by the system at the end of the sales period.

There are two kinds of costs:

$$\text{Cost of excess inventory} = \begin{cases} c_o(Q-D), & \text{if } D < Q \\ 0, & \text{if } D \geq Q \end{cases}$$

$$\text{Cost of stockout} = \begin{cases} 0, & \text{if } D \leq Q \\ c_u(D-Q), & \text{if } D > Q \end{cases}$$

The total expected cost $TC(Q)$ is:

$$TC(Q) = \int_0^Q c_o(Q-D)f(D)dD + \int_Q^\infty c_u(D-Q)f(D)dD$$

The necessary condition for Q^* to be optimal is:*

$$\frac{d[TC(Q)]}{dQ} = c_o \int_0^{Q^*} f(D)dD - c_u \int_{Q^*}^\infty f(D)dD = 0$$

As $\int_Q^\infty f(D)dD = 1 - \int_Q^\infty f(D)dD$, it follows that:

$$\int_0^{Q^*} f(D)dD = \frac{c_u}{c_u + c_o} \tag{25}$$

If we let F(D) be the cumulative probability distribution of D, then (25) yields:

$$F(Q^*) = \frac{c_u}{c_u + c_o} \tag{26}$$

If we check the second derivative it shows that A^* is a global minimum:

$$\frac{d^2[TC(Q)]}{dQ^2} = c_o f(Q) + c_u f(Q) > 0$$

The problem can have alternative formulations by including revenues from sales (Hadley and Whitin [1963, ch. 6-2]) or by using discrete probability distributions (f applicable). In the latter case, a formula similar to (26) can be derived or payoff (or loss) tables can be used for solution (Schlaifer [1959, ch. 4.1, 7.2]).

2.3 Multiple Items Economic Order Quantity Decision Rules

2.3.1 Economic Order Quantity for Multiple Items Sharing the Same Equipment

This problem is also known as the "multi-product cycling problem" or "the economic lot scheduling problem", and is already bordering the production-inventory systems. However, since we admitted that minor production operations can be performed in pure inventory systems, the cycling problem can become

* See Sokolnikoff and Redheffer [1958, pp. 261-262] for the derivative of a definite integral.

relevant to situations facing such systems. Thus, one piece of equipment is used to process m items on a cyclical basis. Unless there is plenty of idle time, the independent lot sizing and scheduling of items for runs on the equipment is likely to lead to interference between different products. Therefore some cycling policy is necessary in order to avoid the risk of an infeasible schedule.

The simplest way to solve the problem is to impose the rule by which every item is produced once in each cycle (Hanssmann [1962]), which is tantamount to requiring that the number of runs per year be the same for all items, say, N. Given the nature of the problem, a finite production rate of P_i units/year has to be assumed for each item i.

The total annual cost under this policy is:

$$TC = N \sum_{i=1}^{m} A_i + \frac{1}{2N} \sum_{i=1}^{m} H_i D_i \left(1 - \frac{D_i}{P_i}\right) \tag{27}$$

where D_i is the annual demand for the i-th item; the other notations are used as before.

In (27) subscript i denotes each of the m items. The derivation of (27) is similar to that of (2) except that the lot size is expressed as $Q = D/N$.

The optimum number of annual production cycles N^* has to satisfy the first order condition:

$$\frac{d(TC)}{dN} = \sum_{i=1}^{m} A_i - \frac{1}{2N^2} \sum_{i=1}^{m} H_i D_i \left(1 - \frac{D_i}{P_i}\right) = 0$$

Then:

$$N^* = \sqrt{\frac{\sum_{i=1}^{m} H_i D_i \left(1 - \frac{D_i}{P_i}\right)}{2 \sum_{i=1}^{m} A_i}} \tag{28}$$

and, of course, the optimum lot size for the i-th item is:

$$Q_i^* = \frac{D_i}{N^*} \tag{29}$$

It is reasonable to require N^* to be an integer; therefore, if it is not, we check the two bracketing integers and choose the one yielding the lowest total cost. It is important to make sure that the cycle time $T = \frac{1}{N}$ is feasible in the sense that it is long enough to allow for setting up and producing one lot of each item.

Clearly, Q_i^* of (29) is in general different from the lot size determined independently by (2). If Q_i^* is much smaller than the independently determines EOQ, then it might be economical not to run item i every cycle. A rule of thumb is proposed by Magee and Boodman [1967, p. 70]: if "the minimum-cost number of runs for the product alone, for any one or more products is less than half the value for all products, the product is a possible candidate for only occasional runs". Such a case may arise, for instance, when the item exhibits a low sales rate and high setup costs. Then, by the rule of thumb mentioned above, the item would be made only occasionally, such as on every second or third cycle.

A related problem is the treatment of a family of m items. Consider that in order to run any one item of the family (or all of them) a major setup cost A (or, if items are purchased, a major ordering cost) has to be incurred. Within the family, however, only a minor cost a_i is involved in changing over from some item to item i. Because of the significant setup for the family, it is reasonable to coordinate the individual lot sizes so that when one item runs out of stock, all other items in the family or most of them also run out of stock; then, the large setup cost A would be justified.

If T years is the family cycle, by the above argument the cycle T_i for the i-th item should be an integral multiple of T:

(30) $$T_i = k_i T, \qquad k_i > 0 \text{ and integer}$$

The total annual cost is (there are 1/T family cycles per year):

(31) $$TC = \frac{1}{T} A + \sum_{i=1}^{m} a_i \frac{D_i}{Q_i} + \sum_{i=1}^{m} H_i \frac{Q_i}{2}$$

By (30) the number of item i cycles per year is $\frac{D_i}{Q_i} = \frac{1}{T_i} = \frac{1}{k_i T}$. It also follows that $Q_i = D_i k_i T$. After substituting in (31):

(32) $$TC = \frac{1}{T}\left(A + \sum_{i=1}^{m} \frac{a_i}{k_i}\right) + \frac{T}{2} \sum_{i=1}^{m} k_i H_i D_i$$

The necessary optimality condition for $T = \frac{d(TC)}{dT} = 0$, from which it follows that:

(33) $$T^* = \sqrt{\frac{2(A + \sum_{i=1}^{m} \frac{a_i}{k_i})}{\sum_{i=1}^{m} k_i H_i D_i}}$$

Brown [1967, p. 48] offers a heuristic iterative procedure to search for the optimum values of k_i and T^*.

Certainly, the above discussion suggests the fact that different cycles for different items can also be used in the first problem presented in this where the m items were not members of a family. Bomberger [1966] has approached this case by requiring that each individual cycle T_i be an integer multiple, k_i, of some fundamental cycle T, and that the sum of the times required to set up and produce a lot of each item be less than the fundamental cycle. Dynamic programming is used to find the multiples.

Other approaches have also been proposed in the literature (see, for instance, Goyal [1974], Silver [1975]), and the interested reader is referred to Elmaghraby's [1978] review of the economic lot scheduling problem for a start.

The work illustrated above has consistently assumed deterministic demand; the difficulty encountered in the optimization process relates to the combinatorial nature of the search for optimal cycle multiples. Consideration of stochastic demand is bound to render the problem even more difficult. The literature is sparse with respect to this issue (Goval [1973], Graves [1977]). Brown [1971, p. 206] takes a pragmatic view of the problem; his rule calls for the production of one item until some maximum inventory is reached, after which the facility is switched over to another item even if it has not yet run short.

2.3.2 Economic Order Quantity for Multiple Items Under Aggregate Constraints

Unconstrained optimization of economic order quantities is often an unrealistic assumption because of such reasons as limited storage space availability, or a fixed budget for inventory investment, etc. Aggregate constraints also arise in hierarchical production planning systems, where the determination of individual production runs has to observe inventory target levels set by the aggregate plan; this issue, however, has been discussed in Hax and Meal [1975] and Bitran and Hax [1977].

Thus, suppose the warehouse capacity is W, and the storage space required per unit of item i is w_i. Assuming that the simple classical inventory system of section 2.2.1 is used, the following constraint has to be observed:

$$\sum_{i=1}^{m} w_i Q_i \leq W \tag{34}$$

where m is the total number of items controlled. By this constraint we make sure that even if all items reached their maximum inventory positions simultaneously, the warehouse would still be able to contain them all.

The problem is then:

$$\text{(35)} \qquad \text{Minimize } TC = \sum_{i=1}^{m} A_i \frac{D_i}{Q_i} + \frac{1}{2} \sum_{i=1}^{m} H_i Q_i$$

subject to:

$$\text{(36)} \qquad \sum_{i=1}^{m} w_i Q_i - W \leq 0$$

$$\text{(37)} \qquad -Q_i \leq 0$$

The objective function is convex. This can be shown by examining its Hessian $\mathcal{H}$. The Hessian is a diagonal matrix with all positive elements on its main diagonal. Therefore, for all nonzero vectors x it is true that $x^T \mathcal{H} x > 0$, hence the Hessian is positive definite and TC is convex (Luenberger [1973, p. 118]). The total cost reaches its minimum within the confines of constraints (36) - (37), and this is a global minimum.

The Kuhn-Tucker conditions (Mangasarian [1969]) are necessary and sufficient for optimality in this case. Let λ be the nonnegative Lagrange multiplier for (36) and γ for (37). The Lagrangian is:

$$\text{(38)} \qquad L = \sum_{i=1}^{m} A_i \frac{D_i}{Q_i} + \frac{1}{2} \sum_{i=1}^{m} H_i Q_i + \lambda \left(\sum_{i=1}^{m} w_i Q_i - W \right) - \gamma Q_i$$

The Lagrangian can be simplified when we realize that, given the nature of the problem, in the optimum solution $Q_i > 0$ and, therefore $\gamma = 0$. Then, last term in L can be dropped and the Kuhn-Tucker conditions are:

$$\text{(39)} \qquad \frac{\partial L}{\partial Q_i} = 0, \qquad i=1,\ldots,m$$

$$\text{(40)} \qquad \sum_{i=1}^{m} w_i Q_i - W \leq 0$$

$$\text{(41)} \qquad \lambda \sum_{i=1}^{m} w_i Q_i - W = 0$$

$$\text{(42)} \qquad \lambda \geq 0$$

The solution can be obtained by trying both $\lambda = 0$ and $\lambda > 0$.

For $\lambda = 0$ the solution is given by:

$$\frac{\partial L}{\partial Q_i} = 0, \qquad i=1,\ldots,m \tag{43}$$

and must be checked against (40).

For $\lambda \neq 0$ the solution can be obtained from:

$$\frac{\partial L}{\partial Q_i} = 0, \qquad i=1,\ldots,m \tag{44}$$

$$W - \sum_{i=1}^{m} w_i Q_i = 0 \tag{45}$$

Obviously, the case $\lambda = 0$ means that the storage capacity limitation is not binding and, therefore, the optimal solution is to use the economic lot sizes determined for each item independently.

For $\lambda \neq 0$:

$$- A_i \frac{D_i}{Q_i^2} + \frac{1}{2} H_i - \lambda w_i = 0, \qquad i=1,\ldots,m \tag{46}$$

$$\sum_{i=1}^{m} w_i Q_i = W \tag{47}$$

After solving (46) for Q_i, and substituting for Q_i in (47) one obtains:

$$Q_i = \sqrt{\frac{2A_i D_i}{H_i - 2\lambda w_i}}, \qquad i=1,\ldots,m \tag{48}$$

$$\sum_{i=1}^{m} w_i \sqrt{\frac{2A_i D_i}{H_i - 2\lambda w_i}} = W \tag{49}$$

Equation (49) has to be solved for λ and, in general, this cannot be done analytically. Holt, et al. [1960, ch. 10], and Candea [1975] provide discussions of the topic and suggest solutions by graphical methods and search techniques.

3 Forecasting

Future demand requirements play a leading role in the process of making decisions. To reach these kinds of decisions, our knowledge about demand had to extend to moderate distances into the future. For facilities design problems, demand information has been again a major input; this time, however, our knowledge about demand had to be much more far-reaching, extending over periods of years.

Information about future requirements is obtained by the process of forecasting. Demand forecasts can span a large variety of planning horizons, ranging from days (e.g. the "newsboy problem") to years and decades (e.g. capital investment problems). Certainly forecasts can be made for many other purposes (such as weather forecasts, costs forecasts, technological forecasts, etc.). In this paper, however, our interest regards only demand forecasts and related techniques for short and medium-range forecasting, to be used in production and inventory control.

Brown [1977, p. 73] considers that there are two broad approaches to forecasting: the "descriptive" approach and the "explanatory" approach.

In the descriptive approach, the basic assumption is that the underlying process that has generated demand in the past continues into the future. Statistical models are applicable in this case. The time series (or the demand history) is analyzed for components (such as the base or constant component, trend, seasonalities) and then it is extraploated into the future. A word of caution, however: forecasts made by use of mathematical techniques should constitute only the starting point for decision making. Managerial inputs, information on special facts (for instance knowledge about an impending promotional campaign by own company or a competitor)

should interact with the forecasting system and appropriately alter future estimates of demand.

The explanatory approach tries to build causal effects into the model. A large variety of models can be found in the literature, ranging from large econometric models (economy wide, sectoral, industry, etc.) to special models such as predicting buying patterns in certain strata of the society.

This paper concentrates on descriptive models that account for the following demand patterns(see Figure 4):

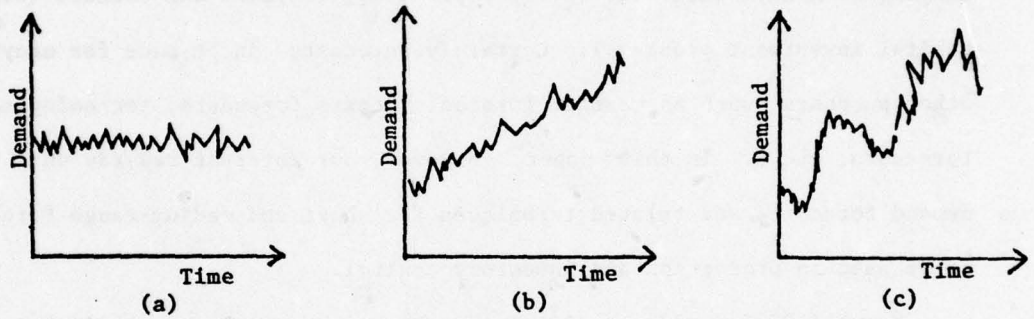


Figure 4: Demand patterns: (a) Base, (b) Trend, (c) Trend and seasonalities

(1) The *base* (or constant) pattern presents a central tendency of the time series at any given time.

(2) The *trend* process exhibits a consistent long-run shift of the average. The linear trend is the most often used assumption.

(3) The *seasonal* pattern shows a cyclic variation, usually with a one year periodicity. Seasonalities can be superimposed upon a background of constant annual average demand, or can be associated with a process in which the annual average features a trend.

In all cases, a random error component (noise) is present, which is the unexplained deviation of the data from the basic generating process.

Other demand patterns exist (Johnson and Montgomery [1974, p. 403]) such as step functions, impulse function, etc., but they are not treated here.

Once an appropriate model is identified, the forecasting problem is threefold:

- estimation of all unknown parameters of the model,
- computation of a forecast by projecting the model into the future,
- updating the parameter estimates as new data become available.

With respect to the task of updating the model parameters by incorporating fresh data two approaches can be pursued: one in which the time origin is fixed and, therefore, every time the forecast is updated, the parameters have to be recomputed from scratch in order to incorporate the latest observations; the other approach, called direct smoothing or adaptive smoothing (Brown [1963, p. 168]) assumes the time origin to be at the end of the current period, and updates the parameters by combining the old estimates and the latest forecast error recorded. Obviously, direct smoothing requires the time origin to be shifted every time the forecast is updated.

To date a sizable number of forecasting techniques have been perfected; it is beyond the scope of this paper to present them all. Books and papers have been published on the topic and some starting references are indicated below:

- *Regression methods* fit some hypothesized model, linear in its coefficients, to the time series. Least square estimators are used for the coefficients. Multiple regression can be used also as an explanatory method of forecasting when a causal relationship between the dependent variable and the independent variables is present. Montgomery and Johnson [1976, ch. 2], Wheelwright and Makridakis [1977, ch. 7].

- *Moving average* computes an average of a constant number N of past observations in order to eliminate the random variations or noise. As a

new observation becomes available, the oldest observation is dropped (in order to keep their number constant) and a new average is computer. Montgomery and Johnson [1976, ch. 2], Makridakis and Wheelwright [1978, ch. 4, 11].

- *Exponential smoothing* has become a very popular method of forecasting. It is present in more detail leter in this paper.

- *Bayesian methods* are useful when one is faced with a lack of historical data at the beginning of the forecasting process. The approach is then to start with some initial subjective estimates of the parameters in the model, and by use of Bayes theorem to modify them in light of the actual data observed. Cohen [1966], Montgomery and Johnson [1976, ch. 10]; for Bayesian theory see Raiffa and Schlaifer [1961], Raiffa [1970].

- *The Box-Jenkins models* are a set of autoregressive models in which successive observations are highly dependent. Three classes of models are considered; the autoregressive process, the moving average (of the errors) process, and the mixed autoregressive-moving average process. The choice of the correct model is made by examining the autocorrelation coefficients. Box and Jenkins [1976], Makridakis and Wheelwright [1978, ch. 10], Wheelwright and Makridakis [1977, ch. 8].

3.1 Exponential Smoothing Methods for Fast Moving Items

Due to the basic difference existing between fast and slow moving items distinct forecasting procedures have to be considered. Slow moving items are treated in a later section. For fast moving items exponential smoothing techniques are presented in this section, with the understanding that forecasts are updated once every time period. The length of the period can vary depending on the characteristics of the controlled items, but one month is a usual period length in many inventory systems.

Out of the multitude of forecasting techniques we choose exponential

smoothing because it is relatively simple, it is reasonably accurate, and efficient from the computational point of view. These reasons certainly account for the widespread use of exponential smoothing.

3.1.1 Exponential Smoothing for the Base (Constant) Pattern

The base pattern (Figure 10a) has a stationary demand distribution. The mean demand is time invariant, although unknown.

Let:

A_t = actual demand materialized in period t

a = the true, unknown, constant demand which generates the observed demand pattern

ε_t = random noise component associated with period t, having $E(\varepsilon_t) = 0$ and variance σ_ε^2.

Thus:

$$A_t = a + \varepsilon_t \tag{50}$$

In order to forecast A_t we have to produce an estimate $\hat{a}$ of the constant component a.

Assume that (T-1) periods have gone by and now we are in a position to produce a forecast for period T. First we try to obtain an estimate of a which we will use as the forecast of next period's demand:

$$F_T^b = \hat{a} \tag{51}$$

Being concerned with the forecast errors, we chose as an estimation criterion the minimization of the sum of weighted squared forecasting errors. Since we think that recent errors are more relevant to the forecasting procedure than earlier errors, we weight the former heavier than the latter:

$$\text{Minimize} \sum_{t=1}^{T-1} \beta^t (A_{T-t} - \hat{a})^2 \quad , \qquad 0 < \beta \leq 1 \tag{52}$$

Clearly the weight β^t decreases with the age of the data.

Setting the first derivative equal to zero one gets:

$$\sum_{t=1}^{T-1} \beta^t A_{T-t} - \hat{a} \sum_{t=1}^{T-1} \beta^t = 0$$

But:

$$\sum_{t=1}^{T-1} \beta^t = \beta + \beta^2 + \ldots + \beta^{T-1} = \frac{1-\beta^T}{1-\beta} - 1 = \beta \frac{1-\beta^{T-1}}{1-\beta}$$

Thus:

$$F_T^b = \hat{a} = \frac{1-\beta}{\beta(1-\beta^{T-1})} \sum_{t=1}^{T-1} \beta^t A_{T-t} \tag{53}$$

Going through a similar process we can reconstruct our forecast F_{T-1}^b for period T-1 made at the end of period T-2:

$$F_{T-1}^b = \frac{1-\beta}{\beta^2(1-\beta^{T-2})} \sum_{t=2}^{T-1} \beta^t A_{T-t} \tag{54}$$

If T is large enough $\beta^{T-1} \to 0$, $\beta^{T-2} \to 0$, and, therefore, F_T^b and F_{T-1}^b get simplified:

$$F_T^b = \frac{1-\beta}{\beta} \sum_{t=1}^{T-1} \beta^t A_{T-t} \tag{55}$$

$$F_{T-1}^b = \frac{1-\beta}{\beta^2} \sum_{t=1}^{T-1} \beta^t A_{T-t} \tag{56}$$

From (55) and (56) it follows:

$$F_T^b = (1-\beta)A_{T-1} + \beta F_{T-1}^b \tag{57}$$

Let $\alpha = 1-\beta$; then:

$$\boxed{F_T^b = \alpha A_{T-1} + (1-\alpha)F_{T-1}^b} \tag{58}$$

F_T^b is the forecast for period T made at the end of period T-1. By the nature of the model it is also the forecast for any other time period in

the future.

Equation (58) can be given an intuitive interpretation: the forecast for period T can be obtained by combining the previous forecast (for period T-1) with the latest actual demand observed. The forecast for period T is thus composed of a fraction $(1-\alpha)$ of the previous forecast plus a fraction α of the actual demand A_{T-1}. The mix is controlled by monitoring parameter α.

The origin of time is continuously updated; thus, the time period for which we have recorded the most recent observation is always labeled T-1, while the upcoming period is always called T.

Equation (58) defines what is called forecasting by exponential smoothing, and α is the smoothing constant. The term "exponential" originates in the exponential decay with time t of the β^t coefficients.

A heuristic intuitive development of (58) can also be conducted (Brown [1959, p. 46]): to get the new forecast of demand, adjust the previous forecast by a fraction α of the amount by which demand exceeded forecast in the last period (i.e., period T-1):

$$(\text{Demand Excess Over Forecast})_{T-1} = A_{T-1} - F^b_{T-1}$$

$$F^b_T = F^b_{T-1} + \alpha(A_{T-1} - F^b_{T-1}) \tag{59}$$

By rearranging the terms:

$$F^b_T = \alpha A_{T-1} + (1-\alpha)F^b_{T-1} \tag{60}$$

which is precisely equation (58). Equation (59), however, allows us to better see the meaning of the smoothing constant. If α is large, the system is "nervous" in reaching to forecast errors, while with a smaller α more of a smoothing effect is obtained (slower response to forecast errors).

Forecast F^b_T is an asymptomatically unbiased estimator of a. To show

this, (59) can be developed recursively, this yielding:

$$F_T^b = \alpha \sum_{t=1}^{T-1} \beta^{t-1} A_{T-t} + \beta^{T-1} F_1^b \tag{61}$$

Then, the expected value of F_T^b in a stabilized system ($T \to \infty$) is:

$$E[F_T^b] = E[\alpha \sum_{t=1}^{\infty} \beta^{t-1} A_{T-t}] = \alpha \sum_{t=1}^{\infty} \beta^{t-1} E[A_{T-t}]$$

As $E[A_{T-t}] = a$, it follows:

$$E[F_T^b] = a\alpha \sum_{t=1}^{\infty} \beta^{t-1} = a\alpha \frac{1}{1-\beta} = a \tag{62}$$

Therefore, our decision to use F_T^b to forecast the unknown "a" appears to have been reasonable.

Two issues have to be resolved before (59) is used to predict demand: the initialization of the model and the appropriate choice of the smoothing constant.

Brown [1963, p. 102] recommends the use of the simple average of the most recent N observations as the initial value of the smoothed statistic:

$$(F_{T-1}^b)_{\text{initial}} = \frac{\sum_{t=1}^{N} A_t}{N} \tag{63}$$

When no demand history is available $(F_{T-1}^b)_{\text{initial}}$ is made equal to a subjective prediction of the average, based on the based judgement of marketing people or derived from similarities with other products.

Call period 1 the first period of the forecasting horizon, period 2 the second period, etc. Then, when we start forecasting for the first time, our forecast F_1^b is precisely the initial value $(F_{T-1}^b)_{\text{initial}}$. At the end of the first period, after the actual demand A_1 will have been observed, (60) will generate a forecast F_2^b for period 2:

$$F_2^b = \alpha A_1 + (1-\alpha) F_1^b$$

It is important to realize that the initial forecast $(F^b_{T-1})_{initial}$ gets soon strongly discounted in the forecasting process by a weight β^t (after t observations), as indicated by (61). Therefore, any reasonable estimate can be used as an initial condition.

The rate of response of the forecasting model is certainly influenced by the choice of the smoothing constant (see (59)). A higher value for α improves the ability of the model to track a fluctuating pattern of demand, but reduces its function of filtering out random variations. The choice of α is also affected by the reliability of the initialization procedure. If the initial estimate $(F^b_{T-1})_{initial}$ is questionable, a larger α will help discount its influence faster. Contrarily, if there is a strong confidence in the prediction of initial conditions, a smaller α will prevent any quick and undesired change by smoothing out the random component of the demand genrating process. The forecast update period is still another factor, in that over a longer period conditions are more likely to change and, therefore, a larger smoothing constant is required in order to incorporate the change into the updated forecast.

From the above discussion it is clear that there are contradictory issues involved in choosing the appropriate value of the smoothing constant. One one hand we want the forecast to be sensitive enough in order to respond to the real changes in the demand pattern. On the other hand forecast stability is desirable as a protection against the system overreacting to random variations in the signal. A third criterion, the accuracy of forecasts (Brown [1963, p. 118]), should also be observed. As a general rule, the literature recommends values for α within the range 0.01 to 0.3, a value of 0.1 being a satisfactory compromise between a very stable system and a "nervous" system (Brown [1963], Johnson and Montgomery [1974]). One way to choose the α is by simulation (Holt, et al [1960]).

Thus, available demand history is partitioned into two sets of data. The first set is used to initialize the model; the model is then run with different smoothing constants over the first part of the series, in order to reduce the effect of initial conditions, and then over the remaining data, considering the second set as fresh demand observations. Select that value of α which optimizes some criterion (e.g. minimizes the sum of squared errors). If a rather large value of α seems to be required (especially if over 0.3) this should be regarded as a strong indication that a different model might be necessary.

It is conceivable that more than one value of the smoothing constant might have to be used if circumstances change. Thus, for instance, if the initial estimate of the average is unreliable we can start out with a large α; after a few periods of running the model, when the effect of initial conditions has worn off and the true process has "taken over", the smoothing constant can be decreased to provide a better stability.

A somewhat similar line of thinking has led to the development of the adaptive control smoothing models for automatically adjusting the value of the smoothing constant. A tracking signal, representative of the forecasting system's performance, is continuously monitored (section 3.5). When the tracking signal exceeds some appropriately chosen bounds, thus indicating unacceptably large forecasting errors, the value of the smoothing constant is increased accordingly, in order to give more weight to recent data and thus bring the system faster in line with the changed pattern of demand. When the out-of-control condition disappears, the smoothing constant can be restored to its "normal" value. Basically this is the essence of the adaptive control model proposed by Trigg and Leach [1967]. Other techniques have been proposed by Eilon and Elmaleh [1970], Montgomery [1970], and Chow [1965] with extensions by Roberts and Reed [1969].

3.1.2 Exponential Smoothing for Demand Patterns with Linear Trend

The linear trend pattern is illustrated in Figure 10b and can be represented by a polynomial of degree one:

$$A_t = a + bt + \varepsilon_t \tag{64}$$

where:

- A_t = materialized (observed) demand in period t
- a = base component, from which the linear trend starts
- b = trend
- ε_t = random noise sample in period t; $E(\varepsilon_t) = 0$ and the variance is σ_ε^2.

Also let F_T^{tr} be the forecast made at the end of period T-1 for period T; the superscript tr denotes the trend model.

We approach the task of producing F_T^{tr} in two ways: by using two smoothing constants, or by using only one smoothing constant but, instead, resorting to double smoothing.

Forecasting the Linear Trend Pattern with Two Smoothing Constants

The first tentative we make is to use the simply smoothed statistic F_T^b of 3.1.1 to forecast the linear trend pattern:

$$F_T^b = \alpha A_{T-1} + (1-\alpha) F_{T-1}^b$$

or, by (61):

$$F_T^b = \alpha \sum_{t=1}^{T-1} \beta^{t-1} A_{T-t} + \beta^{T-1} F_1^b$$

Let us compute the expected value of F_T^b:

$$E[F_T^b] = \alpha \sum_{t=1}^{T-1} \beta^{t-1} E[A_{T-t}] + \beta^{T-1} E[F_1^b] =$$

$$= \alpha \sum_{t=1}^{T-1} \beta^{t-1} [a + b(T-t)] + \beta^{T-1} E[F_1^b] =$$

$$= a\alpha \sum_{t=1}^{T-1} \beta^{t-1} + bT\alpha \sum_{t=1}^{T-1} \beta^{t-1} - b\alpha \sum_{t=1}^{T-1} t\beta^{t-1} + \beta^{T-1} E[F_1^b]$$

Make $T \to \infty$ in order to stabilize the system (i.e., discount to zero initial conditions or any transient response):

$$(65) \qquad E[F_T^b] = a\alpha \sum_{t=1}^{\infty} \beta^{t-1} + bT\alpha \sum_{t=1}^{\infty} \beta^{t-1} - b\alpha \sum_{t=1}^{\infty} t\beta^{t-1} =$$

$$= a\alpha \frac{1}{1-\beta} + bT\alpha \frac{1}{1-\beta} - b\alpha \frac{1}{(1-\beta)^2} =$$

$$= a + bT - b\frac{1}{\alpha}$$

As $E[A_T] = a+bT$, it follows from (65) that in the case of a linear trend pattern the steady state response of the simple exponential smoothing forecast F_T^b for period T lags the expected demand of period T by a constant $b\frac{1}{\alpha}$.

An equivalent result, often found in the literature (Brown [1963, p. 128]), states that simple exponential smoothing develops a lag of $b\frac{1-\alpha}{\alpha}$ behind $a+b(T-1)$, which is immediately apparent from (63)

The larger the smoothing constant the smaller the lag; however, even with a high value of α the bias persists, requiring therefore corrections for trend.

Let $\mathcal{T}_T$ be the estimate of the trend, computed at the end of period T-1, i.e., the projected trend for period T. Since $\mathcal{T}_T$ will also be updated by exponential smoothing, we want to distinguish between the smoothing constant used in computing F_T^b, call it α_1, and the smoothing constant

used in trend calculations, let it be α_2. The trend is smoothed as follows:

$$\mathcal{T}_T = \alpha_2(F^b_T - F^b_{T-1}) - (1-\alpha_2)\,\mathcal{T}_{T-1} \tag{66}$$

which expresses the fact that the trend forecast for period T is obtained as a linear combination between the previous trend forecast (for period T-1) and the current change in the average demand forecast (constant process) $F^b_T - F^b_{T-1}$.

Equation (66) yields:

$$\mathcal{T}_T = \alpha_2 \sum_{t=0}^{T-2} (1-\alpha_2)^t (F^b_{T-t} - F^b_{T-t-1}) + (1-\alpha_2)^{T-1}\,\mathcal{T}_1 \tag{67}$$

The expected value of $\mathcal{T}_T$ is:

$$E[\mathcal{T}_T] = \alpha_2 \sum_{t=0}^{T-2} (1-\alpha_2)^t E[F^b_{T-t}] - \alpha_2 \sum_{t=0}^{T-2} (1-\alpha_2)^t E[F^b_{T-t-1}] + \tag{68}$$

$$+ (1-\alpha_2)^{T-1} E[\,\mathcal{T}_1\,]$$

By (65) we have:

$$E[F^b_{T-t}] = a + b(T-t) - b\,\frac{1}{\alpha_1} \tag{69}$$

$$E[F^b_{T-t-1}] = a + b(T-t-1) - b\,\frac{1}{\alpha_1} \tag{70}$$

Let T grow very large; then the last term in (68) dwindles to zero, so it can be disregarded. From (68), (69), and (70) it follows that:

$$E[\,\mathcal{T}_T\,] = b \tag{71}$$

i.e., the trend computed by (66) is an asymptomatically unbiased estimator of the real trend.

By (64), (65), and (70) we can compute:

$$E[F^b_T + \frac{1}{\alpha_1}\,\mathcal{T}_T] = a + bT - b\,\frac{1}{\alpha_1} + \frac{1}{\alpha_1}\,b = a + bT = E[A_T] \tag{72}$$

and, therefore, expression $F_T^b + \frac{1}{\alpha_1} \mathcal{T}_T$ is an unbiased demand forecast for period T.

The response of the system to the linear trend process is stable, although it tends to overshoot slightly a suddien jump in demand such as an impulse or a step input. Eventually it settles down to the correct value (Brown [1959, ch. 2]).

Here is a summary of the formulae for forecasting a linear trend process:

(73)
$$\boxed{\begin{aligned} F_T^{tr} &= F_T^b + \frac{1}{\alpha_1} \mathcal{T}_T \\ F_T^b &= \alpha_1 A_{T-1} + (1-\alpha_1) F_{T-1}^b \\ \mathcal{T}_T &= \alpha_2 (F_T^b - F_{T-1}^b) + (1-\alpha_2) \mathcal{T}_{T-1} \end{aligned}}$$

F_T^{tr} is the demand forecast made at the end of period T-1 for period T. For any other period more distant in the future, say T+τ, the forecast is obtained with $F_T^{tr} + \tau \mathcal{T}_T$.

The forecasting system (73) is used as follows:

- At the end of the current period (period T-1) a record of the actual sales A_{T-1} of the product is obtained.
- The previous forecast F_{T-1}^b is known, hence F_T^b can be computed.
- The trend is updated to yield $\mathcal{T}_T$. F_{T-1}^b and $\mathcal{T}_{T-1}$ are known from the previous time period, and F_T^b has just been calculated.
- With all elements available, F_T^{tr} is fully determined.

For the choice of the smoothing constants α_1, α_2 the same issues are involved as in the previous section.

The initialization procedure has to provide two starting values: F_1^b and $\mathcal{T}_1$ [or $(F_{T-1}^b)_{initial}$ and $(\mathcal{T}_{T-1})_{initial}$] for the first period of

the forecasting horizon. Suppose that N periods of demand history are available. Fit, by regression, a line a+bt to the data (Figure 11); the regression yields two estimates $\hat{a}$ and $\hat{b}$. Then:

$$F_1^b = (F_{T-1}^b)_{initial} = \hat{a} + \hat{b}N$$

$$\mathcal{T}_1 = (\mathcal{T}_{T-1})_{initial} = \hat{b}$$

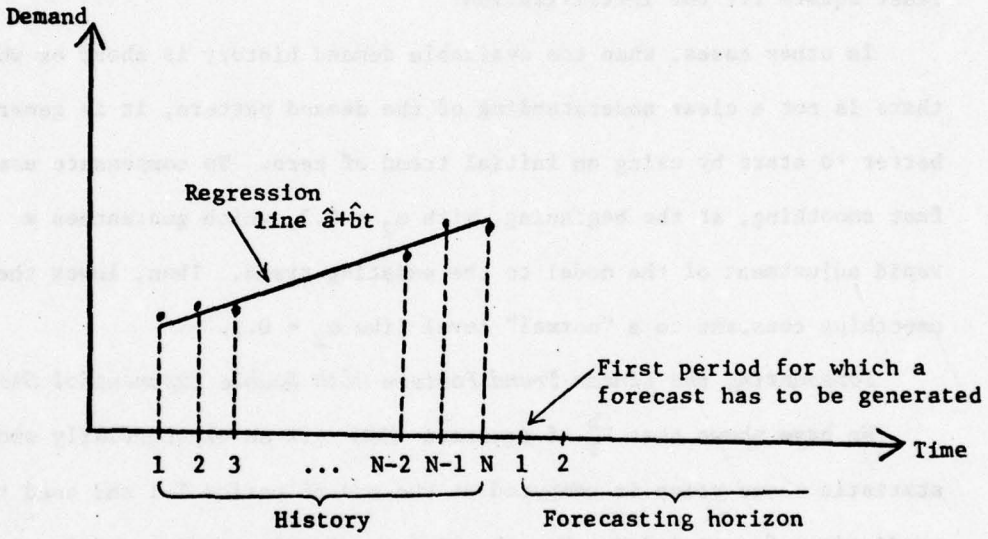


Figure 11: Initialization by the use of regression

The forecast for the first period of the forecasting horizon is: $F_1^{tr} = F_1^b + \mathcal{T}_1$. After the first period is over, the value of the actual demand A_1 becomes known, and the smoothing process can start:

$$F_2^b = \alpha_1 A_1 + (1-\alpha_1)F_1^b$$

$$\mathcal{T}_2 = \alpha_2(F_2^b - F_1^b) + (1-\alpha_2)\mathcal{T}_1$$

$$F_2^{tr} = F_2^b + \frac{1}{\alpha_1}\mathcal{T}_2$$

Obviously, a few periods are necessary for the smoothing process to reach the steady state.

When several years of history are available the initial estimate of the trend can also be computed as follows: calculate the average monthly demand in the first year and in the last year; then, $\mathcal{T}_1$ is the average trend between the first and last years (Holt, et al [1960, p. 265]).

With 7 to 11 months of history Brown [1977, p. 104] recommends the least square fit for initialization.

In other cases, when the available demand history is short or when there is not a clear understanding of the demand pattern, it is generally better to start by using an initial trend of zero. To compensate use fast smoothing, at the beginning, with $\alpha_2 = 0.3$, which guarantees a rapid adjustment of the model to the existing trend. Then, lower the smoothing constant to a "normal" level like $\alpha_2 = 0.1$.

Forecasting the Linear Trend Pattern with Double Exponential Smoothing

We have shown that F_T^b of equation (58) is an exponentially smooted statistic whose value is computed at the end of period T-1 and used to make predictions for period T. For the purpose of this section and for unity of exposure let us think now of F_T^b not as a forecast but rather as a statistical entity whose smoothed value* $S_{[T-1]}$ at the end of period T-1 is obtained by:

$$S_{[T-1]} = \alpha A_{T-1} + (1-\alpha) S_{[T-2]} \tag{74}$$

In (74) $S_{[T-2]}$ is the smoothed statistic computed at the end of period T-2, and A_{T-1} is the actual demand recorded in period T-1; α is the smoothing constant.

---

* Brackets around the subscript indicate that the statistic is computed <u>at the end</u> of the bracketed period; contrast with non-bracketed statistics, e.g. F_T^b, or with observed values <u>during</u> some time period, e.g. A_{T-1}.

If exponential smoothing is applied to the results of smoothing the original data, the second order smoothed statistic $S^{[2]}_{[T-1]}$ is obtained. This process is called double exponential smoothing and is defined by:

$$S^{[2]}_{[T-1]} = \alpha S_{[T-1]} + (1-\alpha) S^{[2]}_{[T-1]} \tag{75}$$

In (75) $S^{[2]}_{[T-1]}$ is the second order smoothed statistic computed at the end of period T-2.

It has been shown earlier that $S_{[T-1]}$ develops a bias of $b\frac{\beta}{\alpha}$ behind A_{T-1}; similarly, the steady state response of $S^{[2]}_{[T-1]}$ is biased by the same $b\frac{\beta}{\alpha}$ behind $S_{[T-1]}$ (see Figure 12).

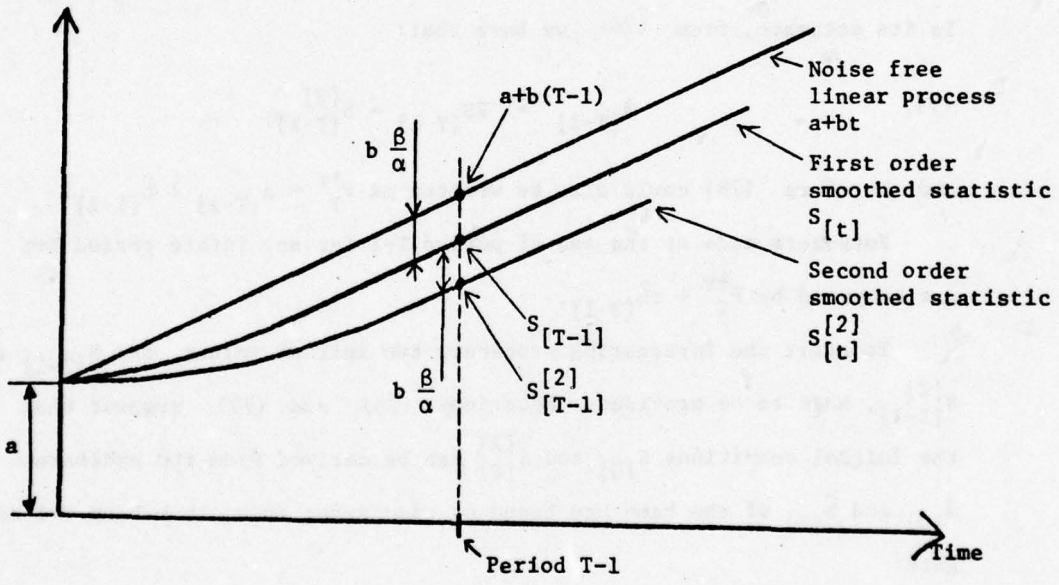


Figure 12: Original data, first order, and second order smoothed statistics for a noise free linear trend process

From Figure 12 it is immediately apparent that:

$$a+b(T-1) = 2E[S_{[T-1]}] - E[S^{[2]}_{[T-1]}] \tag{76}$$

Since the "distance" between $S_{[T-1]}$ and $S^{[2]}_{[T-1]}$ is $b\,\frac{\beta}{\alpha}$, the estimate $\hat{b}_{[T-1]}$ of the trend at the end of period T-1 is:

$$\hat{b}_{[T-1]} = \frac{\alpha}{\beta}\left(S_{[T-1]} - S^{[2]}_{[T-1]}\right) \tag{77}$$

Hence, the unbiased demand forecast F^{tr}_T for period T, produced at the end of period T-1, is given by:

$$F^{tr}_T = 2S_{[T-1]} - S^{[2]}_{[T-1]} + \hat{b}_{[T-1]} \tag{78}$$

The "current" intercept (see Figure 12) is a+b(T-1); if $\hat{a}_{[T-1]}$ in its estimate, from (76) we have that:

$$\hat{a}_{[T-1]} = 2S_{[T-1]} - S^{[2]}_{[T-1]} \tag{79}$$

and therefore (78) could also be written as $F^{tr}_T = \hat{a}_{[T-1]} + \hat{b}_{[T-1]}$.

Forecasts made at the end of period T-1 for any future period T+τ are computed by $F^{tr}_T + \tau\hat{b}_{[T-1]}$.

To start the forecasting procedure two initial values, for $S_{[T-1]}$ and $S^{[2]}_{[T-1]}$, have to be provided. Equations (76) and (77) suggest that the initial conditions $S_{[0]}$ and $S^{[2]}_{[0]}$ can be derived from the estimates $\hat{a}_{[0]}$ and $\hat{b}_{[0]}$ of the base and trend at time zero; thus, by making T=1 one gets:

$$\begin{cases} \hat{a}_{[0]} = 2S_{[0]} - S^{[2]}_{[0]} \\ \hat{b}_{[0]} = \frac{\alpha}{\beta}\left(S_{[0]} - S^{[2]}_{[0]}\right) \end{cases}$$

$\hat{a}_{[0]}$ and $\hat{b}_{[0]}$ can be obtained by fitting a regression line to past data, if available. Otherwise, subjective estimates of the coefficients

have to be produced. Afterwards:

$$\begin{cases} S_{[0]} = \hat{a}_{[0]} - \frac{\beta}{\alpha}\hat{b}_{[0]} \\ S^{[2]}_{[0]} = \hat{a}_{[0]} - 2\frac{\beta}{\alpha}\hat{b}_{[0]} \end{cases}$$

The forecast for the first period of the forecasting horizon is: $F_1^{tr} = 2S_{[0]} - S^{[2]}_{[0]} + \hat{b}_{[0]}$. After period 1 the system of equations (74), (75), (77), and (78) can take over.

Brown [1963, ch. 10] shows that the value of the smoothing constant used in double exponential smoothing, call it α_{double}, should be related to the value of the constant used in simple smoothing, call it α_{simple}, by the following equation:

$$1 - \alpha_{simple} = (1 - \alpha_{double})^2$$

Thus, the equivalent of a small "simple" smoothing constant $\alpha_{simple} = 0.01$ is $\alpha_{double} = 0.005$, while a large smoothing constant equivalent to $\alpha_{simple} = 0.3$ is $\alpha_{double} = 0.163$.

3.1.3 Exponential Smoothing for Demand Patterns with Trend and Seasonalities

Demand patterns with trend and seasonalities (Figure 10c) can exhibit, in general, two seasonal behaviors:

- the multiplicative seasonal effect by which the amplitude of the seasonal swing is proportional to the sales level;

$$\text{Demand}_t = (\text{Base} + \text{Trend}) \cdot (\text{Seasonal factor})_t$$

- the additive seasonal effect, in which case the amplitude of the seasonal pattern does not depend on the level of sales:

$$\text{Demand}_t = \text{Base} + \text{Trend} + (\text{Seasonal factor})_t$$

Since in most cases the amplitude of the seasonal pattern is propor-

tional to the level of sales (Makridakis and Wheelwright [1978, p. 199]), we develop in this section the multiplicative model. Its form is given by Winters [1960]:

$$A_t = (a + bt)c_t + \varepsilon_t \tag{80}$$

where all notations have been explained previously except c_t which is the seasonal factor associated with period t. The amount between parentheses contains no seasonal effects and can be, therefore, forecast by a system similar to (73):

$$F_T^{tr} = F_T^b + \frac{1}{\alpha_1}\mathcal{T}_T \tag{81}$$

$$F_T^b = \alpha_1 \frac{A_{T-1}}{SF_{[T-1-L]}} + (1-\alpha_1)F_{T-1}^b \tag{82}$$

$$\mathcal{T}_T = \alpha_2(F_T^b - F_{T-1}^b) + (1-\alpha_2)\mathcal{T}_{T-1} \tag{83}$$

The above equations are written at the end of period T-1.

In (82) $SF_{[T-1-L]}$ is the smoothed seasonal factor for period T-1 updated L periods ago, where L is the length of the seasonal cycle. In many cases the seasonal cycle is one year, which makes L=12 months or L=13 four week periods. By dividing in (82) the actual demand A_{T-1} by $SF_{[T-1-L]}$ the seasonal component is removed; hence, only the permanent component and the trend enter the smoothing process of F_T^{tr}.

At the end of period T-1 the estimate of the seasonal factor has to be updated. Its new value $SF_{[T-1]}$ is given by:

$$SF_{[T-1]} = \alpha_3 \frac{A_{T-1}}{F_T^b + \frac{1-\alpha_1}{\alpha_1}\mathcal{T}_T} + (1-\alpha_3)SF_{[T-1-L]} \tag{84}$$

This value will be used to deseasonalize the observed data one cycle (L periods) later (e.g., the updated seasonal factor computed at the end of May this year, $SF_{[May]}$, will be used to deseasonalize the actual demand A_{May} recorded during May next year).

In (84) the fraction $(A_{T-1})/(F_T^b + \frac{1-\alpha_1}{\alpha_1} \mathcal{T}_T)$ represents the current observed seasonal variation since, as shown in 3.1.2, the denominator is an asymptomatically unbiased estimate of a+b(T-1).

α_3 is the smoothing constant used for seasonal factor updating.

If F_T^{tr+s} is the demand forecast made at the end of period T-1 for period T, then:

(85) $$F_T^{tr+s} = (F_T^{tr}) \cdot (SF_{[T-L]})$$

For any other more distant period T+τ the forecast is obtained with:

(86) $$F_{T+\tau}^{tr+s} = (F_T^{tr} + \tau \mathcal{T}_T) \cdot (SF_{[T+\tau-L]})$$

Thus, we need the seasonal factor for period T+τ which was last updated at the end of period T+τ-L. If $\tau \geq L$ the appropriate seasonal factor has to be reused cyclically.

Initialization procedures, that are all quite similar, have been proposed by Winters [1960], Johnson and Montgomery [1974, ch. 6-4], Montgomery and Johnson [1976, ch. 5]. Starting values $(F_{T-1}^b)_{initial}$, $(\mathcal{T}_{T-1})_{initial}$, and $(SF_{[T-1-L]})_{initial}$ have to be provided.

Suppose that a demand history for the last N seasonal cycles, of L periods each is available (figure 13). Let $\bar{A}_i$ be the per period average of the observations during the i-th season, i=1,...,N.

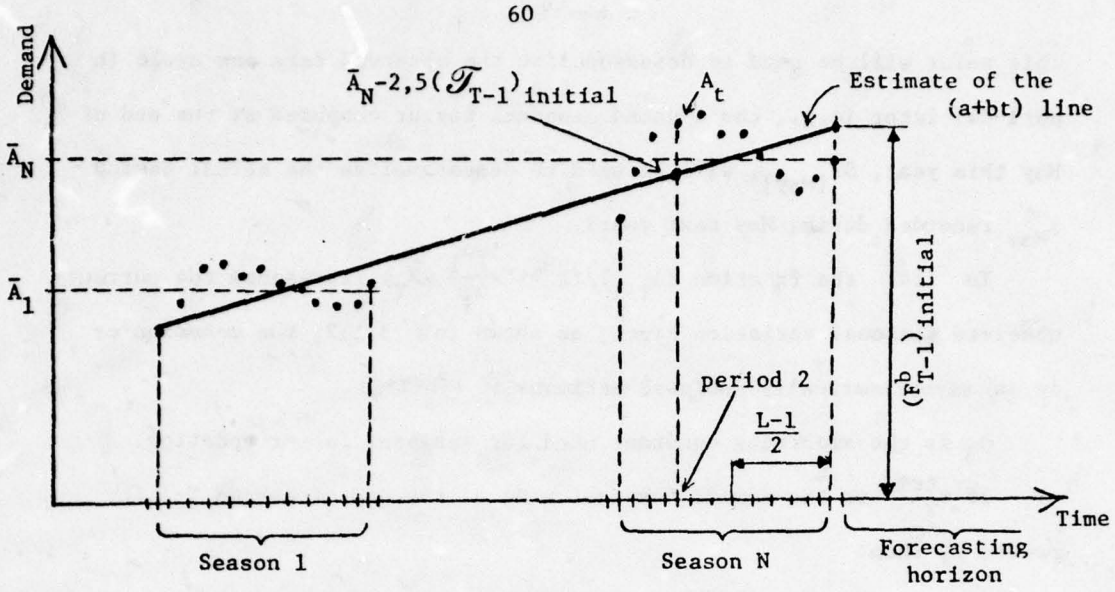


Figure 13: Initialization of the forecasting process for trend and seasonalities

This initial estimate of the trend can be computed from the formula:

$$(\mathcal{T}_{T-1})_{initial} = \frac{\bar{A}_N - \bar{A}_1}{(N-1)L} \tag{87}$$

From Figure 13 it is apparent that (87) gives the slope of the line estimating a+bt.

The starting value for the permanent (base) component can be computed from:

$$(F^b_{T-1})_{initial} = \bar{A}_N + \frac{L-1}{2}(\mathcal{T}_{T-1})_{initial} \tag{88}$$

For each of the N·L periods a seasonal factor is calculated as the ratio of actual sales for the period to the height of the corresponding point on the estimated (a+bt) line:

$$SF_t = \frac{A_t}{\bar{A}_i - [(L+1)/2-j](\mathcal{T}_{T-1})_{initial}} \quad , \quad t=1,\ldots,NL \tag{89}$$

where $\bar{A}_i$ is the average for the season which includes period t, and j is the position of the period within the season. In Figure 13 the case with i=N and j=4 is illustrated.

By (89) N values of the seasonal factor are produced for each period of the season. Their average is then taken to obtain a single estimate for each period:

$$\overline{SFt} = \frac{1}{N} \sum_{k=0}^{N-1} SF_{t+kL}, \qquad t=1,\ldots,L.$$

Finally, the seasonal factors have to be normalized so that they add to L. By this step we make sure that the seasonal factors produce only seasonal adjustments, without increasing or decreasing the average level of demand:

$$(SF_t)_{initial} = \overline{SF}_t \cdot \frac{L}{\sum_{t=1}^{L} \overline{SF}_t}$$

The forecast for the first period of the forecasting horizon is:

$$(90) \qquad F_1^{tr+s} = [(F^b_{T-1})_{initial} + (\mathcal{T}_{T-1})_{initial}](SF_1)_{initial}$$

where $(SF_1)_{initial}$ is the initial estimate of the seasonal factor appropriate for period 1 of the forecasting horizon.

After the actual demand in period 1 A_1 becomes known, the forecasting can be automatically produced by using equations (81) - (86). In equation (84), used here to update the seasonal factor at the end of period 1, the value to use for $SF_{[T-1-L]}$ is $(SF_1)_{initial}$.

An alternative way for initializing the forecasting procedure is shown in the original paper by Winters [1960].

McClain [974], and McClain and Thomas [1973] found that under changing conditions, exponentially smoothing models, like Winter's, can possibly

induce instability of the forecast especially if the trend smoothing constant α_2 is large. In order to avoid the oscillatory behavior Peterson and Silver [1979] recommend for α_2 values not in excess of 0.05; they also suggest that the set of smoothing constants $\alpha_1 = 0.2$, $\alpha_2 = 0.05$, $\alpha_3 = 0.1$ has proven reasonable in many cases.

It is important to mention that special care should be given to the initialization of the seasonal factors because they are re-estimated only once a cycle; therefore the effect of their starting values is felt longer, especially when the smoothing constants are small.

Seasonal demand and more complex shapes of pattern can be described by appropriately combining polynomial, trigonometric, and exponential functions of time.

Thus, a simple sinusoidal model

$$A_t = a_1 + a_2 \sin \frac{2\pi T}{12} + \varepsilon_t$$

can represent a sales pattern observed twelve times during the seasonal cycle. The origin of the sine wave may be shifted t_o periods by using:

$$A_t = a_1 + a_2 \sin \frac{2\pi}{12} (t-t_o) + \varepsilon_t$$

which is equivalent to:

$$A_t = a_1 + a_2' \sin \frac{2\pi t}{12} + a_3 \cos \frac{2\pi t}{12} + \varepsilon_t$$

If, besides seasonalities, a linear trend is also present a simple twelve-point model to reflect this is:*

---

* Unless models are derived from each other, coefficients a_1, a_2, ... used in some expression are in no way connected with coefficients with the same notation involved in another expression.

$$A_t = a_1 + a_2 t + a_3 \sin \frac{2\pi t}{12} + a_4 \cos \frac{2\pi t}{12} + \varepsilon_t$$

In many cases as the level of sales increases with time the seasonal effect also grows, so that:

$$A_t = a_1 + a_2 t + (a_3 + a_4 t) \sin \frac{2\pi t}{12} + (a_5 + a_6 t) \cos \frac{2\pi t}{12} + \varepsilon_t$$

Including harmonics of the basic wave forms allows us to increase the descriptive accuracy of the models.

In dealing with these more complex types of forecasting models, as it was also the case with the simpler ones presented earlier, one has to find an iterative procedure for updating the estimates of the coefficient values with each new observation.

Brown [1963, ch. 11, 12] develops a procedure by which the vector of coefficients in the model is updated directly, by taking linear combinations of the old coefficient estimates (made at the end of the previous period) and the most recently observed one period forecast error. Similarly to the exponential smoothing techniques seen earlier, the approach is to gradually discount the data in time; the criterion observed in estimating and updating the coefficient values is the minimization of squared residuals. Unlike the smoothing presented before, however, the model parameters are smoothed directly rather than through the use of exponentially smoothed statistics.

The discount factor β_k (where k is the number of coefficients in the forecasting model), $0 < \beta_k \leq 1$, is taken as $\beta_k = 1-\alpha_k$ where α_k is the smoothing constant. The larger the number of coefficients (terms) in the model the more descriptive or tracking built-in ability it has and, therefore, the smoothing constant can take on smaller values. If α_1 is the smoothing constant used in simple exponential smoothing the following should hold:

$$(1 - \alpha_k)^k = 1 - \alpha_1 \tag{91}$$

The direct revision of coefficients is shown to be generally appropiate for all forecasting models which consist of time functions generated by a transition matrix, i.e., functions whose values at time period t+1 are linear combinations of the same functions evaluated at the previous time t.*

When the forecasting model contains only polynomial functions of time, and after the transient effect of initial conditions becomes negligible, general exponential smoothing produces forecasts identical to multiple smoothing. Thus, for instance, general exponential smoothing (in steady state) produces for the linear trend model (Brown [1963, p. 172]) the same coefficients we have seen earlier resulting from double exponential smoothing [expressions (77), (79)].

3.2 Forecasting Over Lead Times

In production and inventory systems the lead time starts when a replenishment order is triggered or a shop order is released, and it ends whenever material is received. Since there is always the possibility that available inventory will not last through the lead time, it is important to forecast demand over lead times in order to be able to set appropriate order points.

We have shown in the preceding section how to produce forecasts for future periods. Then, if we are now at the end of period T-1, the forecast of cumulative demand for a lead time of ℓ periods is:

$$(F\ell)_T = \sum_{\tau=0}^{\ell-1} F_{T+\tau}^{\bullet} \tag{92}$$

---

* This class of functions includes polynomials, sinusoids, exponentials, and sums and products of them.

where:

$(F\ell)_T$ = forecast over an ℓ period lead time extending from period T through period $T+\ell-1$

$F^{\bullet}_{T+\tau}$ = forecast made at the end of T-1 for period $T+\tau$. The appropriate forecats have to be used depending on the demand pattern (base, trend, seasonalities).

If the lead time is slightly variable, we should consider an interval long enough to include all reasonable cases, say to 0.8 fractile.

If the lead time is highly and unpredictably variable, we should measure the demand during the lead time directly and then forecast demand per lead time rather than per period.

If the inventory is reviewed after each transaction, the lead time is exactly the time required to deliver the item to stock. However, if the inventory is reviewed on a regular schedule, we must add one review period to the delivery lead time, the review period being the interval between regular reviews.

3.3 Forecast Errors

The random elements in the demand pattern generate errors in the forecast of the expected demand. The noise in the demand process means that the observed demand does fluctuate around the average process and these fluctuations in turn create errors in our estimate of what the average process is. We define the forecast error E_T in period T as the observed discrepancy between the forecast F_T for period T and the actual observation A_T for that period:*

(93) $$E_T = F^{\bullet}_T - A_T$$

* As mentioned earlier, the appropriate forecast $F^{\bullet}_T$ has to be used depending on the underlying demand pattern.

$F_T^{\cdot}$ is the forecast for the next period; therefore we called E_T the one period forecast error.

As forecast errors are a reflection of the uncertainties inherent in the demand process, knowledge of forecast error's characteristics is central to setting safety stocks (order points) to provide protection against stockouts during lead time.

3.3.1 One Period Forecast Errors

In manufacturing environments, where consumer demand has to be fed up several levels of distribution, forecast errors are generally normally distributed. In a retail business, where ultimate consumer demand is served, forecast errors are more likely to be exponentially distributed (Brown [1977, p. 146]).

Since we have repeatedly stressed the use of asymptomatically unbiased statistics in predicting future demand, we should expect the mean of the forecast to be the mean of the observed demand and the mean forecast error to be zero.

The dispersion of the forecast errors is measured by the variance of the distribution of errors. With a mean error of zero, the variance is estimated by the mean square error (MSE). In most applications we may assume the distribution of errors to be stationary and, therefore, we can consider that the squared error fluctuates around a constant level. This immediately suggests using simple exponential smoothing to update the mean square error:

$$(94) \qquad (MSE)_T = \alpha_k E_{T-1}^2 + (1-\alpha_k)(MSE)_{T-1}$$

where:

$(MSE)_T$ = mean square error forecast for period T made at the end of period T-1

E_{T-1} = current forecast error, observed in period T-1

α_k = smoothing constant corresponding to the model used for forecasting; for a model with k coefficients α_k conforms to relation (91).

When a history of demand is available for the item under consideration, an initial value for MSE can be computed by simulating the adopted forecast model over the historical data. The simulated errors (i.e., residuals) are squared and summed. The sum is divided by the number of degrees of freedom to yield the initial MSE. The number of degrees of freedom is given by the difference between the number of errors in the sum and the number of terms in the forecast model.

If the item has no demand history (or the history is too short), an initial value of MSE can be estimated from the variance law (Brown [1977]). The variance law shows that the standard deviation of the forecast errors σ_T and the level of the forecast are related; thus, in the two more common forms of the law the standard deviation is proportional to some power of the forecast: $\sigma_T = a(\dot{F}_T)^b$, with values for b frequently between 0.7 and 0.9, or the variance can be expressed by some polynomial function of the forecast: $\sigma^2 = c(\dot{F}_T) + d(\dot{F}_T)^2$ (Burgin and Wild [1967], Stevens [1974]). Parameters a, b, c, d are particular of all the items that come from a homogeneous population (e.g. items in a product line). Consequently, in systems where forecast models are available and currently used, the variance law can be established by regression by families of items and stored in the data base. When products with no history are introduced, first an initial forecast is generated and then, from the appropriate variance law, the initial estimate of the variance of the forecast errors can be obtained.

For safety stock calculations one needs an updated forecast of the standard deviation of the errors $\hat{\sigma}_T$ for the upcoming period T; this is

given by:

$$\hat{\sigma}_T = \sqrt{(MSE)_T} \tag{95}$$

In earlier treatments of forecasting the standard deviation of errors was estimated by multiplying the mean absolute deviation by a factor of 1.25 (Brown [1963, ch. 19], Montgomery and Johnson [1976, ch. 7]). The mean absolute deviation was updated by simple exponential smoothing.

$$(MAD)_T = \alpha_k |E_{T-1}| + (1-\alpha_k)(MAD)_{T-1}$$

where $(MAD)_T$ is the mean absolute deviation (i.e., the average value of the absolute forecast error) updated at the end of period T-1. For initialization of the smoothing equation, see Montgomery and Johnson [1976, ch. 7-4]. α_k is the smoothing constant corresponding to a k-term model.

Even though the procedure is only an approximation, it was initially used because it held computational advantages relevant at the time when computing equipment was still in an incipient stage (it avoids storing numbers with numerous digits resulting from squaring observed errors, and it does not involve extracting square roots). Later, then, it became rooted as a tradition in the literature and in the computerized inventory systems. However, with the new power acquired by the large modern computers, the quest for this sort of computational simplicty is no longer justified and, therefore, the more accurate procedure based on MSE calculations may be adopted.

3.3.2 Errors in Forecasting Over Lead Times

The value of $\hat{\sigma}_T$ provides a measure of the accuracy of the one period demand forecast. As mentioned earlier, however, our interest is to forecast the demand over the lead time and, therefore, it is the dispersion of the forecast error over the lead time which is significant to measure.

When the lead time is known with certainty, encompassing ℓ periods, the estimated standard deviation $(\hat{\sigma}_\ell)_T$ of the forecast errors over the lead time* is expressed, in most practical applications, in terms of the updated standard deviation estimate $\hat{\sigma}_T$ of the one period errors.

By definition:

$$(96) \qquad (E_\ell)_T = \sum_{\tau=0}^{\ell-1} F^{\cdot}_{T+\tau,[T-1]} - \sum_{\tau=0}^{\ell-1} A_{T+\tau} = \sum_{\tau=0}^{\ell-1} (F^{\cdot}_{T+\tau,[T-1]} - A_{T+\tau})$$

where:

$(E_\ell)_T$ = the cumulative forecast error over an ℓ period lead time extending from period T through period T+ℓ-1

$F_{T+\tau,[T-1]}$ = demand forecast for period T+τ made at the end of period T-1

$A_{T+\tau}$ = actual demand during period T+τ.

If all noise samples ε_t are serially independent (i.e., demands in nonoverlapping time periods are independent) $F^{\cdot}_{T+\tau,[T-1]}$ and $A_{T+\tau}$ are also independent because $F^{\cdot}_{T+\tau,[T-1]}$ is based on demand outcomes up to and including period T-1 and $A_{T+\tau}$ is independent of the demand in any other time period. Therefore:

$$(97) \qquad (\sigma_\ell)^2_T = \mathrm{Var}\left(\sum_{\tau=0}^{\ell-1} F^{\cdot}_{T+\tau,[T-1]}\right) + \ell\sigma^2_\varepsilon$$

Equation (97) shows that the variation in the cumulative forecast error comes from two sources: the variation in the forecast resulting from noise samples in the observations before period T, and the noise that

---

* Estimate $(\hat{\sigma}_\ell)_T$ is computed at the end of period T-1 and refers to a lead time of length ℓ spanning periods T, T+1, ..., T+ℓ-1.

affects the true process throughout the lead time. We should observe that even though the cumulative forecast over the lead time is the sum of ℓ period forecasts, the cumulative forecast variance is not the sum of the period forecast variances. The reason is that all forecasts $F_{T+\tau,[T-1]}$ are based upon the same demand information and therefore are serially correlated.

To illustrate let us consider the constant demand pattern (Figure 10a) for which the forecasting is done with the exponential smoothing model (60):

$$F_T^b = \alpha A_{T-1} + (1-\alpha)F_{T-1}^b$$

Forecast F_T^b is also the forecast for every period of the lead time.

By (61) and for a large enough T:

$$F_T^b = \frac{\alpha}{\beta} \sum_{t=1}^{\infty} \beta^t A_{T-t}$$

The variance of the forecast is then:

$$\text{Var}[F_T^b] = \frac{\alpha^2}{\beta^2} (\beta^2+\beta^4+\beta^6+\ldots)\sigma_\varepsilon^2 = \frac{\alpha^2}{\beta^2} \cdot \frac{\beta^2}{1-\beta^2} \sigma_\varepsilon^2$$

$$\text{Var}[F_T^b] = \frac{\alpha}{1+\beta} \sigma_\varepsilon^2$$

The cumulative forecast is:

$$\sum_{\tau=0}^{\ell-1} F_{T+\tau,[T-1]}^b = \sum_{\tau=0}^{\ell-1} F_T^b = \ell F_T^b$$

Hence, the variance of the cumulative forecast is $\ell^2 \frac{\alpha}{1+\beta} \sigma_\varepsilon^2$.

By (97) the variance of the forecast error over the lead time is given by:

$$(\sigma_\ell)_T^2 = (\ell^2 \frac{\alpha}{1+\beta} + \ell)\sigma_\varepsilon^2 \tag{98}$$

Table 1 lists the values of $(\ell^2 \frac{\alpha}{1+\beta} + \ell)$ for four levels of the smooth-constant α.

Table 1: Variance of the cumulative forecast error as a function of σ_ε^2 for the constant demand pattern

Lead Time ℓ	α=0.3	α=0.2	α=0.1	α=0.05
1	1.1765	1.1111	1.0526	1.0256
2	2.7059	2.4444	2.2105	2.1026
3	4.5882	4.0000	3.4737	3.2308
4	6.8235	5.7778	4.8421	4.4103
5	9.4118	7.7778	6.3158	5.6410
6	12.3529	10.0000	7.8947	6.9231
7	15.6471	12.4444	9.5789	8.2564
8	19.2941	15.1111	11.3684	9.6410
9	23.2941	18.0000	13.2632	11.0769
10	27.6471	21.1111	15.2632	12.5641
11	32.3529	24.4444	17.3684	14.1026
12	37.4118	28.0000	19.5789	15.6923

In order to use Table 1 one has to have an estimate of the noise variance σ_ε^2. First we recall that the mean square error is an estimate of the one period forecast error variance:

(99) $$\hat{\sigma}_T^2 = (MSE)_T$$

Then, (98) provides the relationship between σ_T^2 (make ℓ=0) and σ_ε^2 in the following form:

(100) $$\sigma_T^2 = c_1\sigma_\varepsilon^2$$

where constant c_1 is found in the first line of Table 1.

An estimate of σ_ε^2 can be obtained by:

(101) $$\hat{\sigma}_\varepsilon^2 = \frac{(MSE)_T}{c_1}$$

As mentioned earlier, however, it is customary to compute the standard deviation $(\sigma_\ell)_T$ of the cumulative forecast error in terms of the standard deviation σ_T of the one period errors. The relationship can be expressed

by a coefficient d_ℓ:

$$(102) \qquad d_\ell = \frac{(\sigma_\ell)_T}{\sigma_T} = \sqrt{\frac{\ell^2 \frac{\alpha}{1+\beta} + \ell}{\frac{\alpha}{1+\beta} + 1}}$$

Coefficients d_ℓ are given in Table 2 for various values of the smoothing constant.

Table 2: Standard deviation of the cumultative forecast error as a function of σ_T for the constant demand pattern

Lead Time ℓ	α=0.3	α=0.2	α=0.1	α=0.05
1	1.0000	1.0000	1.0000	1.0000
2	1.5166	1.4832	1.4491	1.4318
3	1.9748	1.8974	1.8166	1.7748
4	2.4083	2.2804	2.1448	2.0736
5	2.8284	2.6458	2.4495	2.3452
6	3.2404	3.0000	2.7386	2.5981
7	3.6469	3.3466	3.0166	2.8373
8	4.0497	3.6878	3.2863	3.0659
9	4.4497	4.0249	3.5496	3.2863
10	4.8477	4.3589	3.8079	3.5000
11	5.2440	4.6904	4.0620	3.7081
12	5.6391	5.0200	4.3128	3.9115

A similar sort of analysis can be conducted for any forecasting model; the algebra, however, tends to become very tedious. The general exponential smoothing approach mentioned earlier, having been developed in matrix form offers the important advantage of a systematic way of proceeding through the computations.*

Thus, for the linear trend model $A_t = a+b_t+\varepsilon_t$ general exponential smooth-

* General exponential smoothing is not reproduced here since the interested reader can find its full development in a number of references like Brown [1963], Montgomery and Johnson [1976], Johnson and Montgomery [1974].

ing (which in steady state yields the same coefficeints as double exponential smoothing) leads to the following variance of the cumulative forecast error:

$$\text{Var}\left(\sum_{\tau=0}^{\ell-1} F^{tr}_{T+\tau,[T-1]}\right) = \ell^2 \frac{\alpha_2}{2(1+\beta_2)^3}\left[5(1+2\beta_2+\beta_2^2) + 4(1-\beta_2^2)\ell + \alpha_2^2\ell^2\right]\sigma_\varepsilon^2$$

where $F^{tr}_{T+\tau,[T-1]}$ is the forecast for period $T+\tau$ made at the end of period $T-1$, and α_2 denotes the smoothing constant to be used in the two coefficient forecasting model. By (97) one can compute the variance of the cumulative forecast error:

$$(\sigma_\ell)^2_T = \left\{\ell^2 \frac{\alpha_2}{2(1+\beta_2)^3}\left[5(1+2\beta_2+\beta_2^2) + 4(1-\beta_2^2)\ell + \alpha_2^2\ell^2\right] + \ell\right\}\sigma_\varepsilon^2$$

From the above equation it is apparent, as it also was in the case of the constant demand pattern, that the variance of the cumulative forecast error is linearly related to the variance of the noise. This holds true, in general, for forecasting models of any complexity if general exponential smoothing procedures can be applied. If c_ℓ stands for the proportionality coefficient, then:

$$(\sigma_\ell)^2_T = c_\ell\sigma_\varepsilon^2$$

Table 3 shows the values of c_ℓ for the linear trend model for various combinations of lead times and smoothing constants, and Table 4 contains the standard deviation of the cumulative forecast error $(\sigma_\ell)_T$ as a function of the standard deviation σ_T of the one period error. The smoothing constant α_2 conforms to relation (91):

Simple smoothing constant α_1	0.3	0.2	0.1	0.05
Equivalent smoothing constant α_2	0.16334	0.10557	0.05132	0.02532

Table 3: Variance of the cumulative forecast error as a function of σ_ε^2 for the linear trend pattern

Lead Time ℓ	α_2=0.16334	α_2=0.10557	α_2=0.05132	α_2=0.02532
1	1.2385	1.1456	1.0672	1.0324
2	3.0215	2.6083	2.2746	2.1309
3	5.4566	4.4285	3.6307	3.2975
4	8.6597	6.6487	5.1445	4.5342
5	12.7554	9.3133	6.8251	5.8432
6	17.8765	12.4691	8.6817	7.2264
7	24.1644	16.1646	10.7239	8.6861
8	31.7688	20.4506	12.9614	10.2242
9	40.8479	25.3801	15.4041	11.8432
10	51.5684	31.0078	18.0624	13.5449
11	64.1054	37.3908	20.9466	15.3319
12	78.6425	44.5882	24.0672	17.2061

Table 4: Standard deviation of the cumulative forecast error as a function of σ_T for the linear trend pattern

Lead Time ℓ	α_2=0.16334	α_2=0.10557	α_2=0.05132	α_2=0.02632
1	1.0000	1.0000	1.0000	1.0000
2	1.5619	1.5089	1.4599	1.4367
3	2.0990	1.9661	1.8445	1.7872
4	2.6443	2.4091	2.1956	2.0957
5	3.2092	2.8512	2.5289	2.3790
6	3.7992	3.2991	2.8521	2.6457
7	4.4171	3.7563	3.1699	2.9006
8	5.0647	4.2251	3.4849	3.1470
9	5.7430	4.7068	3.7992	3.3870
10	6.4527	5.2026	4.1139	3.6222
11	7.1945	5.7130	4.4302	3.8537
12	7.9686	6.2387	4.7488	4.0824

From inspecting Tables 1 - 4 it is apparent that with smaller smoothing constants the effect of the variance of the cumulative forecast is diminished, and the variance of the cumulative error is dominated by the accumulated noise. This is to say that along with the reduction in the serial

correlation of demand forecasts over the lead time, coefficient c_ℓ tends toward the limiting case $c_\ell = \ell$, or $(\sigma_\ell)_T = \ell^{0.5}\sigma_T$.

As mentioned at the beginning, the above results depend on exact independence of the random variation in demand. However, in reality one can expect to find some (unknown) serial correlation in the noise, which makes the variation of the cumulative forecast error depend on lead time somewhat differently than inferred from the previous theoretical results. However, the qualitative findings, like the impact of the smoothing constant on the serial correlation of the forecasts, will still stay valid.

Empirically, the following expression has been found to give a fairly accurate representation in many inventory systems:

$$(\sigma_\ell)_T = \ell^B\sigma_T \tag{103}$$

where B is a power that expresses the relationship between the cumulative forecast error and the duration of the lead time. B is a constant characteristic of all items that come from a homogeneous population, like items in a product line, and can be obtained by regression. Experience shows that for most demand behavior B lies between 0.5 and 1.0. Notice that B=0.5 is the limiting case which implies that the one period forecast errors are independent random variables.

For practical inventory control purposes the theoretical results can also be approximated quite well, in the range $1 \leq \ell \leq 12$, by the following linear function (Brown [1967, p. 144]):

$$(\sigma_\ell)_T = (0.659 + 0.341\ell)\sigma_T \tag{104}$$

For comparison, Figure 14 presents the variation of the $[(\sigma_\ell)_T]/[\sigma_T]$ ratio with lead time ℓ. Brown's function (104) approximates fairly well the theoretical results dervied for the constant demand process (see

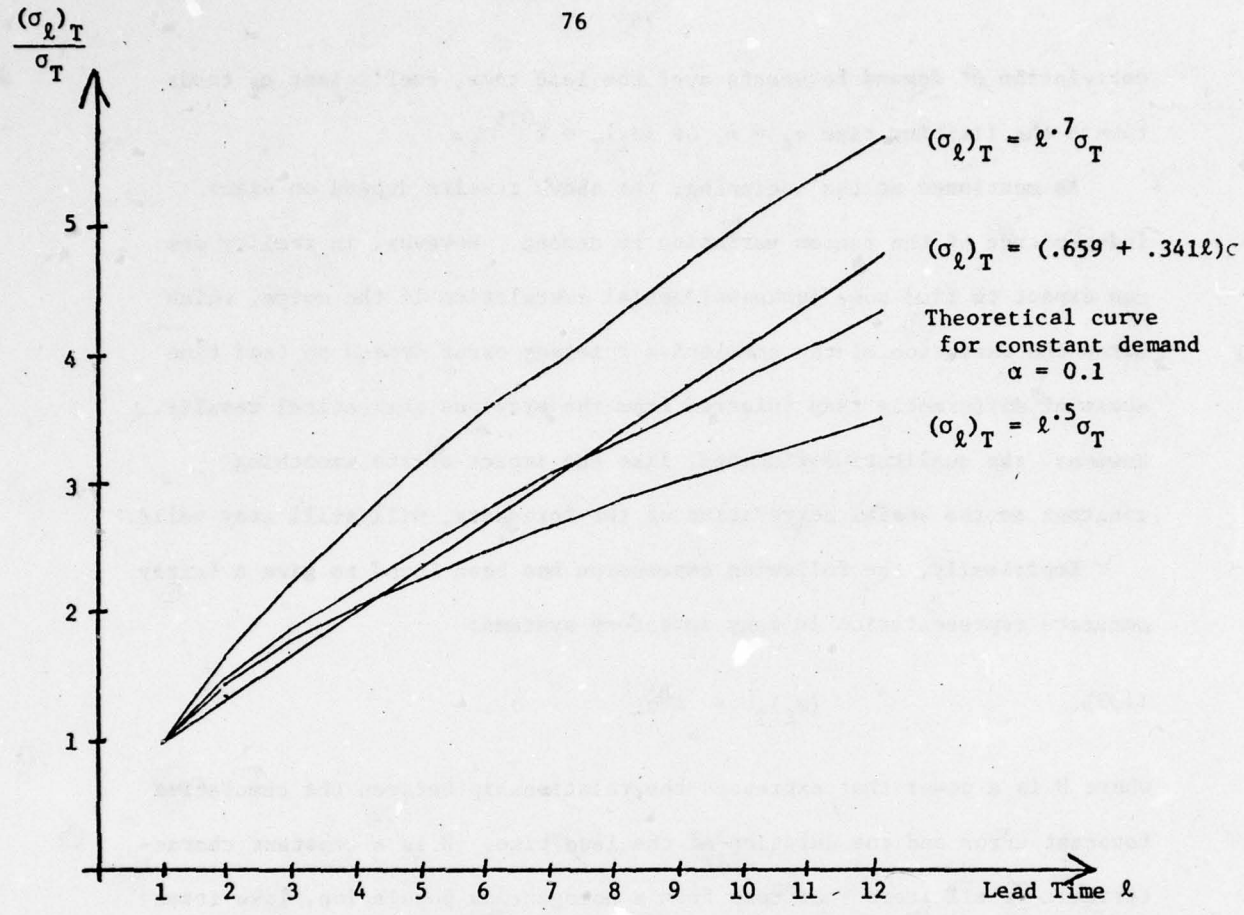


Figure 14: Theoretical and empirical functions for the standard deviation of the cumulative forecast error

Table 2) with a moderate smoothing constant $\alpha = 0.1$ (the approximation is equally good for the linear trend pattern). The limiting case $(\sigma_\ell)_T = \sqrt{\ell}\sigma_T$ is also shown. As the power of ℓ is increased ($\ell^{.7}$ in Figure 14) relation (103) can be used to reflect the effect of the serial correlation of the forecasts upon the variation of the cumulative forecast errors.

When lead times are variable, forecasting should be done per lead time directly rather than by period, and then accumulated. The errors over the lead time can be measured and the mean square error can be updated

by (94) to obtain estimates of the standard deviation of the errors. Starting values in the smoothing equation (94) can be produced by use of the variance law applied to the demand forecast for the upcoming lead time.

3.4 Forecasting Slow Moving Items

Slow moving items are those which have a low level of demand and frequent periods of no usage. It is difficult to define a threshold above which an item would be classified as a fast mover and below which as a slow mover mainly because the threshold depends on the nature of the item: what would be low demand (in units) for some item might very well represent a high demand for some other item. In any case, it seems very unlikely that the separation point between high and low usage should ever be higher than 100 units per year and more likely should be around 50 units per year. Peterson and Silver [1979] recommend classifying items according to their demand over the replenishment lead time: an expected lead time demand of 10 units or larger puts the item in the fast movers' class, while an expected lead time demand of less than 10 units defines a slow mover.

Slow moving items tend to exhibit a lumpy demand pattern; thus, there may be several consecutive periods in which there is no demand at all, followed by one or several periods during which transactions of various sizes take place. Of course, not all lumpy demand items are slow movers and not all slow movers have lumpy demand. In assembly operations most component parts present lumpy requirements induced by the batch type production normally encountered in manufacturing; this does not mean, though, that all those parts are slow movers. Conversely, some very specific foodstuff carried by a supermarket and purchased by the members of an ethnic community might exhibit continuous sales, but in small overall

amounts; thus, that foodstuff can be a slow mover without having a lumpy demand.

In this section we are interested in slow movers with lumpy demand, since the discontinuities in requirements make forecasting methods presented before become inappropriate. Indeed, if ordinary exponential smoothing is used, demand forecasts will tend to be much lower than the average demand per period at the time an order is received, and much higher than the average demand just after an order has been received. This tends to increase the forecast errors.

To illustrate, consider a very simple case in which every third period a demand of size s occurs. If an exponential smoothing forecasting model is used, after the initial conditions wear off it will produce demand forecasts according to a cyclical pattern repeating every 3 periods. Let period zero be the period in which demand occurs. Then, F_0 or F_{before} is the demand forecast made one period earlier for period zero (we denoted it F_{before} to show that the forecast was produced before the demand transaction of size s took place).

At the end of period zero, a forecast F_1 of F_{after} is made for period 1 by exponential smoothing with constant α (we called in F_{after} as opposed to F_{before}):

$$F_1 \equiv F_{after} = (1-\alpha)F_{before} + \alpha s$$

At the end of period 1, with no demand, we forecast:

$$F_2 = (1-\alpha)^2 F_{before} + \alpha(1-\alpha)s$$

At the end of period 2, again without demand, when the cycle closes we produce a forecast F_3 that is equal to F_{before}:

$$F_3 = F_{before} = (1-\alpha)^3 F_{before} + \alpha(1-\alpha)^2 s$$

The graphical representation of Figure 15 shows the cyclical pattern of forecasts for $\alpha = .2$. For this value of the smoothing constant one obtains:

$$F_{before} = \frac{\alpha(1-\alpha)^2}{1-(1-\alpha)^3} s = 0.262s$$

$$F_1 = 0.41s$$

$$F_2 = 0.328s$$

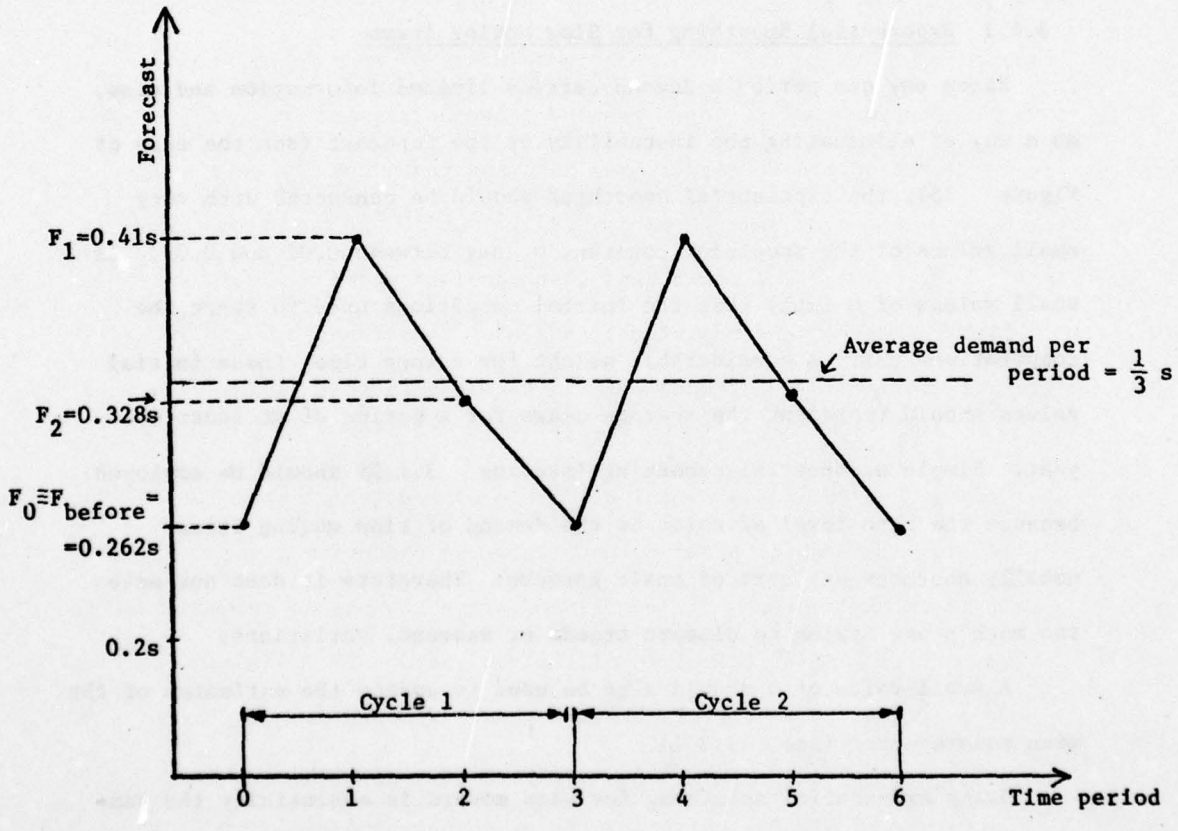


Figure 15: Cyclical pattern of exponential smoothing for a lumpy demand item

In order to establish whether an item is lumpy or not, Brown [1977, ch. 12.3] suggests the following procedure: a forecast model is fit and then simulated over historical data. The residual differences between the data and the model are recorded and their standard deviation is calculated. If that standard deviation is greater than the level in the forecast model, the item is said to have lumpy demand.

Various forecasting methods have been proposed for slow moving items, ranging from simple exponential smoothing to more elaborate techniques. They involve different system costs, and to choose one or another depends very much on the importance and value of the controlled item.

3.4.1 Exponential Smoothing for Slow Moving Items

Since any one period's demand carries limited information and also, as a way of eliminating the instability of the forecast (see the case of Figure 15), the exponential smoothing should be conducted with very small values of the smoothing constant α, say between 0.01 and 0.05. As small values of α imply that the initial conditions used to start the computations carry a considerable weight for a long time, these initial values should represent the average usage for a period of at least one year. Simple exponential smoothing (section 3.1.1) should be employed because the high level of noise in the demand of slow moving items usually obscures any sort of basic pattern. Therefore it does not make too much sense trying to discern trends or seasonal variations.

A small value of α should also be used to update the estimates of the mean square error (see 3.3.1).

Using exponential smoothing for slow movers is essentially the same procedure as for fast moving items. This could constitute an important practical advantage for the implementation stage.

3.4.2 *Forecasting the Demand Transaction Size and the Time Between Transactions*

Croston [1972] argues that simple exponential smoothing in the case of intermittent demand items is biased and has a rather large variance. He proposes that the two components: the magnitude of individual demand occurances and the time between consecutive transactions be forecast.

Suppose that we are at the end of period T, during which a demand of size s_T (possibly zero) has been observed. The time elapsed since the last transaction is p_T periods. For normally distributed demand magnitudes, Croston [1972] proposes the following forecasting procedure:

- if $s_T > 0$ (i.e., in period T a demand transaction occurs), the following updating is performed:

(105) $$\hat{s}_{[T]} = \alpha s_T + (1-\alpha)\hat{s}_{[T-1]}$$

(106) $$\hat{p}_{[T]} = \alpha p_T + (1-\alpha)\hat{p}_{[T-1]}$$

where:

$\hat{s}_{[T]}$ = the estimate, at the end of period T, of the average demand transaction size

$\hat{p}_{[T]}$ = the estimated number of periods until the next nonzero demand occurance, computed at the end of period T.

- if $s_T = 0$ (i.e., in period T no demand occurs), the estimates are kept the same as they were at the end of the previous period.

The procedure yields unbiased estimates and a variance lower than the exponential smoothing.

The updating of the mean square error is done only in those periods in which a nonzero demand occurs; the updating is by exponential smoothing

[see Equation (94)].

3.4.3 Alternative Ways of Estimating the Parameters of Theoretical Probability Distributions of Demand

Because of the relatively few transactions occuring with slow moving items during a year, we can assume, in general, that the probability distribution of demand over the lead time is fairly stationary. Based on the item's history or on some initial subjective assessments, we can estimate the probability distribution of demand and then use it to determine safety stocks and order points (to be seen later).

Empirically the Poisson distribution has been found to represent reasonably the distribution of demand for a slow moving item. The Poisson distribution* is completely defined by a single parameter, the mean value of the random variable. If the mean value is λ and the standard deviation σ, then:

$$\sigma = \sqrt{\lambda}$$

For a slow moving item let the random Poisson variable be the lead time demand. At each receipt of replenishment stock the estimate of the lead time demand can be revised by exponential smoothing:

$$(F\ell)_{new} = \alpha(A\ell) + (1-\alpha)(F\ell)_{old} \tag{107}$$

where:

$(F\ell)_{new}$ = the revised forecast for the lead time demand

$(F\ell)_{old}$ = the old forecast for the lead time demand

$A\ell$ = actual demand during the lead time just ended

α = the smoothing constant.

---

* For a presentation of the Poisson distribution and its properties, see Hastings and Peacock [1975].

The mean square lead time demand (MS_ℓ) can also be updated by exponential smoothing (Brown [1977, ch. 12]):

$$(MS\ell)_{new} = \alpha(A\ell)^2 + (1-\alpha)(MS\ell)_{old} \tag{108}$$

where, again, "new" and "old" denote the revised, respectively, the previous estimate.

The standard deviation of lead time demand is estimated by:*

$$(STD\ell) = \sqrt{(MSL)_{new} - (F\ell)^2_{new}} \tag{109}$$

Another possibility is to estimate the expected lead time demand by an average (or moving average) of historical lead time data; similarly, the standard deviation can be computed from actually observed data rather than from smoothed statistics.

Peterson and Silver [1979, ch. 9] recommend that lead time demand be considered Poisson distributed if:

$$0.9\sqrt{(F\ell)} \leq (STD\ell) \leq 1.1\sqrt{(F\ell)} \tag{110}$$

i.e., if the standard deviation is within 10% of $\sqrt{(F\ell)}$. When (4.110) is not satisfied, the Poisson model is inappropriate and the Laplance distribution** is suggested for use. The Laplace distribution holds the advantage of being a continuous function and it is often easier to work analytically with models of inventory systems if the variables can be treated as continuous. This allows one to eliminate the problems caused by discreteness and to take derivatives instead of working with differences.

---

* The variance σ^2 of a random variable x can be computed as: $\sigma^2 = E(x^2) - (E[x])^2$.

** The Laplace distribution is also known as the "bilateral exponential" (Feller [1971, Vol. II, p. 49]) since it is a two-sided, symmetric exponential function. Some of its important properties are shown by Peterson and Silver [1979, p. 767].

In general, the Poisson distribution can raise computational problems if used for a large number of items. Therefore, Peterson and Silver recommend the use of the Poisson distribution only for expensive items, and the exclusive use of the Laplace distribution for all other items in inventory.

The test by relation (110) is just a crude and simple way of checking on the goodness of the Poisson assumption. In some cases distributions other than Poisson and Laplace might be recommended; for instance, demand transactions can be generated by a Poisson process while the number of units requested when a demand occurs can vary randomly from demand to demand.* Statistical theory provides goodness of fit tests (see Hoel et al. [1971]) to check whether a sample of observed lead time demands are likely to have been generated by some hypothesized stochastic process. However, complicated distributions and sophisticated goodness of fit tests should be used with extreme caution; normally they bring about an increased system cost that might not be justified unless the item is extremely important or expensive, thus presenting the opportunity of a substantial reduction in overall operating costs coming from an improved inventory policy.

When a new item is introduced on the system or when adequate data is unavailable to estimate the probability distribution of lead time demand, Bayesian forecasting can be resorted to. One starts out with some prior assessments in the form of a probability distribution placed over the possible values of one or several parameters that define the stochastic process generating lead time demands. As new data become available, by observation they are blended with prior information, by use of Bayes' rule,

---

* A simple case of such a demand process is presented by Hadley and Whitin [1963, ch. 3-7]; they call it the "stuttering Poisson" process, in which a Poisson distribution generates the demands, and the number of units requested when a demand occurs has a geometric distribution.

to revise our knowledge about the uncertain parameters. The interested reader is directed, as a starting reference, to the book by Montgomery and Johnson [1976, ch. 10].

3.4.4 Empirically Determined Probability Distributions of Lead Time Demand

So far we have been concerned with fitting some theoretical distribution to the available demand data. Consider now the case where we make use of an empirically determined distribution. To keep things simple, assume that the lead time is a constant ℓ. If a demand history is available, divide it into nonoverlapping time intervals of length ℓ and establish how many units have been requested in each such time interval. The range of possible lead time demands has to be partitioned into a number of n classes by the following class limits:

$$A\ell(0) < A\ell(1) < \ldots < A\ell(n-1) < A\ell(n)$$

where $A\ell$ denotes actual demand in a lead time interval of length ℓ. The class limits should be set so that each observation can be assigned to only one class. A histogram can be constructed to represent the frequency with which the observed lead time demands are falling in each class. The frequency thus determined for the i-th class, having limits $A\ell(i-1)$ and $A\ell(i)$, $i=1,\ldots,n$, is a good estimate of the probability $p(i)$ that a lead time demand will lie in this class (Figure 16).

If no adequate history is available one should start with some initial subjective prediction (similar to the Bayesian prior estimate) of the various probabilities $p(i)$, $i=1,\ldots,n$.

Brown [1963, ch. 13] presents a very simple method for updating probabilities $p(i)$ as new data become available, under the assumption that the probability distribution of lead time demands is stationary or changing very slowly with time.

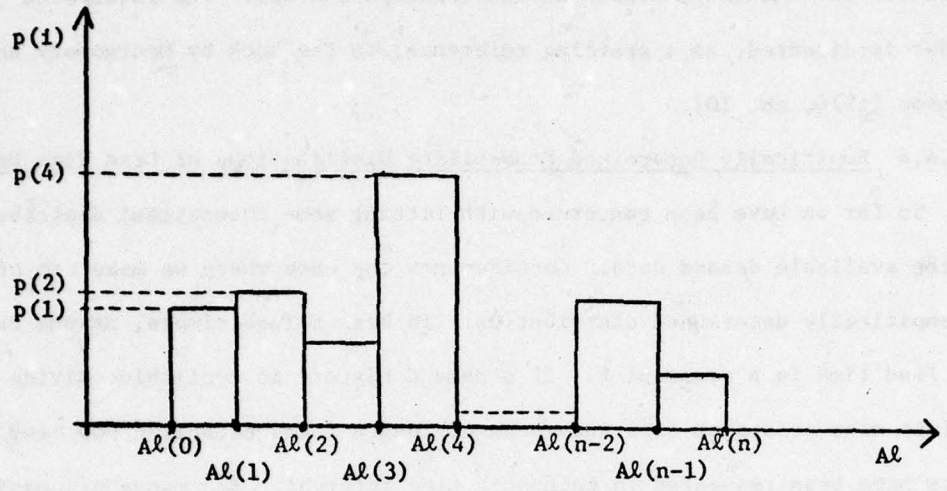


Figure 16: Empirical density

Thus, suppose that a time interval of ℓ periods has just concluded and a demand of Aℓ units has been recorded such that:

$$A\ell(i-1) < A\ell \leq A\ell(i)$$

i.e., Aℓ lies in the i-th class.

At this time old estimates of the probabilities p(i) exist:

$$\mathscr{P}_{old} = \begin{bmatrix} \hat{p}_{old}(1) \\ \hat{p}_{old}(s) \\ \vdots \\ \hat{p}_{old}(n) \end{bmatrix} \tag{111}$$

where $\mathscr{P}_{old}$ is the n×1 column vector of estimates.

Our problem is to revise these estimates in the light of the latest information. To do this define first the following function of Aℓ:

$$u(i) = \begin{cases} 1 & \text{if } A\ell(i-1) < A\ell \leq A\ell(i) \\ 0 & \text{otherwise,} \end{cases} \tag{112}$$

and then construct the n-component column vector $\mathcal{U}$:

$$\mathcal{U} = \begin{bmatrix} u(1) \\ u(2) \\ \vdots \\ u(n) \end{bmatrix}$$

The element of $\mathcal{U}$ whose value is equal to 1 shows the class interval in which the latest recorded actual lead time demand $A\ell$ has fallen; all other elements are zero.

The blending of old and current information is done by vector smoothing according to the rule:

$$\mathcal{P}_{new} = \alpha\mathcal{U} + (1-\alpha)\mathcal{P}_{old} \tag{113}$$

where α is the exponential smoothing constant.

Since in our case $A\ell$ lay in the i-th class, the updated vector $\mathcal{P}_{new}$ has the following components:

$$\begin{aligned} \hat{p}_{new}(1) &= (1-\alpha)\hat{p}_{old}(1) \\ &\vdots \qquad\qquad \vdots \\ \hat{p}_{new}(i-1) &= (1-\alpha)\hat{p}_{old}(i-1) \\ \hat{p}_{new}(i) &= (1-\alpha)\hat{p}_{old}(i-1) + \alpha \\ \hat{p}_{new}(i+1) &= (1-\alpha)\hat{p}_{old}(i-1) \\ &\vdots \qquad\qquad \vdots \\ \hat{p}_{new}(n) &= (1-\alpha)\hat{p}_{old}(n) \end{aligned} \tag{114}$$

All elements of $\mathcal{P}_{new}$ are nonnegative; also, from (114) it follows that $\sum_{k=1}^{n} \hat{p}_{new}(k) = 1$. Therefore, $\mathcal{P}_{new}$ is a probability vector.

It may be shown that the exponential vector smoothing yields unbiased estimates of the true probabilities, and that the variance of the forecast for the i-th probability is:

$$\text{(115)} \qquad \text{Var}[\hat{p}_{new}(i)] = \frac{\alpha}{2-\alpha} p(i)[1-p(i)]$$

From (115) it is obvious that in order to reduce the variance of the forecast, the class limits should be set so that p(i) is either large (close to one) or small (close to zero). Since in most cases in inventory systems the right tail of the distribution is relevant (to be seen later when safety stocks are discussed), Brown recommends that near the tail class limits be set so that class probabilities result in very small values (of the order of 0.02); the rest of the distribution can be considered one event, having a large probability (of the order of 0.9).

The choice of the smoothing constant α goes by reasons similar to the ones put forth in section 3.1.1. If the probability distribution is believed constant in time and the initial conditions are reliable, a small α should be used in order to reduce the variance of the forecast and thus obtain a more stable system. Otherwise, if conditions are likely to change with time and questionable estimates have to be used to initialize the forecasting procedure, a larger smoothing constant is preferred in order to discount starting values faster and to track real changes closer.

The use of empirical distributions looks straightforward and vector smoothing is rather simple to apply. However, because of the scarcity of historical data that plagues many inventory systems (especially manual systems), it might be extremely difficult (if not utterly impossible) to reliably represent the tail of the distribution for the purpose of setting

order points. In such cases the only way out is to use a theoretical distribution whose parameters have to be estimated or guessed from whatever information can be put together.

3.5 Tracking Signals to Monitor the Forecasts

Through the use of unbiased forecast models the mean of the forecast erros should be zero as long as the assumptions on which the forecast is generated are valid. Therefore, so long as the model and the demand generating process are consistent with each other we should expect the forecast errors to fluctuate within reasonable limits around zero. However, if the process changes, thus unvalidating the model, the forecast errors will tend to have repeatedly the same sign, thus moving the average error away from zero. In such cases either automatic procedures are applied to bring the system in line with the changed conditions (see references on adaptive control smoothing, section 3.1.1) or personal intervention is required (to be discussed later).

The purpose of a tracking signal is to detect the situation where the model produces biased forecasts and, thus, to initiate corrective action.

The technique to be presented in this section, suggested by Trigg [1964] and called the smoothed error tracking signal,* computes the expected forecast error by simple exponential smoothing and, then, tests the hypothesis that this value is zero. If the hypothesis is rejected, this constitutes the signal that the forecast model is biased.

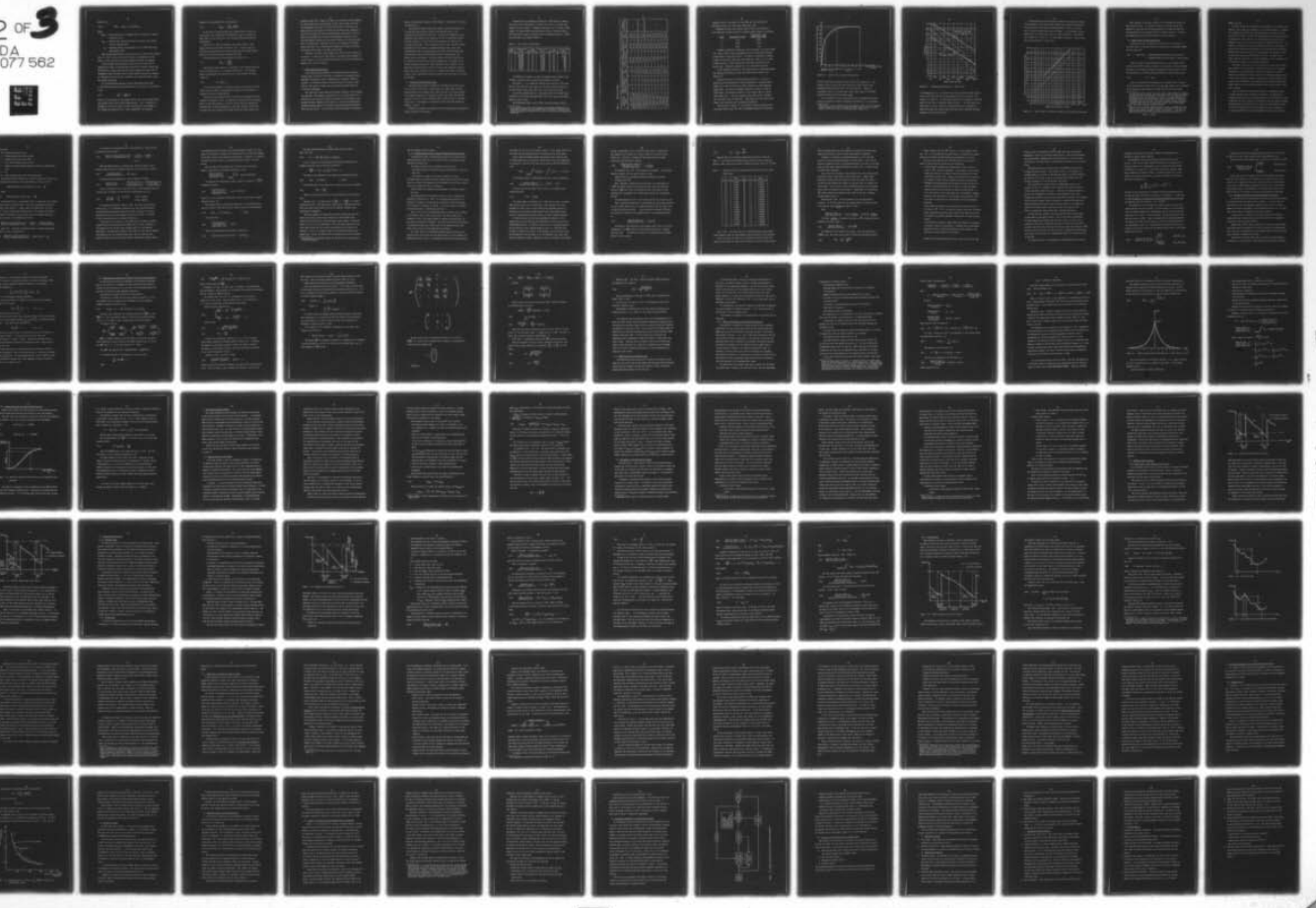
Suppose that a forecast model with k coefficients is currently in use, and we are at the end of period T-1. The smooted error at this point in time, which also represents the error forecast for the upcoming

---

* An alternative technique based on the cumulative forecast error was developed by Brown [1959] but presents weaknesses that make it non-recommended for use; see criticism by Brown [1967, p. 163] himself.

AD-A077 562 MASSACHUSETTS INST OF TECH CAMBRIDGE OPERATIONS RESE--ETC F/G 15/5
INVENTORY MANAGEMENT.(U)
NOV 79 A C HAX , D I CANDEA N00014-75-C-0556
UNCLASSIFIED TR-168 NL

2 OF 3
ADA
077 582



period T, is:

$$(116) \qquad (SE)_T = \alpha_k E_{T-1} + (1-\alpha_k)(SE)_{T-1}$$

where:

$(SE)_T$ = error forecast (i.e., smoothed error) for period T, computed at the end of T-1

E_{T-1} = forecast error actually observed in period T-1 [for definition see Equation (93)]

α_k = smoothing constant corresponding to the k coefficient model [see Equation (91)].

When the smoothing process is started, the initial value of the smoothed error in (116) is set to zero; this reflects our assumption that the model is correct and that the initial parameter estimates are unbiased.

If the forecasts were unbiased one would expect the value of the smoothed error $(SE)_T$ to fluctuate around zero with a constant variance σ^2_{SE}. The derivation of σ^2_{SE} involves tedious algebra because the errors are not independent random variables; this is because demand forecasts are computed as combinations of past data and, therefore, the resulting forecast errors are serially correlated.

For an exponentially smoothed constant demand model Brown [1967, p. 165] derived analytically the expression for the variance of the smoothed error:

$$(117) \qquad \sigma^2_{SE} = \frac{\alpha}{(2-\alpha)^2}\,\sigma^2$$

where σ^2 is the variance of the forecast errors, and α is the simple smoothing constant used for the constant demand pattern. As σ^2 is unknown, an estimate of it has to be used. At the end of period T-1 the estimate is $\hat{\sigma}_T = (MSE)_T$ [see equations (94) - (95)]; consequently, the standard

deviation of the smoothed error is estimated by:

$$(\hat{\sigma}_{SE})_T = \frac{\sqrt{\alpha}}{(2-\alpha)} \sqrt{(MSE)_T} \tag{118}$$

In the case of forecast models with a larger number of terms, the only practical way to approach the estimation of σ_{SE} is by simulation. Brown [1967, ch. 12] did so for a six term model describing trend and seasonalities.

Brown [1977, p. 156] and Montgomery and Johnson [1976, p. 166] suggest that the result given in equation (118) could be used as a reasonable approximation also for models other than the constant demand pattern, for which it had been originally developed.

We can now define the tracking signal for period T as:

$$(TS)_T = \frac{(SE)_T}{(\hat{\sigma}_{SE})_T} \tag{119}$$

The tracking signal measures, in multiples of the standard deviation, how far removed from zero is the updated smooth error. The tracking signal is considered to be satisfactory if it falls between two limits -K and +K:

$$-K \leq (TS)_T \leq K \tag{120}$$

If the tracking signal is outside this range, this should be an indication that the forecast is unacceptably biased and some intervention is required to correct the deficiencies.

The value of the detection limit K is subject to management policies; usually it is between 2 and 5. For expensive or important items the limits have to be set quite tight (i.e., K=2) to be sure that real changes are detected early in their development; in this case, however, one has also to accept the risk of getting a few false alarms when some unlikely (but

possible) large errors trigger the signal, but nothing has really happened. For cheap items, where the processing of false reports might constitute an unacceptable system cost in view of the low value of the items, the limits should be set wider so that when the tracking signal triggered, it would be very likely that some real problems have taken place. Low cost items, that normally do not entail close management attention, can be protected against stockouts by large safety stocks since inventory investment is insignificant. Therefore, even if a real change in the demand pattern has gone on for a while before triggering the signal, the safety stock might still prevent a shortage from occuring.

Certainly, the limits can be changed any time during the operation of the system. If there are too many reports, the limits can be widened; if, on the contrary, it turns out that the system is too slow in detecting real changes, they can be set tighter.

4 Safety Stock Decision Rules

In an inventory system whenever the available inventory reaches a certain level, called the order point, a production or purchasing order, depending on whether the item is manufactured or purchased, should be released. By available inventory we mean the sum of the stock on hand and the stock already on order, less any unfilled customer demand (if it had been backlogged).

The order point should represent sufficient inventory to last through the lead time. When demand is uncertain (as it is in most cases), it is obvious that the order point should be set equal to the forecast of the maximum reasonable demand over the lead time. One way of estimating the maximum reasonable demand is to forecast the expected demand during the lead time (see section 3.2) and then to add an allowance for protection

against the uncertainty inherent in any forecast. This allowance is called the safety stock.

The setting of safety stocks, order points, and other parameters of the inventory system should conform to the characteristics of the controlled items (usage rate and cost). Some items are expensive or are considered very important and, therefore, we like to devote more careful and close attention to them, while low cost products are dealt with on a rather routine basis. From a technical point of view, inventory control procedures that are adequate to manage high usage items, do not work satisfactorily with low usage items and vice versa (we have already made this distinction when we differentiated between forecasting for fast movers and for slow movers). For these reasons, inventory items have to be classified into groups that should be homogeneous with respect to the control procedures that apply. Since investment in the inventory of any given item is proportional to two of the item's most important characteristics, the item's usage and its cost, a commonly used method of classification is the so-called ABC inventory classification according to the annual dollar usage.

4.1 The ABC Inventory Classification

The annual dollar usage is obtained by multiplying the yearly usage of the item by its unit cost. Observations of a large number of multi-item inventories have revealed that a small fraction of the items accounts for a high percentage of the cumulative annual dollar usage while another large percentage of the items represent only a small fraction of the total annual value. This suggests to classify the items into three categories, called A, B, and C.

We present an example to illustrate how the classification is constructed as well as some of its features.

Suppose that we are manging an inventory of 3000 items and, because the system is not computerized, it would be too costly to consider every single record for analysis. Therefore we decide to use a 1 percent random sample,* by picking every one hundredth stock record; from every record the annual dollar usage is read and listed in descending sequence. Assume that the sample is as shown in Table 5.

Table 5: The 1 percent random sample

Item number	Annual dollar usage	Item number	Annual dollar usage	Item number	Annual dollar usage
1	334,369	11	15,835	21	1,495
2	189,094	12	11,271	22	1,451
3	53,104	13	6,634	23	1,394
4	49,514	14	6,374	24	863
5	43,045	15	5,324	25	788
6	33,860	16	4,964	26	384
7	26,903	17	3,533	27	221
8	26,370	18	3,134	28	189
9	18,215	19	2,143	29	122
10	17,501	20	1,926	30	84

To simplify the analysis, we group the sampled items by annual usage intervals, and then compute cumulative percentages as in Table

From Table 6 it is immediately apparent that there are a few items (10% of total inventory) that account for the most substantial portion of the total annual dollar usage (about 67%); these are class A items.

At the other extreme there is a large group (60% of all items) which contributes a very small percentage (about 5%) to the total annual usage; these are class C items.

The intermediate group, class B items, is more balanced in that it

---

* As it becomes clear later, all we need are percentage distributions. As we believe the sample to be representative of the entire inventory population, the sample percentages will be used as estimates of the population percentages.

Table 6: Inventory analysis

Interval of annual dollar usages	Number of items in interval	Cumulative number of items	Cumulative % of items	Annual dollar usage of the items in the interval	Cumulative annual dollar usage	Cumulative % of usage	Inventory classification
Over 200,000	1	1	3.33	334,369	334,369	38.88	A
100,000.01-200,000	1	2	6.67	189,094	523,463	60.87	
50,000.01-100,000	1	3	10.00	53,104	576,567	67.04	
20,000.01-50,000	5	8	26.67	179,692	759,259	87.94	B
10,000.01-20,000	4	12	40.00	62,822	819,081	95.24	
5,000.01-10,000	3	15	50.00	18,332	837,413	97.37	C
2,000.01-5,000	4	19	63.33	13,774	851,187	98.97	
1,000.01-2,000	4	23	76.67	6,166	857,353	99.69	
500.01-1,000	2	25	83.33	1,651	859,004	99.88	
200.01-500	2	27	90.00	605	859,609	99.95	
100.01-200	2	29	96.67	311	859,920	99.99	
100 and less	1	30	10.00	84	860,004	100.00	

contains a fairly large number of items (30%) and also represents an important proportion of total annual usage (about 28%).

Although the break points between these classes vary according to each individual business conditions a common breakdown might be as follows:

Class	Percentage of items	Percentage of total annual dollar usage
A	5-15	50-60
B	20-30	25-40
C	55-75	5-15

In highly technological industries, such as computer or aircraft production, class A tends to have percentagewise very few items while producing a large share of total annual sales. In contrast, inventories of consumer products at the retail level have a more numerous A class, i.e., it takes a larger portion of items to provide the same large fraction of total annual sales. Industrial producers are in an intermediate position.

If the cumulative fraction of total annual usage is represented against the cumulative percentage of items, the points follow a Pareto curve (Figure 17).

An alternative way of plotting the data in Table 6 is to use a hoirzontal axis representing the natural logarithms of the annual dollar usages, and a vertical axis with a normal probability ruling (lognormal graph paper). Two curves are graphed: cumulative fraction of items vs. annual usage, and cumulative fraction of total annual usage vs. annual usage (Figure 18). Each set of points falls nearly along a straight line (the solid lines in Figure 18), which means that the annual dollar usages in our inventory have a lognormal distribution.

The two lines in Figure 18 are parallel, and their properties can be derived from the fact that they are lognormal cumulative distribution

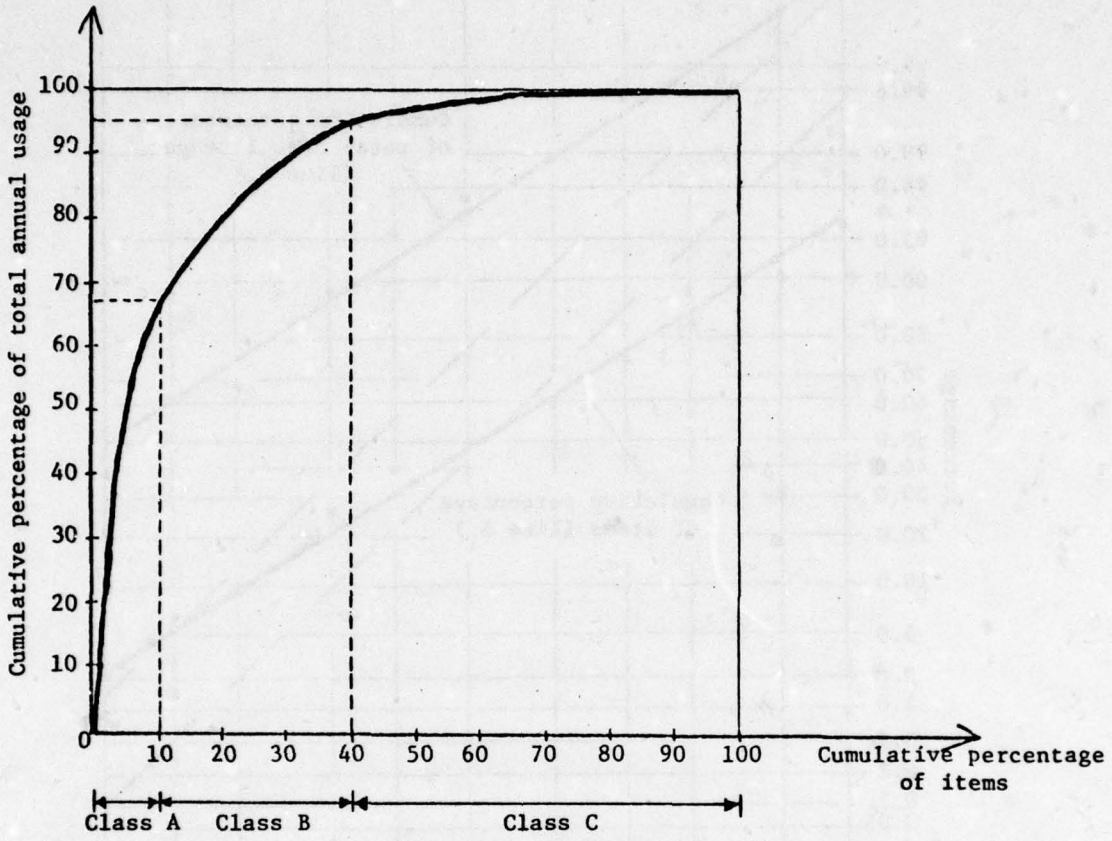


Figure 17: Pareto curve of an ABC classification

functions (Hastings and Peacock [1975], Brown [1959, Appendix C]). For instance line Δ_1 tells us that the items with an annual dollar usage exceeding \$50,000 represent 10% of the total inventory. Similarly, according to line Δ_2, the items whose annual dollar usage is over \$2000 contribute 98.97% to the total annual usage.

It is in general true for all industries that the annual dolar usage across an inventory of items is lognormally distributed.* If an inventory

* Brown [1977, ch. 9.1] shows that not only usage in terms of cost or selling price is lognormal, but other inventory related values as well, such as: material cost, cubic feet of warehousing space, weight, etc.

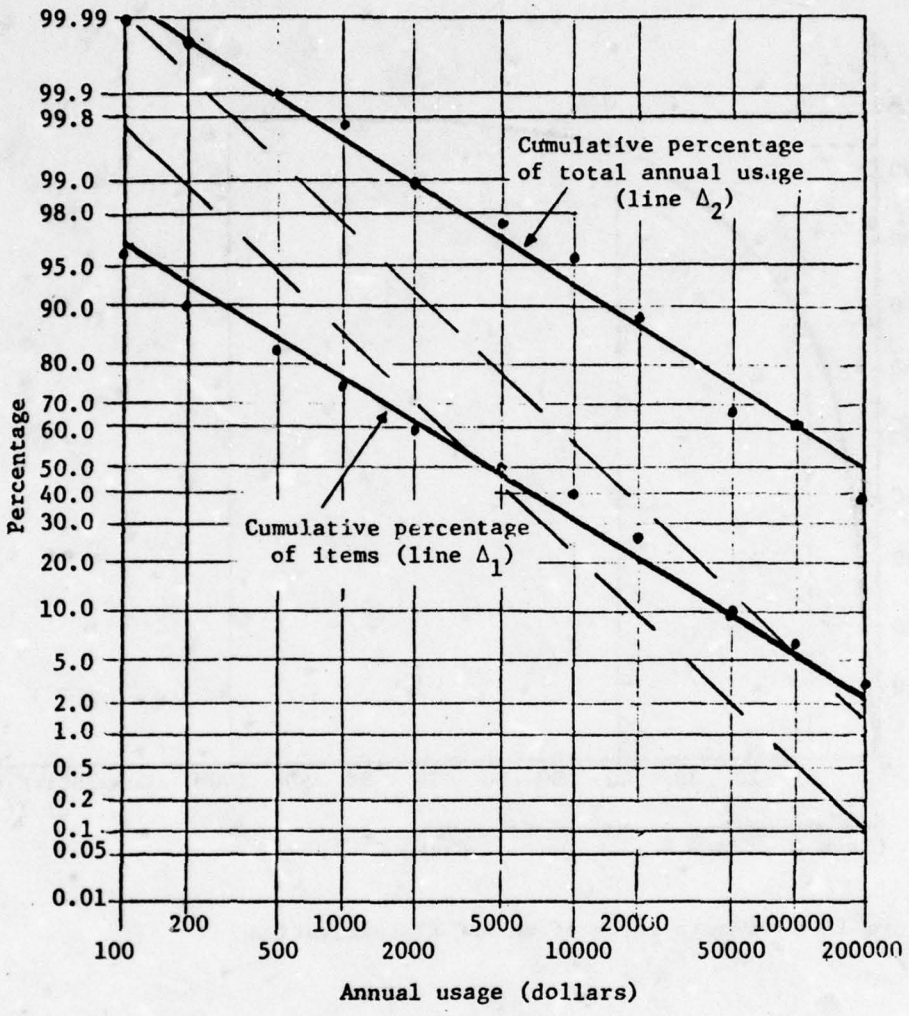


Figure 18: Percentage distributions vs. annual usage

is nonhomogeneous (i.e., it contains two or more separate inventory populations), each population has its own lognormal distribution curves. For instance, in Figure 18 the two dashed parallel lines represent a different group of items than the inventory sample in Table 5 and described by the solid lines Δ_1 and Δ_2. This sort of stratification in an inventory is important because each separate population may require different inventory management and marketing policies.

An alternative way of plotting the distrbution is to graph the cumulative percentage of total annual usage against the cumulative percentage of iterus, using two axes both with normal probability scales. Given that the two cumulative distribution lines in Figure 18 are parallel, the new graph results in another line with a 45° slope (Figure 19). Both inventory populations of Figure 18 are represented in Figure 19 with a solid line and, respectively, a dashed line.

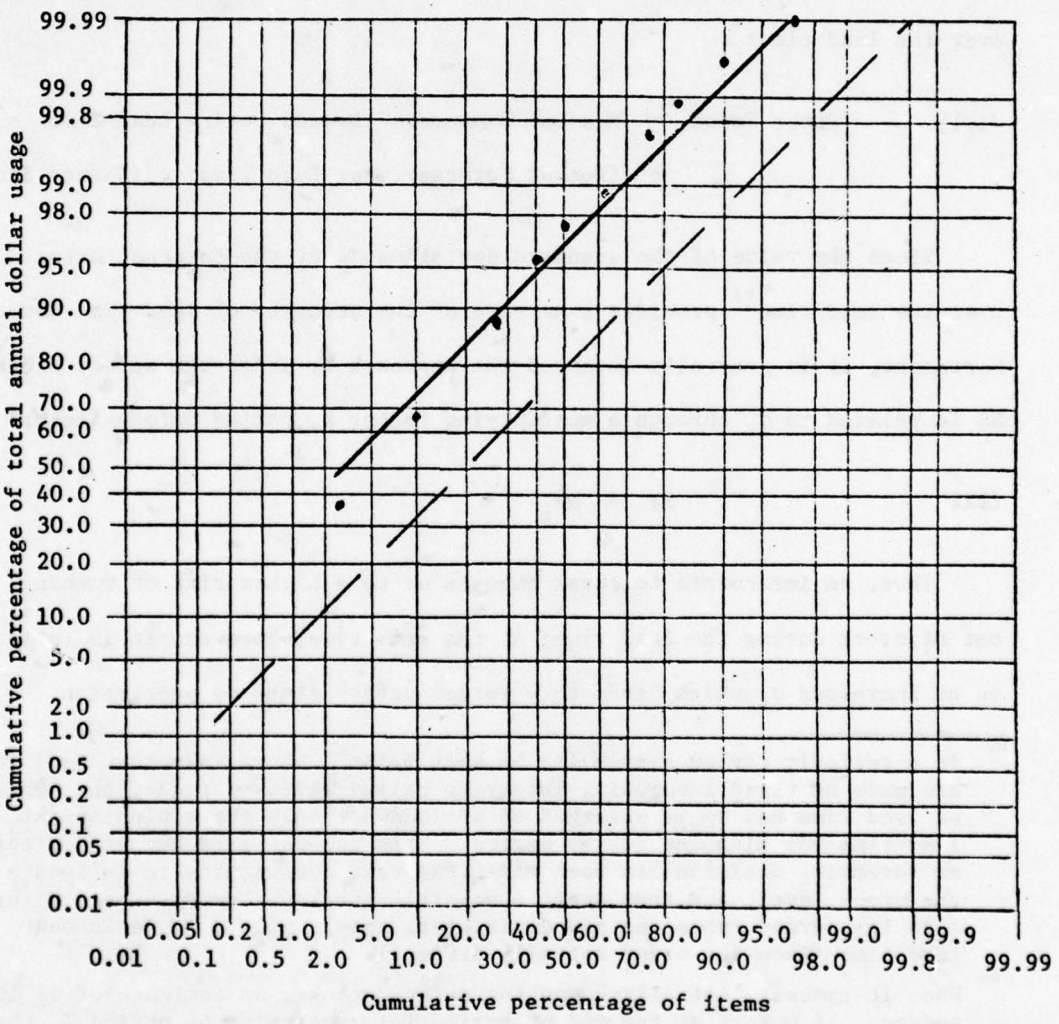


Figure 19: Dollar usage vs. fraction of items for two inventory populations

Thus, whenever the attempt to draw a 45° line through the scatter of points like in Figure 19 results in a bad fit, we would suspect the existence of more than one inventory population; it might also mean that the distribution is not lognormal which is, however, unlikely. A careful investigation should reveal the causes for such an occurrence.

4.2 Safety Stocks for Fast Moving Items

As mentioned earlier, when demand is uncertain the order point has to be set such as to provide sufficient stock to meet any reasonable demand over the lead time:*

(121) Order Point = Maximum Reasonable Demand During Lead Time =

= (Demand Forecast Over Lead Time) + (Safety Stock)

Since the value of the standard deviation σ_ℓ of the forecast errors over the lead time** provides a measure of the accuracy of lead time demand forecasts, it is generally accepted the approach by which the safety stock SS is related to σ_ℓ through a multiplying factor k, called safety factor:

(122) $$SS = k\sigma_\ell$$

Thus, an inaccurate forecast exposes us to a higher risk of running out of stock during the lead time; at the same time, however, it is reflected in an increased σ_ℓ which leads to a larger safety stock as protection

---

* In a periodic review system (to be seen later), where inventory decisions are made at fixed intervals, intervals called review periods, the concept of lead time has to be enlarged to include not only the replenishment lead time but also one review period. This is necessary because, after an inventory decision has been made, the next opportunity to influence the stock level, and thus avert a possible stockout, appears only after a review period (when the new decision is made), plus a replenishment lead time (when the order actually arrives).

** When it comes to actually computing safety stocks, an estimate of σ_ℓ is needed. If we are at the end of period T-1, beginning of period T, the most recently updated estimate is $(\hat{\sigma}_\ell)_T$ (see section 3.3.2) and the safety stock for the ℓ-period lead time starting with period T is:

$$(SS)_T = k(\hat{\sigma}_\ell)_T$$

against the risk.

We should mention that throughout the development of the safety stock decision rules we make the assumption that the average demand is a constant or is changing very slowly with time; this ensures the stability of the order point with time. The issue of time-varying uncertain demand patterns is postponed to a later section. We also consider the lead time to be constant; in case it is not, forecast demand as suggested in section 3.2. Attempts to estimate lead time distributions are bound, in most cases, to remain just academic exercises because of the computational difficulties brought about and because, usually, the additional data required is just not available (Hadley and Whitin [1963, ch. 9-6]). Therefore, with highly variable lead times it is always advisable to try to negotiate with the suppliers in order to cut down the variability.

The safety stock factor k is the control that reflects what management considers to be reasonable in protecting service to the customers. "Reasonable" is understood here in terms of a tradeoff: poor service results in tangible costs (such as the cost of expediting, forgone profit on sales lost because of stockout situations, etc.) and intangible costs (such as loss of goodwill), while improved service means that additional safety stock has to be carried which implies that extra inventory investment is needed.

Thus, the safety stock should correspond to a compromise between to costly shortages and too expensive inventories. One approach to computing safety stocks is to explicitly cost out the shortages and, then, to minimize the total relevant costs associated with the inventory system. Although appealing and rather easy to apply from the optimization point of view, this approach has a serious drawback in that, in general, it is extremely difficult to provide estimates for the cost of stockouts.

Another approach, which avoids this thorny problem, starts from a prespecified service level that has to be achieved by appropriately setting the safety stock. The desired service level is established by management; this implies that some tradeoff between conflicting factors is implicitly worked out and imbedded in the value of the service level. In what follows, examples are presented for both approaches.

4.2.1 Setting Safety Stocks When the Stockout Cost is Proportional to the Number of Units Short

In this case the system incurs a charge c_s for every unit short (i.e., demanded but out of stock) such as, for instance, when items are manufactured on overtime or bought from a competitor at extra cost in order to avert an impending stockout situation.

The model and the derivation to be presented are governed by the following assumptions:

- the demand rate, although uncertain, has its average constant or changing very slowly with time;
- the lead time is deterministic and time invariant;
- the order quantity Q is assumed to have been predetermined;
- an order of size Q is released precisely when available inventory reaches the order point;
- average shortages are considered negligibly small when compared with the average inventory.*

The relevant costs are: the ordering, inventory holding, and stockout costs; we determine the size of the safety stock (i.e., the value of k) such as to minimize the sum of the costs over a time span of one year. As no aggregate constraints are imposed, each item is treated independently

* This assumption is reasonable because practice shows that safety stocks are carried with the purpose of keeping shortages at a low level, which by no means should compare in size with the average level of positive stock.

of the others.

The following notations are use:

D = average demand rate, units per year

A = ordering cost, dollars per order

C = unit cost, dollars per unit of item

r = inventory carrying charge, dollars per dollar of inventory per year

H = rC

TC = total expected annual cost, dollars per year.

The expected annual ordering cost is easily computed given that, with an average annual demand of D and replenishment of size Q, we have:

$$\text{Expected number of replenishments per year} = \frac{D}{Q}$$

Hence,

$$\text{(123)} \qquad \text{Expected annual ordering cost} = A\frac{D}{Q}$$

As mentioned above, a replenishment order is placed when the available inventory reaches the order point. Until the replenishment arrives the inventory level drops by an amount equal to the demand over the lead time. Since we assumed that shortages are, on the average, too small to influence the average on hand inventory level, the expected inventory level just before a replenishment arrives is:

$$\text{Expected on hand inventory just before a replenishment arrives} = (\text{Order point}) - (\text{Expected demand over the lead time})$$

From (121), and from the assumption that an unbiased forecasting system is in use, it follows that:

$$\text{(124)} \qquad \text{Expected on hand inventory just before a replenishment arrives} = \text{Safety Stock} = k\sigma_{\ell}$$

It is obvious that immediately after a replenishment of size Q arrives the inventory level changes:

$$\text{(125)} \qquad \begin{pmatrix}\text{Expected on hand inventory just}\\ \text{after a replenishment arrives}\end{pmatrix} = \begin{pmatrix}\text{Order}\\ \text{Quantity}\end{pmatrix} + \begin{pmatrix}\text{Safety}\\ \text{Stock}\end{pmatrix} = Q + \sigma_\ell$$

Since the demand rate has a constant mean, inventory varies, on the average, linearly between a maximum of $Q+k\sigma_\ell$ and a minimum of $k\sigma_\ell$; therefore:

$$\text{(126)} \qquad \begin{pmatrix}\text{Expected annual}\\ \text{inventory holding cost}\end{pmatrix} = (\frac{Q}{2} + k\sigma_\ell)rC$$

The expected annual shortage cost can be computed as follows:

$$\text{(127)} \qquad \begin{pmatrix}\text{Expected annual}\\ \text{shortage cost}\end{pmatrix} = c_s \cdot \begin{pmatrix}\text{Expected number of}\\ \text{units short per}\\ \text{replenishment cycle}\end{pmatrix} \cdot \begin{pmatrix}\text{Expected number of}\\ \text{replenishment}\\ \text{cycles per year}\end{pmatrix}$$

If $\tilde{d}_\ell$ is the demand over the lead time and d_ℓ is the expected demand over the lead time, the number of units short per replenishment cycle is:

$$\text{(128)} \qquad \begin{pmatrix}\text{Shortage}\\ \text{per cycle}\end{pmatrix} = \begin{cases}\tilde{d} - (d_\ell+k\sigma_\ell) & \text{if } \tilde{d}_\ell > (d_\ell+k\sigma_\ell)\\ 0 & \text{if } \tilde{d}_\ell \leq (d_\ell+k\sigma_\ell)\end{cases}$$

To calculate the expected value of (128) one needs the probability density of $\tilde{d}_\ell$. It is important to mention that, with unbiased forecasts, we see demand $\tilde{d}_\ell$ distributed around the forecast, according to the same distribution function as the forecast errors over lead time. The spread of $\tilde{d}_\ell$ is characterized by the standard deviation σ_ℓ.

In the case of fast movers, if the random noise of demand is normally distributed and the successive noise samples have no serial correlation, the cumulative (over the lead time) forecast errors are also normally distributed. More generally, however, Brown [1963, ch. 19] shows by analysis and simulation that for forecasts based upon a linear, discrete, time-invariant system (as is the case with multiple smoothing or general exponen-

tial smoothing models) forecast errors are approximately normal for a wide range of distributions of demand data (simulations were run with data sampled from normal, uniform, and triangular distributions). Therefore, the normality of lead time forecast errors is a generally accepted practice with fast movers.

Thus let $f(\tilde{d}_\ell)$ be the normal density function for the lead time demand, having mean d_ℓ and standard deviation σ_ℓ. Then,

$$\text{Expected number of units short per replenishment cycle} = \int_{d_\ell+k\sigma_\ell}^{\infty} [\tilde{d}_\ell - (d_\ell+k\sigma_\ell)]f(\tilde{d}_\ell)d(\tilde{d}_\ell)$$

To normalize the distribution we make a chnage of variable: $u = \frac{\tilde{d}_\ell - d_\ell}{\sigma_\ell}$; afterwards we obtain:

$$\text{(129)} \qquad \text{Expected number of units short per replenishment cycle} = \sigma_\ell \int_k^{\infty} (u-k)f(u)du$$

where $f(u)$ is the normalized normal density (with zero mean and a standard deviation equal to 1).

The integral above is the partial expectation of the normally distributed u; we call it G(k) and it will be useful in some later developments:

$$\text{(130)} \qquad G(k) = \int_k^{\infty} (u-k)f(u)du, \qquad u \sim N(0,1)$$

In concise form:

$$\text{(131)} \qquad \text{Expected number of units short per replenishment cycle} = \sigma_\ell G(k)$$

The cost associated with the stockouts is given by:

$$\text{(132)} \qquad \text{Expected annual shortage cost} = \sigma_\ell G(k) \frac{D}{Q} c_s$$

The total expected annual cost results from the sum of (123), (126), and (132):

$$TC = A\frac{D}{Q} + (\frac{Q}{2} + k\sigma_\ell)rC + \sigma_\ell G(k)\frac{D}{Q}c_s \tag{133}$$

As TC is convex in k, the necessary and sufficient condition for a minimum is the null value of the first derivative:*

$$\frac{d(TC)}{dk} = r\sigma_\ell C - c_s\sigma_\ell \frac{D}{Q}\int_k^\infty f(u)du = 0$$

Let P(k) be the complement of the cumulative distribution function:

$$P(k) = \int_k^\infty f(u)du, \qquad u \sim N(0,1) \tag{134}$$

The optimal value of the safety factor k must be set so as to satisfy:

$$P(k) = \frac{QrC}{Dc_s} \tag{135}$$

Values of the P(k) function can be found tabulated in any statistics text.

Equation (135) only makes sense for $\frac{QrC}{Dc_s} < 1$. If $\frac{QrC}{Dc_s} > 1$ or when k results from (135) with some negative value which is considered unacceptable, the safety factor should be set at the smallest allowable value established by management.

When using optimal lot sizing policies faster moving items have Q/D smaller than slower moving itmes; consequently, under similar costs, faster moving items tend to be allotted larger safety stocks. This result, however, is in no way an endorsement of the still widely used practice by which safety stocks are set as equal time supplies for all items in an inventory. First, it is apparent that k is not linearly related to the demand rate. Second, the safety stock is influenced by the cost structure

* See Sokolnikoff and Redheffer [1958, pp. 261-262] for the derivative of a definite integral.

and the accuracy of the forecasts.

4.2.2 Setting Safety Stocks to Achieve a Prespecified Service Level

As mentioned earlier, setting safety stocks so as to achieve a prespecified service level skirts the difficult issue of explicitly determining the shortage cost.

We treat in this section two measures of service level:

- the expected number of stockout occasions (irrespective of size) per year, and
- the fraction of demand to be served routinely from stock.

We choose these two measures of service because practice shows that in many cases they reflect the managerial way of thinking in terms of what constitutes poor or good service to the customer.

Safety Stocks for a Specified Average Number of Stockout Occasions per Year

The significance of this measure of service level is probably best described by Brown [1977, p. 176]: "In any manufacturing process it is practical to expedite one order and get it delivered sooner than normal. It is practical to expedite two orders. It is not practical to expedite half the open orders - nothing gets done then."

Thus, to the extent that expediting does not become a way of living, it is normal to have to move now and then a few orders ahead in the schedule; for this, the average number of shortage occurences has to be kept below a prespecified threshold. The idea is that part of the service is provided by being able to satisfy part of the demand directly from shelf; this service is further improved by expediting.

Let, then, (SO) be the acceptable average number of shortage occasions per year; find a value for the safety factor k so as to achieve the prespecified value of (SO). The assumptions upon which the model is built

are similar to the ones in the previous section 4.2.1, except that in the lost sales case the shortages affect the inventory level (to be seen).

With a lead time demand normally distributed around a mean d_ℓ with a standard deviation σ_ℓ, and having a safety stock of $k\sigma_\ell$, the probability of a stockout at the end of the lead time is given by the P(k) function of (134):

$$P(k) = \int_{d_\ell + k\sigma_\ell}^{\infty} f(\tilde{d}_\ell)d(\tilde{d}_\ell) = \int_{k}^{\infty} f(u)du, \qquad u \sim N(0,1)$$

The expected number of replenishments per year is D/Q; consequently:

$$\text{(136)} \qquad \begin{matrix}\text{Expected number of} \\ \text{shortage occurrances per year}\end{matrix} = \frac{D}{Q} P(k) = (SO)$$

The best value of the safety factor k is so as to conform to the following equation:

$$\text{(137)} \qquad P(k) = \frac{Q}{D} (SO)$$

If a small number of shortages (i.e., high service level) is desired, P(k) should be small and k results large. Thus, decision rule (137) provides the safety stock with the intuitive and desirable feature by which good service requires larger safety stocks, while smaller safety stocks are associated with poorer service, all other things being equal.

We should mention that the above developments leading to (137) assume that any unfilled demand is backordered and served as soon as the replenishment arrives. If, however, in the case of a shortage customers are not willing to wait, unfilled demand is lost. By (129) and (130) an amount of $\sigma_\ell G(k)$ units result, on the average, in lost sales per cycle. The consequence is the following: if in the backorders case, where eventually all demand is served, a demand of Q units is recorded on the average

during a replenishment cycle, in the lost sales case $Q + \sigma_\ell G(k)$ units are demanded on the average in a cycle, of which Q are served and the others lost. Therefore, a system with lost sales works with larger inventories and fewer cycles, namely:

$$\text{(138)} \qquad \begin{matrix}\text{Expected annual number of} \\ \text{replenishment cycles in} \\ \text{the lost sales case}\end{matrix} = \frac{D}{Q + \sigma_\ell G(k)}$$

The decision rule (137) can be modified accordingly. The resulting equation, however, is by no means trivial to solve.

Safety Stocks for a Specified Fraction of Demand to be Served Directly from Stock

Here, management considers the service level to be satisfactory if a $\mathcal{F}$ of the annual demand is served directly from stock; the remaining fraction $(1-\mathcal{F})$ that cannot be satisfied directly from shelf is either backordered or is expedited to be moved faster through the replenishment process.

The same assumptions like in the previous section stay valid here too.

With a lead time demand which is normal (mean d_ℓ, and standard deviation σ_ℓ) and a safety stock of $k\sigma_\ell$, the expected quantity short per order cycle is $\sigma_\ell G(k)$ [see Equation (131)]. During the year an average of $\frac{D}{Q}$ orders are placed and, therefore, the expected amount to be backordered annually is:

$$\text{(139)} \qquad \begin{matrix}\text{Expected number of} \\ \text{units short per year}\end{matrix} = \sigma_\ell G(k)\,\frac{D}{Q}$$

According to the definition of the customer service level it is still satisfactory if $(1-\mathcal{F})D$ cannot be served promptly from stock. Equating this with (139) one obtains the condition to be satisfied by the safety factor k we are seeking:

$$G(k) = \frac{Q}{\sigma_\ell}(1 - \mathcal{F}) \tag{140}$$

Selected values of the partial expectation function are shown in Table 7; they should be sufficient for any practical work in inventory systems. More values may be found in Peterson and Silver [1979, pp. 779-786].

Table 7: Values of the partial expectation function G(k) for the normal distribution

k	G(k)	k	G(k)
0.0	0.3989	1.8	0.01428
0.1	0.3509	1.9	0.01105
0.2	0.3069	2.0	0.008491
0.3	0.2668	2.1	0.006468
0.4	0.2304	2.2	0.004887
0.5	0.1978	2.3	0.003662
0.6	0.1687	2.4	0.002720
0.7	0.1429	2.5	0.002004
0.8	0.1202	2.6	0.001464
0.9	0.1004	2.7	0.001060
1.0	0.08332	2.8	0.000761
1.1	0.06862	2.9	0.000542
1.2	0.05610	3.0	0.000382
1.3	0.04553	3.1	0.000267
1.4	0.03667	3.2	0.000185
1.5	0.02931	3.3	0.000127
1.6	0.02324	3.4	0.000087
1.7	0.01829	3.5	0.000058

From (140) it is clear that, for a specified service level, safety stocks depend on both the order quantity and the accuracy of the forecasts. Thus, the larger the value of Q, the larger the value of G(k), and therefore the smaller the safety factor needed for a given customer service.

This is so because when the order quantity is large there are fewer shipments per year; hence, there are fewer opportunities to backorder.

Similarly, the smaller the standard deviation σ_ℓ of the forecast errors, the large G(k) and the smaller the safety factor. Thus, with an accurate forecasting system, the standard deviation is low and, at the same time, one neds fewer standard deviations as safety stock - a double saving in safety stock investment.

Let us note that according to this decision rule, faster moving items need higher safety stocks than do slower moving items. Indeed, under optimal lot sizing, ratio D/Q tends to increase the higher the usage of the item (Q grows slower than D does); at the same time, the variance low (see section 3.3.1) tells us that the standard deviation of the forecast errors tends to increase with the usage D. Consequently, for a specified $\mathcal{F}$, faster moving items command a lower value for G(k) and, therefore, a larger safety factor k.

Decision rule (140) has been worked out for the backordering situation. In the lost sales case the expected number of cycles per year is not D/Q but rather $\frac{D}{Q+\sigma_\ell G(k)}$ as shown in (138).

Then,

$$\text{Expected number of units short per year} = \sigma_\ell G(k) \frac{D}{Q+\sigma_\ell G(k)} = \sigma_\ell G(k) \frac{D}{Q} \cdot \frac{Q}{Q+\sigma_\ell G(k)}$$

The ratio $\frac{Q}{Q+\sigma_\ell G(k)}$ is precisely the fraction $\mathcal{F}$ of demand satisfied promptly from stock. Hence:

$$\text{(141)} \qquad \text{Expected number of units short per year} = \sigma_\ell G(k) \frac{D}{Q} \mathcal{F}$$

By definition of the service level, (141) has to be equated with $(1-\mathcal{F})D$; then, the safety factor k has to satisfy the following condition:

$$\text{(142)} \qquad G(k) = \frac{Q}{\sigma_\ell} \frac{1-\mathcal{F}}{\mathcal{F}}$$

Since, normally, fraction $\mathcal{F}$ is close to 1, from comparing (140) and (142) it is evident that the safety factor is affected in no significant way bt whether unfilled demand can be backordered or is lost.

The value of the service level $\mathcal{F}$ differs with the category of items for which inventory control has to be implented:

- A items should be handled with relatively low service level (a value of $\mathcal{F} = 0.8$ seems reasonable). This results in lower safety stocks and, therefore, in smaller inventory investments for these relatively expensive items. However, this apparent lack of service is overcome by tight control procedures and efficient expediting throughout the manufacturing and procurement process that will, in practice, increase the aimed service level to a high performance without having to pay the penalty of expensive safety stocks. It is in the control of these items where most of management judgement, attention, and intervention is concentrated, using automatic inventory rules to provide only a sound guideline for these management decisions.
- B items should be handled fairly routinely, with service levels on the order of $\mathcal{F} = 0.95$. The inventory control system has to provide sound replenishment decisions and, normally there should be no need for extensive expediting, management intervention, or tight external control.
- C items should be present in ample supply and handled with a minimum of records, controls, and procedures. Normally, their replenishment decisions should be completely mechanized, having assigned a very high service level (between 0.95 and 0.98).

A number of alternative measures of service level can be found: they

should be used in performance evaluation there where they best parallel management's targets in providing customer service. Some of these measures are presented below accompanied by brief qualifications; for all of them the control is the safety factor k which yields the safety stock in the form of $k\sigma_\ell$.

The expected number of units short per replenishment cycle is given by $\sigma_\ell G_k$ [see equation (131)] and, by Brown's [1977, p. 176] opinion, is a primary measure of service (as the expected number of stockout occasions per year also is) in that it serves in deriving other measures like the fraction of demand to be served routinely from stock.

The probability of stockout during a replenishment cycle can be computed as P(k), the complement of the cumulative distribution function of lead time demand [equation (134)]. This and the previous measure of service focus management's attention on every individual cycle, thus avoiding the pitfalls of service levels calculated as averages. For instance, while the annual average service, expressed as a fraction of demand to be satisfied directly from stock, might look really good, the corresponding value k of the safety factor could still yield, on a cycle by cycle basis, an unacceptably large chance P(k) of shortage occurance.

At the same time, however, the cycle oriented service measures tend to lose sight of the general picture. Thus, if one item is replenished ten times annually and another item only once, and if k is set so as to produce for both items a probability of 0.10 of stockout per cycle, then we expect the first item to run out of stock once a year [see equation (136)] while the second item only once every ten years. Although the safety factor is the same for both items, it is questionable whether the customer service is the same.

For these reasons it is advantageous to simultaneously assess service

performance with measures of different emphasis: replenishment cycle oriented and annual average oriented.

The probability that n_o *cycles develop shortages during a year* ($n_o = 1, 2, \ldots, D/Q$) can be easily computed by defining the occurrence of a stockout situation in a replenishment cycle as a Bernoulli trial having a probability of success (i.e., occurrance) of P(k) [see equation (134)] and a probability of failure [1-P(k)]. A binomial process with a total number of D/Q trials takes place; n_o stockout situations occur during a year with a probability of

$$\frac{(\frac{D}{Q})!}{(\frac{D}{Q} - n_o)!\ n_o!} P(k)^{n_o} [1 - P(k)]^{\frac{D}{Q} - n_o} .$$

The mean of this distribution is a measure of service discussed earlier: the expected number of shortage occurrences per year $\frac{D}{Q} P(k)$. If one wants to know the chance of at least one cycle having shortages during one year the answer is given by computing $1 - [1 - P(k)]^{D/Q}$.

The fraction of time the system has on-hand positive stock is often called the ready rate of the system. As shown earlier, in the backorders case, the average demand over a replenishment cycle is Q, and in the lost sales case is $Q + \sigma_\ell G(k)$. In both instances, the expected number of units short per cycle is $\sigma_\ell G(k)$. Since we assumed that the average demand rate is constant, the ratio of time short in a cycle to the total duration of the replenishment cycle (i.e., the fraction of time the system is out of stock) is the same as the ratio of the expected amount short to the average demand over the cycle:

$$\text{(143)} \qquad \text{Fraction of time the system is out of stock} = \begin{cases} \frac{\sigma_\ell G(k)}{Q}, & \text{backorders case} \\ \frac{\sigma_\ell G(k)}{Q + \sigma_\ell G(k)}, & \text{lost sales case} \end{cases}$$

Since the lost sales case produces higher inventories, the fraction of time out of stock results smaller than in the backorders case.

The ready rate is then computed as 1 - (fraction of time out of stock):

$$(144) \qquad \text{Fraction of time the system has on-hand positive stock} = \begin{cases} \dfrac{Q-\sigma_{\ell}G(k)}{Q}, & \text{backorders case} \\ \dfrac{Q}{Q+\sigma_{\ell}G(k)}, & \text{lost sales case} \end{cases}$$

Notice that under the assumption of unit sized demand transactions, which has been imbedded in the calculation of the expected amount short per cycle, it is apparent from (144) that the ready rate is equivalent to the fraction $\mathcal{F}$ of demand served directly from stock. With non-unit sized demands there is the possibility of overshooting the order point, which leaves the two measures of service only approximately equivalent.

4.2.3 Allocation of Safety Stocks under Aggregate Constraints

In the previous two sections safety stock decisions have been analyzed by considering every inventory item in isolation. It is very often the case, however, that some aggregate constraints apply such as a limited budget for inventory investment, limited warehousing space, etc., in which case items iteract by competing for the scarce resource.

A large variety of decision rules can be developed by optimizing various objective functions within the bounds imposed by the aggregate constraints (Gerson and Brown [1970]).

Suppose then, that we have an inventory of n items, a limited budget for annual inventory investment I; we want to allocate this budget among the n items so as to minimize the total expected number of units short per year; subscript i denotes the i-th item.

Assume that the average demand rate is constant (although demand is probabilistic), the lead time is time invariant and known with certainty,

and an order of size Q_i (which is determined independently and prior to establishing the safety stocks) is released as soon as the available inventory of the i-th item reaches the corresponding order point. One unit of item i costs c_i dollars. Unfilled demand is backordered.

By use of relations (126) and (139) the formal representation o. the problem is:

$$\text{Minimize } Z = \sum_{i=1}^{n} \sigma_{\ell i} G(k_i) \frac{D_i}{Q_i}$$

$$\text{subject to: } \sum_{i=1}^{n} \left(\frac{Q_i}{2} + k_i \sigma_{\ell i}\right) C_i \leq I$$

Two observations are in order here:

- given the nature of the problem, the solution will always have the constraint hold at equality
- as Q_i are predetermined constants, the constraint can be simplified by expressing it only in terms of the required investment in safety stocks I_s.

Hence, the problem we have to solve is:

(145) $$\text{Minimize } Z = \sum_{i=1}^{n} \sigma_{\ell i} G(k_i) \frac{D_i}{Q_i}$$

subject to:

(146) $$\sum_{i=1}^{n} k_i \sigma_{\ell i} C_i = I_s$$

The objective function is convex, which can be shown by looking at the Hessian of Z. As the objective function is separable in the k_i's, the Hessian $\mathcal{H}$ is a nxn diagonal matrix with all positive elements on its main diagnoal of the form:* $a_{ii} = f(k_i)$, i=1,...,n, where $f(k_i)$ is the normal density function (with zero mean, and a standard deviation of 1)

* $\frac{\partial Z}{\partial k_i} = - k_i \int^{\infty} f(u)du$; $\frac{\partial^2 Z}{\partial k_i^2} = f(k_i)$

evaluated at k_i. For all nonzero vectors x, the Hessian satisfies $x^T \mathcal{H} x > 0$, hence it is positive definite, Z is convex (Luenberger [1973, p. 118]), and the Lagrangian multiplier technique can be applied.

The Lagrangian is:

$$L = \sum_{c=1}^{n} \sigma_{\ell i} G(k_i) \frac{D_i}{Q_i} + \lambda \left(\sum_{i=1}^{n} k_i \sigma_{\ell i} C_i - I_s \right) \tag{147}$$

where λ is the nonnegative Lagrange multiplier.

To find the minimum of Z, subject to the constraint on investment, set $\frac{\partial L}{\partial k_i} = 0$ which yields:

$$P(k_i) \cdot \frac{D_i}{Q_i} \cdot \frac{1}{C_i} = \lambda , \qquad i=1,\ldots,n, \tag{148}$$

where $P(k_i)$ is given by (4.134).

Gerson and Brown [1970] solve a similar problem in which the expected value of the total annual shortages $\sum_{i=1}^{n} \sigma_{\ell i} G(k_i) \frac{D_i}{Q_i} C_i$ is minimized subject to (146). The optimal solution satisfies:

$$P(k_i) \frac{D_i}{Q_i} = \lambda , \qquad i=1,\ldots,n, \tag{149}$$

which means that safety factors k_i should be chosen so as to produce the same expected number, λ, of units short per year for each item.

By varying λ in (148) or (149) one obtains different values of the required inventory investment; the solution is that λ which leads to equality in (146).

As mentioned earlier, other objectives can be optimized, subject to aggregate constraints, such as the expected number of shortage occurrences per year $\sum_{i=1}^{n} \frac{D_i}{Q_i} P(k_i)$. For models minimizing the total of ordering, inventory carrying, and stockout costs under aggregate inventory constraint, see Holt, et al. [1960, ch. 12].

4.2.4 Simultaneous Determination of Safety Stocks and Order Quantities

In this section we relax the assumption of predetermined order quantities; our concern is to determine lot sizes and safety stocks jointly. Decision rules are derived for both the case of the individual product and the case of optimization under aggregate constraints.

Order Quantity and Safety Stock for a Single Product

To illustrate, consider the total cost function of section 4.2.1; the optimal Q and k are sought so as to minimize the sum of ordering, inventory holding and stockout costs:

$$\text{(150)} \qquad \text{Minimize} \quad TC = A\frac{D}{Q} + \left(\frac{Q}{2} + k\sigma_\ell\right)rC + \sigma_\ell G(k)\frac{D}{Q}c_s$$

The first order conditions for a minimum are given by $\frac{\partial(TC)}{\partial Q} = 0$ and $\frac{\partial(TC)}{\partial k} = 0$. It is interesting, however, to find out more about the shape of TC; if it is convex, it has a minimum which is a global minimum.

The Hessian of TC is:

$$\mathcal{H} = \begin{pmatrix} \frac{\partial^2(TC)}{\partial Q^2} & \frac{\partial^2(TC)}{\partial Q \partial k} \\ \frac{\partial^2(TC)}{\partial k \partial Q} & \frac{\partial^2(TC)}{\partial k^2} \end{pmatrix} = \begin{pmatrix} \frac{2AD}{Q^3} + \frac{2\sigma_\ell c_s DG(k)}{Q^3} & \frac{\sigma_\ell c_s DP(k)}{Q^2} \\ \frac{\sigma_\ell c_s DP(k)}{Q^2} & \frac{\sigma_\ell c_s Df(k)}{Q} \end{pmatrix}$$

$\mathcal{H}$ is a symmetric matrix of the form $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$; the necessary and sufficient conditions for such a matrix to be positive definite are: $a > 0$ and $\det \mathcal{H} = ac - b^2 > 0$. In our case clearly $a > 0$. Then:

$$\det \mathcal{H} = \frac{1}{Q^4}\left[2A\sigma_\ell c_s D^2 f(k) + 2\sigma_\ell^2 c_s^2 D^2 G(k) f(k) - \sigma_\ell^2 c_s^2 D^2 P(k)^2\right]$$

Q^4 is always positive. With respect to k we notice that:

$$\lim_{k \to \infty} (\det \mathcal{H}) = 0$$

Also:

$$(151) \qquad \frac{\partial(\det \mathcal{H})}{\partial k} = \frac{1}{Q^4}\left[2A\sigma_\ell c_s D^2 f'(k) + 2\sigma_\ell^2 c_s G(k) f'(k)\right]$$

where $f'(k)$ is short for $\frac{df(k)}{dk}$.

Evidently, if $f'(k) < 0$, then (151) is negative. This means that, with increasing k, the determinant of the Hessian tends monotone decreasingly to zero. Hence, det $\mathcal{H}$ is positive.

We conclude that, for $f'(k) < 0$, the Hessian of TC is positive definite, so TC is convex. The restriction $f'(k) < 0$ is satisfied for $k \quad 0$; thus, we would only accept positive safety stocks.

The first order conditions are:

$$(152) \qquad \begin{cases} \frac{\partial(TC)}{\partial Q} = \frac{rC}{2} - \frac{AD + \sigma_\ell c_s DG(k)}{Q^2} = 0 \\ \frac{\partial(TC)}{\partial k} = r\sigma_\ell C - \frac{\sigma_\ell c_s DP(k)}{Q} = 0 \end{cases}$$

Solve to get:

$$(153) \qquad Q = \sqrt{\frac{2[AD + \sigma_\ell c_s DG(k)]}{rC}}$$

$$(154) \qquad P(k) = \frac{QrC}{c_s D}$$

If there is no uncertainty involved (i.e., $\sigma_\ell = 0$), the otpimal order quantity becomes the classical Wilson's formula (1). Otherwise, the presence of uncertainty requires a larger Q in order to diminish the number of replenishment cycles per year and thus reduce the exposure to the risk of running out of stock.

Combining (153) and (154) yields:

$$(155) \qquad \sqrt{2rC[AD + \sigma_\ell c_s DG(k)]} - c_s DP(k) = 0,$$

which can be solved by search technique or by table look-up and trial and error. Once k is known, (153) produces the optimal Q. One can also

find a solution by an iterative scheme that moves between equations (153) and (154) (see, for instance, Hadley and Whitin [1963, ch. 4-4]).

Order Quantities and Safety Stocks under Aggregate Constraints

Here we reconsier the problem of section 4.2.3: find, for the n items in the system, the best order quantities Q_i and safety factors k_i, i=1,...,n, so as to minimize the total expected number of units short per year, subject to a total inventory budget I.

$$\text{(156)} \qquad \text{Minimize } Z = \sum_{i=1}^{n} \sigma_{\ell i} G(k_i) \frac{D_i}{Q_i}$$

subject to:

$$\text{(157)} \qquad \sum_{i=1}^{n} \left(\frac{Q_i}{2} + k_i \sigma_{\ell i} \right) C_i = I$$

In order to apply the Lagrange multiplier method we have to make sure that the Hessian of Z is positive definite (the Hessian of the constraint is a zero matrix, so it cannot affect the second order conditions for optimality, Luenberger [1973, p. 226]).

Z is a function of 2n variables; it separates in n functions, each related to one item only:

$$Z(k_1,Q_1;\ k_2,Q_2;\ \ldots;\ k_n,Q_n) = \sum_{i=1}^{n} z_i(k_i,Q_i)$$

The Hessian $\mathcal{H}$ of Z displays a succession of n symmetric 2 by 2 matrices on its main diagonal; the i-th 2 by 2 matrix is the Hessian h_i of z_i; all other elements in $\mathcal{H}$ are zero.

$$\mathcal{H} = \begin{pmatrix} \frac{\partial^2 Z}{\partial k_1^2} & \frac{\partial^2 Z}{\partial k_1 \partial Q_1} & 0 & 0 & \cdots \\ \frac{\partial^2 Z}{\partial k_1 \partial Q_1} & \frac{\partial^2 Z}{\partial Q_1^2} & 0 & 0 & \cdots \\ 0 & 0 & \frac{\partial^2 Z}{\partial k_2^2} & \frac{\partial^2 Z}{\partial k_2 \partial Q_2} & \cdots \\ 0 & 0 & \frac{\partial^2 Z}{\partial k_2 \partial Q_2} & \frac{\partial^2 Z}{\partial Q_2^2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} =$$

$$= \begin{pmatrix} h_1 & 0 & \cdots & 0 \\ 0 & h_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & h_n \end{pmatrix}$$

We want to show that for all nonzero vectors X it is true that $X^T \mathcal{H} X > 0$. Vector X has 2n elements and we regard it as being made of n vectors with 2 elements each:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Therefore:

$$(158) \qquad x^T \mathcal{H} x = x_1^T h_1 x_1 + x_2^T h_2 x_2 + \ldots + x_n^T h_n x_n$$

We have:

$$h_i = \begin{pmatrix} \frac{\sigma_{\ell i} D_i f(k_i)}{Q_i} & \frac{\sigma_{\ell i} D_i P(k_i)}{Q_i^2} \\ \frac{\sigma_{\ell i} D_i P(k_i)}{Q_i^2} & \frac{2\sigma_{\ell i} D_i G(k_i)}{Q_i^3} \end{pmatrix}$$

Since the upper left element is positive, h_i is positive definite if its determinant is positive.

$$\det h_i = \frac{\sigma_{\ell i}^2 D_i^2}{Q_i^4} [2f(k_i)G(k_i) - P^2(k_i)]$$

$$(159) \qquad \lim_{k_i \to \infty} (\det h_i) = 0$$

$$(160) \qquad \frac{\partial(\det h_i)}{\partial k_i} = \frac{2\sigma_{\ell i}^2 D_i}{Q_i^4} f'(k_i) G(k_i)$$

Under the restriction of using only positive safety factors, we have $f'(k_i) < 0$, the derivative (160) is negative and, by (159), $\det h_i > 0$. Hence, h_i is positive definite, making $x_i^T h_i x_i > 0$.

From (158) it immediately follows that $\mathcal{H}$ is positive definite.

Form the Lagrangian function L using a nonnegative multiplier λ and apply the first order conditions for optimality: $\frac{\partial L}{\partial Q_i} = 0$, $\frac{\partial L}{\partial k_i} = 0$, they yield:

$$(161) \qquad Q_i = \sqrt{\frac{2\sigma_{\ell i} D_i G(k_i)}{\lambda C_i}}$$

$$(162) \qquad Q_i = \frac{D_i P(k_i)}{\lambda C_i}$$

Together (161) and (162) lead to an equation whose solution is the optimal k_i for a given λ:

$$P(k_i) = \sqrt{\frac{2\lambda\sigma_{\ell i}C_i G(k_i)}{D_i}} \tag{163}$$

The Q_i corresponding to the given λ and k_i can be calculated from either (161) or (162).

A search over the values of λ has to be performed in view of the inventory constraint (157).

Developments of decision rules for other objective functions are conducted by Holt, et al. [1960, ch. 13], Gerson and Brown [1970].

It is apparent that consideration of the joint determination of lot sizes and safety stocks brings considerable computational complexity into the system and, therefore, might only the warranted for the most expensive or important of the class A items. For the case of the single product for which Q and k are derived by minimizing the total cost of equation (150), Peterson and Silver [1979, p. 367] find that the cost penalty for computing Q and k independently rather than jointly is larger when ratio Q/σ_ℓ is small. They show that low values of Q/σ_ℓ are more likely with class A items than with B or C items and, consequently, they recommend that use of the more sophisticated approach of joint determination be limited to some A items.

4.3 Safety Stocks for Slow Moving Items

The philosophy behind setting safety stocks for slow movers is the same as in the case of fast moving items: when placing an order sufficient stock should be available (on hand plus already on order) to meet the maximum reasonable demand over the lead time.

If the forecast model in use fits a theoretical distribution to data (e.g.: the Normal, Poisson, Laplace, or Gamma distribution) by estimation of the appropriate parameters, then the safety stock is made equal to some multiple k of the standard deviation σ_ℓ of the forecast errors over the lead time, and the order point follows from (121) Another approach is to work with and determine directly the order point s. Evidently, in view of the definition (121) of the order point, the two approaches are equivalent.

If, alternatively, the empirical cumulative probability distribution of lead time demand is determined (section 3.4.4) the order point is read directly off the cumulative curve.

A few examples and comments are provided below to illustrate the two approaches.

4.3.1 Safety Stocks Based on Theoretical Distributions

This approach is essentially the same procedure used to determine order points for fast moving items except that the expected lead time demand and σ_ℓ would be estimated by procedures specific to slow movers. As mentioned in section 3.4.3, use of complicated probability distributions should be limited to expensive or very important items where the increased system cost is balanced by the possible substantial reduction in inventory investment or by the improved service brought about by the more accurate estimates. For cheap and unimportant items, the normality of forecast errors may still be considered as a reasonable approximation; therefore, the safety factor k should be selected from the same tables for fast moving items, which could constitute an important practical advantage for the implementation stage.

To substantiate the statements made above, consider the case where the system incurs s charge c_s for every unit short, and lead time demand

is generated by a Poisson process.

The following assumptions hold:

- although demand rate is uncertain, its average of D units/year is a constant;
- demand is discrete and units are demanded one at a time;
- an order of size Q is released exactly when available inventory hits the order point;*
- order point s and replenishment quantity Q are discrete;
- lead time is a constant ℓ;
- any unfilled demand is backordered.

The ordering cost is A dollars/order, the cost of the item is C dollars/unit, and the inventory carrying charge is r dollars/dollar·year.

We determine the optimal order size Q and order point s so as to minimize the total of the average annual costs (ordering, holding, and backorders costs).

The basic proofs are derived by Hadley and Whitin [1963, ch. 4-7] and we extract here only the results relevant to our problem.

The expected annual ordering cost is $A \frac{D}{Q}$.

Lead time demand is Poisson with mean (and variance) equal to $D\ell$. Let the mean lead time demand be denoted by $d_\ell = D\ell$, and let $\tilde{d}_\ell$ be the probabilistic demand over the lead time. Then, $p(\tilde{d}_\ell; d_\ell)$ is the Poisson probability that $\tilde{d}_\ell$ units are demanded during a lead time of ℓ periods, given that the process generates on the average d_ℓ units per lead time.

Now, we are interested in the expected on hand inventory $\bar{I}$ that costs

---

* Given the discreteness of demand, if transaction sizes are random overshoots of the order point may occur, and the order point - order quantity system might not be appropriate any longer. Indeed, if an unusually large overshoot takes place, the regular replenishment lot of size Q might not even bring the inventory level back to the order point. Systems that can cope with this sort of situation are discussed in later sections.

us $rC\bar{I}$ to hold. By definition:

$$\begin{pmatrix}\text{Available}\\ \text{inventory}\end{pmatrix} = \begin{pmatrix}\text{On hand}\\ \text{inventory}\end{pmatrix} + \begin{pmatrix}\text{Amount}\\ \text{on order}\end{pmatrix} - \begin{pmatrix}\text{Amount}\\ \text{on backorder}\end{pmatrix}$$

Then:

$$\bar{I} = \begin{pmatrix}\text{Expected available}\\ \text{inventory}\end{pmatrix} - \begin{pmatrix}\text{Expected amount}\\ \text{on order}\end{pmatrix} + \begin{pmatrix}\text{Expected number}\\ \text{of backorders}\\ \text{at any time}\end{pmatrix}$$

We have:

$$\text{Expected available inventory} = \frac{Q+1}{2} + s$$

$$\text{Expected amount on order} = d_\ell = D\ell$$

$$\text{Expected number of backorders at any time} = \frac{1}{Q}\,[b(s) - b(s+Q)]$$

where function $b(v)$ is defined by:

$$b(v) = \frac{(D\ell)^2}{2} P(v-1;\, d_\ell) - (D\ell)vP(v;\, d_\ell) + \frac{v(v+1)}{2} P(v+1;\, d_\ell) \tag{164}$$

In (164) function $P(u;\, d_\ell)$ is the complement of the Poisson cumulative distribution function, i.e.:

$$P(u_\ell;d_\ell) = \sum_{d_\ell=u}^{\infty} p(\tilde{d}_\ell;\, d_\ell) \tag{165}$$

The expected on hand inventory is:

$$\bar{I} = \frac{Q+1}{2} + s - D\ell + \frac{1}{Q}\,[b(s) - b(s+Q)] \tag{166}$$

The last cost component is the backorders cost.

$$\begin{pmatrix}\text{Expected number of}\\ \text{units backordered}\\ \text{per year}\end{pmatrix} = \frac{D}{Q}\,[a(s) - a(s+Q)] \tag{167}$$

where function $a(v)$ is:

$$a(v) = D\ell P(v;D\ell) - vP(v+1;D\ell)$$

The total expected annual cost function TC is constructed from (4.166), (167), and the ordering cost:

$$(168) \quad TC = A\frac{D}{Q} + rc\left\{\frac{Q+1}{2} + s - D\ell + \frac{1}{Q}[b(s) - b(s+Q)]\right\} + c_s\frac{D}{Q}[a(s) - a(s+Q)]$$

When it is rather costly to incur backorders, the minimization of TC acts so as to make shortages improbable. Then the terms a(s+Q) and b(s+Q) become negligibly small, can be removed from the cost function, and TC simplifies substantially.

Minimizing (168) requires a computer and use of a search technique; minimizing the simplified version, with a(s+Q) and b(s+Q) neglected, requires a somewhat less computational effort (Hadley and Whitin [1963, ch. 4-8], but it still takes a number of back and forth iterations between successive values of Q and s.

The complicated way in which total costs depend on Q and s, aggravated by the discreteness of the variables (which precludes use of derivatives in in optimization) explains the reserve management scientists show in recommending the Poisson distribution for practical work. It is not uncommon to find inventory control systems in which for all items lead time demand is assumed to vary normally; the differentiation among A, B, C class, and slow or fast moving items is made when deciding upon the appropriate forecasting technique, the level of service and the amount of managerial attention to be assigned to various groups of items.

It was mentioned earlier [relation (110)] that lead time demand can be considered Poisson distributed if the standard deviation of lead time demand is within 10% of $\sqrt{\text{Lead time demand forecast}}$. Otherwise, Peterson

and Silver [1979, ch. 9] recommend the use of the Laplace distribution.

The Laplace distribution, also called the bilateral exponential, is composed of two symmetric back-to-back exponential functions. Let the Laplace variable be the probabilistic lead time demand $\tilde{d}_\ell$; the distribution is characterized by the mean value d_ℓ and the standard deviation σ_ℓ:

$$f(\tilde{d}_\ell) = \frac{1}{\sqrt{2}\,\sigma_\ell} e^{-\frac{\sqrt{2}}{\sigma_\ell}|d_\ell - d_\ell|} \tag{169}$$

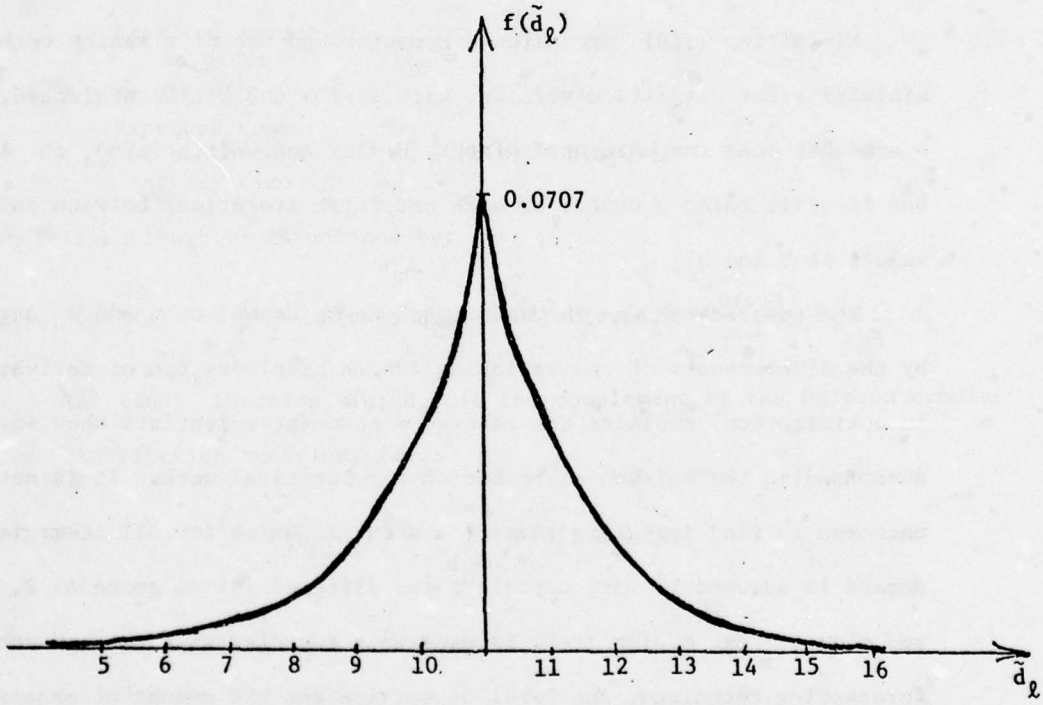


Figure 20: Laplace density function with mean d_ℓ = 8 units and σ_ℓ = 1 unit

Let us reconsider now the problem of section 4.2.1, where the stock-out cost is proportional to the number of units short. The following assumptions hold:

- lead time demand is Laplace distributed;

- the average demand rate is a constant, D units/year
- and order of size Q is released precisely when the available inventory reaches the order point;
- lead time is a constant ℓ;
- unfilled demand is backordered; for each unit backordered a charge c_s is incurred;
- average shortages are considered negligibly small when compared with the average inventory.

Consider that both the safety factor k and the lot size Q are unknown. We restrict our attention only to nonnegative values for k.

Similarly to equation (133) we can write the expression of the total expected annual cost:

$$TC = A\frac{D}{Q} + (\frac{Q}{2} + k\sigma_\ell)rC + c_s\frac{D}{Q}\left(\begin{array}{c}\text{Expected number of}\\ \text{units short per}\\ \text{replenishment cycle}\end{array}\right)$$

$$\begin{array}{c}\text{Expected number of}\\ \text{units short per}\\ \text{replenishment cycle}\end{array} = \int_{d_\ell+k\sigma_\ell}^{\infty} [\tilde{d}_\ell - (d_\ell+k\sigma_\ell)]f(\tilde{d}_\ell)d(\tilde{d}_\ell)$$

Make a change of variable: $u = \frac{\tilde{d}_\ell - d_\ell}{\sigma_\ell}$ to obtain:

$$\begin{array}{c}\text{Expected number of}\\ \text{units short per}\\ \text{replenishment cycle}\end{array} = \frac{\sigma_\ell}{\sqrt{2}}\int_k^{\infty}(u-k)e^{-\sqrt{2}\,u}\,du =$$

$$= \frac{\sigma_\ell}{\sqrt{2}}\int_k^{\infty} u\,e^{-\sqrt{2}\,u}\,du - \frac{\sigma_\ell}{\sqrt{2}}k\int_k^{\infty} e^{-\sqrt{2}\,u}\,du =$$

$$= \frac{\sigma_\ell}{2\sqrt{2}}e^{-\sqrt{2}\,k}(\sqrt{2}\,k+1) - \frac{\sigma_\ell}{2}k\,e^{-\sqrt{2}\,k} =$$

$$= \frac{\sigma_\ell}{2\sqrt{2}}e^{-\sqrt{2}\,k}$$

Hence:

$$(170) \qquad TC = A\frac{D}{Q} + (\frac{Q}{2} + k\sigma_{\ell})rC + c_s \frac{D}{Q} \cdot \frac{\sigma_{\ell}}{2\sqrt{2}} e^{-\sqrt{2}\,k}$$

It is easy to show that TC is convex (its Hessian is positive definite) under our restriction that $k \geq 0$. Setting $\frac{\partial(TC)}{\partial Q} = 0$ and $\frac{\partial(TC)}{\partial k} = 0$ yields:

$$(171) \qquad Q = \sqrt{\frac{2\left(AD + \sigma_{\ell} c_s D \frac{e^{-\sqrt{2}\,k}}{2\sqrt{2}}\right)}{rC}}$$

$$(172) \qquad k = \frac{1}{\sqrt{2}} \ln \frac{c_s D}{2QrC}$$

Again, as seen in earlier developments, the optimal lot size tends to be larger than that given by Wilson's formula (1), thus reducing the exposure to the risk of running out of stock over the year.

As it is often the case, Q can be determined independently of k without incurring a too severe penalty for missing the mathematical optimum. Then, we are left with finding the best safety factor k from (172).

Following the general lines of reasoning used so far, a large variety of results* can be further developed for slow moving items under various service level considerations (see measures of service in section 4.2.2) for either joint or sequential determination of the order quantity and the safety stock. However, as it was already mentioned elsewhere in this book, the choice among models for practical applications has to strike a balance between the need for accurately capturing the observed item's characteristics, the system cost brought about by the model complexity, and the managerial considerations regarding service to customers.

---

* Hadley and Whitin [1963, ch. 4], Peterson and Silver [1979, ch. 7, 9, and 11].

4.3.2 Safety Stocks Based on Empirical Distributions

Suppose that, based on the item's history, we have empirically determined the cumulative probability distribution of lead time demand $\tilde{d}_\ell$.

The order point s is determined by reading it right off the probability curve (Figure 21) so as to satisfy certain service requirements:

$$\text{Prob}\ (\tilde{d}_\ell \leq s) = \text{SERVICE} \tag{173}$$

or

$$\text{Prob}\ (\tilde{d}_\ell > s) = 1 - \text{SERVICE} \tag{174}$$

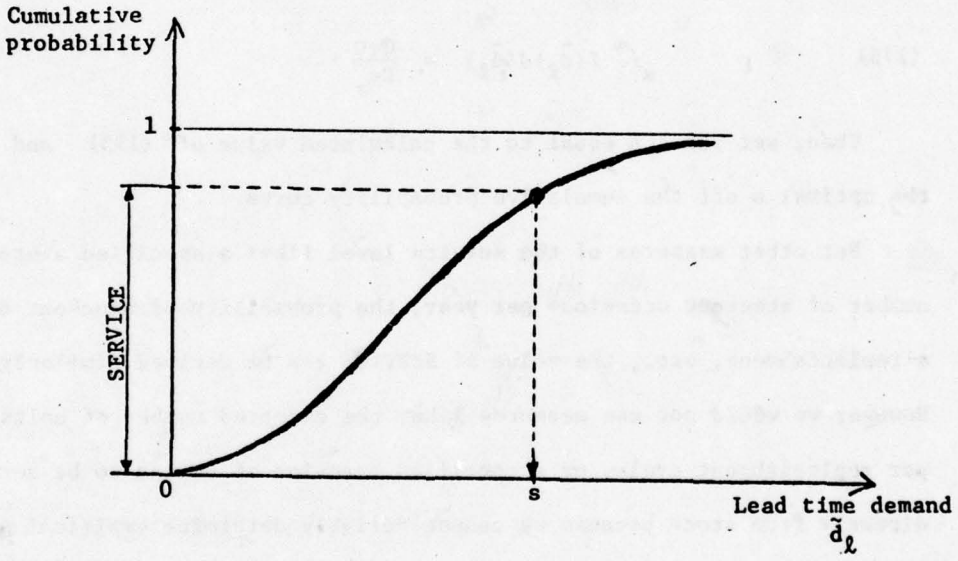


Figure 21: Empirical probability distribution curve of demand over the lead time

The value to be assigned to what we symbolically call SERVICE should result from managerial considerations. Technically, any method that was presented in section 4.2 for setting safety stocks can be used, as long

as it yields a control parameter in the form of either a cumulative probability value or the complement of the cumulative probability.

To illustrate, consider that the cost of backordering is proportional to the number of units short (section 4.2.1). Then, assuming Q and the order point s to be continuous variables, we write the total expected annual cost by similarity to expression (133):

$$TC = A\frac{D}{Q} + (\frac{Q}{2} + s - d_\ell)rC + c_s\frac{D}{Q}\int_s^\infty (\tilde{d}_\ell - s)f(\tilde{d}_\ell)d(\tilde{d}_\ell)$$

Density $f(\tilde{d}_\ell)$ is analytically unknown but this affects us in no way.

For a predetermined Q set $\frac{d(TC)}{ds} = 0$ to get the best s; the first order condition yields:

$$\int_s^\infty f(\tilde{d}_\ell)d(\tilde{d}_\ell) = \frac{QrC}{Dc_s} \tag{175}$$

Then, set SERVICE equal to the calculated value of (175) and read the optimal s off the cumulative probability curve.

For other measures of the service level like: a specified average number of stockout occasions per year, the probability of stockout during a replenishment, etc., the value of SERVICE can be derived similarly. However we would not use measures like: the expected number of units short per replenishment cycle, or a specified fraction of demand to be served directly from stock because we cannot reliably determine empirical partial expectations.

If we have to work with integer numbers for the order point, then s becomes the smallest integer such that $Prob(\tilde{d}_\ell \leq s) \geq SERVICE$.

5 The System Integrative Module

The system integrative module assembles the decision rules derived in the previous chapters into inventory control policies. An inventory control policy should be able to provide answers to two basic questions: WHEN to order a replenishment batch and HOW MUCH to order of the particular item controlled. Throughout the section on safety stocks an order point - order quantity policy (i.e., an order of size Q is released when the available inventory reaches the order point s) has been assumed because it best suited the presentation and development of the topic. Occasionally it was hinted, however, that, for cases not meeting all the assumptions made there, other policies might be appropriate.

Before putting an inventory control system together we still have to deal with the decision whether to make a particular item to order or to stock it.

5.1 Made-to-order vs. Stock Items

If an item is made to order (or purchased to order), its production (or purchase) should be limited to the amount generated by a specific customer order received, and should start at a date compatable with the required or promised delivery date. If, on the contrary, we are dealing with a stock item, we are expected to be able to fill the item's demand directly from stock, at least to the extent made possible by the service level considerations built into the inventory control system.

In section 1.1 several reasons for holding inventories have been discussed, and we have seen that stocks perform certain functions. In order to decide on making to order or stocking an item, we concentrate on two *essential aspects:* inventories are held to attend customer service requirements or to provide cost savings. Consequently, to determine whether or not an item should be stocked we should analyze, first, if service

considerations force us to keep the item in stock, regardless of cost implications, and, second, if this is not the case, whether we should still stock it due to cost considerations.

Made-to-order vs. Stock Classification Based on Service Considerations

We can associate with each product a quantity called "desired maximum promised delivery lead time", which is the maximum period of time we still consider acceptable for the customer to wait for the delivery of the ordered merchandise. This delivery time can be evaluated by taking into account the previous delivery history of the product, the customers' requests, and the competitors' delivery policies. Also, determine for each product the "manufacturing (or purchasing) lead time" as the length of time that elapses from the moment at which a manufacturing order (or purchasing requisition) is released until the product is physically available for delivery. The lead time can be determined from previous history or by estimates provided by foremen and dispatchers (or purchasing agents).

Then, the decision rule is as follows: if the manufacturing (or purchasing) lead time is larger than the desired maximum promised delivery lead time, classify the product as a stock item; otherwise, there is no need to stock the product based on service considerations and an economic analysis of the consequences of the made-to-order versus stock decision has to be performed.

One point to be emphasized: an item may be exceedingly important to make the production or logistics process of a firm feasible. In such a case, although service considerations, measured in terms of the desired delivery lead time, may not force us to stock the item, the constraints imposed by the process itself may require one to do so.

Made-to-order vs. Stock Classification Based on Economic Considerations

We have seen that even if the result of the previous analysis indicates

that the product need not be stocked for service reasons, it is still possible that it would be desirable to stock it for economic reasons. Thus, we have to compare the cost incurred if the item is made to order with the cost resulting if it is stocked.

The assumptions and the data of the problem* are as follows:

- the expected annual demand for the item is D units/year; it is it is expected that a number of N orders are received annually from customers;
- there is a fixed charge of A dollars/order for the manufacturing setup (or for placing a purchasing order);
- if the item is stocked, it is ordered in economic order quantities Q; also, to protect against uncertainties a safety stock (SS) is held;
- it costs C_{STOCK} dollars/unit to procure the item for stock; it costs C_{MTO} dollars/unit when the item is procured to order; the two costs are established by assuming that in the made-to-order case the expected order size is $\frac{D}{N}$ units, and in the stock base it is Q units;
- stock is carried at a charge of r dollars/dollar·year;
- service level is high enough so as to make the backorders cost negligible.

The expected annual made-to-order cost TC_{MTO} is determined by the total ordering cost and the cost of the items themselves:

$$TC_{MTO} = NA + DC_{MTO} \tag{176}$$

When the product is stocked, the expected annual cost TC_{STOCK} is:

$$TC_{STOCK} = \frac{D}{Q} A + [\frac{Q}{2} + (SS)]rC_{STOCK} + DC_{STOCK} + C_{syst}$$

* Under somewhat different assumptions the problem was also worked out by Popps [1965].

where C_{syst}, dollars/year, is the system cost (the item's share) of having the item stocked.

Since, if stocked, the item is order in economic order quantities $Q = \sqrt{\frac{2AD}{rC_{STOCK}}}$ the expression of TC_{STOCK} becomes:

$$(177) \qquad TC_{STOCK} = \sqrt{2DArC_{STOCK}} + (SS)rC_{STOCK} + DC_{STOCK} = C_{syst}$$

Certainly, if quantity discounts are available, items are replenished jointly, or situations other than the conditions of the classical economic lot size model occur, the appropriate formula from section 2 has to be used.

The decision rule to be applied is then: if $TC_{MTO} < TC_{STOCK}$ classify the item as made-to-order; otherwise, classify it as a stock item.

By comparing (176) and (177) some qualitative qualifications can be expressed. As the ordering cost A increases, TC_{MTO} grows faster than TC_{STOCK} and thus the system tends to the stocking situation. A small number of orders per year tends to favor the made-to-order situation; this supports the intuitive feeling that the made-to-order versus stock question is more likely to be raised with slow movers rather than fast moving items. Obviously, if the carrying charge r is large, holding inventories becomes disadvantageous and we are pushed towards the made-to-order decision.

Note that in order to compute the total costs we need estimates for all the parameters involved (yearly demand, number of orders, etc.). The required safety stock can be taken from historical records, if the product was previously stocked; if not, an approximation can be used in the following form:

$$(SS) = k\sqrt{\ell \frac{D}{12}}$$

where k is the safety factor, and ℓ is the lead time, in months. This approximation assumes that demand during the lead time ℓ is Poisson; after all, this assumption is not that bad when we recall that the slow moving items are the ones that are more likely to be involved in made-to-order versus stock decisions.

One final point: because of the statistical variations in the underlying data there are induced fluctuations in the estimates of the parameters in cost expressions (176) and (177). It is conceivable that, when parameters' values are revised, an item which was previously made-to-order might be pushed into the stock items class, or vice versa, solely because of the random fluctuations in data. To prevent this kind of unstable heavior Johnson [1962] suggests that, as long as the difference between the values of TC_{MTO} and TC_{STOCK} does not exceed a certain threshold, the sense of the inequality between TC_{MTO} and TC_{STOCK} may be allowed to change without changing the classification the item had before running the test.

5.2 Continuous vs. Periodic Review Systems

The order point - order quantity control policy presented earlier assumes that the stock level is exactly known at every point in time; this is the only way in which we can tell when the order point s is reached and, then, place a replenishment order for an amount Q. This policy is known under the short name of (s,Q).

In general, a continuous review system is one in which the stock status has to be always known.* In practice, rather than continuously surveying the inventory level, an equivalent approach is adopted: each transaction (taking orders from customers, shipping, receiving stock, placing orders to suppliers) triggers an immediate updating of the inventory

* Zimmermann [1966] calls it "perpetual inventory control system".

the performance of the system, is the current replenishment decision. It follows that, in a periodic review system, the safety stock must be large enough to provide protection for a length of time $\ell+R$. Thus, to achieve the same service level it takes less safety stock in a continuous system than in a periodic system. This is certainly advantageous when:

- inventory carrying charages are high either because the cost of the controlled item is high or because the item requires special conditions that makes stocking expensive;
- the required safety stock is relatively large due to either a high service level desired by management or because of severely fluctuating demand which results in relatively inaccurate forecasts.

If the cost of operating a transactions reporting system is significant, the alternative is a periodic review system. Basically, every R periods the stock status is reviewed and a decision is made about how much to order (possibly nothing). If ordering is expensive compared with the review costs one would place an order only if the inventory level were really low at the time of the review. If, on the contrary, ordering costs are small relative to review costs, one might want to order a positive quantity every time a review is conducted in order to avoid having wasted the costly review.

In what follows several basic policies are examined; each is characterized by a number of control parameters and the purpose of the presentation is to show how to determine their values:

- Continuous review systems
  - (s,Q) policy - when available inventory* reaches level s order Q units

* Recall that the available inventory is defined as the inventory on hand plus the amount on order less the number of units backordered.

status. For this reason the continuous review system is also known as the transactions reporting system.

It is easy to think of situations where a continuous review system is obviously not a good choice. For instance, if the supplier of a line of items only accepts orders once a week there is no reason why we should review the stock of those items more often, and when we review it we should do it right before placing the order. Such a system, as opposed to the previous one, is called periodic review system because the stock status is determined periodically. The time between two stock reviews is the review period; it spans R periods of time.

Before taking a closer look at the various policies that can be operated under the two systems, a few more qualitative considerations are in order here. Besides conditions of the sort shown above, there are also economic considerations that can make one system look more attractive than the other.

If the review costs are small (i.e., the processing of transactions is inexpensive compared with ordering costs and the annual number of transactions is small relative to annual demand for the item) a continuous review system is preferred because it leads to lower overall inventories than a periodic system. Indeed, when working with an (s,Q) policy sufficient safety stock must be provided to offer protection over the ℓ-period replenishment lead time. In a periodic review system the situation is different because replenishment decisions are made R periods apart. Suppose the current decision is made at time t (of course, taking into account all outstanding orders); everything ordered up to and including moment t is delivered to stock until $t+\ell$. The next replenishment decision is made at t+R, with delivery $t+R+\ell$. Clearly, then, in time interval $[t, t+R+\ell]$ our only opportunity to influence the stock level and, hence,

the performance of the system, is the current replenishment decision. It follows that, in a periodic review system, the safety stock must be large enough to provide protection for a length of time ℓ+R. Thus, to achieve the same service level it takes less safety stock in a continuous system than in a periodic system. This is certainly advantageous when:

- inventory carrying charages are high either because the cost of the controlled item is high or because the item requires special conditions that makes stocking expensive;
- the required safety stock is relatively large due to either a high service level desired by management or because of severely fluctuating demand which results in relatively inaccurate forecasts.

If the cost of operating a transactions reporting system is significant, the alternative is a periodic review system. Basically, every R periods the stock status is reviewed and a decision is made about how much to order (possibly nothing). If ordering is expensive compared with the review costs one would place an order only if the inventory level were really low at the time of the review. If, on the contrary, ordering costs are small relative to review costs, one might want to order a positive quantity every time a review is conducted in order to avoid having wasted the costly review.

In what follows several basic policies are examined; each is characterized by a number of control parameters and the purpose of the presentation is to show how to determine their values:

- Continuous review systems
  - (s,Q) policy - when available inventory* reaches level s order Q units

* Recall that the available inventory is defined as the inventory on hand plus the amount on order less the number of units backordered.

- (s,S) policy - when available inventory becomes equal to or less than s order up to level S

- Peridic review systems

  - (nQ,S,R) policy - if, at a review time, the available inventory is less than or equal to s, an amount nQ is ordered (n=1,2,3,...); multiple n should be such that, after the order is placed, the available inventory reaches a level in the interval (s,s+Q]. If available inventory is greater than s no order is placed.

  - (S,R) policy - at each review time a sufficient quantity is ordered to bring the level of the available inventory up to a level S

  - (s,S,R) policy - if, at a review time, the available inventory is less than or equal to s a sufficient quantity is ordered to bring the level of the available inventory up to S; otherwise no order is placed.

Clearly, under the assumption of continuous and deterministic demand there is no difference between the two systems; differences appear, however, for stochastic demands.

Discussions on these policies are conducted under the assumption that the probability distribution of demand is stationary.

Notes on multi-echelon systems and on control policies for the case of dynamic demand patterns are included in section 3.5.

In general, the exact formulation of the operation of an inventory system tends to lead to rather involved models that require, in most cases, a computer and appropriate search routines or iterative schemes to obtain the optimal values of the control variables. We have already seen it in the case of an (s,Q) policy, the simplest of all control policies, where the Poisson generated demands (section 4.3.1) yielded a fairly complicated

cost function. Even if we treat the simpler case of normally distributed demands, the exact formulation still raises computational problems (see Hadley and Whitin [1963, ch. 4-9]); the situation becomes that more serious as these problems are amplified by the number of items in the controlled inventory. Moreover, some problems, like the lost sales case when more than one order is outstanding, still lack a satisfactory exact treatment.

For these reasons we hold the view that, for routine use, heuristic approximate formulations, which yield operational decision rules with minimum cost penalties for missing the mathematical optimum, are recommended. At the same time we have to acknowledge the basic theoretical body of work on inventory control policies since, although the results might be too sophisticated for day-to-day use by most companies, their value is qualitative in that they show the structure of the optimal solution to be followed and how various parameters affect the system's controls.

5.3 Continuous Review Systems

5.3.1 Order Point - Order Quantity, (s,Q), Policy

We are already familiar with this policy since it has been extensively assumed in our treatment of the safety stocks. Figure 22 gives a graphical representation of how the available (AI) and on hand (OH) inventory vary with time under an (s,Q) policy.

The setting of the two control variables s and Q has been discussed in sections 4.2 and 4.3, i.e., either Q is computed as the economic order quantity and the order point s is determined afterwards, or Q and s are calculated jointly.

One form of physical materialization of the (s,Q) model is the "two-bin system", largely applicable to class C items. The stock of an item is stored in two bins. One of them, usually larger, is the working bin

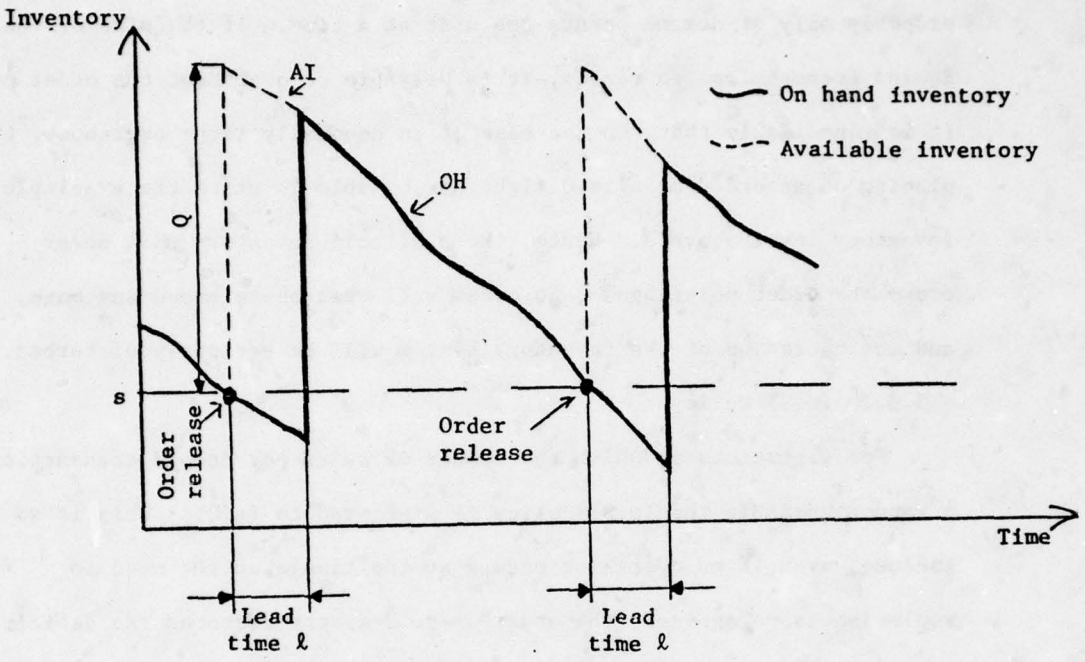


Figure 22: Inventory pattern under (s,Q) policy

from which demand is currently served without requisitions or paper work. The second bin is opened when the working bin becomes empty; at the same time, when this second bin is opened, a replenishment order for the item is released. Clearly, the amount of stock in the second bin materializes the order point. When the order comes in the second bin is refilled first, and the excess amount is put into the working bin. Since this system works properly only if there is no more than one order outstanding at any point in time, it is recommended that the order quantity should last much longer than the replenishment lead time. The order point should be set so as to offer a high service level (recall that C items are cheap) to compensate for the rather loose control.

Before closing this problem we want to reiterate the idea that, when the integrality of demand is taken into account, the (s,Q) policy works

properly only if demand occurs one unit at a time. If the size of the demand transactions is random, it is possible to overshoot the order point. It is conceivable that, in the case of an unusually large overshoot, the placing of an order of size Q might not be able to raise the available inventory level above s. Hence, the available inventory will never cross the order point again, no order will ever be released any more, and the operation of the inventory system will be seriously disturbed.

5.3.2 (s,S) Policy

For situations in which the number of units per demand transaction is a random variable the (s,S) policy is preferred to (s,Q). This is so because, even if an overshoot occurs at the time when the need to replenish is recognized, the order-up-to S system restores the deficit. The replenishment quantity varies depending on how large the overshoot is; in Figure 23 the placing of two replenishment orders, of differing sizes Q_1 and Q_2, is illustrated.

A case in which an (s,S) policy is the obvious choice of a continuous review system is the replenishment control for lumpy items.

The (s,S) policy receives considerable attention by Arrow, et al. [1958]. One of the difficulties associated with modelling it is finding the probability distribution of the overshoots, as this depends on the S-s difference and on the probability distribution of the number of units in a demand transaction. Karlin [1958] derives the distribution of overshoots under the reasonable assumption that the average demand transaction size is substantially smaller than the amount S-s. Once this is established, the two controls s and S can be determined along lines similar to safety stock calculations.* This procedure may be justified when we are dealing with fairly expensive items.

* Peterson and Silver [1979, ch. 14-2] develop an approximate formulation of the problem and its solution.

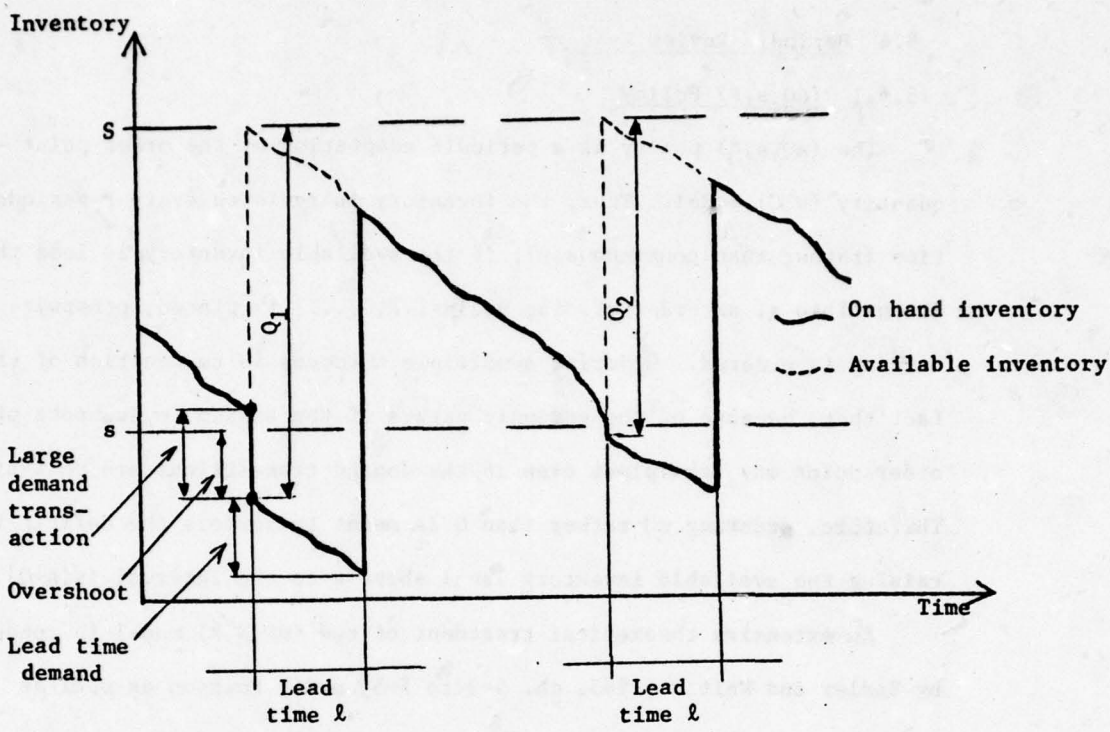


Figure 23: Inventory pattern under (s,S) policy

Real world inventory systems tend to operate heuristic versions of the (s,S) policy. While retaining the essence of (s,S), namely that to make up for overshooting the order point one orders up to S, the two control variables s and S are determined using the (s,Q) model. The basic assumption of this approach, which in most cases turns out to be correct, is that large overshoots are quite improbable and, therefore, neglecting the overshooting phenomenon is a reasonable approximation.

Thus, Q is set equal to the economic order quantity (section 2), then the order point s is calculated according to section 4, and finally the value of S is determined by: S = s+Q. If more precision is required and the added sophistication is warranted, Q and s are computed jointly rather than sequentially.

5.4 Periodic Review Systems

5.4.1 (nQ,s,R) Policy

The (nQ,s,R) policy is a periodic adaptation of the order point - order quantity (s,Q) model. Thus, the inventory is reviewed every R periods of time (rather than continuously); if the available inventory is less than or equal to s, an order of size nQ (n=1,2,3,...) is placed, otherwise nothing is ordered. Ordering a multiple nQ comes in recognition of the fact that, because of the periodic nature of the system, overshoots of the order point may take place even if the demand transactions are unit sized. Therefore, ordering nQ rather than Q is meant to restore the deficit by raising the available inventory level above s in the interval (s,s+Q].

An extensive theoretical treatment of the (nQ,s,R) model is conducted by Hadley and Whitin [1963, ch. 5-3 to 5-5] under Poisson as well as Normally distributed demands. The computational procedures to determine the optimal control parameters Q, s, and R turn out to be quite complicated, requiring a computer and appropriately designed search procedures.

In general, in the real world the (nQ,s,R) policy has gained much less acceptance than the (s,S,R) and (S,R) models. One reason can be that an (s,S,R,) type of control policy, although by no means computationally simple, is expected in general to yield a lower average annual cost than (nQ,s,R). Secondly, the (S,R) policy is easier computationally than both (nQ,s,R) and (s,S,R). Moreover, when ordering costs are relatively low compared to review costs, one would tend to order every time a review is made and, therefore, in such a case the (S,R) policy is essentially optimal besides being computationally easier.

5.4.2 (S,R) Policy

In practical applications (S,R) is the most largely used periodic review policy. It works as follows: every R periods of time the inventory

is reviewed and an order is placed so as to make the available inventory level equal to S.

Its wide acceptance is due to several advantages:

- its operation is simple and, therefore, it is easily understood by clerical personnel;
- the computation of the controls, S and R, is simpler than with other periodic review policies, especially in the heuristic version of the policy;
- it results in a predictable work load on the purchasing or production scheduling departments as opposed to the (nQ,s,R) and (s,S,R) policies where the number of orders released at a review time fluctuates depending on the relative positions of the available inventories with respect to the order points.

As mentioned earlier, however, the (S,R) policy is not universally recommended. Indeed, if ordering cost is high one may save by not placing an order at every review time but rather only then when inventory is low and a replenishment is required in order to avoid a stockout situation.

Figure 24 pictures the time behavior of inventories resulting from an (S,R) policy. Clearly, the size of each replenishment order is equal to the demand during the preceding review interval (see point N).

The order placed at time t (point M) arrives at $t+\ell$. The next order will arrive at $t+\ell+R$. Hence, whether a stockout condition develops between $t+\ell$ and $t+\ell+R$ it all depends on the replenishment decision made at t. The problem is, then, very much similar to the setting of safety stocks in an (s,Q) system, i.e., for a given time R between reviews find the order-up-to-level S which provides adequate protection against stockouts over a time span of $\ell+R$ periods, without building unneeded inventory.

The exact cost equations for the (S,R) policy are formulated by Hadley

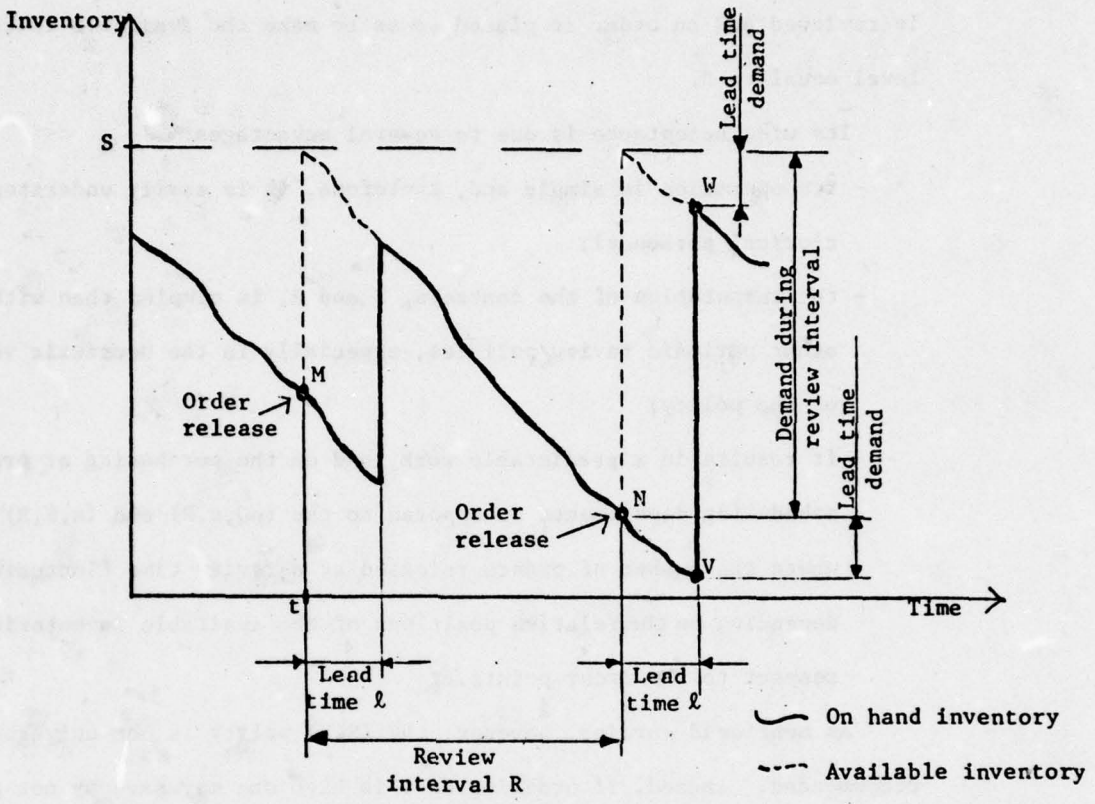


Figure 24: Inventory pattern under (S,R) policy

and Whitin [1963, ch. 5-6] for two cases: Poisson and normally distributed demands. Both cases are shown to be particular instances of the (nQ,s,R) models: for the Poisson demand make Q=1 and s+1=S in the corresponding (nQ,s,R) equations, and for the normal case set s=S and take the limit as Q→0. The resulting relations for the determination of the two controls, S and R, are too complicated to be of real practical use for routine applications.

We present below an approximate heuristic treatment of the (S,R) policy which yields simpler results under a set of reasonable assumptions. Thus, assume that:

- demand is a continuous variable whose probability distribution is stationary;

- the replenishment lead time is a constant;
- units demanded and out of stock are backordered; backorders, however, are expensive and, therefore, average shortages are considered negligibly small relative to the on hand inventory;
- the cost of making a review or, in general, the cost of the control system, is independent of the values of the control variables S and R.

The following notations are used:

D = average demand rate, units per year

A = ordering cost, dollars per order

J = cost of making a review, dollars per review

C = unit cost, dollars per unit of item (no quantity discounts)

r = inventory carrying charge, dollars/dollar/year

c_s = shortage cost incurred by the system for every unit backordered, dollars/unit

$\tilde{d}_{\ell+R}$ = demand over the replenishment lead time plus review interval; it is a random variable having the density function $f(\tilde{d}_{\ell+R})$ with a mean $d_{\ell+R}$ and a standard deviation of the forecast errors over a replenishment lead time plus a review interval of $\sigma_{\ell+R}$. Evidently, density $f(\tilde{d}_{\ell+R})$ depends on both ℓ and R.

The relevant costs are: the ordering, review, inventory carrying and backordering costs; we seek to minimize their total over a span of one year by appropriately setting S and R.

Since demand has a continuous density function the probability of no demand occurring during a review interval is zero; therefore, an order is placed with each review and:

(178) $$\text{Expected annual costs of ordering and review} = \frac{A+J}{R}$$

where R is expressed in years.

As the demand rate has a constant mean, inventory varies on the average linearly between a maximum (just after a replenishment arrives) and a minimum (just before a replenishment arrives).

Point W in Figure 24 illustrates a maximum:

$$\text{(179)} \qquad \begin{matrix}\text{Expected inventory just} \\ \text{after a replenishment arrives}\end{matrix} = S - d_{\ell+R} + D \cdot R$$

where the product D·R yields the demand during the review interval.

Point V illustrates a minimum:

$$\text{(180)} \qquad \begin{matrix}\text{Expected inventory just} \\ \text{before a replenishment arrives}\end{matrix} = S - d_{\ell+R}$$

By neglecting the backorders, according to one of our assumptions, the hali way between the above maximum and minimum can be regarded as closely approximating the average on hand inventory; then:

$$\text{(181)} \qquad \begin{matrix}\text{Expected annual} \\ \text{inventory holding cost}\end{matrix} = (S - d_{\ell+R} + \frac{D \cdot R}{2}) rC$$

Backorders occur whenever demand during the replenishment lead time plus a review interval exceeds S. Then (see section 4.2.1):

$$\text{(182)} \qquad \begin{matrix}\text{Expected annual} \\ \text{backordering costs}\end{matrix} = \frac{c_s}{R} \int_S^{\infty} (\tilde{d}_{\ell+R} - S) f(\tilde{d}_{\ell+R}) d(\tilde{d}_{\ell+R})$$

The total annual cost TC is the sum of (178), (180), and (182). If the review interval R is given, the optimal order-up-to-level S results from:

$$\text{(183)} \qquad \frac{d(TC)}{dS} = rC - \frac{c_s}{R} \int_S^{\infty} f(\tilde{d}_{\ell+R}) d(\tilde{d}_{\ell+R}) = 0$$

Let $\Pi(S) = \int_S^{\infty} f(\tilde{d}_{\ell+R}) d(\tilde{d}_{\ell+R})$, i.e., the complement of the cumulative of $f(\tilde{d}_{\ell+R})$. Then, S_o, the optimal value of S, is the solution of:

(184) $$\Pi(S) = \frac{rCR}{c_s}$$

This result is strikingly similar to (135) in which Q/D, the duration of a cycle, is the analogue of the review interval.

The review interval R could be the result of conditions external to the model; e.g., our vendor can accept our orders only every second Monday.

If this is not the case, R can be derived from the economic order quantity EOQ expressed as a time supply: $R = \frac{EOQ}{D}$. When computing the value of the EOQ the fixed cost J of making a review must be added to the ordering cost A. Of course, the resulting value of R should be adjusted so as to fit the modus operandi of the department which is in charge of inventory control.

Still another alternative is to determine the optimal R and S jointly. This requires the simultaneous solution of (184) and $\frac{\partial(TC)}{\partial R} = 0$. Since desnity $f(\tilde{d}_{\ell+R})$ depends on the review interval R, this case is computationally more involved than the previous situation in which R was exogenously determined. To solve use some numerical technique. A simple approach is the following: for each of a set of values R_1, R_2, R_3, ..., calculate the optimal S_o by use of (184). Plot the total cost TC as a function of R, using the corresponding S_o for the given R to compute TC. From the plot find the optimal R.

To treat the lost sales case we have to rewrite the inventory holding cost equation (181). In the backorders situation, when shortages developed, part of the incoming replenishment was used to satisfy the unfilled orders. This is not the case under the lost sales assumption as shortages are not backlogged; therefore, on hand inventories are larger by the average amount of units out of stock in a review cycle.

(185) $$\text{Expected number of units short per review interval} = \int_{S}^{\infty} (\tilde{d}_{\ell+R} - S) f(\tilde{d}_{\ell+R}) d(\tilde{d}_{\ell+R})$$

(186) $$\text{Expected annual inventory holding cost} = \left[S - d_{\ell+R} + \frac{D \cdot R}{2} + \int_{S}^{\infty} (\tilde{d}_{\ell+R} - S) f(\tilde{d}_{\ell+R}) d(\tilde{d}_{\ell+R}) \right] rC$$

Continue to consider c_s as the cost of one unit out of stock. Then, the total annual cost TC is obtained by summing up (178), (182), and (186). For a given R, the optimal value of S is a solution to the first order condition:

(187) $$\frac{d(TC)}{dS} = rC - RC \int_{S}^{\infty} f(\tilde{d}_{\ell+R}) d(\tilde{d}_{\ell+R}) - \frac{c_s}{R} \int_{S}^{\infty} f(\tilde{d}_{\ell+R}) d(\tilde{d}_{\ell+R}) = 0$$

which yields:

(188) $$\Pi(S) = \frac{rCR}{c_s + rCR}$$

where, as defined earlier, $\Pi(S)$ is the complementary cumulative of $f(\tilde{d}_{\ell+R})$.

In order to avoid costing out shortages explicitly, the order-up-to-level S can be determined by means of prespecified service level considerations. For this purpose it is advantageous to think of the expected inventory just before a replenishment arrives as a safety stock for a time span of +R. Then, from (180) it follows that:

(189) $$S = d_{\ell+R} + SS$$

where SS stands for the safety stock. By analogy with (121), from (189) we conclude that S should be set equal to the maximum reasonable demand during a replenishment lead time plus a review interval.

It is again reasonable to link the safety stock to the standard deviation $\sigma_{\ell+R}$ of the forecast errors over a lead time plus a review interval through a safety factor k:

$$SS = k\sigma_{\ell+R}$$

Thus:

(190) $$S = d_{\ell+R} + k\sigma_{\ell+R} \quad ,$$

and the general expression (185) changes to:

(191) Expected number of units short per review interval $=$

$$= \int_{d_{\ell+R}+k\sigma_{\ell+R}}^{\infty} [\tilde{d}_{\ell+R} - (d_{\ell+R}+k\sigma_{\ell+R})]f(\tilde{d}_{\ell+R})d(\tilde{d}_{\ell+R})$$

For fast moving items under normally distributed forecast errors (see section 4.2.1) we obtain the particular expression:

(192) Expected number of units short per review interval under Normal distribution $= \sigma_{\ell+R}G(k)$

For slow moving items, if the Laplace distribution applies (see section 4.3.1), (191) becomes:

(193) Expected number of units short per review interval under Laplace distribution $= \frac{\sigma_{\ell+R}}{2\sqrt{2}} e^{-\sqrt{2}k}$

By similarity with the developments of section 4.2.2, for a given review interval R and a prespecified measure of the service level, the value of the safety factor k can be determined, and the order-up-to-level S can be set accordingly by equation (190).

Extensions of the (S,R) policy to cover the case of stochastic replenishment lead times are provided by Hadley and Whitin [1963, ch. 5-2] under the assumption that orders cannot cross, i.e., the lower, ℓ_{min}, and the upper, ℓ_{max}, limits to the possible range of lead time values are such that $\ell_{max} < \ell_{min}+R$.

5.4.3 (s,S,R) Policy

When placing of an order is expensive, it may be advantageous not to order at every review time, with the purpose of saving on ordering costs. The way the (s,S,R) policy works is then: at every review time, the level of the available inventory is compared with s; if less than or equal to s a sufficient quantity is ordered to raise the available inventory up to S, if greater than s no order is released (Figure 25).

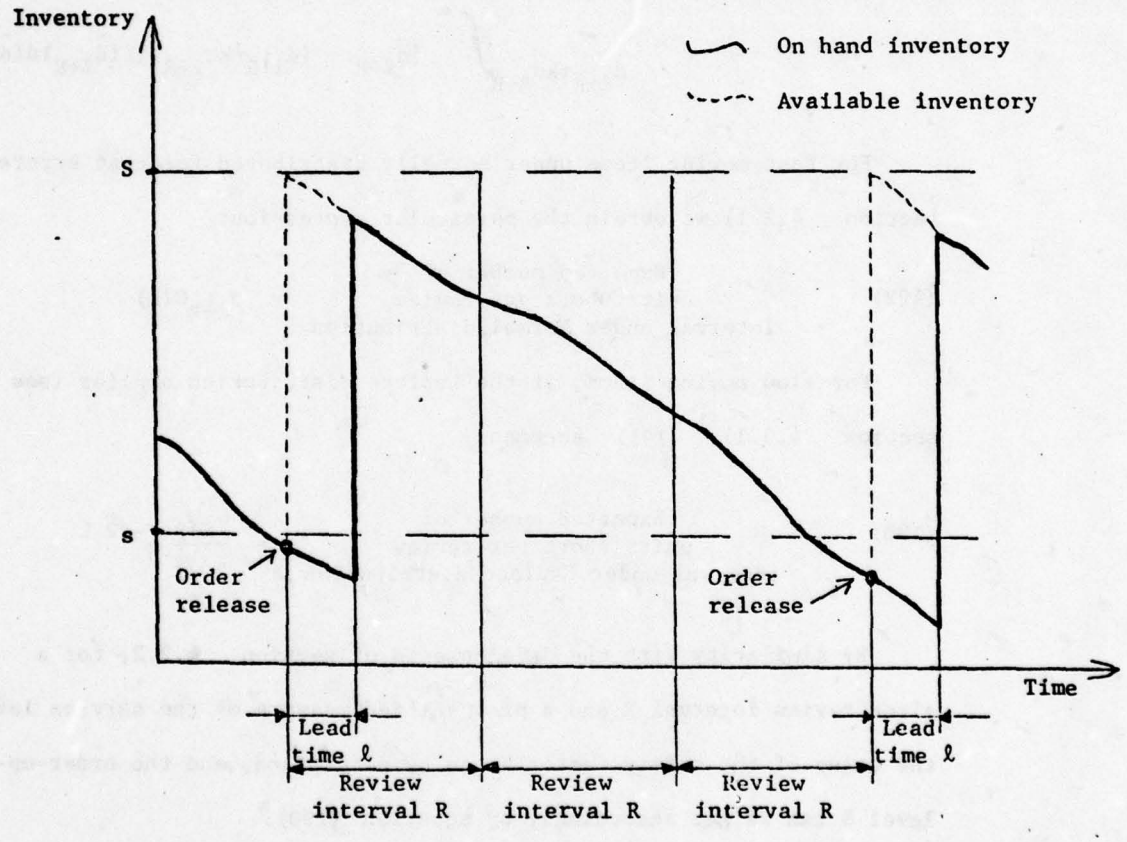


Figure 25: Inventory pattern under (s,S,R) policy

The optimality of this policy is studied by Scarf [1960], following earlier work by Arrow, Harris, and Marschak [1951], and by Dvortezky, Kiefer,

and Wolfowitz [1953]; see also Zabel [1962].

Under the assumptions of a constant replenishment lead time ℓ, backordering of unfilled demand, and constant unit cost for the controlled item, Scarf [1960] proves that the (s,S,R) policy is optimal if function g(x;R) is convex and differentiable everywhere with respect to x. Function g(x;R) is defined as follows: given a given review interval R, if t is a review time for the system and x is the available inventory after ordering a quantity $Q \geq 0$, g(x,R) is the sum of the expected carrying and backorder costs incurred in the period from $t+\ell$ to $t+\ell+R$ and discounted to time t by a positive interest rate. The vehicle of the study is a dynamic programming model developed over an n-period planning horizon (for convenience a period is redefined as spanning the time of a review interval). We present the results briefly below.

Consider, then, that t is a review time for the system; also, t works the beginning of period j.

$$(194) \qquad F_j(y+Q;R) = \min_{Q \geq 0} \left[J + A\delta(Q) + CQ + g_j(y+Q;R) + \right.$$

$$\left. + a \int_0^{\infty} F_{j+1}(y+Q-\tilde{d}_R) f(\tilde{d}_R) d(\tilde{d}_R) \right]$$

with $j=1,2,\ldots,n$; $F_{n+1}(\cdot) = 0$; $\delta(0) = 0$, $\delta(Q) = 1$ when $Q > 0$.

The recurrence relationship (194) calculates the total discounted costs if an optimal quantity Q is ordered as time t and at all future review times. y is the available inventory at time t before any order is placed; after ordering, the inventory level becomes y+Q.

The discount factor a, $0 \leq a \leq 1$, gives the present worth at time t of the costs evaluated at t+R.

$\tilde{d}_R$ is the demand occuring in a period of length R; its probability

distribution is described by the density $f(\tilde{d}_R)$.

All other notations are the same as in section 5.4.2.

Scarf's proof is based on showing inductively that the following function is A-convex if $g(x;R)$ is convex in x and differentiable everywhere:

$$G(x;R) = Cx + g(x;R) + a\int_o^{\infty} F(x-\tilde{d}_R)f(\tilde{d}_R)d(\tilde{d}_R) \tag{195}$$

$G(x;R)$ is A-convex* in x (A is the cost of placing an order) if for any $\alpha \geq 0$:

$$A + G(\alpha+x;R) - G(x;R) - \alpha G'(x;R) \geq 0 \tag{196}$$

where G' indicates the derivative with respect to x.

Notice that (196) does not imply convexity for $G(x;R)$. Therefore, $G(x;R)$ can have a shape as shown in Figure 26, with several local minima. s_o and S_o denote the optimum control variables of the (s,S,R) policy.

In the n-period dynamic programming model a $G_j(x;R)$ function can be written for each period j=1,2,...,n; the optimal policy in period j is characterized by two numbers s_{oj}, S_{oj}. Then, if available inventory y_j prior to the placing of any order is less than or equal to s_{oj} order up to S_{oj} because $G_j(y_j;R) \geq G_j(S_{oj}) + A$; if $y_j > s_{oj}$ do not order since $G_j(y_j;R) < G_j(S_{oj}) + A$.

Although local minima and maxima may exist, an A-convex function may not display a behavior like that of Figure 27. It is easy to show that point M, a local maximum, leads to a violation of the A-convexity property. Indeed, set $x = m$ and $\alpha = S_o - m$; as $G'(m;R) = 0$ relation (196) becomes

---

* A function f(x) is convex if for any x_1, x_2 and any β, $0 \leq \beta \leq 1$, the following holds: $f[\beta x_1 + (1-\beta)x_2] \leq \beta f(x_1) + (1-\beta)f(x_2)$. An equivalent definition of convexity is that for any $\alpha \geq 0$ we have: $f(\alpha+x) - f(x) - \alpha f'(x) \geq 0$, where f' is the derivative of f(x) with respect to x.

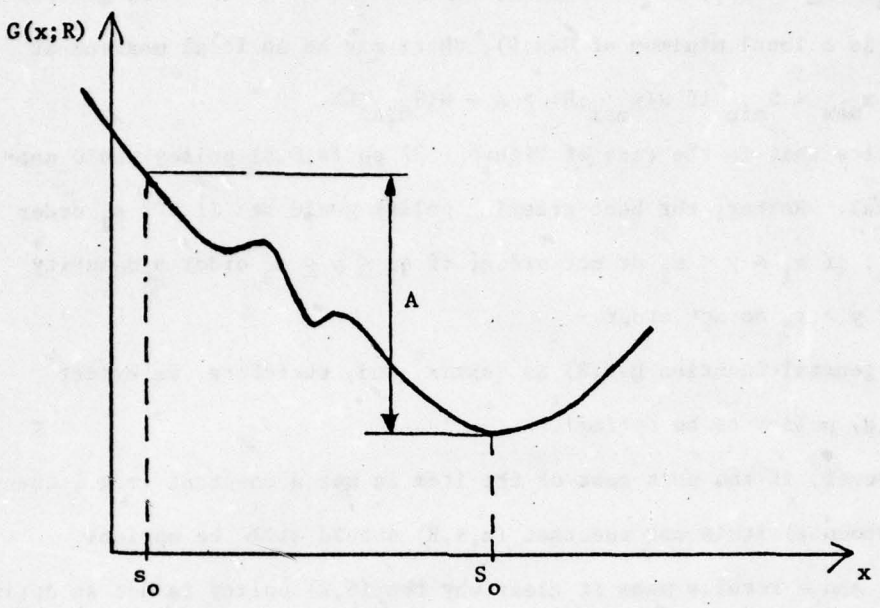


Figure 26: An A-convex curve

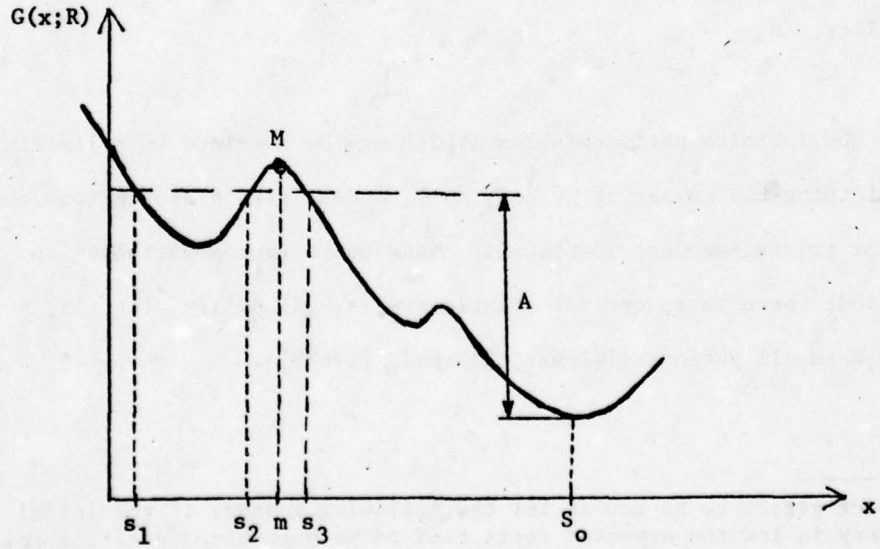


Figure 27: A case where an (s,S,R) policy is not optimal

$A + G(S_o;R) \geq G(m;R)$, which is obviously a contradiction. More generally, if S_{min} is a local minimum of $G(x;R)$, there may be no local maximum at a point $x_{max} < S_{min}$ if $G(x_{max};R) > A + G(S_{min};R)$.

Notice that in the case of Figure 27 an (s,S,R) policy would not be optimal. Rather, the best ordering policy would be: if $y \leq s_1$ order up to S_o; if $s_1 < y < s_2$ do not order; if $s_2 \leq y \leq s_3$ order a quantity $S_o - y$; if $y > s_3$ do not order.

In general function $g(x;R)$ is convex* and, therefore, we expect an (s,S,R) policy to be optimal.

However, if the unit cost of the item is not a constant (e.g., quantity discounts) it is not rue that (s,S,R) should still be optimal

The above results make it clear why the (S,R) policy is not an optimal policy unless $A = 0$. From Figure 26 it is apparent that the optimal S_o is the minimizing value of x in (195) and the reorder point s_o is such that $G(s_o) = G(S_o) + A$. Then, for $A = 0$ we have $s_o = S_o$ which yields an (S,R) policy.

For the infinite period problem (which can be obtained as a limiting case by letting the number of periods $n \to \infty$), under Scarf's assumptions and if demands are independent identically distributed random variables in each period, there is an optimal stiationary (s,S,R) policy, i.e., $s_j = s$ and $S_j = S$ in all periods (Iglehart [1963a], [1963b]).

---

* We expect $g(x;R)$ to be convex for the following reason: if the intial inventory is low the expected costs tend to be high because of the relatively large number of backorders. As x increases, the backorders situation improves and costs go down. But, beyond a certain point, a further rise in x causes the inventory holding charges to go up and eventually outweigh the reduction in backorder penalty, thus inducing an increase in $g(x;R)$.

Theoretically, one would find the solution to the recurrence equation (194) for a given R and discount rate a, to obtain the amount to be ordered Q as a function of the initial inventory y. In the steady state case with an infinite number of periods, the ensuing functional equation has to be solved or, if an approximation is acceptable, an n-period system (with n large) can be used instead. The optimal function Q_o)y;R) yields then the optimal s_o and S_o (see, for instance, Hadley and Whitin [1963, ch. 8-5]). To also optimize with respect to R, different values of the review interval have to be tried. In the real world the choice of values for R is limited by organizational considerations, relations wtih suppliers and customers, etc.

An alternative approach to computing the optimal values of s and S under stationary demand is to minimize the expected costs (review and ordering plus inventory holding and backorder costs) per unit time. Markov chain theory or renewal processes are used by various authors to obtain average cost formulas: Arrow, Harris, and Marschak [1951], Karlin [1958], Wagner, O'Hagan, and Lundh [1965], Veinott and Wagner [1965]. Hadley and Whitin [1963, ch. 5-9] derive the average annual cost by computing the expected cost per replenishment cycle (i.e., the time between the placing of two consecutive orders) and then dividing by the average length of a cycle. Thus they make use of a basic result of renewal theory (Karlin and Taylor [1975, ch. 5]) by which the long run expected costs per unit time = (expected costs per cycle)/(expected length of the cycle).

As a general rule, solving whichever approach is taken is difficult

computationally,* especially when thinking in terms of systems of realistic size (thousands or even more items to be controlled). As with other ordering policies, this has encouraged the development of approximate heuristic approaches: Wagner, O'Hagan and Lundh [1965], Snyder [1974], Naddor [1975], Peterson and Silver [1979, ch. 8.9].

The two control variables s and S should have some safety and economic features and, in principle, can be thought of, from a pragmatic point of view, as follows: the order point s represents an inventory level which, if at the current review time no order is placed, should be large enough to satisfy the maximum reasonable demand during a review interval R plus a replenishment lead time. When an order is released it has to satisfy an economic criterion and, therefore, the amount ordered is some sort of economic order quantity. The time span covered on the average by this quantity has to be an integer number of review intervals. Then, the order-up-to-level S is obtained by adding to s the economic order quantity.

In industrial settings it is often true that review costs are relatively high and, therefore, it is common to find review intervals large enough so that an order is placed at every review time with either an (S,R) or (s,S,R) policy. Hence, also considering the computational advantage, one would expect to find that for practical implementations the (s,S,R) policy is replaced by the (S,R) policy without important deviations from optimality.

On the other hand, if ordering costs are high relative to the review costs, in which case an (s,S,R) policy is definitely more economical than (S,R), the question can still be raised whether it is not advantageous to

---

* Under the assumption of exponentially distributed demand per time period Karlin [1958] obtains simple expressions for s and S. Thus, $S = s+Q$, where Q is given by the classical Wilson lot size formula, and s results from an exponential expression. However, the result presents interest rather from a theoretical point of view, because the exponential distribution is not a good match for the probability distribution of demand per period.

switch over to a continuous review system rather than using periodic review.

5.5 Other Issues in Inventory Control Systems

In the preceding sections we have confined our presentation to some fundamental problems in single item, single stocking point, stationary demand inventory control systems, emphasizing the need for operational replenishment policies which, while skirting undue computational difficulties, would still maintain the essence of theoretically proved optimal decisions.

As soon as the aforementioned assumptions are relaxed other control policies emerge. In section 2.3 we have seen some cases of multiple items sharing the same equipment or replenished under aggregate constraints. Other joint replenishment systems are possible. For example, consider a family of m items; in order to replenish any one item of the family, a major setup cost A is incurred at the family level and a minor setup cost a_i is involved in including item i in the order. Suppose a periodic review system is used, with coordinated (s_i,S_i,R) ordering policies, according to which each item has its individual order point s_i and order-up-to-level S_i, but the review interval R is the same for all items. For any given R the control variables s_i and S_i, i=1,2,...,m, have to be optimal, as discussed in section 5.4.3. The best R is determined by search so as to minimize the total expected costs of the system (review, ordering, inventory holding, and shortage costs).

A special type of continuous review (s,S) system developed for the case of a family of coordinated items is the can-order (s,c,S) policy. (Balintfy [1964], Maher, et al. [1973], Silver [1974]). If an item has to be ordered, the major family setup is incurred in excess of the minor setup a_i charged if item i is included in the replenishment. Each item

has its individual controls $s_i < c_i < S_i$, i=1,2,...,m. (s,c,S) operates as follows: when the available inventory of some item i becomes equal to or smaller than s_i an order is released for an amount sufficient to bring the inventory level up to S_i. At the same time, if any other item j of the family has its available inventory at or below c_j it is included in the order with a quantity large enough to raise the inventory level to S_j. Hence, c_j is called the "can-order point" and the policy is named accordingly. The idea is then evident: if at the time when a major setup is incurred item j's inventory is at c_j or lower, this is a signal that pretty soon item j itself will reach its order point* s_j and trigger a replenishment order and the associated major setup. Thus, in order to avoid an excessive number of major setups the (s,c,S) policy allows an item to be ordered even if it has not reached the order point s.

If the single stocking point assumption is dropped, the multi-echelon situation has to be modelled. Multi-echelon distribution networks and multi-stage production systems are the two broad classes of problems under this heading, and their vastness prohibits any attempt to treat the topic in this chapter; rather, a few notes are made and starting references are provided for the interested reader.

A "stage", "location", or "echelon" is any physical point at which inventories may be held. In the field of production systems, the terms "stage", "facility", and sometimes "station" are interchangeably used. A multi-echelon distribution system involves transfers of goods between locations, exogenous sources, and customers. A multi-stage production system can be thought of as being a production process in which component parts have to be obtained by manufacturing or by purchasing, then assembled

* s_j is also called "must-order point" to contrast with the "can-order point" c_j.

into subassemblies, assemblies, and finally into the finished good. It is useful, for modelling purposes, to conceptualize the system as a network in which a node is a location or stage; an arc connecting two nodes is used to represent an activity involving both nodes, and is usually directed. The node from which an arc leaves is the predecessor; the node where it ends is the successor. In multi-stage models we can identify two kinds of demands for the product stocked or manufactured at a stage: independent demand coming from customers (or market demand), and dependent demand generated by successor stages.

A number of multi-stage configurations may be distinguished:

- serial system - each stage can have no more than one successor and one predecessor;
- parallel system - each stage is single, it serves only independent demands (it has no predecessor or successor); however, stages may share costs;
- pure assembly system - each stage can have any number of predecessor stages, but at most one successor stage; the corresponding network converges twoard a node which, in a manufacturing setting, represents the assemblying of the final product to be delivered to the customer;
- arborescent system - each stage has a single predecessor but any number of successors; the associated network has a single supply location;
- acyclic system - each stage can have any number of predecessors and successors but the network representation contains no cycles, i.e., it is not possible to make a sequence of shipments such that material starts and ends at the same location;
- general system - the associated network allows transfers of goods between any two locations; the structure is arbitrary with no restric-

tions on the relationship between stages.

Because of the dependent demand, the treatment of multi-echelon situations is bound to differ markedly from the single stocking point problem. In general, dependent demand does not lend itself to statistical forecasting and, therefore, the methods of statistical inventory control are no longer appropriate.

In the area of distribution systems, comprehensive surveys have been written by Veinott [1966], Clark [1972], Aggarwal [1974], Karmarkar [1975]. We chose to present briefly in this section an important concept in multi-stage inventory control, developed by Simpson [1958], namely the base stock system.

Consider n stages in series, where the material flow takes place from stage 1 to stage 2, then from stage 2 to stage 3, etc. The characteristic of the base stock system is that information on actual customer demand is fed back directly to each stage. Each stage controls its ordering policy inde-

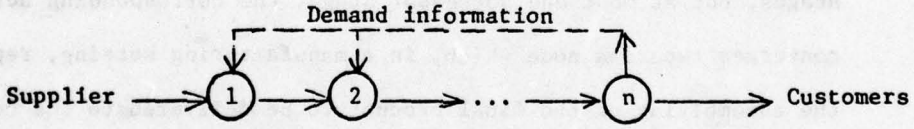


Figure 28: Serial system with n stages

pendently based on demand information rather than reacting to its successor's ordering policy. The system acquires thus much more stability than in the case where each stage sees only the demand generated by its successor and where small changes in end-item demand can lead to large oscillations in the replenishment orders and in the levels of inventories upstream.*

In Simpson's work, a base stock level B_i is associated with each stage

* The phenomenon is described by Forrester [1961, ch. 2].

i=1,2,...,n. When an order is received at one of the stages i a replenishment order is immediately placed with the predecessor stage i-1 for an equal amount. When a stage is out of stock any demand which occurs then at that stage is backordered; however, the reorder is still placed with the predecessor immediately when the demand arrives. Thus, the on hand inventory minus the amount backordered plus the quantity on order is always maintained constant and equal to the base stock level. Note that when end-items are withdrawn from the final stage n, a process of explosion takes place throughout the system.

The performance at stage i is characterized by the service time T_i defined as the maximum time needed to fill an order placed against inventory i. Evidently, if inventories are high, smaller service times are achieved and vice-versa. The determination of the optimum service times and, hence, the optimal base stock levels depends on a policy variable set by management: the probability α_i that the service time at stage i exceeds the value of T_i.

Given the α_i's (i=1,2,...,n) the base stock levels are calculated so as to minimize the expected inventory holding cost per unit time for the entire system. Simpson also imposes the conditions that $T_1=0$, which means that the raw material inventory (stage 1) is never empty, and $T_n=0$, i.e., delivery to the customer should take place without delay. The optimal solution is that each stage should carry either no inventory at all or the maximum reasonable level of stock which assures that the stage's inventory is never empty (the equivalent of $T_i=0$).

In the original development of the base stock system, the ordering or setup cost is left out of consideration. However, when it becomes so important that neglecting it would affect the economics of our decisions, the operation of the base stock control should be modified accordingly.

Certainly one would not place a replenishment order after each demand; rather, ordering should be made in economic order quantities or, if demand is not unit sized, a minimum amount should accumulate before an order is released. Thus, an (s,Q) or (s,S) policy would be used at each stage. However, the essence of the base stock system must still be retained, namely that the system is driven by actual customer demand. Therefore, the values of Q_i and s_i for stage i are based on end-item demand forecasts and the associated forecast errors over the replenishment lead time corresponding to stage i (Peterson and Silver [1979, ch. 12]).

In the area of multi-stage production systems, the assumption about the planning horizon appears to determine two different approaches to the problem. The infinite horizon models are of the economic order quantity type; tney start by evaluating in-process inventories, and then minimize the total of setup and inventory holding costs. In the finite horizon, models demand occurs by time periods, can be deterministic or stochastic, varying or constant in time; usually these models are of the mathematical programming sort. Some features of a general nature are reviewed below; for a more detailed view of the literature, the interested reader is directed to Johnson and Montgomery [1974, ch. 3-8 and 4-7], and Candea [1977].

In the multi-stage lot size problem demand is, in most cases, deterministic and occuring at a constant rate, costs are time-invariant, and the system is assumed to be in operation indefinitely into the future. In this static environment the assumption of time-invariant economic order quantities is reasonable. Schwarz and Schrage [1975] show that under the stated conditions, time-invariant lot sizes are optimal. If production is instantaneous and there are no capacity constraints, Crowston and Wagner [1970], and Crowston, et al. [1973] prove that in pure assembly systems

it is optimal to set the lot size at stage s equal to an integral multiple of the lot size at stage immediately succeeding s. This is not necessarily true when production is non-instantaneous (see Jensen and Khan [1972]), or when the structure of the process is not pure assembly. Candea [1977] finds the optimal lot sizing policy for the latter case, when the process is assembly leading to the production of an end-item; however, a stage is no longer restricted to having at most one successor, it can have several successors (e.g., a component part produced at some stage goes into two or more subassemblies assembled at successor stages). Whenever lot sizing in integral multiples is assumed to be or is optimal, the multiples are computed by dynamic programming procedures, by branch and bound lagorithms, or even by considering all possible combinations of lot sizes if the problem is small enough (Taha and Skeith [1970]). To reduce the search space, many procedures start out by developing heuristic bounds on the range of lot size multiples to be considered.

In the finite horizon models, in every period of the planning horizon a demand for a nonnegative amount of the product (or products) under consideration occurs. If the production process structure under study is facilities in series, the problem can be represented as a network flow problem of the transshipment type. If the cost structure is linear network algorithms, of the dynamic programming type, exist (Zangwill [1969] for the single product case; Veinott [1969] extends Zangwill's approach to arborescent and assembly line structures).

When an assembly system is modelled, the result is no longer a network flow problem; it becomes a combinatorial problem, and either dynamic programming or branch and bound algorithms are fit for solution. The authors who have studied this kind of problems put considerable effort into:

- exploiting the characteristics of the optimal solution in order to provide good formulations for the recurrence equations in the dynamic programming algorithms;
- finding good bounds in the branch and bound algorithms;
- finding "reasonable" restrictive assumptions upon the cost functions, which would thus enable them to develop better bounds and more efficient algorithms.

In general, the multi-stage production problem is still in an incipient stage in terms of results applicable to real world systems. The general case is pruned by assumptions until its structure allows solution by an existing technique. For instance, capacity constraints are more often than not disposed of* in order to allow solution by concave cost network techniques, and setup costs are neglected to permit formulation and solution by linear programming.

While discussing multi-stage systems, and since we have already touched upon the area of production, we should also mention the topic of Material Requirements Planning (MRP). Conceptualized for a manufacturing environment, MRP deals with the problems of determining the requirements of components (parts, subassemblies, assemblies), establishing the points in time when the components are needed, and scheduling the manufacturing or purchasing of the components so that they become available at the time of usage and not much earlier. The starting point for MRP is the assembly schedule, by time period, for the end-items (or final products); this is called the "master schedule". Working backwards from the master schedule, by a process

---

* Capacitated formulations may be found but, usually, other simplifications are brought in: Dorsey, et al. [1974], [1975] study the simpler problem of stages in parallel; von Lanzenauer [1970] waives the setup charges and obtains a linear programming model; Klingman, et al. [1977] develop a formulation for a parallel system with a planning horizon of one time period; Gabbay [1975] tackles the serial system with multiple products under restrictive assumptions on the cost structure.

called "explosion", the planned production quantities of end-product are projected into the appropriate amounts of required components. The book by Orlicky [1975] should make a good starting reference. The increase in computing capability brought about by the advent of the large scale random access computer has encouraged and has made possible the adoption of the method. Its popularity is due, in part, to its straightforwardness and its orderly and systematic approach to timed requirements planning. However, a big problem has still not been convincingly solved: how to draw a good (leave aside optimal) feasible master schedule? MRP assumes the master schedule as given, while we have seen earlier that optimizing efficient techniques for production planning in multi-stage systems are yet to be developed.

As already mentioned at the outset of section 5.5, the inventory replenishment policies developed earlier assumed stationary demand. When this assumption is dropped we have to deal with dynamic models for time-varying demand. Dynamic models usually have a finite planning horizon consisting of T time periods, and the demand is given in the form of d_i, $i=1,2,\ldots,T$ to be served in period i; thus, the demand can vary from period to period. Demand can be deterministic or stochastic.

If demand is deterministic and time invariant, one would order an economic order quantity precisely one lead time before the existing stock would vanish. When demand is deterministic and time-varying, a similar policy should be used, except that the order quantity will also change with time. An extensive treatment of this topic has been given in Hax [1978]. In the single product case, for which the optimizing algorithm of Wagner-Whitin is presented in the aforementioned chapter, a number of heuristics are proposed, coming mainly from the research work done in connection with lot

sizing in an MRP context. Gorham [1968] describes the Part Period Balancing procedure by which the lot size covers the requirements for a number of periods selected in such a way that the total cost of placing orders be as close to the total carrying costs as possible. Silver and Meal [1973] propose a heuristic that sets the economic order quantity such as to minimize total relevant costs per unit time for the duration of the replenishment quantity. Other procedures are presented by Plossl and Wight [1971] and Orlicky [1975, ch. 6]. Berry [1972] suggests an experimental framework for systematically comparing the various lot sizing procedures.

In the case of stochastic demand, any attempt to use inventory control policies of the types described in sections 5.3 and 5.4 should take account of the time variability of demand and, consequently, one would expect that the values of the control parameters (order quantity Q, order point s, order-up-to-level S) also become functions of time. We have seen the computational problems already encountered in designing control policies under stationary demand assumptions. Therefore, with dynamic stochastic demand an optimization of the policy control variables is out of the question for any implementation purposes, and one has to resort to heuristics (see, for example, Peterson and Silver [1979, ch. 8.9]). Unfortunately this might not be of too much help either, as the problem might prove extremely difficult for reasons other than computational. Recall that, to develop a dynamic model with stochastic demand, one needs to know the distribution of demand in each future period. Or, in the real world, simply predicting what the mean demand will be for each future period could be a very tough job, not to speak of forecasting the exact nature of the probability distribution.

6 The System-Management Interaction and Evaluation Module

This module is intended to provide management with such information as to allow the selection of policy variables, to permit an evaluation of system performance and the identification of problem areas where managerial intervention is required.

6.1 Exchange Curves

In section 2.1 a discussion on costs in inventory systems was carried out. If we refer, for instance, to the inventory carrying costs (storage and handling, property taxes, insurance, spoilage, obsolescence, pilferage, rent for storing facilities, capital costs), it is apparent that some components represent out-of-pocket expenditures (e.g., rent for leased storage space, or interest paid for capital borrowed from banks), other components are foregone opportunities for return that could be earned by alternative uses of internal funds (e.g., company owned storage space could be used for other productive activities, or internal funds tied up in inventories might be put to work in alternative investments and produce a certain rate of profit), and some have both features.

Also, it is important to realize that costs used in production and inventory control problems may differ from the accounting costs as the bases of their definition can differ. Finally, many cost elements cannot be determined accurately, so accounting rules have to be employed which may change as a matter of company policy.

The idea emerging from these considerations is that some costs, to the extent that they include foregone opportunities or their definition is subject to arbitrary rules, can be adjusted so as to help achieve certain objectives. Therefore, they can be thought of as representing management policy variables.

Other parameters involved in designing an inventory control system, without being costs, are also policy variables, such as the prespecified service measures used in setting safety stocks (section 4.2.2).

An exchange curve shows the tradeoffs that can be achieved between two or more aggregate measures of performance of the inventory system as the policy variable's value is varied. Thus, for instance, a large value for the inventory carrying charge r generates smaller inventories* with more frequent replenishments and higher total ordering costs. A small r encourages the buildup of inventories and reduces the annual replenishment expenses.

If for a population of n stocked items we consider that the charge r is the same for all of them, although unknown yet, we can graph an exchange curve in terms of the total (i.e., across all items) average inventory investment vs. the total annual ordering cost, having r as a parameter and assuming that all items are replenished in economic order quantities.

The economic order quantity Q_i^* for the i-th item is given by Wilson's lot size formula (4.1). It is straightforward to derive the two aggregate measures mentioned above:

$$Y = \begin{matrix}\text{Total average} \\ \text{inventory investment}\end{matrix} = \sum_{i=1}^{n} \frac{1}{2} Q_i^* C_i = \frac{1}{\sqrt{2r}} \sum_{i=1}^{n} \sqrt{A_i D_i C_i}$$

$$X = \begin{matrix}\text{Total annual} \\ \text{ordering cost}\end{matrix} = \sum_{i=1}^{n} \frac{A_i D_i}{Q_i^*} = \frac{\sqrt{r}}{\sqrt{2}} \sum_{i=1}^{n} \sqrt{A_i D_i C_i}$$

---

* By inventories we are refering here to cycle or working stock. Safety stocks, for a given level of customer service, tend to increase as EOQ decreases as a result of a high r (see section 4).

The equation of the exchange curve is an hyperbola:

$$XY = \frac{1}{2}\left(\sum_{i=1}^{n} \sqrt{A_i D_i C_i}\right)^2$$

It is also true that:

$$\frac{Y}{X} = \frac{1}{r}$$

and this helps us pinpoint various values of the inventory carrying charge in the exchange curve.

If management does not want a total investment in inventory larger on the average than, say, \$650,000 the exchange curve (Figure 29) shows that the inventory carrying charge should not exceed r = 0.27; for this

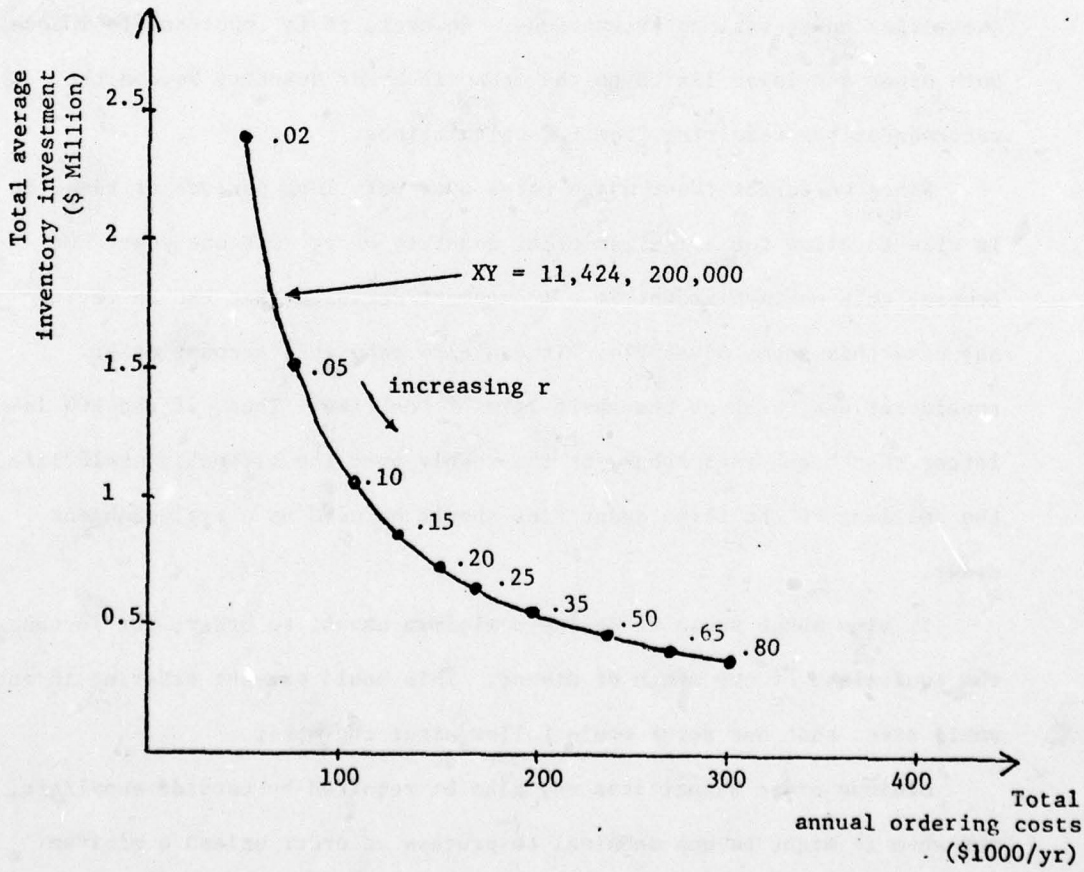


Figure 29: Exchange curve for a value of $\sum_{i=1}^{n} \sqrt{A_i D_i C_i}$ = 478,000, and replenishment by EOQ

value of r the total annual ordering cost is $175,628. A value of r = 0.20 (which is most commonly cited in the literature on inventory control as being typical of the American manufacturing industry) implies a total average inventory investment of $755,784 and ordering costs of $151,157.

When exchange curves for safety stocks are derived, the policy variable can be, for example, the safety factor k (see section 4) and the aggregate measures may be chosen: the total safety stock investment vs. the expected annual value of backordered demand. For details see Brown [1967, ch. 14 and 17] or Peterson and Silver [1979, ch. 6 and 7].

6.2 How Much to Order?

We have presented in section 2 decision rules on economic order quantities under various assumptions. However, it is important to impose both upper and lower limits to the economic order quantity beyond the recommendations resulting from EOQ calculations.

Since we cannot trust usage rates over very long periods of time, it is wise to allow for a maximum order quantity of at most one year. Of course, this maximum amount is a management decision that can be revised any time this seems advisable. It can also take into account other considerations, such as the shelf life of the time. Thus, if the EOQ is larger than a one year supply or the supply over the allowable shelf life, the smallest of the three quantities should be used as a replenishment order.

It also makes sense to define a minimum amount to order; for instance, the equivalent of one month of demand. This would prevent ordering in such small sizes that one setup would follow after the other.

Minimum order quantitites may also be required by outside suppliers, for whom it might be uneconomical to process an order unless a minimum amount is requested.

In some production processes, such as in the chemical industry, the size of the production batch could be imposed by the capacity of certain reaction tanks or of some special containers.

Finally, for lumpy demand items, Brown [1977, p. 248] recommends that the economic order quantity should be rounded upward to be at least as large as the standard deviation of the lead time demand.

6.3 Updating Frequency of System Parameters

The frequency of updating the system parameters depends on several factors such as whether the item is fast moving or slow moving, whether it is A, B, or C class item.

A very important parameter is the length of the review period; that is, the length of the time interval between regular reviews of the available inventory. For the important A items, continuous close surveillance is recommended, i.e., stock status should be updated after each transaction. The C class items warrant the least attention and, therefore, large safety stocks and a simple control system (e.g., the two-bin system) should be employed. For all other items it seems reasonable to recommend the use of one week as a review period, as a proxy for a continuous review system.

Since demand forecasts are used primarily to define order points and order quantities, the updating frequency of the forecasting is directly related with the updating of EOQ's and order points. Monthly updating of order points and quarterly updating of EOQ's for fast moving items are normally considered appropriate. This implies that demand forecasting should be updated monthly and that the lead times should be expressed in months. However, for important A items, the EOQ and order point should be reviewed every time an order is placed.

In the case of slow moving items, if demand forecasts and order

points are based on empirical distributions, we suggest that the order points and EOQ's be updated only once a year. However, when exponential smoothing with a very small α is used for forecasting, it is necessary to update demand forecasts and order points once a month, and FOQ's once a quarter.

For lumpy demand items, where an (s,S) ordering policy is appropriate, the maximum and minimum levels are revised once a year or after at least 30 demand transactions, whichever occurs earlier (Brown [1977, p. 248]).

6.4 Actions to be Taken When the Tracking Signal is Triggered

Section 3.5 has dealt with the issue of tracking signals to monitor forecast errors, and a test based on the smoothed forecast error, to check whether the forecast is biased, was developed there. Discussions were also presented in section 3 on adaptive control techniques, by which the value of the smoothing constant was regarded as a parameter and its value changed automatically in response to indications from the tracking signal of an out-of-control situation.

We should be aware, however, that forecasting may not be left entirely under the supervision of automated systems, as sometimes fundamental changes occur in the demand pattern that require human intervention. It is also true that if we are dealing with a large number of items, it could be impractical to study the demand behavior of every item whose tracking signal triggers and, therefore, some mixed system, that combines automatic control with managerial action, is needed.

The procedure we present here attempts to correct the disturbance the first time it occurs by automatically increasing the smoothing constant, thus making the forecast respond faster to actual demand changes. After a specified number of months, the computations switch back to the normal value of α, if the forecast agrees with the demand. But, if the

tracking signal is triggered while using the fast smoothing constant, a report is printed containing the demand history for the past 12 months (or more if available and desired) and external intervention is requested to correct the forecast. Since in many cases the increase of α is sufficient to correct the forecasting inaccuracies,* this procedure only signals for management analysis of those cases requiring external judgement.

Particular values of α which correspond to normal smoothing have been discussed in section 3. To decide whether to use normal smoothing or fast smoothing, we need to keep track of a smoothing rate counter, COUNT, for each stock item. If the forecast model was initiated by using one year or more of previous data, or if we are fairly confident of the values used to initialize the forecasting model, the counter COUNT is set to zero. If, however, we are dealing with a new item or one with a very short history, the counter is set equal to the number of months we think is needed for the forecast to track the demand properly, by using the fast smoothing coefficient. Let us say, for the sake of being specific, that we let COUNT = 6. Actually, any number from 4 to 9 seems to be appropriate, depending on the item's demand pattern. Each month, in the process of revising the forecast, the counter value is reduced by one, until it reaches zero.

Whenever the tracking signal is triggered and the counter is zero, the normal smoothing coefficient should be used; and whenever the counter

---

* Sometimes errors in recording data may slip through and, if large enough, when processed in revising the forecasts might trip the tracking signal. These one-time errors are normally corrected by the automatic procedure outlined above. However, to prevent this sort of situation or at least to detect the big outliers, the concept of "demand filters" was developed. Thus, any actual demand which is more than 4 or 5 current standard deviations away from the forecast should be called to the attention of an appropriate management person, as a possible outlier.

is positive, the fast smoothing constant should be used.

Whenever the tracking signal is triggered and the counter is zero, automatically the counter is increased, say to COUNT = 6, and the smoothing rate takes on the fast value. In this form, during the coming six months the forecast will be uased more on current demand and less on past history.

Whenever the tracking signal is triggered and the counter has a positive value, this indicates that the faster smoothing is not adequate to compensate for the change in demand pattern, and management should exercise judgement to incorporate the proper corrections into the model. The computer should generate a report containing the item code number, its cost (to measure the importance of its contribution to inventory investment), the value of the tracking signal detection limit, the type of forecasting model used (with or without trend), and historical information of the past 12 months regarding actual demand, forecast level, trend (if available), demand forecast, forecast error, and the standard deviation or variance of the forecast errors. When the necessary corrections have been made, the smoothed error is set to zero and the smoothing rate counter is set to zero or six (say), depending on whether normal or fast smoothing is used, respectively.

The types of corrective actions management can take to improve the forecasting procedures are, basically the following:

- Change the forecast level,
- Change the forecast trend, if corrections for trends are used,
- Incorporate trend, if corrections for trends are not used,
- Cancel trend, if corrections for trend were used but that seem inappropriate,
- Modify the value of the smoothing constants,

- Modify the value of the tracking signal factor.

These modifications are self explained. It may take some time until management develops the proper skill to deal with this problem. At least until the learning process is well established, it could pay off to keep track of the forecast errors resulting from the modified forecasts and the unchanged forecasts, in order to encourage the type of changes which seem to result in significant improvement.

6.5 Management Adjustments to Statistical Forecasts

The statistical forecasting methods that we have incorporated in the inventory control system are designed to detect and extrapolate consistent patterns in past demand data including base, trends, and seasonalities. However, there are many instances where information is available about changes in the demand pattern that have not yet taken place, that could effectively improve the statistical forecast provided by the system. Price changes development of new items, advertising promotions, competitor policies known in advance, changes in the national economy, and opening of new markets are just a few examples of events that could greatly affect the demand of all or a few items. These events can be recognized by management long before they actually modify the demand pattern of a given item. Thus, it becomes important that management introduces these subjective elements into the forecasting system, by adjusting the statistical forecast provided by the computer. Figure 30 illustrates the demand forecasting system when external management adjustments are combined with the statistical forecast computations to create the final forecast.

The actions that management could take to adjust the statistical forecast are essentially the same actions that are available when the demand tracking signal is triggered, namely:

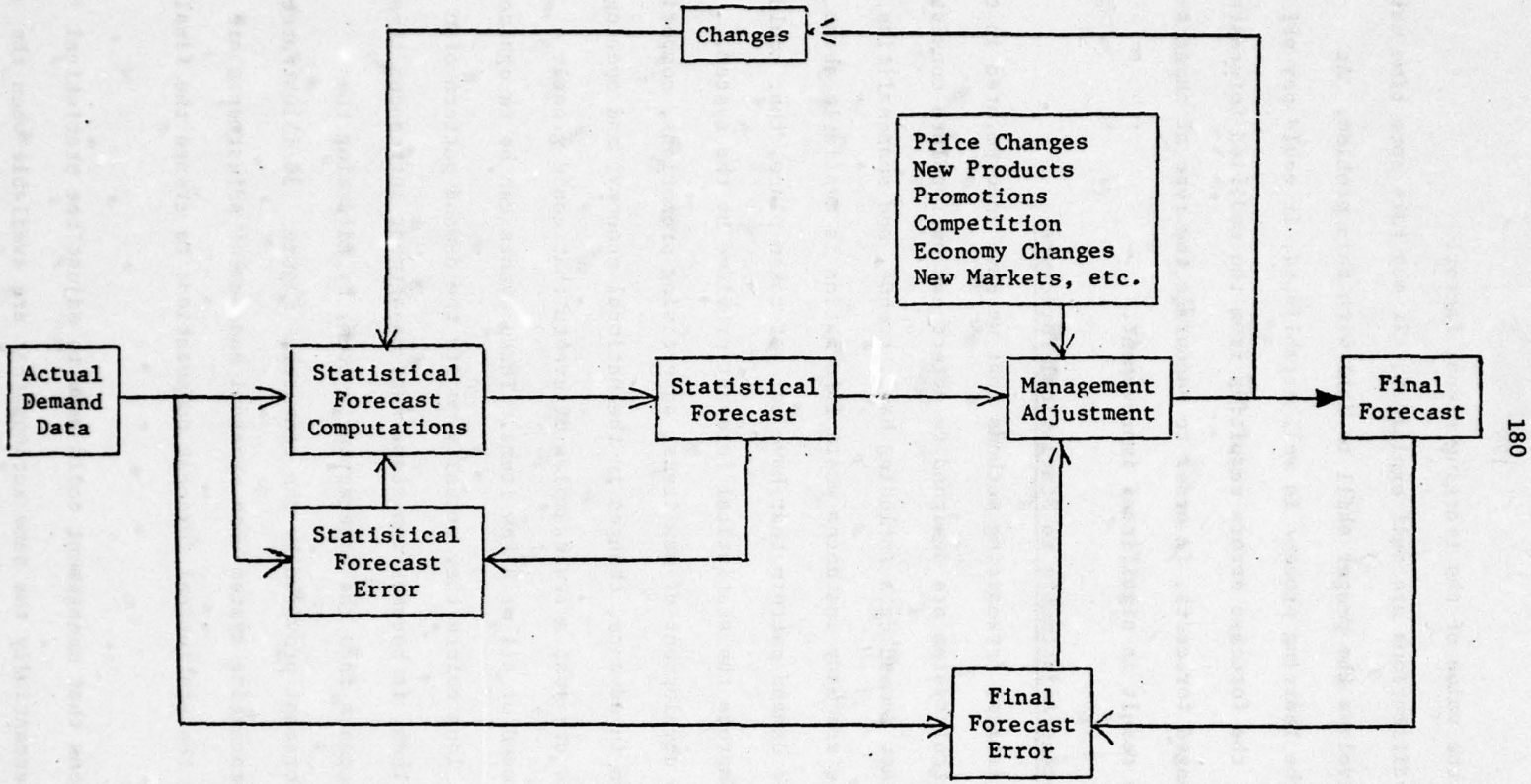


Figure 30: Demand forecasting system with management adjustment

- Change the value of the forecast for the present period with or without carrying this change into the forecast model.
- Modify some or all of the parameters involved in the forecasting model, i.e., smoothing coefficients, level, trend, tracking signal factor.
- Modify the type of forecasting model used, i.e., forecasting with or without trend, consideration of seasonal cycles.

Whatever the change may be, when a forecast adjustment results exclusively from management intervention, it is important to keep track separately of the forecast errors incurred by the adjusted forecast and by the unchanged statistical forecast. In this way, a measure of performance can be attached to external adjustments and a learning process can be established, by encouraging the continuation of those kinds of changes which usually improve the forecast and by discouraging changes in the wrong direction.

6.6 *Production and Inventory Control System Outputs*

The purpose of this section is to provide a very general discussion on the type of outputs we could expect to generate from a production and inventory control system. The system outputs can be divided into four categories:

A. Operating documents which include shop orders, purchase requisitions, and finished product final assembly orders,

B. Routine status reports,

C. Management exception reports,

D. Analysis reports.

The operating documents are those which direct the production operations. The primary purpose of the production and inventory control system is to

issue these documents at the appropriate times for the required quantities. The routine status reports are used for reference purposes and, to some extent, for management review to check that the system is operating satisfactorily. The management exception reports are intended to draw attention to particular problem areas and usually requires some management intervention or judgement. The analysis reports periodically summarize various operating statistics and, therefore, permit an evaluation of the operating performance in the recent past. These reports can be used to identify more general problem areas than are specified in the management exception reports and also provide data which serves as a basis for management selection of system parameters.

We now provide brief comments regarding each type of output we may consider appropriate in a production and inventory control system:

A. Operating Documents

1. *Shop Orders*. Shop orders are the documents which direct the in-plant production of manufactured products.
2. *Purchase Requisitions*. These documents generate a request to initiate the replenishment of a particular purchased item for a specific quantity.

B. Routine Status Reports

1. *Inventory Status*. The inventory status report is prepared once every review period and shows the various status information for each item for which there is some physical inventory or for any item which is designated as a stocked item but which may not have any inventory at that time.
2. *Non-Stock Items Requirement Status*. The non-stock items requirement status report shows the level of known requirements based on open orders by the date of requirements in the future. It shows whether a shop order or a purchase requisition has been released as yet to

cover the requirements or if not, the date on which the document will be generated.

3. *Shop Order and Purchase Requisition Status.* This report is primarily intended to show the degree of completion of partially completed open orders.
4. *Finished Product Requirement Status.* This report shows the level of future customer order commitments for each finished product. It shows the number of customer orders for which shipment is past due for each finished product.
5. *Open Customer Order Status.* This report lists each open customer order, i.e., all of the items included in the order which have not been shipped and the status of each item.

C. Management Exception Reports.

1. *Customer Order Edit Rejects.* This report lists the customer orders and order changes which were entered into the system, but which were rejected because of some type of validity check error.
2. *Inventory Shortage.* It may be decided to have inventory shortages reported in two different ways. One report lists the inventory shortages by items, i.e., it identifies each shop order number for which each item is backordered, the date the backorder occurred and the quantity unfilled. This report is a quick indication of how great the need is for a particular item. The other version of the inventory shortage report would organize the information by shop order, listing all components required for the shop order which are not available and will also specify the quantity required of each. This report is useful to determine which items are holding up production of the particular shop order.
3. *Past Due Report.* This report lists all of the open shop orders which

have not been completed by the specified due date, the open purchase requisitions for which the material has not been received by the specified due date, and the open customer orders which have not been shipped by the specified promised shipping date.

4. *Forecast Tracking Error Signals.* Whenever the forecasting procedures are not tracking demand in a satisfactory way automatically a message is released indicating the forecast tracking error. According to the magnitude of this error and the frequency in which this error occurred, an automatic correction is made by the system or a notice is released indicating the need of outside management intervention (just discussed in section 6.4).

D. Analysis Reports

1. *Shipping on Time/Delay Performance.* This report analyzes the shipping performance for customer orders.
2. *Vendor on Time Delivery/Quality Performance.* This report would contain an analysis of the purchase orders placed.
3. *Vendor and Shop Lead Time Analysis.* This report contains an analysis of the manufacturing lead time on shop orders and purchase delivery lead times on purchase requisitions on all such actions in the past quarter.
4. *Production Rate Report.* The production rate for an item is the rate at which units are finished once the first unit in a lot has been completed. It is determined by dividing the total number of units in a lot by the elapsed time between the date the first unit is completed and the date the last unit is completed.
5. *Inventory Service Performance.* This report shows for each stocked item the number of purchase/manufactured lots completed in the past quarter, the completion performance on this lot, i.e., the number

that were completed on time and the number of days late for any late orders, the total usage over the past quarter and the total number of units backordered over the period.

6. *Usage Cost Rank Listing.* This report lists all products ranked in order of their usage during the past year, to be used as the basis for the ABC item classification (see section 4.1).
7. *Inactive Stock Report.* This report lists all of the items which did not have any usage during the past year, but for which there is some stock in inventory.
8. *Suggested Classifications for Made-to-Order Versus Stock.* This report is the result of the made-to-order versus stock analysis performed on each finished product (section 5.1). For each finished product the report indicates one of the following four classifications as the result of the analysis:
   - Stock because of the delivery service requirement
   - Stock because of economic justification
   - Made-to-order because of economic justification
   - No clear economic advantage
9. *Finished Product Minimum Promised Delivery Time.* This report specifies the minimum delivery time that can be provided for each finished product based on the total manufacturing lead time of the finished product.

AD-A077 562 MASSACHUSETTS INST OF TECH CAMBRIDGE OPERATIONS RESE--ETC F/G 15/5
INVENTORY MANAGEMENT.(U)
NOV 79 A C HAX , D I CANDEA N00014-75-C-0556
UNCLASSIFIED TR-168 NL

3 OF 3
ADA
077 562



END
DATE
FILMED
1 -80
DDC

Bibliography

Aggarwal, S. C., "A Review of Current Inventory Theory and Its Applications", International Journal of Production Research, No. 12, 1974, pp. 443-472.

Arrow, K. J., T. E. Harris, and J. Marschak, "Optimal Inventory Policy", Econometrica, XIX, 1951, pp. 250-272.

Arrow, K. J., S. Karlin, and H. Scarf (editors), Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, Stanford, California, 1958.

Balintfy, J. L., "On a Basic Class of Multi-Item Inventory Problems", Management Science, Vol. 10, No. 2, January 1964, pp. 287-297.

Bitran, G. R., and A. C. Hax, "On the Design of Hierarchical Production Planning Systems", Decision Sciences, Vol. 7, No. 1, January 1977.

Bellman, R., I. Glicksberg, and O. Gross, "On the Optimal Inventory Equation", Management Science, Vol. 2, No. 1, October 1955, pp. 83-104.

Benli, O., and P. Nanda, "A Solution Procedure for Location-Allocation-Production Problems", Syracuse, New York, Department of Industrial Engineering and Operations Research, College of Engineering, Syracuse University, 1977.

Berry, w. L., "Lot Sizing Procedures for Requirements Planning Systems: A Framework for Analysis", Production and Inventory Management, Vol. 13, No. 2, 1972, pp. 19-34.

Bomberger, E. E., "A Dynamic Programming Approach to a Lot Size Scheduling Problem", Management Science, Vol. 12, No. 11, July 1966, pp. 778-784.

Box, G. E. P., and G. M. Jenkins, Time Series Analysis, Holden-Day, San Francisco, 1976.

Brown, R. G., Decision Rules for Inventory Management, Holt, Rinehart and Winston, New York, 1967.

Brown, R. G., "Forecasting" in J. J. Moder and S. E. Elmaghraby (editors), Handbook of Operations Research - Models and Applications, Van Nostrand Reinhold Company, New York, 1978b.

Brown, R. G., "Inventory Control" in J. J. Moder and S. E. Elmaghraby (editors), Handbook of Operations Research - Models and Applications, Van Nostrand Reinhold Company, New York, 1978a.

Brown, R. G., Management Decisions for Production Operations, The Dryden Press, Inc., Hinsdale, Illinois, 1971.

Brown, R. G., Materials Management Systems, John Wiley & Sons, New York, 1977.

Brown, R. G., Smoothing, Forecasting and Prediction of Discrete Time Series, Prentice Hall, Englewood Cliffs, New Jersey, 1963.

Brown, R. G., Statistical Forecasting for Inventory Control, McGraw-Hill, New York, 1959.

Buffa, E. S., Modern Production Management, Wiley, New York, 1969.

Buffa, E. S., and W. H. Taubert, Production-Inventory Systems: Planning and Control, Irwin, Homewood, Illinois, 1972.

Burgin, T. A., and A. R. Wild, "Stock Control-Experience and Usable Theory", Operational Research Quarterly, Vol. 18, No. 1, March 1967, pp. 35-52.

Candea, D. I., "A Comparative Study of Solutions to the Holt, Modigiliani, Muth and Simon Disaggregation Model by Search Techniques", Sloan School of Management, M.I.T., WP814-75, Cambridge, MA, October 1975.

Candea, D. I., "Issues of Hierarchical Planning in Multi-Stage Production Systems", Technical Report No. 134, Operations Research Center, M.I.T., Cambridge, MA, July 1977.

Chow, W. M., "Adaptive Control of the Exponential Smoothing Constant", Journal of Industrial Engineering, Vol. 16, No. 5, 1965, pp. 314-317.

Clark, A. J., "An Informal Survey of Multi-Echelon Inventory Theory, Naval Research Logistics Quarterly, Vol. 19, 1972, pp. 621-650.

Cohen, G. D., "Bayesian Adjustment of Sales Forecasts in Multi-Item Inventory Control Systems", Journal of Industrial Engineering, Vol. 17, No. 9, 1966, pp. 474-479.

Croston, J. D., "Forecasting and Stock Control for Intermittent Demands", Operational Research Quarterly, Vol. 23, No. 3, September 1972, pp. 289-303.

Croston, J. D., "Stock Levels for Slow-Moving Items", Operations Research Quarterly, Vol. 25, No. 1, March 1974, pp. 123-130.

Crowston, W. B., and M. Wagner, "Lot Size Determination in Multi-Stage Assembly Systems", Alfred P. Sloan School of Management, M.I.T., WP508-71, Cambridge, MA, September 1970.

Crowston, W. B., M. Wagner, and J. F. Williams, "Economic Lot Size Determination in Multi-Stage Assembly Systems", Management Science, Vol. 19, No. 5, January 1973, pp. 517-527.

Dorsey, R. C., T. J. Hodgson, and H. D. Ratliff, "A Network Approach to a Multi-Facility, Multi-Product Production Scheduling Problem without Backordering", Management Science, Vol. 21, No. 7, March 1975, pp. 813-822.

Dorsey, R. C., T. J. Hodgson, and H. D. Ratliff, "A Production Scheduling Problem with Batch Processing", Operations Research, Vol. 22, No. 6, November-December 1974, pp. 1271-1279.

Dvoretzky, A., J. Kiefer, and J. Wolfowitz, "The Inventory Problem: I, Case of Known Distributions of Demand; II, Case of Unknown Distributions of Demand", Econometrica, XX, 1952, pp. 187-222.

Dvoretzky, A., J. Kiefer, and J. Wolfowitz, "On the Optimal Character of the (s,S) Policy in Inventory Theory", Econometrica, XXI, 1953, pp. 586-596.

Dyer, D., "To Stock or Not to Stock? That is the Question", Modern Distribution Management, Vol. 7, No. 5, March 23, 1973, pp. 3-7.

Eilon, S., and J. Elmaleh, "Adaptive Limits in Inventory Control", Management Science, Vol. 16, No. 8, April 1970, pp. B533-B548.

Eilon, S., and J. Elmaleh, "An Evaluation of Alternative Inventory Control Policies", The International Journal of Production Research, Vol. 7, No. 1, 1968, pp. 3-14.

Elmaghraby, S. E., "The Economic Lot Scheduling Problem (ELSP): Review and Extensions", Management Science, Vol. 24, No. 6, February 1978, pp. 587-598.

Feller, W., An Introduction to Probability Theory and Its Applications, Vol. II, John Wiley and Sons, New York, 1971.

Forrester, J. W., Industrial Dynamics, M.I.T. Press, Cambridge, MA, 1961.

Gabbay, H., "A Hierarchical Approach to Production Planning", Operations Research Center, M.I.T., TR-120, Cambridge, MA, December 1975.

Geisler, M., "A Test of a Statistical Method for Computing Selected Inventory Model Characteristics by Simulation", Management Science, Vol. 10, No. 4, July 1964, pp. 709-715.

Gerson, G., and R. G. Brown, "Decision Rules for Equal Shortage Policies", Naval Research Logistics Quarterly, Vol. 17, No. 3., 1970, pp. 351-358.

Gorham, T., "Dynamic Order Quantities", Production and Inventory Management, Vol. 9, No. 1, 1968, pp. 75-79.

Goyal, S. K., "Analysis of Joint Replenishment Inventory Systems with Resource Restrictions", Operational Research Quarterly, Vol. 26, No. 1, April 1975, pp. 197-203.

Goyal, S. K., "Determination of Optimum Packaging Frequency of Items Jointly Replenished", Management Science, Vol. 21, No. 4, December 1974, pp. 436-443.

Goyal, S. K., "Lot Size Scheduling on a Single Machine for Stochastic Demand", Management Science, Vol. 19, No. 11, July 1973, pp. 1322-1325.

Graves, S. C., "The Multi-Product Production Cycling Problem", unpublished dissertation, University of Rochester, October 1977.

Groff, G. K., and J. F. Muth, Operations Management: Analysis for Decisions, Richard D. Irwin, Homewood, Illinois, 1972.

Hadley, G., and T. M. Whitin, Analysis of Inventory Systems, Prentice-Hall, Englewood Cliffs, NJ, 1963.

Hanssmann, F., Operations Research in Production and Inventory Control, John Wiley & Sons, New York, 1962.

Hastings, N. A. J., and J. B. Peacock, Statistical Distributions, London Butterworths, London, 1975.

Hax, A. C., "Aggregate Production Planning", in J. Moder and S. E. Elmaghraby (editors), Handbook of Operations Research: Models and Applications, Van Nostrand Publishing Co., 1978. pp. 127-172.

Hax, A. C., "The Design of Large Scale Logistics Systems: A Survey and an Approach", in W. H. Marlow (editor), Modern Trends in Logistics Research, M.I.T. Press, Cambridge, MA, 1976.

Hax, A. C., "Integration of Strategic and Tactical Planning in the Aluminum Industry", Operations Research Center, Working Paper 026-73, M.I.T., Cambridge, MA, September 1973.

Hax, A. C., and H. C. Meal, "Hierarchical Integration of Production Planning and Scheduling", in M. A. Geisler (editor), Studies in Management Sciences, Logistics, Vol. I, North Holland-American Elsevier, 1975.

Hoel, P. G., S. C. Port, and C. J. Stone, Introduction to Statistical Theory, Houghton Mifflin Company, Boston, 1971.

Holt, C. C., F. Modigliani, J. F. Muth, and H. A. Simon, Planning Production, Inventories, and Work Force, Prentice-Hall, Englewood Cliffs, NJ, 1960.

Iglehart, D., "Dynamic Programming and Stationary Analysis of Inventory Problems", in H. E. Scarf, D. M. Gilford, and M. W. Shelly (editors), Multistage Inventory Models and Techniques, Stanford University Press, Stanford, CA, 1963b.

Iglehart, D., "Optimality of (s,S) Policies in the Infinite Horizon Dynamic Inventory Problem", Management Science, Vol. 9, No. 2, January 1963a, pp. 259-267.

Jensen, P. A., and H. A. Khan, "Scheduling in a Multi-Stage Production System with Setup and Inventory Costs", AIIE Transactions, Vol. 4, No. 2, 1972, pp. 126-133.

Johnson, J., "On Stock Selection at Spare Parts Stores Sections", Naval Research Logistics Quarterly, Vol. 9, No. 1, March 1962, pp. 49-59.

Johnson, L. A., and D. C. Montgomery, Operations Research in Production Planning, Scheduling, and Inventory Control, John Wiley and Sons, New York, 1974.

Karlin, S., "The Application of Renewal Theory to the Study of Inventory Policies", in K. J. Arrow, S. Karlin, and H. Scarf (editors), Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, Stanford, CA, 1958.

Karlin, S., "Steady State Solutions", in K. J. Arrow, S. Karlin, and H. Scarf (editors), Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, Stanford, CA, 1958.

Karmarkar, U. S., "Multilocation Distribution Systems", Technical Report No. 117, Operations Research Center, M.I.T., Cambridge, MA, September 1975.

Klingman, D. D., R. M. Soland, and G. T. Ross, "Optimal Lot-Sizing and Machine Loading for Multiple Products", in The Problems of Disaggregation in Manufacturing and Service Organizations - Conference Proceedings, The Ohio State University, Columbus, Ohio, March 10-11, 1977.

Krauss, G. H., "An Approach to the Analysis of Integrated Production-Distribution Systems", in The Problems of Disaggregation in Manufacturing and Service Organizations - Conference Proceedings, The Ohio State University, Columbus, Ohio, March 10-11, 1977.

Levy, J., "Optimum Inventory Policy When Demand Is Increasing", Operations Research, Vol. 8, No. 6, November-December 1960, pp. 861-863.

Luenberger, D. C., Introduction to Linear and Nonlinear Programming, Addison Wesley, Reading, MA, 1973.

Magee, J. F., Industrial Logistics, McGraw-Hill, New York, 1968.

Magee, J. F., and D. M. Boodman, Production Planning and Inventory Control, McGraw-Hill, New York, 1967.

Maher, M., J. Gittins, and R. Morgan, "An Analysis of a Multi-Line Re-Order System Using a Can-Order Policy", Management Science, Vol. 19, No. 7, March 1973, pp. 800-808.

Makridakis, S., and S. C. Wheelwright, Forecasting - Methods and Applications, John Wiley and Sons, New York, 1978.

Makridakis, S., and S. C. Wheelwright, Interactive Forecasting, Holden-Day, San Francisco, 1978.

Mangasarian, O. L., Nonlinear Programming, McGraw-Hill, New York, 1969.

McClain, J. O., "Dynamics of Exponential Smoothing with Trend Seasonal Terms", Management Science, Vol. 20, No. 9, May 1974, pp. 1300-1304.

McClain, J. O., and L. J. Thomas, "Response Variance Tradeoffs in Adaptive Forecasting", Operations Research, Vol. 21, No. 2, March-April 1973, pp. 554-568.

McGarrah, R. E., Production and Logistics Management: Text and Cases, John Wiley and Sons, New York, 1963.

Montgomery, D. C., "Adaptive Control of Exponential Smoothing Parameters by Evolutionary Operation", AIIE Transactions, Vol. 2, No. 3, 1970, pp. 268-269.

Montgomery, D. C., and L. A. Johnson, Forecasting and Times Series Analysis, McGraw-Hill, New York, 1976.

Morgan, J. I., "Question for Solving the Inventory Problem", Harvard Business Review, Vol. 41, No. 4, July-August 1963, pp. 95-110.

Naddor, E., "Optimal and Heuristic Decisions in Single- and Multi-Item Inventory Systems", Management Science, Vol. 21, No. 11, July 1975, pp. 1234-1249.

Oral, M., M. S. Salvador, A. Reisman, and B. V. Dean, "On the Evaluation of Shortage Costs for Inventory Control of Finished Goods", Management Science, Vol. 18, No. 6, Feburary 1972, pp. B344-B351.

Orlicky, J., Material Requirements Planning, McGraw-Hill, New York, 1975.

Peterson, R., and E. A. Silver, Decision Systems for Inventory Management and Production Planning, John Wiley and Sons, New York, 1979.

Plossl, G. W., and O. W. Wight, "Material Requirements Planning by Computer", Special Report of the American Production and Inventory Control Society, Washington, D.C., 1971.

Popp, W., "Simple and Combined Inventory Policies, Production to Stock or to Order", Management Science, Vol. 11, No. 9, July 1965, pp. 868-873.

Pratt, J. W., H. Raiffa, and R. Schlaifer, Introduction to Statistical Decision Theory, McGraw-Hill, New York, 1965.

Raiffa, H., Decision Analysis - Introductory Lectures on Choices Under Uncertainty, Addison-Wesley, Reading, MA, 1970.

Raiffa, H., and R. Schlaifer, Applied Statistical Decision Theory, Harvard University Press, Cambridge, MA, 1961.

Roberts, S. D., and R. Reed, "The Development of a Self-Adaptive Forecasting Technique", AIIE Transactions, Vol. 1, No. 4, 1969, pp. 314-322.

Scarf, H., "The Optimality of (S,s) Policies in the Dynamic Inventory Problem", in K. J. Arrow, S. Karlin, and P. Suppes (editors), Mathematical Methods in the Social Sciences, Stanford University Press, Stanford, CA, 1960.

Scarf, H., "A Survey of Analytical Techniques in Inventory Theory", in H. E. Scarf, D. M. Gilford, and M. W. Shelly (editors), Multistage Inventory Models and Techniques, Stanford University Press, Stanford, CA, 1963.

Scarf, H. E., D. M. Gilford, and M. W. Shelly (editors), Multistage Inventory Models and Techniques, Stanford University Press, Stanford, CA, 1963.

Schlaifer, M. J., "The Use of an Economic Lot Range in Scheduling Production", Management Science, Vol. 5, No. 4, July 1959, pp. 434-442.

Schwartz, B. L, "A New Approach to Stockout Penalties", Management Science, Vol. 12, No. 12, August 1966, pp. B538-B544.

Schwartz, B. L., "Optimal Inventory Policies in Perturbed Demand Models", Management Science, Vol. 16, No. 8, April 1970, pp. B509-B518.

Schwarz, L. B., and L. Schrage, "Optimal and System Myopic Policies for Multi-Echelon Production/Inventory Assembly Systems", Management Science, Vol. 21, No. 11, July 1975, pp. 1285-1294.

Silver, E. A., "A Control System for Coordinated Inventory Replenishment", International Journal of Production Research, Vol. 12, No. 6, 1974. pp. 647-671.

Silver, E. A., "Modifying the Economic Order Quantity (EOQ) to Handle Coordinated Replenishments of Two or More Items", Production and Inventory Management, Vol. 16, No. 3, 1975, pp. 26-38.

Silver, E. A., and H. C. Meal, "A Heuristic for Selecting Lot Size Requirements for the Case of a Deterministic Time-Varying Demand Rate and Discrete Opportunities for Replenishment", Production and Inventory Management, Vol. 14, No. 2, 1973, pp. 64-74.

Simpson, K. S., "In-Process Inventories", Operations Research, Vol. 6, No. 6, November 1958, pp. 863-873.

Snyder, R., "Computation of (S,s) Ordering Policy Parameters", Management Science, Vol. 21, No. 2, October 1974, pp. 223-229.

Sokolnikoff, I. S., and R. M. Redheffer, Mathematics of Physics and Modern Engineering, McGraw-Hill, New York, 1958.

Solomon, M. ..., "The Use of an Economic Lot Range in Scheduling Production", Management Science, Vol. 5, No. 4, July 1959, pp. 434-442.

Stevens, C. F., "On the Variability of Demand for Families of Items", Operational Research Quarterly, Vol. 25, No. 3, September 1974, pp. 411-419.

Tana, H. A., and R. W. Skeith, "The Economic Lot Sizes in Multi-Stage Production Systems", AIIE Transactions, June 1970, pp. 157-162.

Trigg, D. W., "Monitoring a Forecast System", Operational Research Quarterly, Vol. 15, No. 3, September 1964, pp. 271-274.

Trigg, D. W., and A. G. Leach, "Exponential Smoothing with an Adaptive Response Rate", Operational Research Quarterly, Vol. 18, No. 1, 1967, pp. 53-59.

Veinott, A. F., "Minimum Concave-Cost Solution of Leontief Substitution Models of Multi-Facility Inventory Systems", Operations Research, Vol. 17, March-April 1969, pp. 262-291.

Veinott, A. F., Jr., "On the Optimality of (s,S) Inventory Policies: New Conditions and a New Proof", SIAM Journal on Applied Mathematics, Vol. 14, No. 5, September 1966b.

Veinott, A. F., Jr., "The Status of Mathematical Inventory Theory", Management Science, Vol. 12, No. 11, July 1966a, pp. 745-777.

Veinott, A. F., Jr., and H. Wagner, "Computing Optimal (s,S) Inventory Policies", Management Science, Vol. 11, No. 5, March 1965, pp. 525-552.

von Lanzenauer, C. H., "Production and Employment Scheduling in Multi-Stage Production Systems", Naval Research Logistics Quarterly, Vol. 17, No. 2, June 1970, pp. 193-198.

Wagner, H., M. O'Hagan, and B. Lundh, "An Empirical Study of Exactly and Approximately Optimal Inventory Policies", Management Science, Vol. 11, No. 7, May 1965, pp. 690-723.

Wheelwright, S. C., and S. Makridakis, Forecasting Methods for Management, John Wiley and Sons, New York, 1977.

Whitin, T. M., The Theory of Inventory Management, Princeton University Press, Princeton, NJ, 1953.

Winters, P. R., "Forecasting Sales by Exponentially Weighted Moving Averages", Management Science, Vol. 6, No. 3, November 1960, pp. 324-342.

Zabel, E., "A Note on the Optimality of (S,s) Policies in Inventory Theory", Management Science, Vol. 9, No. 1, October 1962, pp. 123-125.

Zangwill, W. I., "A Backlogging Model and a Multi-Echelon Model of A Dynamic Economic Lot Size Production System - A Network Approach", Management Science, Vol. 15, No. 9, May 1969, pp. 506-527.

Zimmermann, H. J., "Periodic vs. Perpetual Inventory Control Systems", Production and Inventory Management, Vol. 7, No. 4, October 1966, pp. 66-79.

Zimmermann, H. J., and M. G. Sovereign, Quantitative Models for Production Management, Prentice-Hall, Englewood Cliffs, NJ, 1974.